Working Paper

n°: 2013-61-01

Respect the Data!

Gilles Laurent [a]

a INSEEC Business School, 27 avenue Claude Vellefaux 75010 Paris, France

January 2013

An ulterior version of this article appeared in *International Journal of Research in Marketing*, vol 30, Issue 4, December 2013, ISSN 0167-8116

Abstract

I believe consumer research suffers from a lack of respect for the data. For example, we routinely do not report full experiments when they do not produce the results we expected and we often eliminate alleged "outliers" on the basis of inappropriate rules, all this leading to biased test reports. We rely less and less on non-experimental data, while the serious limitations of experimental data may create structural discrepancies with the other, non-experimental instances of the phenomenon or processes we research, and make it impossible to study a number of major consumer phenomena. We tend to accept lessons from empirical data only when we can describe them as confirming pre-existing conceptual frameworks. This paper discusses multiple forms of lack of respect for the data and proposes recommendations. Overall, in contrast with Stapel, I argue that the data should not play a subordinate role.

(143 words)

1. Introduction

Two recent and widely discussed scandals provided evidence of complete forgeries of data sets by consumer researchers: the Stapel case (investigated in depth by the Levelt Committee et al. 2012) and the Smeesters case (Retractionwatch 2012). But the lessons to be derived should not be limited to the avoidance of this extreme form of wrongdoing. Lack of respect for the data may take many other forms and this article attempts to draw the attention of consumer behavior colleagues and doctoral students on the important dangers this creates for consumer behavior research.

In a citation that rightfully entailed the consternation of the Levelt Committee et al., Stapel wrote (2000): “The freedom we have in the design of our experiments is so enormous that when an experiment does not give us what we are looking for, we blame the experiment, not our theory. (At least, that is the way I work.) Is this problematic? No.” I agree with the Levelt Committee et al. that this citation expresses Stapel’s “primacy of the theory – and therefore the subordinate role of the data?” (2012, p.40). I argue in this article that we should not relegate the data to a subordinate role. This introduction illustrates the dangers of not respecting the data through a few examples. Longer discussions and recommendations appear in the following sections of the paper.

1.1 We should not mutilate our data set

A first danger comes from unwarranted modifications of our data set after we have collected it. More than with the (hopefully) rare forgery of some or all observations, I am concerned about data elimination, discarding observations or even full data sets. Consider a first example, the “best of” tactic: it has become very common to run multiple experiments and to report only some of them, those that support the researcher's argument; or to collect multiple measures of a dependent variable, and to report only a selected subset of the results. This leads to dangerous biases in reported hypothesis tests. Another dangerous behavior is, before performing an ANOVA on experimental data, to eliminate alleged "outliers" on the basis of inappropriate rules which, as shown below, are likely to bias upwards the F tests while not ensur that the reduced data set follows a Gaussian distribution. The unwarranted discarding of experiments or of alleged "outliers" amounts to analyze and present mutilated data sets, rather than the original ones.

1.2 We should avoid discrepancies between our experimental data sets and the other instances of the phenomenon or process we investigate

Another danger comes from forgetting that the phenomena and underlying processes we research pre-existed our investigations, occur in many other settings while we investigate them in one specific setting, and will exist (hopefully) long after we are done with our project. Given that experimental research has become extremely dominant in consumer research (as evidenced below), we should be concerned about possible discrepancies between the data in our experiments and the data relative to the non-experimental instances of the phenomena or processes we investigate. Discrepancies may appear because consumer behavior experiments are *myopic* in the temporal sense, mostly analyzing short-run answers to short-run stimuli, whereas consumers typically develop knowledge and choice strategies over time and repeated choice occasions. Other discrepancies may appear when results have been obtained in the narrow setting of a lab, sampling from a very specific student population, and no effort is made, nor encouraged by review teams to confirm these results by replications outside that setting. Equally, some major aspects of the behavior of consumers cannot be studied in a lab setting, e.g.

the changes occurring after Chinese consumers move to a large city while enjoying a largely increased income.

1.3 We should let the data speak out of their pre-defined script

A third danger appears if we listen to the data only when we can describe them as consistent with some pre-defined schema, confirming hypotheses based on previous research or an a priori conceptual framework. I argue we should listen to the data even (or particularly) when they give us unexpected lessons, especially when we have no solid pre-existing theory (a case not infrequent in consumer behavior or more broadly in marketing). In such cases, we should be open to learning new lessons and keep an abductive approach. For example, if we observe a lack of convergence among different measures of a theoretically unidimensional construct, we may reconsider this unidimensionality rather than eliminate divergent measures. Equally, we should be cautious about possible non-linearities, and collect data in a manner that permits to detect them.

Some of the problems I discuss in this article are driven by certain practices or requests from review teams and I therefore discuss this specific aspect when appropriate.

Finally, it would be nice to justify this article by providing empirical evidence on the wrongdoings I criticize. However, the guilty seldom report their wrongdoings. For example, authors typically do not mention the experiments they ran but which did not produce the expected results, except if pressured to do so (see a case in section 2.2). Thus, to present evidence on wrongdoings, one needs exceptional circumstances such as the Stapel scandal to see an independent authority collect and report systematically a compendium of examples of inappropriate behavior (Levelt Committee et al. 2012), or to see the guilty publicly discuss their own wrongdoings (Bhattacharjee 2013). While I can mention specific examples of the practices

I criticize, I clearly cannot assess their frequency statistically. I would be delighted to learn that they are, or have become, exceedingly rare or extinct.

2. We should not mutilate our data set

This section discusses three forms of lack of respect for the data at the analysis stage: the hidden or unwarranted elimination of observations, the selective reporting of experiments, and "best of" tactics, before suggesting better reciprocal controls by co-authors. I conceal.

2.1 The concealed or unwarranted elimination of observations

As evidenced in the next paragraphs, it is frequently necessary to "clean" a data set, to eliminate certain observations before performing the statistical analysis. However, we should do it openly and for appropriate reasons.

To support this section by empirical data, I surveyed all 72 experimental papers in the most recent volume of *Journal of Consumer Research* (vol.39, June 2012 to April 2013). I counted 53 instances in which the authors reported eliminating observations for a specific reason (details in the web appendix). The allocation of these instances among articles reveals a troubling pattern. Since the average frequency is 53/72 = 0.74 instance per article, if the instances followed a Poisson process, we would expect to observe (out of 72 articles), 34.49 articles with no instance, 25.39 with a single instance, and 12.12 articles with two or more instances. In fact we observe a very different pattern: too many papers with no instance of elimination (51), too few papers papers with a single instance (4), too many papers with two instances or more (17). These differences are very significant ($\chi^2 = 27.89$ with 2 d.f., $p = 4.4\ 10^{-7}$).

Why so many articles that do not report any data elimination at all (51 out of 72, i.e.71%)? Since there is a very large diversity of factors that may lead to eliminate observations because they correspond to unforeseen incidental deviations from the scheduled data collection protocol (e.g., computer problem, a failure to adequately follow instructions, refusal to perform experimental tasks, missing answers, sickness, as illustrated in the web appendix), I think unlikely that all the observations collected for all those 51 articles were immediately perfect, especially as each of them typically includes four experiments or more. Rather, I conclude that at least some of these authors eliminated some of their original observations but did not report how and why in the paper. My conclusion is reinforced by the fact that, among articles that report at least one instance of eliminating an observation, more than 80% (17 out of 21) report two instances or more (up to five instances in two articles). Thus, when authors correctly decide to report the elimination of observations, they find most often more than one instance to report. I conclude that the elimination of data is very common but mostly hidden. My conclusion is reinforced by the contrast between the frequent occurrence in this sample of articles that do not report data elimination and multiple informal discussions with colleagues who indicated orally that they considered this preliminary data cleaning as both routine and necessary (e.g., to detect respondents who did not read instructions carefully). .

My first recommendation is to respect the data by always reporting clearly whether observations were eliminated and, if so, how many and for which specific reason.

The elimination of an observation is legitimate when it is based on a documented incident in the data collection such as the ones mentioned above. But observations are also eliminated for the sole reason that they are "outliers," i.e. that they take extreme values on the dependent variable. These eliminations are supposed to bring the updated data set (after elimination) close to a normal (Gaussian) distribution, and to make possible an ANOVA. Two rules are often mentioned to do so: eliminate observations that are more than 3 standard deviations away from (or above) the sample mean of the original distribution; eliminate observations that are more than

1.5 Interquartile Ranges (IQR) above the upper quartile of the original distribution or more than 1.5 IQR below its lower quartile. Before discussing these two rules, let me stress that non-parametric statistics would often offer a robust, easy to use, alternative approach for data analysis while not requiring to eliminate observations with extreme values.

Consider first the rule of eliminating observations that are more than 1.5 IQR beyond the quartiles. It relies on a statistical tool introduced by Tukey: the "box-and-whiskers" plot (1977 pp.43 seq.). Let me describe it using Figure 1, panel A as an example. To visualize the distribution of a numerical variable, Tukey first computes its median and quartiles and plots a "box" that contains the middle 50% of the distribution (from top quartile to bottom quartile, with a horizontal bar indicating the median). Then Tukey defines two "whiskers" or "inner fences" at 1.5 IQR beyond the top and bottom quartiles, marked in the plot by two horizontal lines above and below the box[1], and two "outer fences" located a further 1.5 IQR beyond the whiskers (not represented on the plot). Observations between the whiskers and the "outer fences" are "outside" and observations beyond the "outer fences" are "far out." Most often in recent practice names are slightly different: e.g. in SPSS the "box-and-whiskers" plot is called the "boxplot," "outside" observations are called "outliers" and "far out" observations are called "extreme cases."

When reading Tukey (1977, chapter 2), it is clear that he designed this innovative plot to provide an intuitive visualization of the distribution of a variable. For example, in Figure 1, it is immediately obvious that the distribution in panel A (performance of 167 subjects on the classical "speed of treatment" test) is symmetrical, close to a Gaussian, while the distribution is highly skewed in panel B (quantities of an industrial product purchased by different customers).

[1] There is an exception however: if there is no observation beyond one the "whiskers" defined as above, this whisker is moved closer to the median, at the level of the extreme observation.

Nowhere in his book does Tukey suggest that one should eliminate observations located "outside" or "far out." Nevertheless, in multiple informal discussions I have heard colleagues refer to Tukey to justify eliminating observations for the *sole* reason that they were located beyond the whiskers on the dependent variable, calling them "outliers."

In a boxplot, an observation located beyond the whiskers is by definition called an "outlier." An outlier is indeed "a statistical observation that is markedly different in value from the others of the sample" (Merriam-Webster dictionary online). This does not imply automatically that it does not belong in the data set and should therefore be discarded. There are several reasons for that.

The two main reasons are that this procedure strongly biases ANOVA results and that it does NOT bring the updated data set (after elimination) to a normal (Gaussian) distribution. Consider the two possible horns of the alternative: when the raw data set is Gaussian, and when it is not Gaussian.

If the population of interest is perfectly Gaussian, one expects about 4.3% of the observations to be beyond the whiskers[2] and there is no reason to eliminate automatically all of them. Indeed, what would be surprising would be to observe no such "outlier[3]." Eliminating these observations implies that the updated distribution is no longer Gaussian, having lost its two tails. Hence, the estimator of the within-group variance, an essential component of the ANOVA

[2] In a standardized Gaussian distribution, the quartiles are at -.675 and .675, the IQR at 1.35, the whiskers at -2.025 and 2.025; so we expect about 4.3% of the observations to be outside the whiskers without being in any sense outliers from the Gaussian distribution. The "outer fences" are at -3.375 and 3.375 and, in contrast, we expect only 0.074% of observations to be beyond them.

[3] In a sample of n =120 from a Gaussian distribution, the probability of having at least one observation beyond the whiskers is 99.5% (1 minus 95.7% at the power of 120).

procedure, will be biased downwards, by more than 20%. Since the resulting estimate is at the denominator of the F statistics, the suppression of the two tails leads to overestimate the F by about 25%, changing marginal tests from “non-significant” to “significant.”

What if the distribution is not Gaussian, but lognormal or exponential, a frequent case in many real-world populations? Literature provides numerous examples of naturally occurring lognormal distributions taken from “across the sciences[4]” (Atchison and Brown 1963, Johnson, Kotz, and Balakrishnan 1994, Limpert, Stahel and Abbt 2001). Since firm sizes are known to follow lognormal distributions, consider the distribution in Figure 1, panel B (quantities purchased by different customers in an industrial market) to illustrate the absurd consequences of removing observations beyond the whiskers when the raw distribution is lognormal and therefore skewed to the right. The distribution in panel A has a measured skewness of .19 (SD = .19), consistent with the zero skewness of a theoretical Gaussian distribution, while the distribution in panel B has a measured skewness of 2.66 (SD=.21), very significantly above zero. Now eliminate all the observations beyond the whiskers in the data from panel B. The distribution of the remaining observations (panel C) remains clearly not Gaussian and inappropriate for ANOVA (skewness at 1.68 with SD=.22). This leads again to underestimate the within-group variance: in this case, the estimated variance drops from 15.2 million in the original data to 2.8 million, a reduction of more than 80% that would lead to a massive upward bias in the estimated F statistics.

[4] Particle physics, medicine and physiology, economics and sociology (income, inheritance, wealth, bank deposits, industry and firm size, town size, expenditures), biology, anthropometry, ecology, industrial processes, philology, psychology, agriculture, entomology, geology, insurance, and psychology.

Thus, in both cases (Gaussian and non-Gaussian), removing observations located beyond the whiskers leads to underestimate within-group variance and overestimate F statistics in ANOVA[5]. If we wanted to use Swiftian irony, we would encourage authors to set aside as "outliers" all observations outside the quartiles: this would even more underestimate within-group variance, overestimate the F statistics, and increase the chances of getting "significant" and therefore publishable results.

Another reason for not removing observations located beyond the whiskers is that this leads to absurdities for many distributions. Consider again the lognormal distribution in figure 1, panel B. The upper whisker is located at 7,382, which leads to eliminate systematically from the sample all the major purchasers in this market. Even more absurdly, the computed value of the lower whisker is *negative*, so we eliminate no small purchaser. Combining these effects, the procedure creates a major bias: the mean drops by 42%, from 2,479 in the raw data to 1,446!

A last reason for rejecting this procedure is the authoritative argument: as indicated above, Tukey, who developed this definition of "outliers" on the basis of boxplots, never recommended to chop those "outliers" off. [John Tukey is not Dr. Guillotin.] Rather, he recommended a totally different approach, the "re-expression" of the original variable, which I describe later.

The rule of eliminating observations that are more than 3 standard deviations above the sample mean of the original distribution does better on Gaussian distributions (it removes the top 0.13% of the distribution, only slightly biasing the sample). But it may still lead to absurd consequences for other distributions. In a lognormal distribution like the one in panel B, the rule

[5] Lack of space prevents a similar discussion for other non-Gaussian distributions. Sim, Gan & Chang (2005, p.650) demonstrate that, for detecting outliers, the usual values of 1.5 or 3 IQR are "completely inappropriate for a skewed distribution, such as the exponential distribution."

eliminates the top 4% of the sample and still heavily biases the mean which drops by 22% from 2,479 to 1,934. More importantly, the distribution remains non-Gaussian, heavily skewed (2.42 with a SD at .21) and the estimated variance is biased downwards by 52% from 15.2 to 7.3 million. Thus the F statistics obtained from this updated sample are still erroneously overestimated by more than two. In addition, the "3 SD" rule suffers from circular reasoning: the mean and standard deviation used to decide which observations to eliminate are estimated on the original sample, including extreme observations that will be eliminated later.

Some colleagues further argue in informal discussions that alleged "outliers" should be identified and eliminated separately for each experimental condition. I think this is logically inconsistent. In an experiment, one wishes to test by an ANOVA the *hypothesis* H_0 that the distribution of the dependent variable is the same for all conditions. Now, to treat separately each condition in the preliminary step (elimination of alleged "outliers") one has to *assume* that the conditions lead to different distributions of the dependent variable, i.e. to assume a negative conclusion (reject) to the test of H_0 that will be performed in the second step. It is logically inconsistent to test a hypothesis on a data set that has been previously modified (through the elimination step) by assuming that the hypothesis is rejected.

It happens (it happened to me recently) that a review team suggests to apply the same procedure, regarding "outliers," to all studies in an article. I think this is inappropriate. Different studies in the same article may have different types of dependent variables, for example the average of Likert scales (with likely a Gaussian distribution), an amount subjects would be willing to pay (with likely a lognormal distribution), and a duration (likely with an exponential distribution). The procedure to handle alleged "outliers" should not be the same in these different cases.

Finally, the removal of "outliers" offers ample opportunities to create different variants of a data set: the original data set, the set obtained by removing "outliers" identified over the full sample, or the set obtained after removing "outliers" identified within each experimental condition. Researchers can perform ANOVA on these three sets and report the results that fit best with the conclusions they want to reach. This is a typical example of the "best of" tactics denounced by the Levelt Committee et al. (2012), which I shall discuss in section 2.3.

Overall, respect for the data leads to several recommendations. First, while we can eliminate observations if we can justify it by one of the specific incidents described above, we should NOT remove observations for the sole reason that they take extreme values without assessing the overall shape of the raw distribution, which should always be the first step. If the distribution is close to a Gaussian, then we should analyze the raw data without eliminating all the observations beyond the whiskers (eliminating observations at more than 3 SD from the mean or beyond the "outer fences" could be argued). If the data are far from a Gaussian, we should always consider non-parametric tests as an alternative approach. For a good example, see Frederick (2012 p.2). Another option is to follow Tukey's recommendation (1977 chapter 3) and "re-express" the original variable, i.e. transform it such that the distribution of the transformed variable is as symmetrical and close to a Gaussian as possible. Tukey devotes that full chapter to describe possible transformations: logarithms, power functions, reciprocals, etc. For a good example, see Frederick (2012 p.17). Figure 1, panel D, provides another example: it shows the boxplot of the logarithm of the variable described in raw form in panel B, i.e. the logarithms of the quantity purchased by different purchasers. In the raw data there were many high value outliers in Tukey's sense: 8.7% beyond the whiskers, including 5.8% beyond the "outer fences." After re-expression, the distribution is symmetrical and appropriate for an ANOVA. The percentage of outliers in Tukey's sense drops to 1.4%, all located at the bottom end of the distribution as there are no longer high-value "outliers." Company sizes are known to follow lognormal distributions and panel D clearly shows that the major purchasers on this market belong in the distribution and in the analysis.

A final recommendation is that, whenever we eliminate observations, a web appendix should describe precisely the specific argument for the elimination, the full distribution of observations before and after elimination (indicating the values taken by the discarded observations); we should avoid succinct and non-informative statements such as "For the pencils condition, two outliers were excluded" (Frederick 2012 p.5).

2.2 Do not hide experiments

On the basis of numerous discussions with colleagues, I believe that it has become common, when preparing an article, to run many more experiments than what will appear in the article. In her ACR Presidential address, Kahn (2006) cited a study that assumed that an A paper requires running three times as many experiments as what appears in the published paper. This is illustrated by Simmons et al. (2011): Francis (2012) had asserted that the results presented in Galak and Meyvis (2011), seven positive studies out of eight, were "suspicious," because, given the size of the effect, there should have been more negative studies. The latter authors replied that indeed they had obtained five additional negative studies but had kept them in their "file drawer" rather than reporting them in their article; and that their results were no longer suspicious if one took into account all 13 studies. Discussing with colleagues, I feel this is due in part to the fact that many authors believe that review teams require uniformly perfect results, a belief that discourages authors to report non significant or even marginally significant studies.

This concurs with Stapel's statement "that journal editors preferred simplicity. 'They are actually telling you: 'Leave out this stuff. Make it simpler'.' " (Bhattacharjee 2013, p.4).

Respect for the data should lead us to mention in the article the experiments that did not produce the expected results, rather than to discard them inconspicuously. They provide very rich information: empirical evidence designed and collected under the full control of the experimenter but that contradict predictions. This is the exact opposite of serendipity. If an effect is so weak that it is significant in only four experiments out of eight, this is informative and should be reported. If the effect appears only with certain manipulations, certain measures, certain populations, certain experimental settings, etc., this is informative and should be reported. This is the place for an abductive approach. We should develop hypotheses on why these carefully designed and well-controlled experiments did not produce the expected results, and test them. Reporting all experiments would avoid the situation described by Simmons et al. (2011) above. Note that Galak and Meyvis (2012) further replied that Francis should have asked them about their "file drawer" studies instead of writing an article criticizing them. It would have been simpler to report all 13 studies in the initial article, my personal recommendation.

I think part of this problem has to do with the blurring of the distinction between pilot tests and full-scale experiments, due to the wide acceptance[6] (if not recommendation!) of student

[6] Among the 72 experimental articles that appeared in vol.39 of *Journal of Consumer Research*, 90% used student samples, 26% used MTurk or another unspecified "online panel," providing minimal statistics on respondents' age and gender, and 14% used MTurk or another unspecified "online panel," with no indication whatsoever on the demographics of the respondents.

samples or Mechanical Turk[7] samples which have made the cost of experiments so low[8]. As a consequence, researchers can just run an experiment, examine the results, and decide after the fact whether they will consider it as a study appearing in the article (when the results support the argument) or they will mention it briefly as a pre-test or even discreetly discard it (when the results do not support the argument). I propose, at least in the case of experiments using student labs or Mechanical Turk or similar online panels, to develop norms linked to sample sizes. We could agree that, say, any study up to n = 15 per cell should be considered as a pilot study: a researcher could do an unlimited number of those but they could not appear as experiments in the paper; and that experiments should have a sample size of at least 30 per cell and should all be reported in the paper, even if they do not provide the expected results. Of course, this proposition does not apply to populations for which samples are costly or very hard to get, e.g. for an f-MRI (functional Magnetic Resonance Imagery) study or an experiment run on CEO's or on subjects above 90 years old; nor to cases where the population of interest is very small and the sample covers all or most of it, e.g. all M.D.'s with a specific qualification.

2.3 Avoid "best of" and "verification bias" tactics

It is unacceptable for researchers to fabricate full data sets, e.g. by typing in numbers in an excel file while pretending they were obtained from experiments, or by having some simulation software create a multivariate data set that has pre-defined relationships across variables and pretending they were survey data. The Stapel and Smeesters scandals at least (at last?) should

[7] The Amazon Mechanical Turk allows, among other things, to run social science experiments online. The choice of the name "Mechanical Turk" is strange since the original 18th century "Mechanical Turk" was a fraud, in that the machine that apparently played chess well was in fact only hiding a midget-size human player.

[8] According to Wikipedia (2013), "the cost of MTurk [is] considerably lower than other means of conducting surveys, with workers willing to complete tasks for less than half the US minimum wage."

convince colleagues who could be tempted by such forgeries that there is a high risk that they will be detected at the review stage or later due to logical inconsistencies or implausible patterns in the results or in the fabricated data set.

But we should not limit our concerns to such extreme cases of complete forgery. Respect for the data should lead us also to avoid several forms of unethical behavior which hide or manipulate only part of the collected data. There is a symmetrical lack of respect for the data between forgery, creating observations that never existed, and hiding or manipulating actual observations: In both cases, there is a discrepancy between actual data and analyzed data. The joint report of the three Committees investigating the Stapel case (Levelt Committee et al., 2012) provide an impressive series of examples.

a) Multiplying statistical tests to increase the chances that at least one of them ends up significant, and reporting selectively this significant test. This includes collecting multiple measures of dependent or mediating variables and reporting only one or a subset or a well-chosen combination of them; collecting data on multiple experimental conditions, multiple manipulations, multiple moderators, multiple product categories, and reporting only some of them.

b) Checking, after the collection of each observation, whether the result is "significant" (at 5%) and stopping immediately data collection when this is the case, for fear that the result would no longer be significant after the next observations.

c) Comparing the experimental group from one experiment with the control group of another experiment without mentioning it, because the control group of the first experiment did not produce the desired contrast.

d) Removing certain experimental conditions or subgroups of respondents once the results are known.

e) Merging data from multiple experiments without mentioning it, to increase the number of subjects to arrive at significant results. Of course, this should not be confused with a meta-analytic approach, which clearly acknowledges that it combines results from multiple studies.

f) Reporting reliabilities in a misleading manner (unreported values, erroneous values, values computed on subsamples, different ad hoc selection of items for the same scale in different studies, reference to a standard scale while using a nonstandard form).

g) Erroneously reporting p values.

h) Adding fictitious observations to those that had been actually collected, or selectively modifying the values of certain variables.

i) Replacing missing data by estimated data without mentioning it.

A series of interesting recent papers have discussed such statistical manipulations (e.g., Simmons, Nelson, and Simonsohn 2011 with the telling title "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" or Wagenmakers et al. 2011 who discuss the dangers of an exploratory approach and suggest guidelines for confirmatory research).

Note that the Levelt Committee et al. observe (2012, p.53) that it happens that journal reviewers encourage, or require, some of these behaviors, e.g. removing certain experimental conditions. The first *IJRM* editorial by Jacob Goldenberg and Eitan Muller (2012) provides an excellent example of requirements that authors have to follow in order to ensure integrity, and to allow for better checks by the review team: make their data publicly available, report all measures and conditions, present results with and without covariates, with and without "outliers," explain choice of sample sizes. A *JCR* editorial (Luce, McGill & Peracchio 2012) warns against "best of" tactics and stresses the importance of full present and permanent disclosure: other researchers should be able to understand fully how the research was conducted; data and material should be preserved for possible future re-investigation. Similarly, *Marketing Science* (Desai 2013) now asks authors to submit data sets and estimation codes, to ensure that papers are replicable, while allowing for exceptions decided by the editor in well-specified cases (protected or compiled data sets, big data).

2.4 Use better control by co-authors

Apparently, several co-authors of our recently exposed fraudulent colleague Stapel had satisfied themselves with the story told by that colleague, namely that he had collected, coded, and analyzed (with perfect results!) the data. These co-authors had never checked the original filled questionnaires, and often had never even asked for a copy of the data file or of the computer output. The outburst of the Stapel scandal has demonstrated that all co-authors of an article have a major interest in checking possible ethical problems. If the article ends up withdrawn or retracted, not only will all co-authors suffer from this mechanical reduction in their publication record, but worse also from suspicions, likely unwarranted, derived from their association with a tainted colleague, an "element of stigmatization that may persist long into their future career" (Levelt Committee et al., 2012, p.34).

Data collection and data analysis are something too important to be left blindly to a single data analyst per study. My recommendation is simple: co-authors should organize some systematic reciprocal checking of their data collection, manipulation, and analysis. If there is a possible ethical problem, it is much better to identify it and find a correct solution among authors than to wait for the reviewers to expose it. Common practice by good market research companies provides two simple recommendations: always give precise indications on the data collection: subjects, date, location, etc. so that later control remains possible at the lab or with the online panel; keep in some dropbox the data, programs, and output, so that they remain available for random verifications by any of the co-authors.

In addition, without denying the benefits of a division of labor between co-authors following each person's specific competence, I believe that, independently of possible ethical problems, there are benefits to involve co-authors in data decision at all stages: developing the manipulations and instruments, running pilots, deciding on the possible elimination of observations, on possible transformations, and even analyzing the data. The cost of having two co-authors work in parallel at each of these stages is minimal compared with the potential benefits of avoiding errors and of exploring possible variants in the research process, not to speak of improving the consistency between the conceptual framework, the data collection, the analysis, and the final write-up when those are handled by different co-authors.

3. We should avoid discrepancies between our experimental data sets and the other instances of the phenomenon or process we investigate

Recent years have seen an increase in the frequency of experimental research in consumer behavior: more and more articles based on an experimental approach (they represent 86% of the 84 articles published in vol.39 of *Journal of Consumer Research*), more and more experiments per article (3.6 in 2010 JCR vs. 1.7 in 1990), more and more complex interactions per experiment: 95% articles with 2-way interactions and 33% with 3-way in 2009 JCR vs. 27% and 13% in 1979 JCR (Hamilton 2012). In this section, I discuss four problems that may create a discrepancy between the data in an experiment and the non-experimental instances of the phenomenon and underlying processes we wish to investigate.

These criticisms of experimental research should not blind us to the limits of other data collection methods. Surveys, e.g., also suffer from problems (Rindfleisch et al. 2008): low response rates; disagreements among multiple respondents from a same organization, which cast doubt on the reliability of surveys relying on a single informant; response styles and halo effects, etc. Marketing research relies on multiple forms of data: scanner data from store panels or censuses, online or mobile phone records, social media data, consumer panels, archival data from companies, field experiments and quasi-experiments, time series, etc. Of course, we should respect the data whatever the form of data we use. My focus on experimental research in this article is due to space constraints.

3.1 Look beyond the constraints of experimental myopia

Whatever the pros of experimental research, we should always remember that it remains a fundamentally **myopic** method in the conditions in which it is used in consumer behavior.

3.1.1 Forward and backward myopia.

In our domain, experimental research is almost always myopic in a longitudinal sense: studying short-term reactions to short-term manipulations. Typically, our subjects are exposed for a few minutes to a manipulation, and their reactions are assessed immediately, sometimes after a “filler task” that lasts less than one hour. In the better but rare cases, an effect will be measured one week or a few weeks later. Mechanical Turk and similar online panels raise this to an acme: one reason for their attractiveness is that they can deliver results in a few days, sometimes overnight. In contrast with, say, medicine or educational research, and in contrast with CRM professionals, we almost never apply manipulations and assess results over long periods such as several weeks, months, years (for an exception see Townsend and Liu 2012). This restricts us to short run causes and short run consequences, whereas it should be essential, when considering a phenomenon or a theoretical process, to assess whether it is transitory or durable, and whether the first reaction to a stimulus may trigger a compensating, negative mediating mechanism that could counteract the initial reaction (such as when a higher immediate clicking probability may lead to lower clicking probabilities later, or when an icy road leads to more careful driving).

Consumers live a long life. They are exposed repeatedly to situations and stimuli. Over very long periods, often over their lifetime, they make repeated decisions, develop knowledge of and attitudes toward categories and brands, establish habits, build heuristics. Because of the constraints of our lab experiments, respondents are not given the time they would normally have to do that. Subjects have to develop new solutions to new problems, in contrast to what they do when having for the n^{th} time to choose a toothpaste, a perfume, a car, a radio station, a movie. Decision processes in short-run experiments are not necessarily representative of consumers' long-run decision processes.

Further, these short-term constraints force us to transpose real-life marketing stimuli and consumer responses in a minute form that fits in the ephemeral time frame of an experiment. The study of long-run phenomena that cannot fit in that time frame should not be rejected by editors and therefore abandoned by researchers. What if medicine had studied factors causing cancer in that time frame?

3.1.2 Sideways myopia.

Experimental myopia is also cross-sectional: a number of important consumer behavior problems are beyond the sight of experimental research on student subjects (used, remember, by 90% of the experimental articles published in vol.39 of *JCR*). Is it possible to extend to nonstudent samples results obtained on student samples? I refer the reader to the fundamental reference by Sears (1986) and to the classical meta-analysis by Peterson (2001, p.450) who concludes that caution must be exercised when attempting such an extension and emphasizes the importance of "replicating research based on college student subjects with nonstudent subjects before attempting any generalizations." (See also Henrich, Heine, and Norenzayan 2010; Hooghe, Stolle et al. 2010; Henry 2008 and for a counter-example Voelckner and Sattler 2007). I discuss three specific limitations.

There are limits to what can be manipulated. Consider several major changes in worldwide consumer behavior. In advanced countries, financial difficulties due to unemployment force a growing number of persons to change their consumption behavior. In emerging countries, many persons change their consumption behavior because they benefit from markedly increased economic resources, or because they move from a traditional village to a large city. One can certainly manipulate student subjects' perception of wealth by modifying the scale on which they are asked to report the balance of their bank account, and assess the impact of this manipulation on some hypothetical (or even actual in-lab) consumption behavior. But is there any link between the processes involved there and those entailed by the deep and enduring changes described above? Similarly, the changes in consumption entailed by the recent deep changes in family structures seem hard to study.

Besides, we know that cognitive ability declines consistently and steadily beginning in the 20's and continuing through old age (see, e.g. results on 11 different measures in Park et al 2002). If an experiment on undergraduates reveals a phenomenon and underlying processes, can we generalize these results to consumers with lower abilities, cognitive or otherwise? To consumers outside of the Western world? To poorly educated consumers? Henrich, Heine, and Norenzayan (2010, p.61) caution about samples drawn from "Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies." Their findings suggest that, in a large variety of domains, members of these societies are "among the least representative populations one could find for generalizing about humans."

In a domain for which we have a duty, experimental research on students is especially myopic: the development of public policy recommendations regarding the rapidly growing older population, who suffers from decreases in multiple abilities. This population needs to be protected and helped to take full advantage of consumption opportunities, especially of opportunities offered by new technology. To develop these policy recommendations, we need to study, by experiments and other approaches, older subjects with the complex syndrome that characterizes them.

While we should not stop experiments on students, I recommend that we should not restrict our view of the world to this small, carefully insulated cell and that we should have as a priority to show that our results equally apply to them, as well as to other geographical regions and cultures. What would we say if, in the sports domain, academic experiments on the impact of different training schedules on performance were tested only on subjects aged 60 and above? Do we think coach "practitioners" would conclude that our results also apply to teenagers? The limitations entailed by student subjects should be especially bothersome as, compared to social psychologists who study human behavior in general, we are more interested in marketing in the behavior of specific consumer subpopulations (segments), defined e.g. according to age, education, income, cultural background, etc. From a marketing perspective, exploring the impact of such variables is thus much more important than a simple statistical control of covariates.

We may get inspiration from medical research by using more frequently alternative methodological approaches such as real-life quasi-experiments, meta-analyses, epidemiological studies, econometric analyses of behavioral or survey data, etc. If medicine had followed the current preponderance of experiments in Consumer Behavior research, it would have been impossible to do research on, say, the impact of asbestos or cigarettes on cancer, because exposition to asbestos or cigarettes does not result from a random assignment of respondents.

Finally, on a pedestrian level, when we use Mechanical Turk or a similar online panel, there should be a precise description of the recruitment process, resulting demographics but also controls of such basic qualifications as a good understanding of the questionnaire language.

3.2 Provide replications

As the Roman legal adage says, "one witness is no witness.[9]" Respect for the data implies that we should not accept the existence of a newly described phenomenon or process on the basis of a unique piece of empirical evidence in a single setting with a single operationalization, what we could call a "singleton" (Ioannidis, 2005, 2012; Roediger, 2012). There are unfortunately several examples of controversies about articles describing a new, spectacular effect that a number of colleagues have not been able to replicate. The two most famous cases are the controversy between Bargh et al. (1996) who showed that priming undergraduates with the concept of "old age" led them to walk more slowly when exiting the lab and Doyen et al. (2012) who could not replicate the experiment identically; and the controversy following the article by Bem (2011) in which he showed, among other "psi" processes, a "retroactive facilitation of recall" (pp.419-420). I refer the reader to Yong (2012) for a description of those controversies.

[9] "testis unus testis nullus"

In contrast with medicine, journals in our field have been known not to be keen to publish replications. However, the very low rate of replication in consumer behavior and marketing, as well as in related behavioral sciences, has been a subject of revived concern recently. As indicated by the title of the recent special section on replicability in *Perspectives in Psychological Science* edited by Pashler and Wagenmakers (2012), this expresses a crisis of confidence. Before turning to my personal recommendations, let me refer the reader to that special section, and especially to Iaonnidis (2012), as well as to the useful *in fine* remarks by Albers (2012) on the difficulty of reproducing published results; let me commiserate with Evanschitzky et al. (2011) and Makel et al. (2012) about the extremely low rate of replication research in our domain; and let me applaud to the initiative of this journal, *IJRM*, which has opened a new "Replication Corner" to encourage replication work by making its publication easier (http://www.journals.elsevier.com/international-journal-of-research-in-marketing/news/ijrm-replication-corner-structure-and-process/).

We need two types of replications to answer two different questions. First, to be sure the outcome is not due to random chance, we need to obtain the same results in an identical replication, run in another lab by another researcher. (This requires that the original article contains a full disclosure of the original experiment.) No less an authority than Kahneman (2012) suggests, on the specific case of priming research, to organize a chain of different labs to perform identical replications of their respective experiments.

We also need "conceptual" replications in markedly different settings. As my professor of Linear Algebra at MIT said, "Anyone who describes a general theory and then provides a single application is cheating you" (Abelson, 1974, private communication). We should beware of articles in which complex interactions between several conceptual variables are tested using a single operationalization of each concept and samples from the same subject pool. These replications could differ from the initial study by sampling from another population (as discussed above), relying on different research methods, using different operationalizations of the conceptual variables. It should be especially important to provide a replication for a new phenomenon if it was first evidenced in a sample of students

I propose therefore we should require authors who first describe a new phenomenon on a student or MTurk sample to provide themselves a conceptual replication in a markedly different setting, what we could call a "self-replication." (Of course, this proposition does not apply to cases where samples are costly or very hard to get, see section 2.2.) An additional benefit would be to force authors to delineate more precisely the concepts under study by developing two different operationalizations: the risk of confounds decreases with the number of different operationalizations. Of course, this proposition concurs with Winer's (1999) plea in favor of combining in-lab experimental research with modeling approaches based e.g. on scanner data[10]. It also concurs with Ehrenberg's lifelong insistence on "empirical generalisations" i.e. regularities in results obtained in a variety of settings (Uncles, Ehrenberg, & Hammond, 1995).

Equally, if an article follows up on previous research, respect for the data should lead us to start with a replication of the previous evidence as Study Zero. The replication could be identical if the original article contained a replication in a different setting, but conceptual if the original article had used a single setting, procedure, population, operational definition, and especially if that article is the sole previous evidence of the phenomenon. Obviously the description of the replication in the new article could be very brief or in a web appendix, but I

[10] This is consistent with my earlier plea to improve the validity of marketing models by collecting preliminary qualitative input (Laurent, 2000).

think it is important that the replication has taken place. This is also the interest of the new investigator. There is no point for a researcher to embark in a project that builds on a previous paper if the phenomenon investigated in that paper cannot be replicated.

Of course, exceptions to these requests for replications could be granted by editors: for example, if the problem tackled is urgent or has important policy implications (such as protecting children, older consumers, addicted persons, or car drivers); if the contribution is a new methodology (the data set serves only an example); if the data collection is exceptionally time-consuming or costly (e.g. for a B2B study with multiple informants in many companies and the merging of several data bases); if extensive data bases are involved (e.g. for scanner panels or store censuses); or for confidentiality or privacy reasons..

3.3 Consider the phenomenon and its underlying processes in their entirety

Respect for the data implies that, when researching a phenomenon we should try to consider it in its entirety, rather than in the narrowly limited setting of a specific data set.

It is more convincing to demonstrate a phenomenon or process with multiple operational manipulations of a conceptual independent variable (IV) or moderator (as in Wan & Rucker 2013 who use three different manipulations of high vs. low confidence level) than with the same operational manipulation in all studies of an article. A single operational definition runs a much higher risk of confounding the conceptual variable of interest with other conceptual variables, and the result could be due some other aspect of the manipulation. If an operational definition makes its début in the article, or has been seldom used previously, the author should discuss extensively why it is a valid and reliable implementation of the conceptual variable. This is not necessary when the operational definition has been extensively validated by previous research (e.g. when a moderator is assessed by a test included in Wechsler’s Adult Intelligence test).

While it is good to use multiple measures of the dependent variable, the goal should be to verify that these measures are reasonably correlated and not to pick the measure that best supports the authors' argument once the results are known. All implemented measures should be reported and their links. Again, unless the measures have been extensively used previously, the authors should justify why each one is a valid measure. Equally, while the choice of the mediators and moderators considered in a study is always argued, it would be interesting to add a discussion of mediators and moderators that could have been considered but were not included.

Many experimental articles report results only on the variables they manipulate. To avoid what econometricians call a "specification error" (omitting an explanatory variable, which can lead to biased and inconsistent estimators), we should not restrict ourselves to the variables we are interested in and manipulate; we should rather analyze and report, as much as possible, the effect of all the variables that may impact the dependent variable. Specifically, authors should report the impact of natural (measured) covariates or moderators such as gender, age, level of education, including possible interactions with manipulated variables. It is more efficient statistically to control the impact of such characteristics by including relevant variables in the equations than by considering their impact is part of the random term.

3.4 Report the size and strength of phenomena

Too often in published articles, the existence of an effect is evidenced only by a significance test (most often t or F). This is not enough evidence of its importance. We should also report the strength and size of the effect in absolute terms: how large is the adjusted R^2 or ω^2? how strong is the elasticity? And we should compare them with the effect of covariates. Is the elasticity of increased shelf space larger or smaller than elasticity to local advertising? If priming subjects with old age leads them to walk more slowly, is the impact stronger or weaker than the impact of actual physiological age? Effect strengths and sizes should be reported in a standardized form that allows for comparisons across studies and meta-analyses, e.g. in the form of elasticity coefficients (Albers 2012).

I propose that editors should consider the strength of the demonstrated effects as an important criterion to evaluate a research contribution. Respect for data implies that we learn more by revealing a strong effect than a weak effect, a high elasticity rather than a small elasticity. We should not limit ourselves to evidencing effects qualitatively. Even if an effect has been known qualitatively for a long time, it remains important to learn when it is strong and when it is weak (e.g. the work of Lodish and colleagues on advertising impact 1995 and follow-ups). If we use different operationalizations of a conceptual variable, it is important to know whether they have the same impact and, if not, which one is stronger. Practical implications by governmental authorities and companies will be more likely if an effect is important. And asking for strong effects reduces the risk of accepting an article with false positives.

Respect for the data should lead to be cautious about another tendency, namely for review teams to ask for "non obvious" results. "Hindsight effects" (Slovic and Fischhoff 1977; Bernstein et al. 2011) may lead to reject a paper because the review team think, once the results are known, that the hypothesized relationships and results are "obvious," i.e. correspond to their intuition. The problem is that, without reading the manuscript, their intuition might have been opposite. A feeling of "obviousness" is not data. Of course, it is reasonable to reject an article if hypotheses and results are very similar to what has been demonstrated by previous research and bring few marginal insights. But this should not be confused with cases where, in the absence of previous research, the main argument for rejection is that the results appear, after the fact, "intuitive" or "obvious" to the review team.

4 We should let the data speak out of their pre-defined script

Published research in consumer behavior shows a predominance of the hypothetico-deductive approach, presenting experimental results based on preexisting theory. Respect for the data should make us ready to also accept unexpected results. We should not put ourselves in a circular trap by refusing to accept results if they are not based on a preliminary theory.

This important question has been addressed recently by several senior colleagues. Alba (2012), "in defense of bumbling," states that it is "not illegitimate to engage in abduction … or to acknowledge that an if-then statement can be valuable even if the intervening causal link has not or cannot be identified." Lynch (2012) encourages us to "more often look to the substantive domain as inspiration for our research" and to "start with the consumer phenomenon and then try to explain it rather than always starting with concepts in the literature and then thinking of where they might apply." Lynch et al. (2012) argue that we should be open to non-deductive as well as to deductive routes. Park (2012) argues that we should be open to what he calls "incomplete" or "cute" research, which includes novel and interesting empirical findings, even if the authors have not identified the underlying processes. Albers (2012, p.121) writes that "reviewers should accept more studies that are descriptive and not necessarily test a theory." In this section, I focus briefly on three specific personal examples of a non-deductive approach.

4.1 Revealing multidimensionality

When trying to measure a conceptual variable, the empirical need for a multidimensional measure may reveal that this conceptual variable is multi-faceted. When Jean-Noël Kapferer and I embarked into the development of an empirical measure of involvement (1985), we thought it would be a unidimensional scale. It was through the iterative process between qualitative interviews, items wording, data collections, and analyses that we concluded that it was necessary to measure consumer involvement profiles (pleasure value, symbolic value, risk importance, probability of error) rather than a single unidimensional involvement score. And this led us to identify conceptually multiple facets of involvement.

The current predominant reliance on confirmatory statistical analyses may prevent us from the rich potential benefits of exploration. It is dangerous to try to confirm the possible theoretical relationships involving a construct designated by a word without being sure that the word indeed denotes a single, unidimensional construct. An exploratory factor analysis approach of multiple items supposed to measure a conceptual variable may reveal whether or not there is indeed a unique concept under the name. In reverse, in a confirmatory approach, imposing a priori unidimensionality on a concept may lead us to discard items because they do not fit in the pre-determined schema while they may represent another important facet of the concept. Similarly, it is useful to test alternative manipulations of a conceptual variable with the same manipulation checks. (And of course we should not apply to formative constructs confirmatory methods designed for reflective constructs.)

4.2 Checking for nonlinearities

Respect for the data should lead us to explore possible nonlinearities in relationships between conceptual variables.

Experimental designs most often manipulate each explanatory variable (IVs, moderators) at only two levels, high and low. Results are typically presented in two traditional simple forms: the sample means corresponding to each combination of the manipulated levels (at best with confidence intervals around these averages), or straight lines connecting those sample means. This assumes the effect is monotonic and implicitly assumes it is linear while many other shapes are possible (U-shaped, inverted U-shaped, ceiling effects, etc.). The lack of intermediate manipulations makes it impossible to assess whether these assumptions hold. The authors should therefore either provide strong theoretical arguments for monotonicity and linearity or introduce one or more intermediate levels in the manipulation. In addition, when only two levels are used it impossible to interpolate between them nor to extrapolate beyond them.

Besides, authors seldom justify the choice of the specific levels they implement, whereas other choices would lead perhaps to different conclusions on the direction, size or significance of the effects. Figure 2, an actual example, illustrates these problems. The exploratory variable (abscissa) is age. We collected data on respondents of all ages, rather than contrasting a homogeneous sample of “young” Ss (e.g. 19 to 22) against a homogeneous sample of “old” respondents (e.g. 59 to 62). We could not have detected the nonlinearity with such a two-level design. Besides, if we had contrasted a group of “young” Ss (point A on figure 2) vs. a group of “old” Ss, our conclusions would have depended on our (arbitrary) definition of the “old.” If we had defined “old” as being around 60 year old (point B on figure 2) we would have concluded that age has no impact; if we had defined “old” as being around 85 year old (point C) we would have concluded that age has a very strong effect. In both cases (B or C), the traditional presentation of experimental results would have given the impression of a linear relationship: a flat line from A to B, or a steep line from A to C. None of these two lines would match the actual non-linear relationship apparent on figure 2. This example illustrates how a graphical presentation of detailed data provides deeper information than the simple presentation of the sample means of each condition, and should therefore be encouraged (Smith et al., 2002).

Relatedly, note that the reliance in experiments on just two or three levels per variable can be considered a remnant of ancient, pre-word processing, pre-computer days when experimental material had to be typed and xeroxed, which led to create only a limited number of versions of a questionnaire. Now that experiments are run on high-tech supports, there is not obstacle to using a large number of different values for many experimental variables (e.g., manipulated fake score on a test, time of exposure, stimulus-response interval, degree of contrast, strength of distractor, length of filler task); at the limit, no obstacle to using a different random value for each respondent. Should we restrict ourselves forever to 2 x 2 or 2 x 2 x 2 plans? Should we keep confusing the distinction between factors and covariates with the distinction between categorical and continuous variables?

Finally, even if an effect is linear, the choice of the levels will impact the statistical results: if the random effects are homoskedastic, the strength and significance of the results will increase with the distance between the two levels: pick them close to have a small or no effect, pick them far away to have a strong effect.

4.3 Identifying complex functional forms

When an empirical scatter plot indicates that the relationship between two numerical variables x and y is not linear, respect for the data implies that we should attempt to identify the appropriate functional form. I refer to cases where the full set of observations diverges from a linear relationship, not where only one or a few observations really stand out from an otherwise linear relationship (this can be assessed e.g. by observing whether outliers with similar values of the explanatory variable tend to have similar values of residuals or "deleted residuals[11]"). To try

[11] In addition to traditional regression residuals, standard software packages (e.g., SPSS) offer more complete "influence diagnostics" (Belsley, Kuh, and Welsch 1980) for each observation in the data set: its "deleted residual" (prediction error for one observation if the regression is estimated without that observation), dfbetas (how much the estimated beta coefficient would be changed if this observation were eliminated), etc. As the name indicates, these tools are very useful to detect which observations have a strong influence on the results of the regression. This by no means implies that such observations should be dismissed as this may result from a non-linear relationship.

to identify the functional form, one can use two approaches. If the research project focuses on one specific data set, one can use the approach described by Albers (2012, pp.112-115) to identify a non-linear function of x that fits well the values of y. It has two key components: consider functional forms that respect logical constraints (e.g. include diminishing returns at least above certain levels) and use an exploratory approach based on computing moving averages of y and observing how they change as a function of x ("visual inspection"). If the project rather focuses on a large number of parallel data sets (e.g., coming from different categories or different countries), the goal is to find a single functional form that fits all the data sets (with at most a change in parameters), while respecting logical constraints. Tukey himself, in the same book on "Exploratory Data Analysis" in which he introduces the boxplot, suggests (1977, chapters 5 and 6) to try to identify "re-expressions" of either x or y or both that "straighten out the plot" (to use the title of Tukey's chapter 6), i.e. to find a transformation of the variables so that the relationship between the transformed variables becomes linear. The key concept, here, is that identifying the re-expression that straightens up the plot reveals the underlying structure of the relationship. Obvious examples are exponential growth and decline, for which the plot of y against time (x) is highly non-linear but the plot of the logarithm of y against time is linear, revealing the multiplicative impact of time. Tukey suggests again a variety of possible re-expressions: logarithms, reciprocals, powers, etc. To cite a personal example (Laurent et al., 1995), after observing that, in each of 39 different product categories, there was an almost perfect but highly non-linear relationship between brand recognition scores y ("aided awareness") and brand recall scores x ("spontaneous awareness"); and that these non-linear relationships differed widely across categories, we showed that the relationship could be straightened up in all categories by plotting the log-odds of y against the log-odds of x, the only difference across categories being the value of a single parameter. This revealed that the relationship between brand recall and recognition can be modeled as a Rasch process.

Of course, when we linearize a relationship, we have to be careful about the distribution of residuals (and therefore of random error terms) around both the original non-linear and the linearized relationships: Is it Gaussian? Is it homoskedastic? What are the consequence for estimation? etc.

Overall, respect for the data recommends to let the data speak even in the absence of a pre-defined script as we did in these three examples.

Conclusion

As I said in the introduction, the lessons to be derived from the recent Stapel and Smeesters scandals should not be limited to cautiousness about the sheer forgery of full data sets. We need to be more broadly respectful of our data.

Table 1 summarizes the discussions in the article, in terms of specific actions we should do or not do. Beyond these specific recommendations, let me conclude with three general pleas.

We should respect our own data set, collected following our design and under our control. Data should not be subordinate to researchers. We should not feel free to mutilate our data at will and without reporting it. Observations should be set aside only if we have a good, specific reason, not just because they take extreme values, and this should occur rarely if our experiments and measures have been well designed. Equally, we should not hide completely in our "file drawer" the experiments that did not produce the expected results. Transparency is essential. Conversely, editors should stop requiring fairy-tale reports in which each single experiment works as expected.

We should respect the vast data beyond our narrow experimental data sets and we should be very concerned about the possible discrepancies between them. Is it reasonable to practically exclude alternative, non-experimental methods which have proved so useful in medicine like quasi-experiments, meta-analyses, epidemiological studies, econometric analyses of behavioral or survey data? Is it reasonable to rely almost exclusively on either student samples or samples from Mechanical Turk or similar online panels? Is it reasonable to consider that the vast data outside the lab is totally subordinate to the data collected in that narrow, insulated cell? Never in the history of consumption have we seen worldwide so many diverse groups of consumers going through so many structural changes, due to, e.g., massive income increases in many developing countries, increasing unemployment in many developed countries, increased number of old consumers, worldwide changes in distribution and information networks. Should we remain blind to this and keep looking only at our WEIRD labs?

We should respect the data when they tell us something unexpected. In contrast to Stapel's statement cited in the introduction, when an experiment does not give us what we are looking for, I think we should question the theory as much as the data. Further, without abandoning experiments and the hypothetico-deductive approach, we should be open also to discovering new, unexpected relationships from data collected with the alternative methods mentioned above.

We should respect the data because we should respect the phenomena and underlying processes we study. We should keep a balance between the researcher, the theory, and what is being researched. The latter should not be subordinate. Data should not be subordinate.

Table 1

Respect the data: Recommendations

Do not mutilate your data set

1. Do not hide experiments that did not “work” (did not produce the expected results). Speculate, and if possible develop and test hypotheses on why these experiments did not produce the expected results.

2. Clearly decide, before collecting a data set, whether this is to be a pilot test or an experiment.

3. If you discard alleged “outliers,” mention this briefly in the text and justify it in a web appendix: precise reason and raw value for each discarded observation, description of distributions before and after the removal of outliers (including estimated within-group variance).

4. Do not remove alleged “outliers” for the sole reason that they are located beyond the whiskers in a boxplot (more than 1.5 IQR beyond the quartiles).

5. Use different procedures to handle alleged “outliers” in the same article if the distributions of the variables differ across studies.

6. Check the distribution of the DVs before performing an ANOVA. If it is far from a Gaussian distribution, consider a non-parametric analysis or re-express the variable to bring it as close as possible to a Gaussian distribution before performing an ANOVA.

Avoid discrepancies between your experimental data sets and the other instances of the phenomenon or process we investigate

7. If you reveal a new phenomenon or a new process, provide a conceptual replication yourself with a different operationalization in a different setting.
8. If you follow up on previous research, begin by a replication of previous results.
9. Fully disclose research procedures in order to permit replications by other researchers in a reasonable time..
10. Avoid the “best of” or “verification bias” tactics revealed by the Levelt Committee et al. (2012) and listed in section 2.3 of this article.
11. Use multiple manipulations and measures unless the manipulation or measure has been extensively validated by previous research. If not, justify why they are valid and reliable

operationalizations of the conceptual variables. Report all the measures of a variable and check their convergence.

12. Do not restrict subject pools to undergraduate students or Mechanical Turk samples. In a given article, draw samples from a reasonable variety of populations, rather than from a single student population or a single Mechanical Turk service; if possible, from different geographical regions and cultures.
13. If using Mechanical Turk, control demographic descriptors and subjects' understanding of the language.

Let the data speak out of their pre-defined script

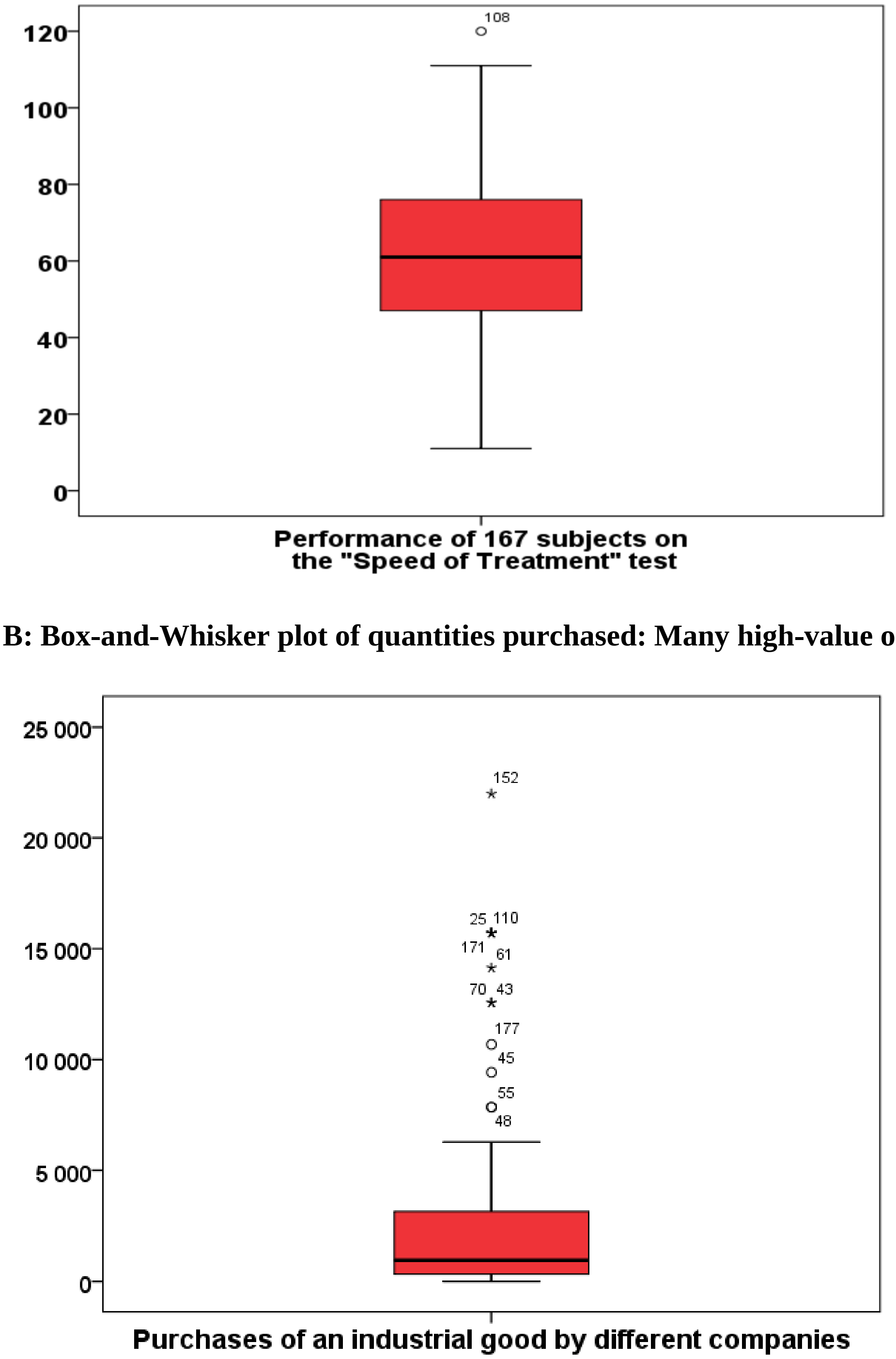
14. Be open to research that does not follow the hypothetico-deductive approach.
15. Be open to accept results even in the absence of antecedent theory, unexpected discoveries and relationships.
16. Do not discard results because they do not fit in your pre-defined schemas.
17. Besides lab experiments, be open to alternative data collection methods, such as real-life quasi-experiments, meta-analyses, epidemiological studies, econometric analyses of behavioral or survey data.
18. Research important topics in consumer behavior (from public policy, theoretical, or managerial points of view) even if these topics cannot be studied through laboratory experiments using students or through Mechanical Turk or similar panels.
19. Be cautious about hindsight bias when tempted to reject "obvious" results.
20. Avoid specification errors by including all available variables (and interactions when appropriate) in the analysis.

Sundry

21. Report effects strength (e.g. adjusted R^2) and size (e.g. elasticity coefficient). Consider the strength of the effects as an important criterion when evaluating an article.
22. Arrange for co-authors to check each other's work at each stage.

Figure 1

Examples of “Box-and-Whiskers” plots



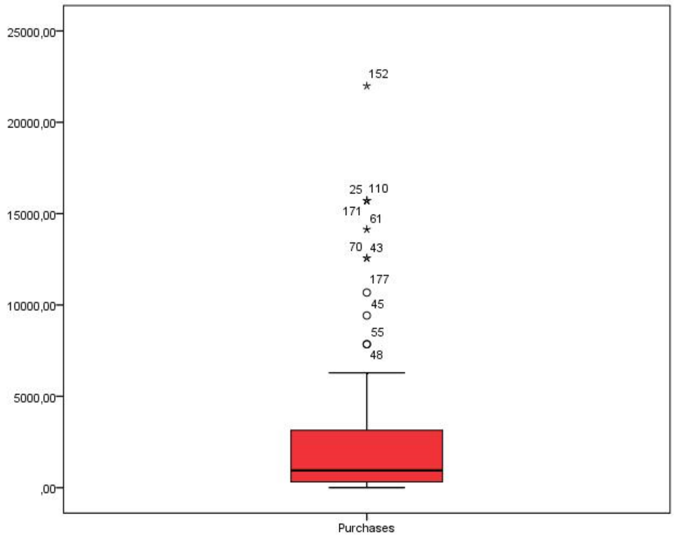
Panel A: Box-and-Whisker plot of subjects’ performance on a psychological test: A single outlier

Panel B: Box-and-Whisker plot of quantities purchased: Many high-value outliers

Figure 1 (continued)

Examples of “Box-and-Whiskers” plots

Panel C: Box-and-Whisker plot quantities purchased AFTER application of the alleged Tukey rule: Still many high-value outliers



Panel D: Box-and-Whisker plot of the natural logarithm of quantities purchased: No high-value outliers

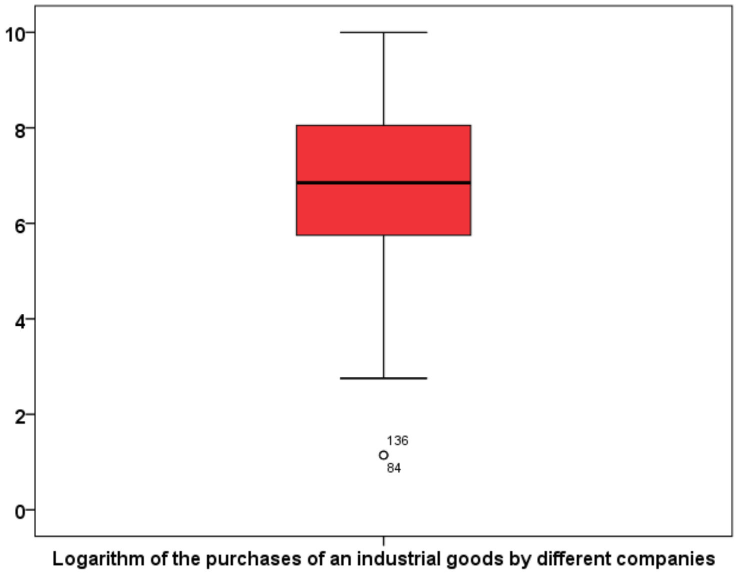
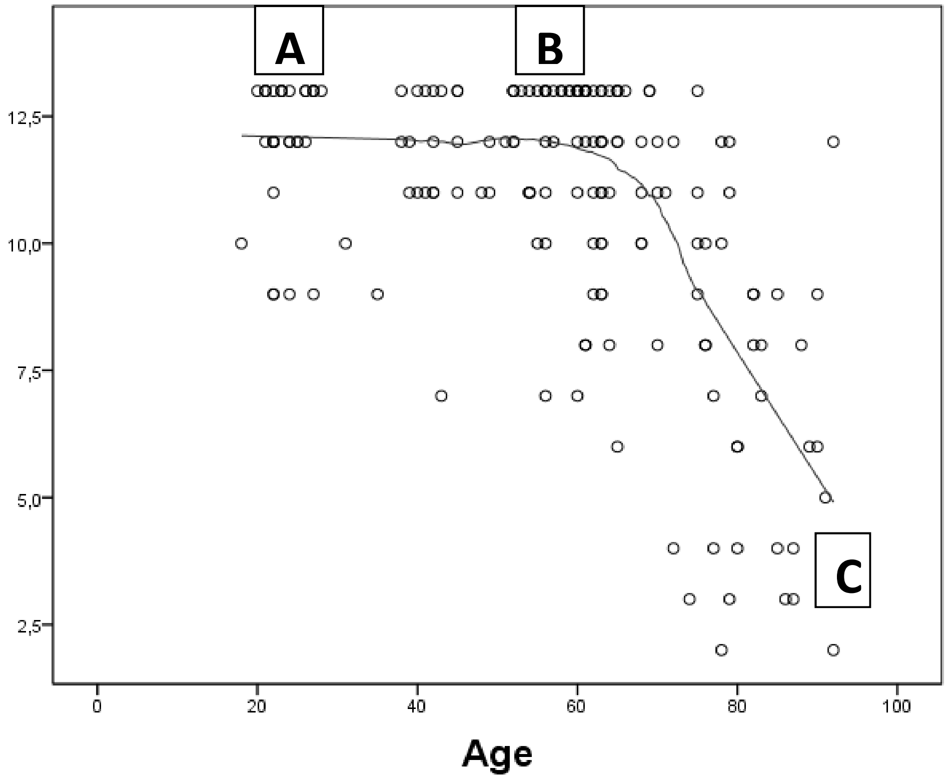


Figure 2

Example of non-linear impact of age on a variable



The real impact of age on the dependent variable is highly non-linear. If we had used a two-level design and simply contrasted a group A of “young” subjects (19 to 22) against a single group of “old” subjects, our conclusions would have depended on our (arbitrary) definition of the “old.” If we had defined “old” as being around 60 year old (point B) we would have concluded that age has no impact; if we had rather defined “old” as being around 85 year old (point C) we would have concluded that age has a very strong effect. In both cases (B or C), the traditional design and presentation of two-level experimental results would have given the impression of a linear relationship: a flat line from A to B, or a steep line from A to C. None of these two lines would match the actual non-linear relationship.

References

Alba, J.W. (2012). In Defense of Bumbling. *Journal of Consumer Research* 38(6), 981-987.

Albers, S. (2012). Optimizable and Implementable Aggregate Response Modeling for Marketing Decision Support. *International Journal of Research in Marketing*, 29(2), 111-122.

Bargh, J.A., Chen, M., Burrows, L. (1996). Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Priming on Action. *Journal of personality and Social Psychology* 71, 230-244.

Bem, D.J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology* 100(3), 407-425.

Bernstein, D.M., Erdfelder, E., Meltzoff, A.N., Peria, W., & Loftus, G.R. (2011). Hindsight Bias From 3 to 95 Years of Age. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(2), 378-391.

Bhattacharjee, Y. (2013). The Mind of a Con Man. *The New York Times* available at http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all&_r=0

Desai (2013). *Marketing Science* Replication and Disclosure Policy. *Marketing Science* 32(1), 1-3.

Doyen, S., Klein, O., Pichon, C.L., & Cleeremans, A. (2012). Behavioral Priming: It's All I the Mind, but Whose Mind? *PLoS ONE* 7, e29081, http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0029081.

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J.S. (2007). Replication Research's Disturbing Trend. *Journal of Business Research*, 60, 411-415.

Francis, G. (2012). The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science* 7: 585-594,doi:10.1177/1745691612459520

Frederick, S. (2012). Overestimating Others' Willingness to Pay. *Journal of Consumer Research*, 38, 1-21.

Galak, J. & Meyvis, T. (2011). The Pain was Greater if it Will Happen Again: The Effect of Anticipated Continuation on Retrospective Discomfort. *Journal of Experimental Psychology: General*, *140*, 63–75.

Galak, J. & Meyvis, T. (2012). You Could Have Just Asked: Reply to Francis (2012). *Perspectives on Psychological Science* 7: 595-596,doi:10.1177/1745691612463079

Goldenberg, J. & Muller, E. (2012). Editorial. *International Journal of Research in Marketing*, http://portal.idc.ac.il/en/main/research/IJRM/Documents/Editorial_Statement_2012k.pdf

Hamilton, R. (2012). Trends and Countertrends in Consumer Research. Presentation at the Haring Symposium, personal communication from Professor Hamilton.

Henrich, J., Heine, S.J., and Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences* 33(2-3), 61-83.

Henry, P.J. (2008). The College Sophomore in the Laboratory Redux: Influences of a Narrow Data Base on Social Psychology's View of the Nature of Prejudice. *Psychological Inquiry* 19(2), 49-71.

Hooghe,M., Stolle, D., Maheo, V.A., and Vissers, S. (2010). Why can't a student be more like an average person? Sampling and attrition effects in social science field and laboratory experiments. *Annals of the American Academy of Political and Social Science*, 628(1), March, 85-96.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoSMed* 2(8): e124, doi :10.1371/journal.pmed.0020124.

Ioannidis, J.P.A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science* 7: 645-654,doi:10.1177/1745691612464056

Kahn, B. (2006). "Moving the Needle:" Can ACR Help Increase our Research Productivity? ACR Presidential Address, available at http://www.acrwebsite.org/research_resources/Kahn%20slides%20acr%20speech%202006.pdf

Kahneman, D. (2012). A proposal to Deal with Questions about Priming Effects. Open letter, available on http://www.decisionsciencenews.com/2012/10/05/kahneman-on-the-storm-of-doubts-surrounding-social-priming-research/

Laurent, G. (2000). Improving the External Validity of Marketing Models: A Plea for More Qualitative Input. *International Journal of Research in Marketing* 17, 177-182.

Laurent, G. & Kapferer, J.N. (1985). Measuring Consumer Involvement Profiles. *Journal of Marketing Research* 22(1), 41-53.

Laurent, G., Kapferer, J.N., & Roussel, F. (1995). The Underlying Structure of Brand Awareness Scores. *Marketing Science*, 14 (3, Part 2 of 2), G170-G179.

Levelt Committee, Noort Committee, & Drenh Committee (2012). Flawed Science: The Fraudulent Researh practices of Social Psychologist Diederik Stapel. Tilburg University.

Lodish, L.M., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M.E. (1995). How Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments. Journal of Marketing Research 32(2), 125-139.

Luce, M.F., McGill, A., & Peracchio, L. (2012). Promoting an environment of Scientific Integrity: Individual and Community Responsibilities. *Journal of Consumer Research* 39(3), iii-viii.

Lynch, J.G., Jr. (2012). Substantive Consumer Research. *Advances in Consumer Research* 38: Association for Consumer Research.

Lynch, J.G., Jr., Alba, J.W., Krishna, A., Morwitz, V.G., & Gürhan-Canli, Z. (2012). Knowledge Creation in Consumer Research: Multiple Routes, Multiple Criteria. *Journal of Consumer Psychology*, xxx.

Makel, M.C., Plucker, J.A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science* 2012(7) 537

Park, C.W. (2012). Two Types of Attractive Research: Cute Research and Beautiful Research. *Journal of Consumer Psychology* 22, 299-302.

Pashler, H. & Wagenmakers, E.J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science* 7: 528-530,doi:10.1177/1745691612465253

Peterson, R.A. 2001). On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-analysis. *Journal of Consumer Research* 28(3), December 2001, 450-461.

Retractionwatch (2012). http://retractionwatch.wordpress.com/2012/06/25/following-investigation-erasmus-social-psychology-professor-retracts-two-studies-resigns/ consulted on 26 May 2013.

Rindfleisch, A., Malter, A.J., Ganesan, S., & Moorman, C. (2008). Cross-Sectionalversus Longitudinal Survey Research: Concepts, Findings, and Guidelines. *Journal of Marketing Research* 45(2), 261-279.

Roediger, H.L., III (2012). Psychology Woes and a Partial Cure: The Value of Replication. *Observer* 25(2), Feb 2012.

Sears, D.O. (1986). College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology* 51(3), 515-530.

Sim, C.H., Gan, F.F., & Chang, T.C. (2005). Outlier Labeling with Boxplot Procedures. Journal of the American Statitical Association 100 (470), 642-652.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22(11), 1359-1366.

Slovic, P. & Fischhoff, B. (1977). On the Psychology of Experimental Surprises. *Journal of Experimental Psychology: Human Perception and Performance* 3(4), 544-551.

Smith, L.D., Best, L.A., Stubbs, A., Archibald, A.B., & Robertson-Nay, R. (2002). Constructing Knowledge: The Role of Graphs and Tables in Hard and Soft Psychology. *American Psychologist* 57 (10), 749-761.

Stapel, D.A. (2000). Moving from fads and fashions to integration: Illustrations from knowledge accessibility research. *European Bulletin of Social Psychology*, 12, 4-27

Townsend, C. & Liu, W. (2012). Is Planning Good for You? The Differential Impact of Planning on Self-Regulation. Journal of Consumer Research 39(4), 688-703.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Uncles, M., Ehrenberg, A.S.C., & Hammond, K. (1995). Patterns of Buyer Behavior: Regularities, Models, and Extensions. *Marketing Science* 14(3, part 2 of 2), G71-G78.

Wan, E.W. & Rucker, D.D. (2013). Confidence and Construal Framing: When Confidence Increases versus Decreases Information Processing. *Journal of Consumer Research* 39(5), 977-992.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., and van der Maas, H.L.J. (2011). Why Psychologists must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* 100(3), 426-432.

Wikipedia (2013). wikipedia.org/wiki/Amazon_Mechanical_Turk. Consulted on 8 May 2013.

Winer, R. (1999). Experimentation in the 21st Century: The Importance of External Validity. *Journal of the Academy of Marketing Science*

Yong, E. (2012). Replication Studies: Bad Copy. *Nature* 485, 298-300.