J. of the Acad. Mark. Sci.
DOI 10.1007/s11747-017-0528-7

METHODOLOGICAL PAPER

On the significance of statistically insignificant results in consumer behavior experiments

Robert A. Peterson[1] · **U. N. Umesh**[2]

Received: 11 August 2016 / Accepted: 17 March 2017

Abstract Experimentation is the sine qua non of consumer behavior research, and much of what is thought to be known about the behavior of consumers is based on findings from experiments. However, many articles that report consumer behavior experiments contain one or more results that are *significantly insignificant*. That is, one or more experimental results are so unusually weak or minuscule that they are unlikely to have come about by chance. As such, significantly insignificant results can be due to the "failure" of the theory underlying an experiment and/or the flawed design or implementation of an experiment. Consequently, significantly insignificant results have implications for the theories and methodologies employed in consumer behavior experiments, the quality of conclusions drawn from the experiments, and the credibility of the consumer behavior research discipline as a whole.

Keywords Consumer behavior experiments · F-statistics · Insignificant results · Experimental failure · Theory failure

Introduction

Experimentation is the sine qua non of consumer behavior research. To illustrate, in two recent volumes of the *Journal of Consumer Research* (volumes 40 and 41), of the 179 research articles, 151 (or 84%) reported the results of experiments. Moreover, a majority of the articles reporting experimental research results contain multiple experiments.

The prototypical experiment in consumer behavior research consists of (1) specifying (usually null) hypotheses based on some theory, (2) creating an experimental design, (3) implementing the experimental design (including empirical data collection), (4) assessing the hypotheses by means of statistical analysis, and (5) drawing inferences. Statistical analysis commonly involves an analysis of variance (ANOVA) in which a treatment effect—an experimental manipulation—is compared to an estimate of experimental "error" by means of an F-statistic.

The significance, or lack of significance, of a calculated F-statistic at some value of p is the fundamental basis used to test the success or failure of an experimental treatment or manipulation. If the F-statistic is significant, the null hypothesis is deemed to be rejected, and the treatment or manipulation is considered to be successful. If the F-statistic is not significant, typically a researcher mentions in passing that it was not significant, moves on to F-tests of other treatments or manipulations in the experiment, and rarely, if ever, discusses the implications of a non-significant treatment or manipulation. An exception to this common treatment was reported by Duclos et al. (2013, p. 130). Although they acknowledged that their two-way ANOVAs did not reveal any significant main or interactive effects (F-statistics were respectively .64, .05, .08, .14, 2.35, and .00 [some of which were significantly insignificant]), they noted that "While we could have stopped here, we nonetheless proceeded with a follow-up ANCOVA …."

Stated somewhat differently, testing a null hypothesis by means of analysis of variance can result in one of the three outcomes. One outcome is that the F-statistic is "statistically significant" in that it is greater than some theoretical value.

John Hulland served as Area Editor for this article.

✉ Robert A. Peterson
rap@austin.utexas.edu

U. N. Umesh
umesh@wsu.edu

[1] The University of Texas at Austin, Austin, TX 78712, USA

[2] Washington State University, Vancouver, WA 98686, USA

Published online: 29 March 2017

This would lead to not accepting (rejecting) the null hypothesis. A second outcome is that the F-statistic is "statistically nonsignificant" or "ambiguous" in that it is less than some theoretical value such that the null hypothesis cannot be rejected unambiguously. A third outcome is that the F-statistic is so small that it is "significantly insignificant" such that the probability of it occurring by chance (e.g., $p < .05$) is small, even if the null hypothesis of no effect were true. Most consumer behavior experiments focus on only the first two outcomes and ignore the implications of significantly insignificant F-statistics.

Ignoring the third outcome reflects what Greenwald (1975) addressed in his article "Consequences of Prejudice against the Null Hypothesis." In Greenwald's 1975 survey, reviewers and researchers mentioned that if a null hypothesis is not rejected, it would be highly inadvisable to try to publish the results of a study (only 6% recommended doing so), and more studies on the subject would be warranted. In contrast, if a null hypothesis is rejected, the advice was to publish without any further extensions or replications (50% of the surveyed individuals recommended doing so). Hence, it is not surprising that small F-statistics are not considered worthy of comment and are ignored in the literature, whereas when a study reports one or more significant F-statistics, it sometimes gets published and only the significant F-statistics are discussed in the article, with small ("insignificant") F-statistics merely presented as an afterthought or ignored. In other words, there seems to be a bias against the null hypothesis.

Much has been written about the analysis of data derived from experiments, and especially experiments incorporating human subjects. For example, numerous articles and books have been written about the importance of calculating measures of the variance accounted for by treatments, undertaking power analyses, properly interpreting *p*-values, and computing confidence intervals around various point estimates (e.g., Bakker and Wicherts 2011; Steiger 2004). However, there is a lacunae in the literature about the analysis and implications of experimentally derived results that do not lead to rejecting a null hypothesis.

The purpose of the present research is fourfold. The first purpose is to provide a concise and nontechnical context for understanding the concept of statistically significant insignificant results in consumer behavior experiments. To reiterate, a "statistically significant insignificant result" is defined as an observed F-statistic derived from an experiment that is so small that there is only a small probability (e.g., .05) that it occurred due to chance when the null hypothesis of no effect is true. In other words, the observed F-statistic is "significantly insignificant." The second purpose is to present a metric for objectively determining whether an experimental result (i.e., an observed F-statistic) is significantly insignificant. The third purpose of the present research is to document the incidence of significantly insignificant results reported in a consumer behavior journal to illustrate the nature and scope of the phenomenon. The final purpose is to describe possible causes of significantly insignificant results and make recommendations for addressing these causes. In brief, there are two possible causes of significantly insignificant results: "failure" of the theory underlying an experiment and "failure" in the design and/or implementation of an experiment.

The F-statistic

Two essential activities comprise a consumer behavior experiment: manipulating something, usually information (a treatment), and randomly assigning the manipulation (treatment) to research subjects. Evaluating whether an experimental manipulation—a main effect, a simple main effect, or an interaction effect—is "successful" in null hypothesis testing consists of comparing the effect of the treatment or manipulation to the effect of the randomization. This is accomplished through a ratio in which the manipulation or treatment effect is the numerator and the randomization effect is the denominator.

As noted in virtually all statistics textbooks (e.g., Hicks 1964), in a traditional fixed-effects ANOVA this ratio is termed an F-value or F-statistic. The numerator of this ratio is the sum of squares due to the manipulation (treatment) divided by its associated degrees of freedom, or SS_{tr}/df_{tr}, which has a chi-square distribution and is referred to as the mean square due to treatment (MS_{tr}). The denominator of this ratio is the sum of squares due to the randomization ("error") divided by its associated degrees of freedom, or SS_e/df_e, which also has a chi-square distribution and is referred to as the mean square due to error (MS_e). Whereas some treatises refer to the treatment sum of squares as the *between* sum of squares, and the error sum of squares as the *within* sum of squares, the quantities are respectively identical regardless of semantics.

The ratio MS_{tr}/MS_e is an F-statistic that follows an F-probability distribution if the null hypothesis of no treatment effect (and certain statistical assumptions) is correct. The F-statistic is an omnibus, unimodal, non-directional statistic. If the null hypothesis is correct, the expected value of the F-statistic is n/(n-2), where n is the total number of research subjects. If n is "reasonably large," the expected value of the F-statistic is approximately 1.0, and 1.0 is typically used as a heuristic when evaluating, interpreting, and communicating the results of a consumer behavior experiment. For instance, if there are 200 research subjects, the expected value of the F-statistic is 200/198 = 1.01. An F-statistic of 1.0 implies that the treatment effect is equal to the error effect since $MS_{tr} = MS_e$. However, if MS_{tr} is "significantly" larger than MS_e, where significance is determined by comparing the F-statistic with a theoretical value drawn from an F-probability distribution,

the null hypothesis is deemed to be rejected, and the manipulation or treatment is considered a success.

The distribution of theoretical F-values is a two-parameter right-skewed probability distribution based on the ratio of two independent chi-square distributions, with the two parameters being the degrees of freedom associated with the respective chi-square distributions. Figure 1 presents a typical F-probability distribution curve. In the context of consumer behavior experiments, if p is the probability of making a type I error (e.g., $p = .05$), the right-hand shaded portion of the curve ($F_{1\text{-}p}$ with df_{tr} and df_e) would have associated with it the "critical value" of F (F_c) against which the test F-statistic (F_t) would be compared. If $F_t > F_c$, a significant treatment effect is assumed to exist for a given value of p.

The assumptions underlying an F-test are fairly straightforward. Research subjects have to be selected randomly and treatments randomly assigned. Research subjects must be independent and come from a normally distributed population for each treatment. The population variances of the response distributions in each treatment condition must be equal. In general it is best if sample sizes are approximately equal across treatments. Finally, the residuals in each treatment condition should be normally distributed.

Significantly insignificant effects

Perusal of the consumer behavior experimentation literature reveals that while test F-statistics less than 1.0 are common, they are frequently not reported or simply reported as "$F_t < 1.0$" and ignored. However, some $F_t < 1.0$ values may be so small that they fall into the left-hand shaded portion of the theoretical F-distribution curve presented in Fig. 1. If so, they reflect a *significantly insignificant* treatment effect for a given p-value. Technically (e.g., Guenther 1964, p. 19),

$$F_{(1-p),df1,df2} = 1/F_{p,df2,df1}$$

This relationship and the general nature of an F-probability distribution can be demonstrated with an example. The standard F-test to determine whether two population variances, variance one and variance two, are equal is to compute the ratio of the variances. If the two variances are similar, the ratio will be close to 1.0. If the variance of the sample drawn from population one is the numerator and the variance of the sample drawn from population two is the denominator, then very large values of F (i.e., $F >> 1.0$) indicate that population one possesses a significantly larger variance than population two. Conversely, if this ratio is much smaller than 1.0 (i.e., $F << 1.0$), then population two possesses a significantly larger variance than population one. Thus the null hypothesis of equality of variances would be rejected if either the calculated F-statistic is significantly larger than 1.0 or significantly smaller than 1.0. This analogy illustrates that the "inverse F" metric merely reflects a technical extension of the standard two-variance F-test to an ANOVA application.

In brief, very small values of F_t that are often considered nuisance statistics by consumer researchers when undertaking null hypothesis testing may turn out to be *significantly insignificant*. As such, consumer behavior experiments that have insignificant values of F_t in the context of null hypothesis testing but are so small as to be significantly insignificant need to be examined to determine *why* they occurred and *what* their potential implications are (as they are, in fact, statistically insignificant). Thus, for instance, consider the finding of Jin, Huang, and Zhang (2013, p. 721) that their experiment produced two significant main effects "but no interaction effect ($F(1, 215) = 0.01$, NS)." Whereas the absence of a significant interaction effect permits an inference regarding the underlying null hypothesis test, the fact that this effect was significantly insignificant should have been a signal that the theoretical foundation of the experiment as well as its design and implementation require an assessment.

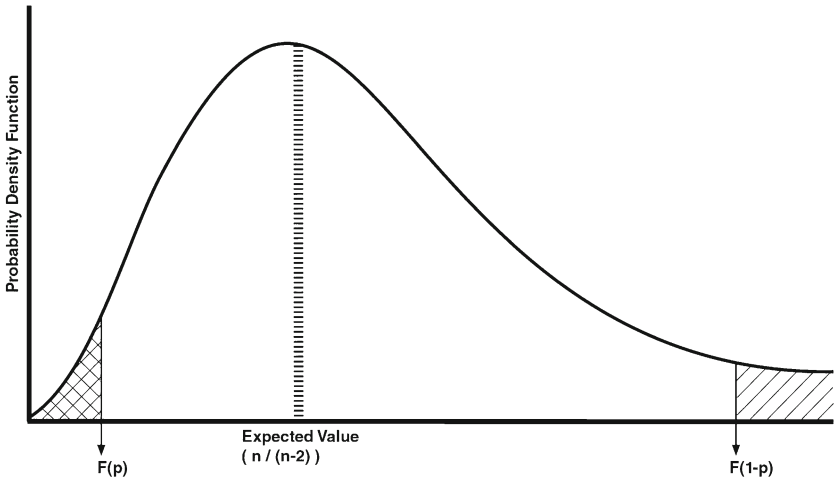
When the inverse of the traditional F-statistic is very (significantly) large, it implies that MS_e is so much larger than MS_{tr} that it simply cannot be attributed to a null model with no treatment effect. Its interpretation is that even by chance one cannot expect such a large (insignificant) error effect, at least relative to the treatment effect. Thus, such an instance raises a "red flag" that suggests potential problems with the theory underlying the experiment and/or the design or implementation of the experiment itself.

Statistically insignificant $F_t < 1.0$ in the literature

To provide an example of the incidence of $F_t < 1.0$ statistics in consumer behavior research experiments, every issue of the *Journal of Consumer Research* (*JCR*) in volumes 1–41 was searched. The investigation was limited to *JCR* because it is the premier academic journal reporting consumer research experiments, and if significantly insignificant results are found in articles appearing therein, it is likely that similar results will be found elsewhere. Only those $F_t < 1.0$ that were contrary to or non-supportive of a hypothesis or expected treatment effect were harvested. $F_t < 1.0$ that were associated with manipulation checks were not harvested, nor were $F_t < 1.0$ values in pretests or pilot studies. Because an F-statistic with $df_{tr} = 1$ and $df_e = m$ (for a two group or cell comparison) is equivalent to a squared t-statistic (i.e., t^2) with df_m, in certain instances reported t-statistics were converted to F-statistics and harvested.

Across the *Journal of Consumer Research* issues searched, an estimated *minimum* of 771 (or 60.8%) of the articles reporting the results of experiments contained *one or more* F_t statistics that were less than 1.0. Indeed, numerous articles reporting experiments contained several F_t statistics that were less than 1.0. For example, Calder and Burnkrant (1977) reported that 29 of the 45 main and interaction effects they

Fig. 1 Theoretical F-distribution



investigated had values less than 1.0, and Schlosser (2003) reported that 26 of her 51 F_t values were less than one.

To reiterate, it is important to recognize that the existence of F_t statistics less than 1.0 in and of itself is not intrinsically concerning. Sampling variation alone would produce, due to chance, some F_t statistics above and others below the expected value of 1.0 under a null hypothesis of no treatment effect. Indeed, Voelkle et al. (2007) found that 51% of the psychology studies they investigated had at least one $F_t < 1.0$. The use of $F_t < 1.0$ as a decision and reporting heuristic is somewhat arbitrary and subjective. Given randomness, sometimes the numerator will be larger than the denominator and at other times the denominator will be larger than the numerator. Only when the less-than-one F_t statistics are so small as to be "significantly small" and likely not "due to chance" should there be cause for alarm.

It is also important to reiterate that the 771 articles reporting consumer behavior experiments are the *minimum* number of articles reporting F_t statistics less than 1.0. It is not possible to determine the exact number of *JCR* articles reporting $F_t < 1.0$ statistics. This is because there are articles that only report that certain effects were "not significant" but do not record actual F-statistics, or report "all F_t statistics were 1.3 or less." Thus, the estimate of 60.8% of *JCR* articles reporting consumer behavior experiments with one or more F_t statistics less than 1.0 must be considered a lower bound on the estimate.

Among the 771 articles in *JCR* reporting experimentally derived F_t statistics less than 1.0, a *minimum* of 235 articles contained F_t statistics that were so small as to be significantly insignificant (using a *p*-value of .05). Thus, a *minimum* of 18.5% of all articles reporting the results of an experiment in *JCR*, and 30.5% of the articles in *JCR* reporting at least one experimentally derived F_t statistic less than 1.0 also reported at least one F_t statistic that was significantly insignificant at $p < .05$. Consider the following examples as illustrative of articles reporting significantly insignificant F_t statistics:

- In their Study 3, White, Argo, and Sengupta (2012, p. 712) concluded that "The main effects for priming ($F(1, 202) = .00$, NS) and social identity threat ($F(1, 202) = .004$, NS) did not reach significance."
- In their experiment 2C, Ma and Roese (2013, pp. 1223–1224) reported statistics that included t-values of 0.12 ($p = .91$) and .00 ($p = 1$), and F-values of .02 ($p = .89$), .00 ($p = .96$), and .08 ($p = .78$).
- In their Study 2A, Di Muro and Murray (2012, p. 579) reported that "neither of the main effects were significant (level of arousal: $F(1, 122) = .46$, $p = 50$; mood valence: $F(1, 122) = 0.003$, $p = .96$)."
- In their Study 4, Norton et al. (2013), p. 250) reported that "there was no significant effect of competitor type in either the ambiguous ($F(2,59) = .02$, $p = .98$) [condition]…[and] no difference in selling prices across the ambiguous and similar seller conditions ($F(1, 62) = 0.01$, $p = .91$."

Analogous to the number of articles reporting F_t statistics less than 1.0, the exact number of articles reporting significantly insignificant F_t statistics will never be known. This is because a sizable number of the articles reporting F_t statistics less than 1.0 merely reported "$F_t < 1.0$" or grouped results together under the rubric of "not significant" or "NS." Thus 30.5% is the lower bound on the percentage of *JCR* articles reporting one or more F_t statistics less than 1.0 *and* having some of these F_t statistics be significantly insignificant, and 18.5% is the lower bound on all *JCR* articles reporting the results of an experiment and having one or more F_t statistics that are significantly insignificant at $p < .05$.

Significantly insignificant F_t statistics occurred for main effects, simple main effects, and interaction effects; research subjects who were college students and non-students; experiments with relatively large and small sample sizes; and laboratory experiments and field experiments. Some articles reported exact F_t statistics that were less than 1.0 as well as

the p-values associated with these F_t statistics. Since $p + (1-p) = 1.0$, in such cases it is possible to immediately identify a small F_t statistic that is significantly insignificant (e.g., if $(1-p)$ is .973, p must be .027).

Reasons for significantly insignificant F_t statistics

Technically, a significantly insignificant F_t statistic means that the treatment or manipulation effect is very small relative to the error effect, or conversely the error effect is very large relative to the treatment or manipulation effect. In other words, a significantly insignificant F_t statistic could be due to issues affecting its numerator and/or its denominator. More generally, as Fig. 2 illustrates, a significantly insignificant F_t statistic could be caused by issues relating to the theory underlying an experiment ("theory failure") or issues relating to the flawed design and/or implementation of an experiment ("experimental failure"), such as experimental design–related causes, statistical model–related causes, and/or research subject–related causes.

Theory failure

Simply stated, a "theory is a systematically related set of statements, including some lawlike generalizations, that is empirically testable" (Rudner 1966, p. 10). The sources of theories range from casual observations to rigorous logic to mathematical models to empirical research. Empirically testing a consumer behavior theory by means of an experiment is most frequently accomplished by testing null (and alternative) hypotheses derived from the theory.

Sometimes theories are incorrect: what is propounded or postulated may simply not be true, may be contrary to reality, or may lack generality. Theory failures occur frequently but are sometimes misunderstood or even ignored; indeed, studies that fail to "prove" a theory are seldom published. Only "high profile" theory failures, such as one proposing a nonexistent link between MMR vaccine and autism (Wakefield et al. 1998) or the lack of generality of the "sleeper effect" (Greenwald et al. 1986) have received widespread recognition.

There is nothing intrinsically wrong with theories failing their testing. The role of experimentation is to empirically test a theory and determine whether the results support the theory. When an observed effect in an experiment is significantly insignificant, this may be because the underlying theory is incorrect. More specifically, if an F_t statistic is zero, this means that its numerator must also be zero, which in turn means that there is no treatment effect whatsoever. Consequently, it is possible to speculate that an F_t statistic of zero would more likely seem to be prima facie evidence of a theoretical failure as opposed to an experimental failure because of the nature of treatments—typically carefully crafted and tested by a researcher prior to the experiment and generally void of nuisance factors such as measurement error. As such, significantly insignificant F_t statistics are consistent with Popper's (1959) notion of falsifiability.

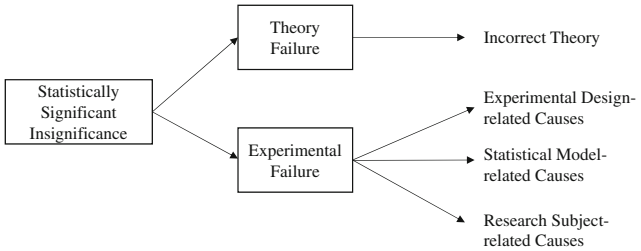


Fig. 2 Reasons for significantly insignificant F-statistics

Experimental failure

An experiment can also produce significantly insignificant results if it was improperly designed and/or implemented. More specifically, experimental failure can result from one or more of three broad (albeit somewhat related) classes of causes: (1) experimental design–related causes, (2) statistical model–related causes, and (3) research subject–related causes. Although these causes can affect the numerator of an F_t statistic, as discussed in the following paragraphs, more likely they artificially inflate its denominator by adding experimental error or noise, leading to a small F_t statistic. Regardless of the cause, any F_t statistic that is significantly insignificant should serve as an indication that something is amiss with an experiment.

Experimental design–related causes Significantly insignificant F_t statistics may be produced if an experimental manipulation is flawed. The manipulation might be so trivial or transparent that it is obvious, or so nuanced or ambiguous that it might not be attended to, comprehended by, or cognitively processed by the research subjects. The latter consequence would seem to potentially be true for three-way, four-way, or higher-level interaction manipulations consisting of numerous (often minor or subtle) cues. Also, a manipulation or even the instructions given to the research subjects might differentially affect the variances of a treatment as well as the means of experimental cells, leading to a violation of the statistical model being applied. Brownie et al. (1990), Bryk and Raudenbush (1988), and Louviere (2001), among others, have written about issues relating to increased response heterogeneity among research subjects resulting from experimental manipulations. While no two research subjects are ever identical and will likely respond a little differently even when in the same treatment condition, response heterogeneity exists when treatments affect not only means but also variances. Response heterogeneity can influence F_t statistics such that "abnormally

low" F_t statistics can occur in some instances due to artificially inflated error terms.

For example, the variance in each treatment condition could be artificially inflated due to non-obvious factors. Large within-treatment variance, with no change in mean effects, will tend to produce low F_t statistics. Hence, it is recommended that the equality of variances across treatment conditions be routinely tested and appropriate action (e.g., rerunning the experiment or transforming the variances) taken if necessary.

It is also possible that an experimental design was underpowered when testing a null hypothesis. Several researchers, including Baroudi and Orlikowski (1989), Maxwell (2004), and Voelkle et al. (2007), have discussed the implications of small sample sizes on the power of tests in experiments as well as the lack of precision in measurement instruments leading to F_t statistics less than one. Baroudi and Orlikowski (1989) in particular have stressed the need for statistical power analyses when interpreting instances in which the phenomenon being investigated does not exist (such as when significantly insignificant results arise due to a theory failure).

Due to the concerns expressed by Cohen (1992) and others, the American Psychological Association (APA) convened a Task Force on Statistical Inference (Wilkinson and Task Force on Statistical Inference 1999). One of the recommendations for statistical power and sample sizes was to "Document the effect sizes, sampling and measurement assumptions, as well as the analytical procedures used in power calculations." Small F_t statistics (and unusually small F_t statistics) are more common when experimental designs lack power. (Low power exists when there are minimal differences between treatment conditions, leading to a small value for the numerator of an F_t statistic and subsequently a small F_t statistic.) Calculating and documenting statistical power facilitates understanding the resulting F_t statistics, particularly when they are less than 1.0.

Both MS_{tr} and MS_e can be affected by measurement error, and approaches to ameliorate such error are commonplace. However, to the extent that measurement error is likely to inflate MS_e more than MS_{tr} (in part due to the manner in which factor levels are crafted, tested, and implemented), the resulting F_t statistic may become significantly insignificant. Moreover, treatment variances might be artificially inflated due to design problems related to method bias (e.g., Podsakoff et al. 2012). Method bias refers to methodological decisions such as the type, context, and wording of a rating scale used or the mode of scale administration employed. If methodological decisions negatively impact the reliability or validity of experimentally derived data, they can lead to significantly insignificant F_t statistics.

Statistical model–related causes There has been a plethora of research on the use of proper statistical models when analyzing experimentally derived data (e.g., Christensen 2003; Glass et al. 1972). Much of this research addresses the impact of omitting sources of systematic variation when analyzing experimental data, the consequences of model assumptions being violated (e.g., correlated observations), or a misspecified model (e.g., including linear but not nonlinear terms), all of which could impact both the numerator and denominator of an F_t statistic and hence decrease its magnitude.

All of these problems have the potential to produce small F_t statistics. However, it would be incorrect to say that such problems tend to decrease F_t statistics in every circumstance. Even so, a significantly insignificant F_t statistic should be viewed as an indicator of model assumptions possibly being violated or an instance of model misspecification.

Research subject–related causes Much has been written about how the characteristics and behaviors of research subjects can lead to demand artifacts that can negatively affect the results of a study (cf. Simonson et al. 2001). Because of the nature of an experiment (treatment manipulation and random assignment of treatments to research subjects), it is likely that research subjects participating in experiments literally construct their responses "on the spot" when answering questions rather than retrieving answers from memory or going through some intensive cognitive process (Peterson 2005). This means that research subject responses are susceptible to a variety of potentially contaminating factors that might contribute to artificially inflating the denominator of an F_t statistic observed in an experiment and hence lead to a significantly insignificant result. Among such contaminating factors are response styles including acquiescence response style, extreme response style, and midpoint response style (Weijters et al. 2008), to mention just a few.

Further, research subjects and even experiment administrators may not pay close attention to experimental stimuli or completely follow instructions when respectively participating in or implementing an experiment. In extreme cases, research subjects may attempt to sabotage an experiment, correctly or incorrectly guess or game the purpose of an experiment (perhaps by talking with other research subjects), or simply speed through an experimental task with minimal attentiveness or commitment. Few of the articles examined in the present research reported a structured debriefing process to discern research subject–related issues that might affect the overall validity of an experiment as well as produce significantly insignificant results.

From a practical standpoint in data collection, violations of an experimental design or model assumptions due to research subject behaviors should be viewed as a distinct possibility given the many implementation issues relating to research subjects. Consider the case where research subjects are asked to rate widely varying stimuli. When research subjects give the same rating for different stimuli, due to inattention, desire

to quickly complete the rating task, or even undermine an experiment, stimuli that should be rated highly end up with negative residuals, and stimuli that should be rated as low end up with positive residuals (i.e., rating everything as a "3" will show a − 1 residual value for something that is expected to be a "4" and a + 1 for something that is expected to be a "2"). If the experiment administrator is not careful, the same research subject may end up in more than one treatment cell, or colluding with friends in the same or a different treatment cell ("you give a '5' and I will give a '1' so on an average the answer will be '3'"). Research subjects might ignore some of the features of a stimulus and focus on just one attribute to complete their task quickly, whereas the model used by the researcher might have many attributes. All of these potential issues regarding research subjects, whether data are collected in a controlled laboratory setting, in a field study, or online, can affect the magnitude of an F_t statistic and thus the interpretation of an experiment's outcomes.

Discussion

Experimentation is the coin of the realm in consumer behavior research, and properly designing and implementing experiments is fundamental for furthering knowledge of consumer behavior. In the first two volumes of the *Journal of Consumer Research*, out of 56 research articles (excluding editorials and such), 15 (or 27%) reported experimental research, whereas, as previously mentioned, 151 articles (or 84%) of the research articles in Volumes 40 and 41 reported experimental research. Thus, both the absolute and relative number of articles reporting experiments have increased dramatically over time and reflect the importance of experimentation in consumer behavior research.

Especially challenging is designing manipulations that rigorously test hypotheses of theoretical interest while not signaling the true intent of an experiment to research subjects who have an abundance of "real world" experience with consumption and other behaviors of interest. In essence, experimental researchers must somehow negate research subjects' real world knowledge to achieve meaningful results. Otherwise, this knowledge may lead to many of the problems discussed earlier: misspecified models, response heterogeneity, non-independence of research subjects and residuals, and so forth. If so, the results of an experiment cannot be interpreted using traditional norms in which small F_t statistics are ignored.

Manipulations must be designed that constitute valid representations of the theoretical constructs being investigated; they cannot simultaneously be representations of other constructs besides those of theoretical interest. Sometimes manipulations may result in confounding constructs in a theory with constructs or variables not in the theory. In such situations experimental tests of the theory may result in significantly insignificant results unless the confounding constructs are identified and controlled. Further, the experimental task and procedure must be communicated with clarity and precision to all participants, research subjects and experiment administrators alike. Accomplishing all of this in a manner that produces an internally valid as well as externally valid test of theoretically interesting hypotheses requires both art and science. The present research offers a metric to systematically and quantitatively diagnose the validity of consumer behavior experiments as well as an approach for identifying possibly contaminating factors that can decrease study reliability and validity.

Because designing and implementing robust consumer behavior experiments is challenging, it should not be surprising that manipulations do not always "work" in the sense that F_t statistics are not always statistically significant. Moreover, it should also not be surprising that F_t statistics less than the expected value of 1.0 (actually n/(n-2)) are commonplace. However, the fact that at least 30.5% of the *JCR* articles containing one or more $F_t < 1.0$ statistics also contain significantly insignificant F_t statistics is somewhat disconcerting. These significantly insignificant F_t statistics may be due to inadequate or incorrect theory, experimental design and/or implementation flaws, inappropriate statistical models used to analyze the experimental data, and/or characteristics of the research subjects.

Close examination of the F_t statistics that are significantly insignificant and the context in which they occur suggests that they may sometimes arise from an incorrect theory or the improper evaluation of a theory: theory failure. This occurs when a significantly insignificant F_t statistic is zero, which means that the numerator of the F_t statistic must also be zero. This "absolute zero" in turn implies that the underlying theory or its hypothesis being tested may be "wrong" to the extent that the theoretical effect being sought may not exist.

At the same time, a flawed experimental design, a lack of control over the experimental process, or the use of misspecified or inappropriate statistical models tends to artificially inflate F_t statistic denominators (e.g., Christensen 2003; Meek et al. 2007), thus potentially leading to small but non-zero F_t statistics: experimental failure. Likewise, permanent or transitory characteristics of individuals comprising the research subjects of an experiment may artificially inflate the denominator of an F_t statistic, again leading to a small but non-zero F_t statistic. Regardless of their nature or source, though, significantly insignificant F_t statistics need to be carefully assessed to monitor and evaluate the integrity of an experiment, especially where there are multiple instances of significantly insignificant F_t statistics produced by the experiment.

Implications

The implications of significantly insignificant F_t statistics in consumer behavior experiments are straightforward: the results of and inferences drawn from such experiments are suspect and call into question the quality and credibility of the experiments. At best, the existence of a significantly insignificant F_t statistic (or statistics) in an experiment raises issues regarding the internal as well as external validity of the experiment and reduces the level of confidence in any inferences emanating from the experiment. At worst the existence of a significantly insignificant F_t statistic (or statistics) implies that the underlying theory guiding the experiment may be wrong or the results are incorrect and the inferences improper. Note that these implications apply not just to a particular experiment containing significantly insignificant F_t statistics; there are also secondary and even tertiary implications if the results serve as the basis of further research or are incorporated into meta-analyses intended to generalize research findings. To the degree that significantly insignificant results are permitted to stand, subsequent research or knowledge based on the results will be tainted.

There is evidence that suggests F_t statistics less than one should lead to a careful review of all aspects of an experiment. Recently a group of 270 researchers (Open Science Collaboration 2015) attempted to directly replicate experimental and correlational research reported in 100 articles appearing in three psychology journals in 2008, *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*. The methodology and findings of each replication attempt were duly recorded and made publicly available. Although (1) there can never be a perfect or an absolute replication, and (2) some of the conclusions were that an original study or treatment effect was "partially replicated," it was possible to classify 90 of the 100 articles as reporting experimental research that was either successfully or unsuccessfully replicated according to criteria set forth in Open Science Collaboration (2015). Ten of the 100 articles reported research that was not experimental or for which there was no clear statement as to whether the replication was deemed successful.

These 90 articles were re-examined to document the magnitude of the reported F_t statistics and determine if there was a relationship between the magnitudes of the F_t statistics and the replication outcomes. Of the 90 articles reporting the results of one or more experiments, 66 (or 73%) reported one or more F_t statistics less than one, a percentage somewhat higher than that observed for *JCR* articles (60.8%). Of those articles reporting the results of an experiment that was not replicated, 87% reported at least one F_t statistic less than one. Of those articles reporting the results of an experiment that was replicated, 51% reported at least one F_t statistic less than one. These percentages are significantly different at $p < .05$. Stated somewhat differently, of the articles reporting experiments wherein at least one F_t statistic was less than one, 73% contained results that were *not replicated*. Simultaneously, of the articles reporting experiments that did not have F_t statistics less than one, 71% contained results that *were replicated*. Thus, it appears that the results of experiments with F_t statistics less than one have a higher probability of not being replicated than the results of experiments wherein there were no F_t statistics less than one.

For a variety of reasons this evidence must be considered anecdotal: conclusions regarding replications tended to be subjective and at times equivocal; replication quality varied; some replication attempts were very specific whereas others might be termed generalization attempts rather than replication attempts; and some articles contained more than one experiment (of which only one might have been the subject of a replication attempt). Indeed, the Open Science Collaboration (2015) replication effort has been criticized by Gilbert et al. (2016), but it has also been defended by Anderson et al. (2016). Even so, the evidence identifies a possible impact of F_t statistics being less than one. (See also, for example, Camerer et al. (2016) for attempted replications of experimental research in economics.)

Observations

The review of experiments reported in *JCR* revealed numerous instances of selective reporting of results, inconsistent statistics, an absence of measures of variance accounted for and confidence intervals, a lack of power analyses, and misinterpretations of statistics and *p*-values. Although there was no attempt to formally document these instances, they appear to be of the same order of magnitude as the findings of Bakker and Wicherts (2011). In a review of 281 psychology articles published in 2008, Bakker and Wicherts found that a reporting or calculation error occurred in 15% of these articles. It is important to study the conditions under which such instances occur and how best to detect them and mitigate their effects.

Recommendations

Certain recommendations for consumer behavior experiments follow from the present research. First, and perhaps most obviously, researchers should report all F_t statistics associated with an experiment, along with their actual *p*-values. Simply reporting $F_t <$ or $>$ than some value is not being transparent. Likewise, simply reporting that a result or F_t statistic is not significant or that $p > .xx$ is not being transparent. For instance, Samper and Schwartz's finding that "There was no main effect of price ($F(1, 107) = 0.00$, NS)…" (2013, p. 1347) arguably should have led them to evaluate both the

theoretical underpinnings of their experiment as well as its methodological characteristics.

Ignoring or glossing over significantly insignificant F_t statistics should be avoided. There should be sufficient information reported for not only the reviewers of a submitted manuscript but also the readers of an article such that they can make informed judgments as to the quality of the reported research and attempt replications if so desired. This information should include the numerators and denominators of all calculated F_t statistics (especially those F_t statistics that are significantly insignificant) in an experiment (perhaps through a standard ANOVA table, although doing so would likely add several pages to an article) so that insights could be gleaned as to the factors contributing to the magnitudes of the reported F_t statistics. It should also include ANOVA tables from earlier attempts at data collection that may not be discussed in the study. Not providing such information precludes even a cursory examination of possible causes of a significantly insignificant F_t statistic and benchmarking against other studies. As such, not providing this information can be viewed as an abrogation of a researcher's responsibility.

Second, significantly insignificant F_t statistics should send a strong signal to researchers that there is likely something amiss with their theory and/or their experiment. When a significantly insignificant F_t statistic is identified, hopefully in a pretest, pilot study, or manipulation check, steps should be taken to determine why it occurred. Actions should include careful review of the underlying theory and close inspection of the experimental design and its implementation, especially the manipulation(s), research subjects and debriefing activities, as well as reanalysis and ancillary analyses of the statistical model and data. For example, techniques such as those of Simonsohn (2013) or Van der Linden and Guo (2008) might be applied to data to estimate whether data fabrication or falsification is an issue.

More generally, individual responses of research subjects should be examined for unusual patterns and outliers. Did a research subject provide the same rating for all scale items (i.e., "straight-lining")? Did a research subject respond 1, 2, 3, 4, and 5 for five items in a row on a five-point scale? Did a research subject provide the same rating (e.g., "2") for all or almost all scales even when some scales were designed to be reverse coded (e.g., a research subject rating a stimulus as "2" for both "do you like" scales and "do you dislike" scales). Some of these behaviors, when repeated over even a subset of research subjects, can cause MS_{tr} to be "unusually small" or MS_e to be "unusually large," resulting in small and potentially significantly insignificant F_t statistics.

Frequently, experimental data are collected in a group context, such as 20 students being exposed to different treatments at the same time in the same room. If so, there is the risk of research subjects influencing each other orally, copying others' responses, clarifying questions with each other, or the like. This may result in observations not being independent and in violation of a statistical model. Where possible, ancillary data should be collected, reported, and analyzed so that individual differences (e.g., gender) or treatment-specific characteristics can be analyzed as covariates or moderator variables.

Third, while it is always good research practice to replicate experimental results prior to arriving at conclusions or submitting a manuscript for publication consideration, it would seem especially imperative to replicate experiments that result in F_t statistics that are significantly insignificant. By implication, replications are less likely to be successful in such cases than when there are no significantly insignificant F_t statistics.

Concluding note

The goal of this article is to alert researchers, journal editors and reviewers, and journal readers to the concept and existence of significantly insignificant results emanating from consumer behavior experiments, and to present this information in a readable and descriptive style so that it is widely applied by all consumer behavior researchers. Although the proffered metric has been alluded to in the consumer behavior and marketing literatures (e.g., Monroe 1976; Peterson and Cagley 1973), for whatever reason consumer behavior researchers do not seem to be aware of the metric or how it can be used to improve consumer behavior research. The fact that nearly one-fifth of the articles that have appeared in a prestigious consumer research journal and that report the findings of experiments contain one or more significantly insignificant results should be of concern. Such results undermine both the quality and credibility of consumer behavior experiments and potentially have theoretical, methodological, reputational, and knowledge implications for the discipline. Therefore, significantly insignificant results in consumer behavior experiments must be publicly acknowledged and addressed. They cannot be simply ignored.

Over time a rich literature on the need for replicating the findings of consumer behavior research has evolved (e.g., Hunter 2001; Lynch et al. 2015; Lehmann and Bengart 2016; Raman 1994). Simultaneously, there have been discussions regarding the use of college students as research subjects in consumer behavior experiments (e.g., Peterson 2001; Peterson and Merunka 2014). Peterson and Merunka (2014) go as far as to recommend that every manuscript that reports empirically based research and is submitted to a top-tier journal must justify the theoretical relevance of the research subjects for the specific research questions. Online experiments have the potential for similar problems as it is difficult to monitor the research subjects or validate their responses. Recently, p-hacking has been identified as a major problem in social science research. In p-hacking researchers

manipulate factors in a research study repeatedly, or repeat an experiment until significantly positive results are obtained (e.g., Simonsohn et al. 2014). Although replications are necessary and research subjects should be scrutinized, the present manuscript offers a systematic approach and metric that can be used to objectively and quantitatively assess the efficacy of any consumer behavior experiment. This approach and metric should be part of every consumer behavior experimentalist's toolkit.

It is important to note, though, that not all consumer behavior experiments contain significantly insignificant results, and only experiments reported in the *Journal of Consumer Research* served as the database in the present study. Thus, it is not possible to generalize beyond these experiments or this journal. Even so, because *JCR* is the premier journal for reporting consumer behavior experiments, there is no reason to believe that different results would be found for other consumer behavior experiments or in other journals reporting consumer behavior experiments. Moreover, given the present results and anecdotal evidence presented here, significantly insignificant F_t statistics likely occur in other types of behavioral experiments.

Acknowledgements The authors would like to express their appreciation to Dwight Merunka and Steven P. Brown for their insights and suggestions.

References

Anderson, C. J., et al. (2016). Response to comment on "estimating the reproducibility of psychological science.". *Science, 351*(6277), 1037–103b.

Bakker, M., & Wicherts, J. M. (2011). The mis(reporting) of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666–678.

Baroudi, J. J., & Orlikowski, W. J. (1989). The problem of statistical power in MIS research. *MIS Quarterly, 13*(1), 87–108.

Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modifying the *t* and ANOVA *F* tests when treatment is expected to increase variability relative to controls. *Biometrics, 46*(1), 259–266.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: a challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396–404.

Calder, B. J., & Burnkrant, R. E. (1977). Interpersonal influence on consumer behavior: an attribution theory approach. *Journal of Consumer Research, 4*(1), 29–38.

Camerer, C., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6277), 1433–1436.

Christensen, R. F. (2003). Significantly insignificant F tests. *The American Statistician, 57*(1), 27–32.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Di Muro, F., & Murray, K. B. (2012). An arousal regulation explanation of mood effects on consumer choice. *Journal of Consumer Research, 39*(3), 574–584.

Duclos, R., Wan, E. W., & Jiang, Y. (2013). Show me the honey! Effects of social exclusion on financial risk-taking. *Journal of Consumer Research, 40*(1), 122–135.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. W. (2016). Comment on "estimating the reproducibility of psychological science.". *Science, 351*(6277), 1037–103a.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1–20.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*(2), 216–229.

Guenther, W. C. (1964). *Analysis of variance*. Englewood Cliffs: Prentice-Hall, Inc.

Hicks, C. R. (1964). *Fundamental concepts in the design of experiments*. New York: Holt, Rinehart and Winston.

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research, 28*(1), 149–158.

Jin, S. E. L., Huang, S.-C., & Zhang, Y. (2013). The unexpected positive impact of fixed structures on goal completion. *Journal of Consumer Research, 40*(4), 711–725.

Lehmann, S., & Bengart, P. (2016). Replications hardly possible: reporting practice in top-tier marketing journals. *Journal of Modeling in Management, 11*(2), 427–445.

Louviere, J. J. (2001). What if consumer experiments impact variances as well as means? Response variability as a behavioral phenomenon. *Journal of Consumer Research, 28*(3), 506–511.

Lynch, J. G., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: in praise of conceptual replications. *International Journal of Research in Marketing, 32*(4), 333–342.

Ma, J., & Roese, N. J. (2013). The countability effect: comparative versus experiential reactions to reward distributions. *Journal of Consumer Research, 39*(6), 1219–1233.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

Meek, G. E., Ozgur, C., & Dunning, K. A. (2007). Small F-ratios: red flags in the linear model. *Journal of Data Science, 5*, 199–215.

Monroe, K. B. (1976). The influence of price differences and brand familiarity on brand preferences. *Journal of Consumer Research, 3*(1), 42–49.

Norton, D. A., Lamberton, C. P., & Naylor, R. W. (2013). The devil you (don't) know: interpersonal ambiguity and inference making in competitive contexts. *Journal of Consumer Research, 40*(2), 239–254.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac41–8.

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), 450–461.

Peterson, R. A. (2005). Response construction in consumer behavior research. *Journal of Business Research, 58*(3), 348–353.

Peterson, R. A., & Cagley, J. W. (1973). The effect of shelf-space upon sales of branded products: a reappraisal. *Journal of Marketing Research, 10*(1), 103–104.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research, 67*(5), 1035–1041.

Podsakoff, P. M., MacKenzie, S., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.

Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.

Raman, K. (1994). Inductive inference and replications: a bayesian perspective. *Journal of Consumer Research, 20*(4), 633–643.

Rudner, R. (1966). *Philosophy of social science*. Englewood Cliffs: Prentice Hall.

Samper, A., & Schwartz, J. A. (2013). Price inferences for sacred versus secular goods: changing the price of medicine influences perceived health risk. *Journal of Consumer Research, 39*(6), 1343–1358.

Schlosser, A. E. (2003). Experiencing products in the virtual world: the role of goal and imagery in influencing attitudes versus purchase intentions. *Journal of Consumer Research, 30*(2), 184–198.

Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(11), 1875–1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547.

Simonson, I., Carmon, Z., Dhar, R., Drolet, A., & Nowlis, S. M. (2001). Consumer research: in search of identity. *Annual Review of Psychology, 52*, 249–275.

Steiger, J. H. (2004). Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182.

Van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384.

Voelkle, M. C., Ackerman, P. L., & Wittmann, W. W. (2007). Effect sizes and F ratios < 1.0: Sense or nonsense? *Methodology, 3*(1), 35–46.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Retracted: ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet, 351*(9103), 637–641.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409–422.

White, K., Argo, J. J., & Sengupta, J. (2012). Dissociative versus associative responses to social identity threat: the role of consumer self-construal. *Journal of Consumer Research, 39*(4), 704–719.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.