EMAC Distinguished Marketing Scholar

Respect the Data!

Gilles Laurent

Professor
INSEEC Business School
27, avenue Claude Vellefaux
75010 Paris (France)
glaurent@groupeinseec
(33)786848100

==================================================================================================

ARTICLE INFO

Article history:

First received in January 7, 2013 and was under review for \( \frac{3}{4} \) months.

Area Editor: Berend Wierenga

==================================================================================================

Acknowledgements:

I thank Soenke Albers, Gary Lilien, Berend Wierenga, and Eitan Muller for their extremely useful comments and suggestions on previous versions of this article. I also thank Jean-Louis Chandon, Joe Lajos, Raphaëlle Lambert-Pandraud, Don Lehmann, and Mark Uncles for their suggestions. This article benefited from discussions at two panels, held at HEC Paris in September 2012 and ACR in October 2012. Ultimately though, the opinions and recommendations suggested herein are solely my own responsibility.
Respect the Data!

Abstract

Consumer research suffers from a lack of respect for data. Researchers routinely fail to report full experiments that do not produce expected results and often eliminate alleged “outliers” on the basis of inappropriate rules, leading to biased test reports. Scholars appear to be relying less on non-experimental data, even as the serious limitations of experimental data may create structural discrepancies with the other, non-experimental cases of a phenomenon or process, such that it becomes impossible to study some major consumer phenomena. The lessons from empirical data get accepted only when they can be described as confirming preexisting conceptual frameworks. This article presents an extensive analysis of multiple forms of a lack of respect for the data and proposes some remedies. Overall, data should never play a subordinate part.
1. **Introduction**

Two recent, widely reported scandals involved complete forgeries of data sets by consumer researchers: the Stapel case (investigated in depth by the Levelt Committee, Noort Committee, & Drenh Committee, 2012) and the Smeesters case (Retractionwatch, 2012; Simonsohn, 2013). But the data-related lessons to be derived from these extreme cases should not be limited just to avoiding obvious forgeries. Rather, a lack of respect for data takes various, often more subtle forms but can be just as dangerous. With this article, I attempt to draw the attention of my consumer behavior colleagues and doctoral students to the threat that insufficient respect for data creates for consumer behavior research.

In a citation that rightfully caused great consternation for the Levelt Committee et al. (2012), Stapel (2000, p.6) claimed: “The leeway, 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.” This citation clearly expresses Stapel’s belief in the “primacy of the theory—and therefore the subordinate role of the data?” (Levelt Committee et al., 2012, p. 40). But data are not subordinate. In this introduction, I illustrate the dangers of treating data as subordinate with a few examples. I reserve the in-depth discussions and recommendations for subsequent sections.

1.1 **Do not mutilate the data set**

We must beware unwarranted modifications of a data set after it has been collected. Beyond the (hopefully) rare forgery of observations, I am concerned about data elimination practices that discard certain observations or even full data sets. Consider the “best of” tactic: A common research approach is to run multiple experiments but report only some of them, namely,
those that support the researcher’s argument, or else collect multiple measures of a dependent variable and then report only a selected subset of the results. This selective reporting tactic leads to dangerous biases in reported hypothesis tests. Another dangerous behavior eliminates alleged outliers, before performing analyses of variance (ANOVA) on experimental data, using inappropriate rules that likely produce an upward bias in the F-tests without ensuring that the reduced data set follows a Gaussian distribution. Both “best of” tactics and unwarranted discarding of alleged outliers produce articles that analyze and report mutilated, rather than the original, data sets.

1.2 Avoid discrepancies between experimental data sets and other instances of the phenomenon or process under investigation

The phenomena and underlying processes we research existed prior to our investigations, occur in many other settings than the ones in which we investigate them, and will continue long after we are done with our project. These are truisms that researchers must keep in mind. Considering the dominance of experimental research in consumer research, we need to be concerned about possible discrepancies between the data in our experiments and data relative to non-experimental instances of the phenomena or processes being investigated. Discrepancies might arise because consumer behavior experiments tend to be myopic in a temporal sense, mostly analyzing short-run answers to short-run stimuli, whereas consumers typically develop knowledge and choice strategies over time and multiple, repeated choice occasions. Other discrepancies may occur when the study results come from a narrow laboratory setting or a very specific student population, without any confirmation by replications in other settings, whether undertaken by the author or encouraged by reviewers. Some central aspects of consumers’ behaviors cannot be studied in a lab. Consider, for example, the changes that have occurred as...
Chinese consumers increasingly have moved to large cities and found ways to earn higher incomes.

1.3 Let the data speak outside of their predefined script

If we listen to data only when they speak consistently with some predefined schema or confirm hypotheses that are based on previous research or an a priori conceptual framework, researchers encounter another danger. Instead, we need to listen to the data even (or particularly) when they offer unexpected lessons, especially if we have no solid preexisting theory (which is not infrequent in consumer behavior or, more broadly, in marketing). In such cases, we must remain open to learning new lessons and maintain an abductive approach. For example, if we observe a lack of convergence across different measures of a theoretically unidimensional construct, we may reconsider this unidimensionality rather than eliminate divergent measures. Similarly, we should be cautious about possible nonlinearities and collect data in a manner that reveals them.

1.4. Finding ways to address these threats

The problems I highlight in this article sometimes are driven by specific practices or requests made by review teams; I therefore discuss this specific aspect when appropriate. Of course, it also would be nice to justify my assertions by providing empirical evidence of these wrongdoings. But the guilty seldom report their crimes. Authors rarely mention experiments that they ran but did not include in the article because they did not produce the expected results, unless pressured to do so (see Section 2.2). To present evidence of wrongdoings, exceptional circumstances are necessary, such as the Stapel scandal, in which independent authorities collected and reported a systematic compendium of inappropriate behavior (Levelt Committee et al., 2012) and the author himself acknowledged his questionable practices (Bhattacharjee 2013).
I can cite specific examples of the practices I criticize, but I cannot assess their frequency with statistical methods. John, Loewenstein, and Prelec (2012) offer an interesting estimation of the prevalence of questionable research practices though.

2. **Mutilated data sets**

Three forms of lack of respect for the data occur at the analysis stage: hidden or unwarranted elimination of observations, selective reporting of experiments, and “best of” tactics. These concerns suggest the need for better reciprocal controls by coauthors.

2.1 **Concealed or unwarranted elimination of observations**

It is frequently necessary to “clean” a data set and eliminate certain observations before performing the statistical analysis. However, this process should be transparent and undertaken for appropriate reasons. To provide empirical support for the claims in this section, I reviewed all 72 experimental articles in the most recent volume of *Journal of Consumer Research* (*JCR* Vol. 39, June 2012–April 2013). I counted 53 instances in which the authors reported eliminating observations for a specific reason (for details, see the Web Appendix). The allocation of these instances across articles reveals a troubling pattern: The average frequency is $53/72 = .74$ instances per article, so if they follow a Poisson process, we should expect to observe 34.49 articles with no instance, 25.39 with a single instance, and 12.12 articles with two or more instances. But the pattern in reality is very different, featuring too many papers with no instances (51), too few papers with a single instance (4), and too many with two instances or more (17). These differences are significant ($\chi^2 = 27.89, 2 \text{ df}, p = 4.4 \times 10^{-7}$).

Why do so many articles report no data elimination at all (51 out of 72, or 71%)? Diverse factors might lead to the elimination of observations, often corresponding to unforeseen, incidental deviations from the scheduled data collection protocol (e.g., computer problem, failure
to follow instructions, refusal to perform experimental tasks, missing answers, sickness; see the Web Appendix), so it seems unlikely that all the observations collected for all 51 articles were immediately perfect—especially considering that they generally include at least four experiments each. Rather, at least some of these authors likely eliminated some original observations but did not report how or why. This conclusion is reinforced by the realization that, among articles that report at least one instance of eliminating an observation, more than 80% (17 of 21) reported two instances or more (and up to five in two articles). That is, when authors correctly report that they eliminated observations, they usually find more than one instance to report. Therefore, I conclude that the elimination of data is very common but mostly hidden. This conclusion is reinforced by the contrast between my review of articles that did not report any data elimination and my multiple informal discussions with colleagues who indicated that they considered preliminary data cleaning both routine and necessary (e.g., to detect respondents who did not read the instructions carefully). Thus, the first recommendation is to respect the data by always reporting clearly whether observations have been eliminated and, if so, how many and for what specific reason.

Eliminating an observation is legitimate if it is based on a documented incident in the data collection, as listed previously. But observations also might be eliminated solely because they appear to be “outliers,” in that they show extreme values on the dependent variable. These eliminations reportedly seek to bring the updated data set (i.e., after elimination) close to a normal (Gaussian) distribution and enable ANOVA. Two common rules guide these eliminations: Eliminate observations that are more than 3 standard deviations away from (or above) the sample mean of the original distribution and 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 nonparametric statistics would offer a robust, easy-to-use, alternative approach for data analysis that does not require eliminating observations with extreme values.

The rule that suggests eliminating observations that are more than 1.5 IQR beyond the quartiles relies on a statistical tool introduced by Tukey (1977): the “box-and-whiskers” plot, an example of which appears in Figure 1, Panel A. To visualize the distribution of a numerical variable, Tukey recommends computing its median and quartiles to plot a “box” that contains the middle 50% of the distribution (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 as well as 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,” while observations beyond the outer fences are “far out.” Recent applications change these names slightly, such that in SPSS, the box-and-whiskers plot is called the “boxplot,” outside observations are “outliers,” and far out observations are called “extreme cases.”

Tukey’s (1977, chapter 2) original text makes 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 (of the performance of 167 respondents on a classical “speed of treatment” test) is symmetrical, close to a Gaussian, whereas the distribution in Panel B (of the quantities of an industrial product purchased by different customers) is highly skewed. But at no point does Tukey suggest that observations located outside or far out should be eliminated. In multiple informal discussions, I have heard colleagues

---

1 There is an exception: If there is no observation beyond one of the whiskers defined as above, it moves closer to the median, at the level of the extreme observation.
refer to Tukey to justify eliminating observations that they considered outliers, solely because
they located beyond the whiskers on the dependent variable.

By definition, an outlier is “a statistical observation that is markedly different in value
from the others of the sample” (Merriam-Webster dictionary online). In a boxplot, an
observation located beyond the whiskers also gets referred to as an outlier. But this similarity in
terms does not imply that such a particular observation does not belong in the data set or should
be discarded, and two primary reasons argue against such actions: This procedure strongly biases
ANOVA results, yet it does not bring the updated data set (after elimination) into a normal
(Gaussian) distribution. Consider the two possible horns of this alternative: when the raw data set
is Gaussian and when it is not.

First, if the population of interest is perfectly Gaussian, about 4.3% of the observations
should be beyond the whiskers,2 and there is no reason to eliminate all of them automatically. A
more surprising observation would be if there were no outliers.3 Eliminating such observations
implies that the updated distribution is no longer Gaussian, because it will have lost its two tails.
The estimator of the within-group variance, an essential component of the ANOVA procedure,
thus gets biased downward by more than 20%. Because the resulting estimate is the denominator
of the F-statistics, the suppression of the two tails would lead to overestimates of F by about
25%, changing marginal tests from non-significant to significant.

Second, if the distribution is not Gaussian, it might be lognormal or exponential, as is
frequent in real-world populations. Various examples of naturally occurring lognormal

---

2 In a standardized Gaussian distribution, the quartiles are at -.675 and .675, the IQR is at 1.35, and the whiskers
appear at -2.025 and 2.025. Therefore, about 4.3% of the observations should fall outside the whiskers, without
being outliers in the Gaussian distribution. The outer fences are at -3.375 and 3.375, so only about .074% of
observations should be beyond them.

3 In a sample of n =120 with Gaussian distribution, the probability of having at least one observation beyond -2.025
or 2.025 is 99.5% (i.e., 1 minus 95.7% at the power of 120).
distributions come from across the sciences\footnote{Examples include particle physics, medicine and physiology, economics and sociology (i.e., 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.} \citep{Atchison1963, Johnson1994, Limpert2001}. Firm sizes follow lognormal distributions, for example, so 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 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; the distribution in Panel B reveals a measured skewness of 2.66 (SD = .21), which is significantly above zero. Eliminating all the observations beyond the whiskers in the data from Panel B produces a distribution of the remaining observations (Panel C) that is clearly not Gaussian and still inappropriate for ANOVA (skewness = 1.68, SD = .22). And this study would again underestimate the within-group variance, with a decrease in estimated variance from 15.2 million in the original data to 2.8 million, or more than 80%, prompting a massive upward bias in the estimated F statistics. Thus, in both Gaussian and non-Gaussian cases, removing observations located beyond the whiskers leads to underestimates of within-group variance and overestimates of F statistics in ANOVA.\footnote{Space considerations prevent a similar discussion for other non-Gaussian distributions. However, Sim, Gan, and 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.”}

Another reason not to remove observations located beyond the whiskers is that it leads to absurdities for many distributions. Consider again the lognormal distribution in Figure 1, Panel B. The upper whisker is located at 7,382, so all major purchasers in this market might be eliminated systematically. Even more absurdly, the computed value of the lower whisker is
negative, so a researcher applying this rule would eliminate no small purchasers. Together, the procedure creates a major bias: The mean drops by 42%, from 2,479 in the raw data to 1,446.

Finally, this procedure should be rejected on the basis of an authoritative argument: Tukey (1977), who developed the definition of outliers on the basis of boxplots, never recommended excluding them. John Tukey is not Dr. Guillotin. Instead, he recommended a totally different approach for the “re-expression” of the original variable, which I describe later.

Moving on to the next rule, which recommends eliminating observations that are more than 3 standard deviations above the sample mean of the original distribution, the Gaussian distributions are better, because this procedure removes only the top .13% of the distribution and just slightly biases the sample. But it still can lead to absurd consequences for other distributions. In a lognormal distribution similar to the one in Panel B, the rule eliminates the top 4% of the sample and still heavily biases the mean, which drops by 22% from 2,479 to 1,934. Moreover, the distribution remains non-Gaussian and heavily skewed (2.42, SD = .21), and the estimated variance gets biased downward by 52% (from 15.2 to 7.3 million). The F statistics obtained from this updated sample thus continue to be erroneously overestimated. In addition, the 3 SD rule suffers from circular reasoning, in that the mean and standard deviation used to decide which observations to eliminate get estimated from the original sample, including the extreme observations that they serve to eliminate subsequently.

In informal discussions, I have heard colleagues argue that alleged outliers should be identified and eliminated separately for each experimental condition, a position I consider logically inconsistent. Researchers running an experiment hope to test, using ANOVA, the null hypothesis that the distribution of the dependent variable is the same for all conditions. To treat each condition separately in a preliminary step requires the assumption that the conditions lead to
different distributions of the dependent variable—that is, a rejection of the test of \( H_0 \) that will be performed in the second step. It is logically inconsistent to test a hypothesis using a data set that has been modified already (in the elimination step) by the assumption that this self-same hypothesis is rejected.

In some cases, a review team suggests applying the same procedure regarding outliers to all studies in an article; it happened to me personally recently. I consider this recommendation inappropriate. Different studies in the same article may have unique dependent variables, such as the average of Likert scales (likely a Gaussian distribution), the amount consumers would be willing to pay (likely a lognormal distribution), and a duration (likely an exponential distribution). The procedure to handle alleged outliers cannot be the same in these distinct cases.

Finally, the removal of outliers offers ample opportunities to create different variants of a data set: the original, the set obtained after removing outliers in identified over the full sample, or the set obtained after removing outliers identified within each experimental condition. Researchers could perform ANOVAs on all three sets and simply report the results that fit best with the conclusions they want to reach—a typical example of the “best of” tactics denounced by the Levelt Committee et al. (2012) and that I discuss further in Section 2.3.

In this context, respect for the data suggests several recommendations. First, researchers can eliminate observations if they can justify doing so by specific incidents that affected the data collection. They may not remove observations solely because they indicate extreme values, especially before the researcher assesses the overall shape of the raw distribution, which should always be the first step. If the distribution is close to Gaussian, the raw data should be analyzed without eliminating observations beyond the whiskers (though eliminating observations at more than 3 SD from the mean or beyond the outer fences might be reasonable). If the data are far
from Gaussian, we should consider nonparametric tests as an alternative approach (for a good example, see Frederick, 2012, p.2). Another option would be to follow Tukey’s (1977) recommendation and “re-express” the original variable by transforming it, such that the distribution of the transformed variable is as symmetrical and close to Gaussian as possible. Tukey devotes a full chapter (Chapter 3) to describing possible transformations: logarithms, power functions, reciprocals, and so forth. For a good example, see Frederick (2012 p.17). In Figure 1, Panel D, I provide another example, namely, the boxplot of the logarithm of the variable described in raw form in Panel B. These raw data contained 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 emerged as symmetrical and appropriate for an ANOVA. The percentage of outliers dropped to 1.4%, all located at the bottom end of the distribution, because no more high-value “outliers” existed. Company sizes follow lognormal distributions, so Panel D clearly shows that the major purchasers on this market belong in the distribution and in the analysis.

In practice, whenever researchers eliminate observations, they should include an appendix that describes precisely their argument for the elimination, as well as the full distribution of observations before and after elimination (i.e., values taken by the discarded observations). We should avoid succinct, non-informative statements such as, “For the pencils condition, two outliers were excluded” (Frederick, 2012, p. 5).

2.2 Hidden experiments

Discussions with colleagues suggest that it has become common, when preparing an article, to run many more experiments than appear in the final article. In her ACR Presidential address, Kahn (2006) relied on an estimation that indicated an A-level paper requires running
three times as many experiments as appear in the published version. Simmons, Nelson and Simonsohn (2011) describe how Francis (2012) argued that results presented by Galak and Meyvis (2011), in which seven of eight studies were positive, seemed “suspicious,” because the size of the effect indicated there should have been more negative studies. The latter authors replied that they actually had obtained five additional negative studies but kept them in their “file drawer” rather than reporting them; thus their results were no longer suspicious, taking all 13 studies into account. It appears that many authors believe that journal review teams demand uniformly perfect results—a belief that discourages authors from reporting 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 instead requires mentioning experiments that did not produce the expected results, rather than discarding them. They provide rich information, namely, empirical evidence designed and collected under the full control of the experimenter but contradictory with the predictions. It is the 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, measures, populations, experimental settings, and so forth, this too is informative and should be reported. Thus, an abductive approach may be the key. Researchers should develop hypotheses about why carefully designed, well-controlled experiments fail to produce the expected results, then test them. Reporting all experiments would help avoid the situation described earlier by Simmons et al. (2011). Galak and Meyvis (2012) asserted that Francis (2012) should have asked them about their “file drawer” studies instead of
writing an article criticizing them. Ultimately, it would have been simpler and more straightforward for them to have reported all 13 studies in the initial article.

This problem also might reflect the blurred distinction between pilot tests and full-scale experiments, due to the wide acceptance\(^6\) (or even recommendation!) of the use of student samples or Mechanical Turk (MTurk)\(^7\) samples that lower the cost of experiments.\(^8\) As a consequence, researchers can quickly run an experiment, examine the results, and decide later whether they will include it in their article (if the results support the argument) or mention it briefly as a pretest or even discreetly discard it (if the results do not support the argument). For experiments using student labs or MTurk or similar online panels, I propose we need norms linked to sample sizes. For example, perhaps any study with up to 15 respondents per cell should be a pilot study, such that researchers can conduct an unlimited number of them, but they do not appear as full-fledged experiments. Rather, experiments might require a sample size of at least 30 respondents per cell, and every experiment conducted should be reported, even if they do not provide the expected results. Such propositions would not apply to populations for which samples are costly or very hard to get, such as an f-MRI (functional magnetic resonance imagery) study, an experiment run with CEOs, or an investigation of consumers older than 90 years. Nor does it apply to cases in which the population of interest is very small, such that the sample covers all or most of it (e.g., all doctors with a specific specialization and qualification).

2.3 Avoiding “best of” and “verification bias” tactics

---

\(^6\) Among the 72 experimental articles that appeared in Vol. 39 of *Journal of Consumer Research*, 90% used student samples, 26% used online panels with minimal statistics on respondents’ age and gender, and 14% used online panels with no indication of demographics.

\(^7\) Amazon’s Mechanical Turk supports social science experiments run online. The choice of the name “Mechanical Turk” is rather strange; the original eighteenth-century “Mechanical Turk” was a fraud: a machine that apparently played chess well but actually was hiding a small 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.”
It is totally unacceptable for researchers to fabricate full data sets of any sort, whether by typing numbers into an Excel file and pretending they were obtained from experiments (Bhattacharjee 2013) or by any other method, such as having simulation software create a multivariate data set with predefined relationships across variables and then pretending they are survey data. The Stapel (Levelt Committee et al., 2012) and Smeesters (Simonsohn, 2013, pp.18 seq.) scandals at least (or at last?) should convince colleagues who could be tempted to engage in such forgeries that there is a high risk they will be detected during the review stage or later, due to logical inconsistencies or implausible patterns in the fabricated data set.

But my concerns are not limited to such extreme cases. Respect for the data demands addressing and avoiding several forms of unethical behavior that hide or manipulate part of the collected data. In these cases, whether obvious or subtle, a discrepancy arises between the actual data and the analyzed data. The joint report of the three Committees investigating the Stapel case (Levelt Committee et al., 2012) provides an impressive series of examples:

1. Multiplying statistical tests to increase the chances that at least one is significant, then reporting this significant test selectively. This forgery requires collecting multiple measures of dependent or mediating variables and reporting only one, a subset, or a well-chosen combination. Another method collects data on multiple experimental conditions, manipulations, moderators, or product categories and reports only some of them.

2. Checking, after the collection of each observation, whether the result is significant (at 5%) and then stopping the data collection immediately, for fear that the result might no longer be significant after additional observations.
3. Comparing an experimental group from one experiment with the control group of another experiment, because the control group in the first experiment did not produce the desired contrast.

4. Removing experimental conditions or subgroups of respondents after the results are known.

5. Merging data from multiple experiments without mentioning it, to increase the number of respondents and arrive at significant results. This fraud should not be confused with a meta-analytic approach, which clearly acknowledges that it combines results from multiple studies.

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

7. Erroneous reports of $p$ values.

8. Adding fictitious observations to those actually collected or selectively modifying the values of certain variables.

9. Replacing missing data with estimated data, without mentioning it.

A recent series of interesting articles has discussed such statistical manipulations: Simmons et al. (2011) offer the telling title “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,” and Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) discuss the dangers of an exploratory approach and suggest guidelines for confirmatory research.

Finally, the Levelt Committee et al. (2012, p. 53) observe that journal reviewers encourage or even require some of these behaviors, such as removing certain experimental conditions or subgroups after the results are known.
conditions. An *International Journal of Research in Marketing (IJRM)* editorial by Jacob Goldenberg and Eitan Muller (2012) provides an excellent example of the requirements that authors should follow to ensure integrity and enable better checks by the review team, such as making their data publicly available, reporting all measures and conditions, presenting results with and without covariates and with and without outliers, and explaining their 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, so that other researchers can understand fully how the research was conducted; their data and material also 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 the research is replicable, while allowing for exceptions in well-specified cases (protected or compiled data sets, big data), as determined by the editor.

2.4 Better control by coauthors

Apparently, several of Stapel’s coauthors satisfied themselves with the story told by their colleague, namely, that he had collected, coded, and analyzed the data (with perfect results!). These coauthors never checked the original questionnaires and often did not ask for a copy of the data file or the computer output. The aftermath of the Stapel scandal has demonstrated that all coauthors of an article have an interest in and responsibility to check for possible ethical problems. If an article is withdrawn or retracted, all coauthors suffer from the reduction in their publication record, as well as from suspicions, possibly unwarranted, due to 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 too important to be left to a single data analyst. My recommendation here is simple: Coauthors should organize systematic, reciprocal checking of their data collection, manipulation, and analysis. If a possible problem exists, it is much better to identify it and find a solution among the author team, rather than to wait for reviewers to expose it. Common practice by good market research companies also offers two simple, exemplary recommendations: (1) Always give precise details about the data collection, including respondents, dates, and locations, to enable subsequent controls in the lab or with the online panel and (2) maintain, in some dropbox, the data, programs, and output so that they remain available for random verifications by any coauthors.

Without denying the benefits of a division of labor among coauthors, according to each person’s specific competence, I believe that independent of possible ethical problems, there are benefits of involving coauthors in data decisions at all stages (i.e., developing the manipulations and instruments, running pilot tests, deciding on possible eliminations of observations, choosing possible transformations, even analyzing the data). The cost of having two coauthors work in parallel during each stage is minimal compared with the potential benefits of avoiding errors and exploring possible variants in the research process—not to mention improving the consistency across the conceptual framework, the data collection, the analysis, and the final write-up of the findings.

3. Discrepancies between experimental data sets and instances of the phenomena

In recent years, we have seen an increase in the frequency of experimental research in the consumer behavior field, with more articles using experimental approaches (i.e., 86% of the 84 articles published in Vol. 39 of *JCR*), more experiments per article (in *JCR*, 3.6 in 2010 vs. 1.7 in 1990), and more complex interactions per experiment (95% articles with two-way interactions
and 33% with three-way in 2009 vs. 27% and 13%, respectively, in 1979 in JCR; Hamilton 2012). In this section, I discuss four potential discrepancies between the data in an experiment and non-experimental instances of the phenomenon or underlying processes.

3.1 Constraints of experimental myopia

Whatever the positive benefits of experimental research, it remains a fundamentally myopic method, as used for consumer behavior studies.

3.1.1 Forward and backward myopia

Experimental research is almost always myopic in a longitudinal sense, because consumer behavior researchers study short-term reactions to short-term manipulations. That is, in most cases, respondents are exposed for a few minutes to a manipulation, then provide immediate reactions—sometimes after a “filler task” that lasts less than one hour. In better but rare cases, an effect might be measured a week or a month later. Mechanical Turk and similar online panels intensify this concern, in that one of the reasons for their attractiveness is their ability to deliver results in a few days, sometimes overnight. In contrast with medical or educational research, as well as unlike customer relationship management professionals, as consumer behavior researchers, 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). Thus we remain restricted to short-run causes and short-run consequences. Yet it is essential, when considering a phenomenon or theoretical process, to assess whether it is transitory or durable, and whether the first reaction to a stimulus triggers a subsequent, compensating, negative mediating mechanism that might counteract the initial reaction (e.g., a higher immediate clicking probability leads to lower clicking probabilities later; an icy road prompts more careful driving).
Consumers are long-lived. They receive repeated exposures to situations and stimuli. Over very long periods, often their entire lifetimes, they make repeated decisions, develop knowledge of and attitudes toward categories and brands, establish habits, and build heuristics. With the constraints of lab experiments, respondents do not have the time they normally would take. Instead, they have to develop new solutions to new problems immediately, in contrast with their actions when choosing, for the \( n^{th} \) time, a brand of toothpaste, a perfume, a car, a radio station, or a movie. Decision processes in short-run experiments thus are not necessarily representative of consumers’ long-run decision processes.

Furthermore, short-term constraints force researchers to transpose real-life marketing stimuli and consumer responses into a miniaturized form that fits in the ephemeral time frame of an experiment. Studying long-run phenomena that cannot fit in such time frame should not be rejected by editors or abandoned by researchers. What if medical researchers studied the factors that cause cancer only in a short time frame?

3.1.2 Lateral myopia

Experimental myopia is also lateral: Important consumer behavior problems simply are beyond the reach of experimental research using student subjects (which, as I noted previously, were used by 90% of the experimental articles published in Vol. 39 of \( JCR \)). Is it possible to extend results obtained with student samples to other samples? In a fundamental reference by Sears (1986) and a classical meta-analysis by Peterson (2001, p. 450), the recommendation is for caution when attempting such an extension, with an emphasis on the importance of “replicating research based on college student subjects with nonstudent subjects before attempting any generalizations” (see also Henrich, Heine, & Norenzayan 2010; Henry, 2008; Hooghe, Stolle,
Maheo, & Vissers, 2010; and for a good example of non-student samples Völckner & Sattler, 2007). Three specific limitations seem critical.

First, there are limits to what can be manipulated. Consider some major changes in consumer behavior worldwide. In advanced countries, financial difficulties due to substantial unemployment rates force changes to consumption; in emerging countries, many consumers change their behavior when they benefit from markedly increased economic resources or move from a traditional village to a large city. Researchers might manipulate student subjects’ perceptions of wealth, by modifying the scale on which they report the balance of their bank account, and then assess the impact of this manipulation on some hypothetical (or even actual laboratory-based) consumption behavior. But is there any link between the processes the students undertake and those imposed by deep and enduring life changes? Similarly, changes in consumption entailed by deep shifts in family structures seem hard to study.

Second, we know that a continuous, regular decline occurs for processing-intensive tasks (e.g., speed of processing, working memory, and long-term memory) beginning when people are in their 20s and continuing through old age, whereas verbal knowledge increases across the life span (Park et al., 2002). If an experiment with 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 et al. (2010, p. 61) caution about depending too much on samples drawn from “Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies,” because in many domains, the members of these societies are “among the least representative populations one could find for generalizing about humans.”
I do not mean to suggest a ban on experiments with students. Instead, I recommend that we take care not to restrict our view of the world to this small, carefully insulated cell. A research priority should be to show that results equally apply to them as they do to people in other geographical regions and cultures. Imagine the obverse: That sports researchers conducted experiments on the impact of different training schedules on performance only with participants aged 60 years and above? Would coaching practitioners conclude that the results probably apply to teenagers? The limitations entailed by student subjects should be especially bothersome for marketing researchers who, compared with social psychologists who study human behavior in general, are more interested in the behavior of specific consumer subpopulations (segments), defined according to age, education, income, cultural background, and so forth. From a marketing perspective, exploring the impact of such variables is much more important than imposing a simple statistical control of covariates.

Furthermore, experimental research on students is especially myopic in a domain with intense implications: the development of public policy recommendations for the rapidly growing elderly population, who suffer from decreases in multiple abilities. This population needs protection and assistance to take full advantage of consumption opportunities, especially those offered by new technology. To develop such policy recommendations, we need to study, using experiments and other approaches, older subjects who experience the complex syndromes that characterize this cohort.

Inspiration is available in the medical research field; we should use more alternative methodological approaches, such as real-life quasi-experiments, meta-analyses, epidemiological studies, econometric analyses of behavioral or survey data, and so on. If medicine had adopted the preponderance of experiments that appear in consumer behavior research, it would have been
impossible to determine the impact of asbestos or cigarettes on cancer, because exposures to these carcinogens does not result from a random assignment of respondents (nor, returning to my first point in this section, do the effects arise after brief exposures).

Third, on a pedestrian level, when researchers use MTurk or a similar online panel, they should offer a precise description of the recruitment process, resulting demographics, and controls on basic qualifications, such as a good understanding of the questionnaire language.

3.2 Replications

As a Roman legal adage cautions, “one witness is no witness.” Respect for the data means we cannot accept a newly described phenomenon or process on the basis of a single, unique piece of empirical evidence in a specific setting with one operationalization—that is, a “singleton” (Ioannidis, 2005, 2012; Roediger, 2012). Yet several controversial examples come to mind, in which articles describe a new, spectacular effect that proves hard to replicate. In two famous cases, Bargh, Chen, and Burrows (1996) showed that priming undergraduates with the concept of “old age” led them to walk more slowly when exiting the lab, but Doyen, Klein, Pichon, and Cleeremans (2012) could not replicate the experiment identically. Bem (2011) asserted, among other “psi” processes, a “retroactive facilitation of recall” (pp. 419–20). Yong (2012) provides a clear description of these controversies.

In contrast with medical research, journals in our field have not been keen to publish replications, a preference of revived concern recently. As indicated by the title of a recent special section on replicability in Perspectives in Psychological Science (Pashler & Wagenmakers, 2012), it implies a crisis of confidence. Before turning to my recommendations, I note that this special section is insightful—especially Ioannidis’s (2012) contribution, as well as the useful in

---

9 Testis unus testis nullus.
fine remarks by Albers (2012) on the difficulty of reproducing published results; I also commiserate with Evanschitzky, Baumgarth, Hubbard, and Armstrong (2011) and Makel, Plucker, and Hegarty (2012) about the extremely low rate of replication research in our domain and applaud the initiative of IJRM, which opened a new “Replication Corner” to encourage its publication (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 confirm outcomes are not due to random chance, it is necessary to obtain the same results in an identical replication, run in another lab by another researcher. (In turn, the original article must provide a full disclosure of the original experiment.) No less an authority than Kahneman (2012) suggests, for the specific case of priming research, organizing a chain of labs that can perform identical replications of their respective experiments. Second, we need “conceptual” replications in markedly different settings. As my linear algebra professor at MIT once said, “Anyone who describes a general theory and then provides a single application is cheating you” (Abelson, 1974, private communication). Articles that test complex interactions between several conceptual variables using a single operationalization of each concept and samples from the same subject pool are suspect. Replications should differ from the initial study by sampling another population (as discussed previously), relying on different research methods, or using different operationalizations of the conceptual variables. It is especially important to provide a replication for a new phenomenon that has first been discovered among a sample of students.

Therefore, I propose requirements that mandate that authors who first describe a new phenomenon with a student or MTurk sample provide a conceptual replication in a markedly different setting, or a “self-replication.” (This proposition does not apply to cases in which
samples are very costly or hard to get, as outlined in Section 2.2.) As an additional benefit, this requirement would force authors to delineate more precisely the concepts they study by developing two different operationalizations; the risk of confounds decreases with more distinct operationalizations. This proposition concurs with Winer’s (1999) plea to combine lab-based experimental research with modeling approaches based on scanner or other types of data. It also reflects Ehrenberg’s lifelong insistence on “empirical generalizations,” that is, regularities in results obtained in various settings (e.g., Uncles, Ehrenberg, & Hammond, 1995).

If an article follows up on previous research, respect for the data means it should start with a replication of the previous evidence, as “Study 0.” The replication could be identical if the original article contained a replication in a different setting or conceptual if the original article used a single setting, procedure, population, or operational definition—and especially if that prior article offers the sole previous evidence of the phenomenon. The description of the replication in the new article could be very brief or even relegated to a web appendix, but I consider it important that the replication takes place. This recommendation is in the interest of the new investigator: There is no point in embarking on a project that builds on previous work if the phenomenon investigated cannot be replicated.

Exceptions to these requests for replications could be granted at the discretion of editors, such as when the problem being tackled is urgent or has important policy implications (e.g., protecting children, older consumers, addicted persons, or drivers); the contribution features a new methodology (e.g., the data set serves only as an example); the data collection is exceptionally time consuming or costly (e.g., a business-to-business study with multiple

---

10 It also is consistent with a previous plea to improve the validity of marketing models by collecting preliminary qualitative input (Laurent, 2000).
informants in many companies that merges several databases); extensive databases are involved (e.g., scanner panels, store censuses); or because confidentiality or privacy demands it.

3.3. Considering the phenomenon and its underlying processes in their entirety

When researching a phenomenon, it needs to be considered in its entirety, rather than in a narrowly limited setting or for only a specific data set. In particular, it is far more convincing to demonstrate a phenomenon or process with multiple operational manipulations of a conceptual independent variable (IV) or moderator (e.g., Wan & Rucker, 2013, who use three different manipulations of high and low confidence levels) than with the same operational manipulation across all studies in an article. A single operational definition runs a much higher risk of confounding the conceptual variable of interest with other conceptual variables, and the results easily could be due to 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 valid and reliable. Such efforts are less necessary when the operational definition has been extensively validated by previous research (e.g., a moderator assessed by a test included in Wechsler’s Adult Intelligence test).

Multiple measures of the dependent variable are helpful, but the ultimate goal should be to verify that the measures are reasonably correlated, not just pick the measure that best supports the authors’ argument after the results emerge. Implemented measures should be reported, along with their links. Again, unless the measures have been extensively used previously, the authors should justify why each one is valid. Although discussion of the choice of the mediators and moderators contained in a study is common, it also would be interesting to add a discussion of mediators and moderators that might have been but were not included.
Many experimental articles also report the results related only to the variables they manipulate. To avoid what econometricians call a “specification error” (i.e., omitting an explanatory variable, which can lead to biased and inconsistent estimators), researchers should not restrict themselves only to the variables in which they are interested and manipulate; instead, we should analyze and report, as much as possible, the effect of all the variables that might affect the dependent variable. Specifically, authors should report the impact of natural (measured) covariates or moderators such as gender, age, and 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 simply to assume their impact is part of the random term.

3.4 Size and strength of phenomena

In published articles, the existence of an effect is too often evidenced by just a significance test (t or F). This is not enough. Researchers should also report the strength and size of the effect in absolute terms: How large is the adjusted $R^2$ or $\omega^2$? How strong is the elasticity? In addition, the reports should include comparisons with the effect of covariates. Is the elasticity of increased shelf space greater or smaller than the elasticity of 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? Reports of effect strengths and sizes also should be standardized, allowing for comparisons across studies and meta-analyses, such as in the form of elasticity coefficients (Albers, 2012).

At the same time, editors need to 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, or a high elasticity rather than a small
elasticity. We cannot limit ourselves to qualitative evidence. Even if an effect has strong qualitative support, it continues to be important to learn when it is strong and when it is weak (e.g., Lodish et al., 1995, who study advertising effects and their follow-ups). After using different operationalizations of a conceptual variable, it is important to indicate whether they have the same impact or which is stronger. Practical implications for government authorities and companies are more likely when the effect is important. Asking for strong effects also reduces the risk of accepting an article with false positives.

Respect for the data requires caution about another tendency too, namely, the habit of review teams to ask for “non-obvious” results. These hindsight effects (Bernstein, Erdfelder, Meltzoff, Peria, & Loftus, 2011; Slovic & Fischhoff, 1977) could lead to a rejection, because the review team thinks, once the results are known, that the hypothesized relationships and results are “obvious,” in that they correspond to intuition. But without having read the manuscript, their intuition might have been different. A feeling of “obviousness” is not data. It certainly is reasonable to reject an article if its hypotheses and results are very similar to what has been demonstrated by previous research and offer few or marginal insights. But this situation differs from cases in which, absent previous research, the main argument for rejection is that the results appear, after the fact, intuitive or obvious to the review team.

These criticisms of experimental research also should not blind us to the limits of other data collection methods. Surveys suffer from problems (Rindfleisch, Malter, Ganesan, & Moorman, 2008), such as low response rates; disagreements among multiple respondents from the same organization, which cast doubt on the reliability of surveys relying on single informants; or response styles and halo effects. Marketing research relies on multiple forms of data, including scanner data from store panels or censuses, online or mobile phone records,
social media data, consumer panels, archival data from companies, field and quasi-experiments, and time series. As these examples show, we must respect data, whatever form they take. My focus on experimental research here is due to space constraints.

4 Let the data speak, beyond predefined scripts

Published consumer behavior research reveals the predominance of hypothetico-deductive approaches, presenting experimental results based on preexisting theory. Respect for the data should make us ready to accept unexpected results too. We cannot risk the circular trap of refusing to accept results if they are not based on a preliminary theory. This important question already has been addressed by several senior colleagues. Alba (2012, p.985), “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 (2011, pp.1 and 4) also encourages researchers to “more often look to the substantive domain as inspiration for our research” and “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.” According to Lynch, Alba, Krishna, Morwitz, and Gurhan-Canli (2012), we should be open to both non-deductive and deductive routes. Park (2012) also argues for “incomplete” or “cute” research, which features novel, interesting empirical findings, even if the authors have not identified the underlying processes. As Albers (2012, p. 121) writes, “reviewers should accept more studies that are descriptive and not necessarily test a theory.” In this section, I offer three brief, personal examples of a non-deductive approach.

4.1 Revealing multidimensionality
To measure a conceptual variable, the empirical need for a multidimensional measure may reveal a multifaceted conceptual variable. When Jean-Noël Kapferer and I embarked on developing an empirical measure of involvement (1985), we thought it would be a unidimensional scale. Through an iterative process, tacking across qualitative interviews, item wording, data collections, and analyses, we concluded instead 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. This realization led us to identify multiple conceptual facets of involvement.

An overreliance on confirmatory statistical analyses may limit the rich potential benefits of exploration. It is dangerous to try to confirm possible theoretical relationships involving a construct designated by a single word without being sure that that very word denotes a single, unidimensional construct. An exploratory factor analysis approach with multiple items that reportedly measure a conceptual variable may reveal whether there is a unique concept aligned with that name. In a confirmatory approach, imposing a priori unidimensionality on a concept instead may lead to the elimination of items that do not fit the predetermined schema, even though they actually represent an important facet of the concept. Similarly, it is useful to test alternative manipulations of a conceptual variable with the same checks (though we should not apply confirmatory methods designed for reflective constructs to formative constructs).

4.2 Checking for nonlinearities

Respect for the data demands exploring possible nonlinearities in relationships of conceptual variables. Experimental designs often manipulate each explanatory variable (IVs, moderators) at two levels, high and low. The results are presented in two traditional, simple forms: the sample means corresponding to each combination of the manipulated levels
(sometimes with confidence intervals around these averages) or straight lines connecting the sample means. These forms assume the effect is monotonic and implicitly linear, yet many other shapes are possible (e.g., U-shaped, inverted U-shaped, ceilings). The lack of intermediate manipulations makes it impossible to assess whether these assumptions hold. Authors therefore should provide strong theoretical arguments for monotonicity and linearity or introduce one or more intermediate levels in the manipulation. In addition, with only two levels, it becomes impossible to interpolate between them or extrapolate beyond them.

Yet authors seldom justify the choice of the specific levels they implement, even though other choices could lead to different conclusions regarding the direction, size, or significance of 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” (e.g., 19–22 years) against a homogeneous sample of “old” (e.g., 59–62 years) respondents. We could not have detected the nonlinearity with such a two-level design. In contrasting the “young” group (point A in Figure 2) with the “old” group, our conclusions would have depended on our (arbitrary) definition of “old.” If we had defined “old” as 60 years or older (point B in Figure 2), age would not have appeared to exert an impact, but if “old” meant being around 85 years of age (point C), we would have concluded that age had 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. Neither matches the actual nonlinear relationship apparent in Figure 2. As this example illustrates, a graphical presentation of detailed data provides deeper information than the simple presentation of the sample means of each condition and should be encouraged (Smith, Best, Stubbs, Archibald, & Robertson-Nay, 2002).
In a related note, reliance on just two or three levels per variable is likely a remnant of the ancient, pre-word processing, pre-computer days, when experimental material had to be manually typed and copied, which limited the number of versions of a questionnaire in a practical sense. Now that experiments are run with high-tech support, there is no obstacle to using vast numbers of different values for many experimental variables (e.g., manipulated fake scores on a test, time of exposure, stimulus-response interval, degree of contrast, strength of distractor, length of filler task); at the limit, there is no viable obstacle to using a different random value for each respondent. Should we restrict ourselves forever to $2 \times 2$ or $2 \times 2 \times 2$ designs? Doing so would mean 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 levels affects the statistical results: If the random effects are homoskedastic, the strength and significance of the results increases with the distance between the two levels. If the researcher picks close-by levels, they likely have a small or no effect; another researcher who picks levels farther apart should find a stronger 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 the need to identify the appropriate functional form. I refer here to cases in which the full set of observations diverges from a linear relationship, not when one or a few observations truly stand out from an otherwise linear relationship (e.g., observing whether outliers with similar values of the explanatory variable tend to have similar values of residuals or “deleted residuals”\(^{11}\)). To identify the functional form, two

\(^{11}\) In addition to traditional regression residuals, standard software packages (e.g., SPSS) offer more complete “influence diagnostics” (Belsley, Kuh, & Welsch 1980) for each observation in the data set, such as its deleted residual (prediction error for one observation if the regression is estimated without that observation), dfbetas (how
approaches are viable. If the research project focuses on a specific data set, the researcher can use the approach described by Albers (2012, pp.112-115) to identify a nonlinear function of \( x \) that fits well with the values of \( y \). First, consider the functional forms that respect the logical constraints (e.g., include diminishing returns at least above certain levels). Second, use an exploratory approach that computes the moving averages of \( y \) and observes how they change as a function of \( x \) (visual inspection). If the project instead focuses on many, parallel data sets (e.g., different categories, different countries), the goal is to find a single functional form that fits all data sets (allowing at most for a change in parameters) while respecting the logical constraints. Tukey (1977, chapters 5 and 6), in the same influential book I cited previously, not only introduced the boxplot but also suggested ways to try to identify “re-expressions” of either \( x \) or \( y \), or both, that “straighten out the plot” (the title of his Chapter 6) and identify a transformation of the variables that allows the relationship between the transformed variables to become linear. Identifying the re-expression that straightens the plot should reveal the underlying structure of the relationship. Obvious examples are exponential growth and decline, for which the plot of \( y \) against time (\( x \)) is nonlinear, but the plot of the logarithm of \( y \) against time is linear, which indicates the multiplicative impact of time. Tukey again suggests a variety of possible re-expressions: logarithms, reciprocals, powers. In my own research (Laurent, Kapferer, & Roussel, 1995), after observing that in each of 39 different product categories, there was an almost perfect but highly nonlinear relationship between brand recognition scores \( y \) (aided awareness) and brand recall scores \( x \) (spontaneous awareness), but that these nonlinear relationships also differed widely across categories, we determined that the relationship could be straightened in all categories by plotting the log-odds of \( y \) against the log-odds of \( x \). The only difference across much the estimated beta coefficient would be changed if this observation were eliminated), and so on. As the name indicates, these tools can detect which observations have strong influences on the regression results. Such observations should not just be dismissed, because they may result from nonlinear relationships.
categories was the value of a single parameter. Thus, we could model the relationship between brand recall and recognition as a Rasch process.

Linearizing a relationship still demands caution though, especially with regard to the distribution of residuals (and therefore random error terms), around both the original nonlinear and the linearized relationships: Is it Gaussian? Is it homoskedastic? What are the consequences for estimation?

Overall, respect for the data recommends letting the data speak, even in the absence of a predefined script, as in these three examples.

5. Conclusion

As I wrote in the introduction, the lessons to be derived from recent scandals should not be limited to cautiousness about utter forgery. We need to be more broadly respectful of our data. Therefore, in Table 1 I summarize this discussion, in terms of specific actions that I believe researchers, editors, and journals should (or should not) undertake. Beyond these specific recommendations, I conclude with three general pleas.

First, respect our data sets, collected following our design and under our control. Data cannot be subordinate to researchers. We are not free to mutilate our data at will and without reporting our actions. Observations may be set aside only for good, specific reasons—not just because they take extreme values. Such actions also should be rare, assuming our experiments and measures have been well designed. Nor may we hide experiments in file drawers when they do not produce the expected results. Transparency is essential. At the same time, editors and reviewers must stop requiring fairy-tale reports in which each experiment works as expected.
Second, respect vast data, beyond our narrow experimental data sets, and address possible discrepancies among them. Is it reasonable to exclude alternative, non-experimental methods that have proved so useful in medical fields, such as quasi-experiments, meta-analyses, epidemiological studies, and econometric analyses of behavioral or survey data? Is it reasonable to rely almost exclusively on student samples or samples from Mechanical Turk or similar online panels? Is it really reasonable to assume that the vast data available outside the lab are inferior to data collected in that narrow, insulated cell? Never in the history of consumption research have we seen so many diverse groups of consumers undergoing so many structural changes worldwide, due to massive income shifts in both developing and developed countries, substantial increases in the number of older consumers, and worldwide innovations in distribution and information networks. How can we remain blind to these developments and keep looking only at our WEIRD labs?

Third, respect data that tell us something unexpected. In contrast with Stapel’s claim, cited in the introduction, when an experiment does not give us what we are looking for, we need to question the theory as much as the data. Without completely abandoning experiments and the hypothetico-deductive approach, we need to be more open to discovering new, unexpected relationships from data collected with alternative methods.

Ultimately, we should respect the data because we should respect the phenomena and underlying processes we study. We need a balance among the researcher, the theory, and that which is being researched. The latter should never be subordinate, and nor should data ever appear in a subordinate role.
Table 1

Respect the Data: Recommendations

Do not mutilate data sets

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

2. Clearly decide, before collecting data, whether they will serve a pilot test or an experiment.

3. When discarding alleged “outliers,” mention them briefly in the text and justify the choice, perhaps in a web appendix: Give precise reasons and the raw values of each discarded observation, along with a description of the distributions before and after removing these outliers (including estimated within-group variance).

4. Do not remove alleged “outliers” solely because they locate beyond the whiskers in a boxplot (i.e., 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 dependent variables differ across studies.

6. Check the distribution of the dependent variables before performing an ANOVA. If far from Gaussian, consider a non-parametric analysis or re-express the variable to bring it closer to a Gaussian distribution before performing an ANOVA.

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

7. When revealing a new phenomenon or new process, provide a conceptual replication with a different operationalization in a different setting.

8. When following up on previous research, begin with a replication of the previous results.

9. Fully disclose research procedures to enable replications by other researchers.

10. Avoid “best of” or “verification bias” tactics (Levelt Committee et al., 2012), as 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. Otherwise, justify why they are valid and reliable
operationalizations of the conceptual variables. Report all the measures of a variable and check for convergence.

12. Do not restrict subject pools to undergraduate students or Mechanical Turk samples. In any given article, draw samples from a reasonable variety of populations, rather than from a single student population or MTurk service; if possible, draw them from different geographical regions and cultures.

13. If using MTurk, control for and report demographic descriptors and participants’ understanding of the language.

**Let the data speak outside of their predefined 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, or unpredicted relationships.

16. Do not discard results because they do not fit predefined schemas.

17. Beyond lab experiments, be open to alternative data collection methods, such as real-life quasi-experiments, meta-analyses, epidemiological studies, and 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 they cannot be studied in laboratory experiments using students or online 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 analyses.

**Miscellaneous**

21. Report effects strength (e.g., adjusted R²) and size (e.g., elasticity coefficient). Consider the strength of the effects as an important criterion when evaluating an article.

22. Arrange for coauthors to check one another’s work at every stage.
23. Figure 1

Examples of “Box-and-Whisker” Plots

A. Performance on a psychological test: Single outlier

B. Quantities purchased: Many high-value outliers
C. Quantities purchased after application of the alleged Tukey rule: Many Still many high-value outliers

D. Natural logarithm of quantities purchased: No high-value outliers
Notes: The real impact of age on the dependent variable is highly nonlinear, but an arbitrary two-level design contrasting A (“young” participants, 19–22 years of age) with B (“old” participants, 60 years of age) or with C (“old participants, 85 years of age) would reveal different results (i.e., no effect or an effect of age, respectively). Yet it also would have implied linear relationships in both cases (flat line from A to B, steep line from A to C), rather than the actual nonlinear relationship.
References


