Augmenting null hypothesis significance testing in marketing research

Steven A. Taylor
Illinois State University

Myoung Jin Kim
Illinois State University

Chiharu Ishida
Illinois State University

Jamie R. Mulligan
Illinois State University

ABSTRACT

Issues related to null hypothesis significance testing (NHST) are well known in marketing research. In light of recent developments with the American Psychological Association (APA) Publication Manual (2010) we attempt to stimulate discussion concerning the argument for marketing researchers to begin moving more aggressively toward statistical estimation of effect size and confidence intervals to supplement the traditional use of NHST that emphasizes dichotomous, accept-reject (or whether a relationship exists-doesn't exist) outcomes. We review the literature concerning the current issues and known remedies related to NHST, and then present a methodological framework for marketing researchers' consideration to augment current reporting practices. The proposed framework is demonstrated in a study related to the well-known expectancy disconfirmation theory of satisfaction.

Keywords: Null hypothesis significance testing, Effect size, Confidence intervals, Satisfaction, Meta-analysis

INTRODUCTION

Null hypothesis statistical testing (NHST) is arguably the most widely used approach to hypothesis evaluations among behavioral and social scientists despite the many objections raised over several decades (Nickerson 2000; Levine et al. 2008, Bakker et al. 2012; Cummings 2012; Sawyer & Peter 1983). Greenwald et al. (1996) account for the popularity of NHST because it: (1) provides a dichotomous outcome that can be used for decision making, (2) relies on *p* as a common-language translation for a variety of statistics, and (3) relies on *p* as a measure of confidence in the replicability of null hypothesis rejection. Nord (2012 p. 444) echoes the benefit of NHST to academic disciplines because it serves to legitimize social scientific efforts to "be objective." Therefore, it is not surprising that negative results are disappearing from published scientific research (Fanelli 2012).

A review of marketing journals suggests that marketing researchers generally continue to rely on null-based inferential statistical evidence to empirically convince relevant stakeholders that the effects they report are real, reliable, replicable, and therefore worthy of consideration. A reliance on null hypothesis (H_O) underpinnings is recognized through hypotheses stated in terms of the statistically validated *existence* of (1) hypothesized relationships, or (2) generally described "better" (e.g., greater or more) or "worse" associations. Davis et al. (2013) describe a disturbing trend in methods diversity across research that may exacerbate this phenomenon toward increasing reliance on one of two methods, experiments and modeling.

We agree that NHST has helped support the rapid advancement of social science theory and practice and is not yet ready to be abandoned (see Frick 1996; Mulaik et al. 1997; Nickerson 2000; Wagenmakers 2007). However, methodological advancements have emerged over time in virtually all phases of the marketing research process. The objectives of the current study are to (1) revisit the issues inherent in NHST and identify known remedies, (2) demonstrate that the identified issues are relevant in that they can affect today's marketing research outcomes, and (3) advocate adopting tools and techniques that have emerged over the past three decades that can enhance the confidence that marketing researchers have in reported results.

THEORY -- POTENTIAL ISSUES & REMEDIES ASSOCIATED WITH NHST

Table 1 presents a synopsis of some of the issues known to potentially compromise confidence in reported results based upon NHST. Table 2 presents some of the available remedies discussed in the literature. In short, an opportunity exists to augment traditional NHST practices in marketing research practices. Wetzels et al. (2011) argue that psychological science stands at a similar three-way fork in the road: (1) continuing to rely almost exclusively on p-values as a measure of the efficacy of statistical evidence; (2) embracing the relatively modest change of placing a greater emphasis on the standard reporting of additional information provided by effect sizes (e.g., r, Cohen's d, etc) and confidence intervals; or (3) embracing a more radical change by aggressively moving toward Bayesian approaches. Consequently, a framework is proposed to assist interested marketing researchers in augmenting their traditional NHST results.

A PROPOSED FRAMEWORK TO AUGMENT NHST METHODS AND RESULTS

Table 3 proposes a research framework to assist marketing researchers in augmenting traditional NHST-based evidence supporting research conclusions. The proposed framework is directed primarily toward quantitative human subject research emphasizing self-reported measurement of latent concepts in descriptive/predictive models of marketing constructs. Importantly, it is not designed to be exhaustive, rather, represents a practical and easy-to-implement guide for moving in the research directions advocated herein.

Step 1: Crafting Appropriate Research Inquiries

Cumming (2012) argues that NHST represents a form of dichotomous thinking (reject H_0 or fail to reject H_0). For example, NHST would imply questions like, “Is there a statistically significant relationship between disconfirmation and consumer satisfaction?” In fact, the crux of many criticisms associated with NHST identified above relate to this dichotomous approach to framing research questions. Unfortunately, a significant p-value cannot be taken as proof of a relationship nor a non-significant p-value cannot prove that Ho is true (Gelman 2013).

The proposed research framework starts with recommendations related to the framing of our research questions. Marketing researchers are encouraged to begin by explicitly considering whether their research question(s) represent theory validation, theory broadening, or theory deepening because this determines the kind of evidence necessary to form an appropriate conclusion based on empirical evidence (a criterion). Theory validation might include replication studies or model generalizations. Marketing researchers might ask both NHST-framed questions and effect size questions for theory validation. Anticipated a priori effect sizes could be identified based upon previous research findings summarized via meta-analyses. Perugini & Bagozzi (2001) differentiate theory broadening (adding an independent variable as a parallel predictor with established predictors to increase the explained variance of dependent variables) from theory deepening (the introduction of a new variable that explains how existing predictions function to influence a dependent variable). Both can benefit from evidence of estimation thinking: (1) theory broadening (e.g., “How much does R^2 of the dependent variable increase based on the addition of the parallel exogenous variable?”); (2) theory deepening similarly (e.g., “How much do known theoretical relationships from the literature vary in the presence of a newly introduced influence?”).

Step 2: Validate Obtained Data Prior to Analyses

An emphasis on NHST has also had an impact on conventional approaches taken in data validation (e.g., testing of measurement scale reliability and validity). Mackenzie et al. (2011) also provide a scale development procedure that we recommend considering. Many of the recommendations for additional considerations presented herein are consistent with and built upon their recommended method. The first additional consideration involves the use of exploratory structural equation analysis (ESEM) as a means of item evaluation based upon the method developed by Asparouhov & Muthén (2009). This can help researchers avoid the potential misspecification related to true non-zero factor cross-loadings (see Marsh et al., 2009). Second, assessment of measurement scale invariance is advocated as a regular practice of scale validation in any research (Vandenberg & Lance 2000) if the tests of substantive hypotheses involve multiple groups. Measurement invariance refers to measurement equivalence across different populations (and their subgroups). Finally, the use of marker variable is advocated to assess potential common method variance (CMV) associated with self-reports as a measurement model, representing about ¼ of the variance in a typical research measure. Such variance can arise from respondent's consistency motifs, transient mood states, illusory correlations, item similarity, and social desirability. Podsakoff et al. (2012) assert that the recommended remedy, if the specific source of method bias is unknown or valid measures of the sources of bias are not available, is the CFA marker technique advocated by Williams et al. (2010). These additional considerations allow marketing researchers to be more confident in hypothesis testing results and estimations, reducing the chance that the "significant" results are due to some form of measurement bias.

Step 3: Considerations in Analyses

The framework proposes that marketing researchers consider employing latent modeling techniques (e.g., structural equation modeling, SEM) whenever possible for substantive analyses (Steenkamp & van Trijp 1991; Iacobucci et al. 2007). Second, a Bayesian approach affords a number of advantages in analysis, such as (1) the ability to produce evidence that can strengthen either H_0 or H_a (Nickerson 2000; Kruschke 2011), (2) help learn more about parameter estimates and model fit, (3) perform better with small samples, (4) reduce computational demands, and (5) allow for analyses of new kinds of models (Muthén & Asparouhov 2012). Importantly, we make no claim for novelty in the recommendation of embracing Bayesian theory and practice within the marketing discipline. For example, Lenk & Rao (1990) advocated a forecasting adoption model based on hierarchical Bayes procedures. Arora, Allenby & Ginter (1998) proposed a hierarchical Bayes model of primary and secondary aspects of consumer demand. Park & Kim (2013) use a Bayesian network approach to examine key success factors of mobile gaming. Rossi & Allenby (2003) offer a strong general call for greater use of Bayesian statistics in marketing; and Terui, Chun & Ogawa (2011) discuss the use of Bayes analyses vis-à-vis consumer satisfaction data. What has changed is the ease with which to include Bayesian analyses in recent years. For example, M*Plus* can assist researchers in conducting Bayesian analyses with relative ease (Muthén & Asparouhov 2012). Marketing researchers are advised to consider existing limitations of Bayesian analyses (see Jackman 2009).

Step 4: Reporting Results

For traditional NHST results, there is consistent guidance in the literature as well as the APA manual (2010) to report *exact p*-values rather than exceeding or not some standard (Cummings 2012; LeCroy & Krysik 2007; Thompson 2007). This guidance appears to make moot the need for an a priori standard for *p*-values associated with traditional NHST results as the burden of interpretation shifts from the researcher to the reader. Also, the p_{rep} statistic can be reported that changes to interpretation of the *p* value to the probability of replicating an effect (Killeen 2005).

Marketing researchers are encouraged to explicitly state a priori interpretive standards for reported results. That said, identifying a priori standards for effect sizes can sometimes be a bit problematic. Researchers must first select the specific effect size to report that is most interpretable given their specific inquiry. Then they must specify the a priori expectation of the size of the effect they anticipate. For the most interpretable sizes of effects, Cohen's (1988) reference values can be used (Cummings, 2012). A priori effect size estimates can also be made based on prior similar studies (Sawyer & Ball 1981) or through a formal meta-analysis. Borenstein et al. (2009) also call for reporting a standard for the level of precision for observed effect sizes – a confidence interval for effect sizes (also see Cumming, 2012). The practices of (1) always reporting inter-construct correlation matrices associated with the study to support potential meta-analyses for future research, and (2) reporting results associated with Bayesian analyses in addition to NHST-based results whenever possible are further encouraged.

Step 5: Encourage Replication

Finally, Cumming (2012) makes a strong argument for greater replication in support of emerging estimation emphases. According to Popper (1959) the strongest confirmatory evidence for a scientific hypothesis is failure of concerted efforts of competent researchers to falsify it. Madden et al. (2000) conducted a study of replication research and concluded that natural science editors are generally more favorable toward replication studies than social science editors (also see Easley & Madden 2013). Ioannidis (2012) states that it is difficult to even try to reproduce publish non-significant results today based on current reporting protocols. This may also hold true today for the marketing discipline.

RESULTS -- ILLUSTRATION USING CONSUMER SATISFACTION

Our proposed framework purports to offer advantageous extensions for marketing researchers over traditional NHST-specific approaches, including the framing of more relevant research questions, stronger validation of obtained empirical data, and incorporation of emerging estimation alternatives. The following section demonstrates these claims vis-à-vis the well-known Expectancy Disconfirmation Model of Satisfaction (EDMS Oliver, 2009); see Figure 1.

Step 1: Crafting Appropriate Marketing Research Inquiries

A first step in the framework involves crafting dual NHST and effect size research hypotheses. This requires identification of any existing effect size evidence in order to form a priori hypotheses reflecting anticipated effect sizes. We therefore first conduct a meta-analysis of the expectancy disconfirmation-based underpinnings of consumer satisfaction as suggested by Aquinis et al. (2011b), which updates a previous meta-analysis conducted by Szymanski & Henard (2001). This provides some guidance in forming a more precise research hypothesis (i.e., extend beyond typical existence and direction NHST-based hypotheses) and associated confidence intervals. The current inquiry asks "*How much* do expectations, performance perceptions, and disconfirmation uniquely contribute to consumer satisfaction judgments?"

A Meta-Analysis of Disconfirmation Studies from 2000 to Present

We considered the recommendations of Card (2012) in articulating a sampling frame, including the inclusion and exclusion criteria. The desired sampling frame for the current research involves all published articles, unpublished (rejected, under review, or work-in-progress with sufficient information) articles, and dissertations. Excluded models include purely theoretical explanations (Phillips & Noble 2007), qualitative inquiries, quantitative studies that do not provide sufficient information to code for meta-analyses (after sending requests to study authors), studies designed to explain dissatisfaction (Anderson 1973; Kim & Smith 2005), and those studies that employ non-consumer satisfaction (e.g., B2B satisfaction, employee satisfaction).

Card (2012) further recommends planning the search strategy to identify potential sources of studies. Significant improvements have occurred since Szymanski & Henard's (2001) original meta-analysis in online libraries and article databases, many of which now include published dissertations, To capture published works we conducted individual searches of the ABI-Inform, Article First, Psych Info, Academic Search, and Emerald Full-Text online article databases using the keywords "expectancy disconfirmation," "disconfirmation," "disconfirmation and satisfaction," "expectation and satisfaction," and "performance and satisfaction." Only studies related to consumer satisfaction were included in further analyses. To capture unpublished works, a call for unpublished work was made using ELMAR, a popular listserv for business-related social scientists. We sent a personal email to the lead author or study contact for studies with no reported correlation matrix. Finally, several PhD-trained subject matter experts reviewed our list as the process progressed. We employed an individual level analysis consistent with Szymanski & Henard (2001).

The comprehensive Meta-Analysis 2.0 software was used to derive effect sizes, *r*, (see Table 4) and a series of meta-analyses are conducted, one for each of the relationships identified as H1-H5 in Figure 1. The random effects model was employed for calculations since it is unlikely that all studies are functionally equivalent (Borenstein et al. 2009; Cummings 2012). The values in Table 4 represent our a priori expectations of anticipated effect sizes and their associated confidence intervals to augment the traditional NHST approach and are incorporated into the research hypotheses for the quantitative study to follow as non-NHST (i.e., more precise) research prediction (see Table 7). Readers will note that it is reasonably straightforward to augment traditional NHST hypothesis tests with expected effect sizes in marketing research.

Step 2: Validate Obtained Data Prior to Analyses (Study 2)

We next conducted a second study to assess the research model in Figure 1.

Study 2 Methods

Student participants were solicited from an introductory marketing course at a large, public university in the Midwest of the United States. Students were awarded extra course credit for completing two surveys over the course of the semester. Students were first asked about their predictive expectations of a shopping experience at Wal-Mart, the largest discount retailer in the world. The student respondents were then instructed to go to Wal-Mart and purchase a holiday gift for themselves or someone else and return with the receipt to complete a post-consumption survey. 211 students completed the data gathering exercise, exceeding the required sample size of 121 for desired power. ESEM and CFA were employed using M*Plus*7. The appendix presents the measures, including their reliability and validity scores.

Minimum Required Power for Analyses

We utilized G*Power version 3.1.3 as recommended by Cumming (2012) to conduct a priori power analyses. The default value of effect size (f^2) = 0.15 corresponds to a small effect size given Cohen's (1988) heuristic and represents a conservative estimate. The results indicated that a minimum sample size of 121 was required to obtain 80% power to detect a conservative effect size (0.15) with three predictor variables (expectations, performance, and subjective disconfirmation) at a level of significance of 0.05. Note that other considerations may also apply such as minimum sample sizes to conduct structural equation analyses (see Bagozzi & Yi 2012).

Common Method Variance (CMV)

Following Williams et al.'s (2010) method, the factor loadings and error variances associated with a marker variable from a traditional CFA are comprised in a Baseline Model, which is then compared to an orthogonal model wherein the marker variable was associated with the remainder of the model predictors as fixed indicators (Model-C). A χ^2 difference test provides evidence for the presence of CMV with the marker variable. A comparison of Model-C with an unconstrained set of factor indicators, Model-U, provided a test of whether the marker variable is differentially related to the substantive variables. Finally, Model-R fixed the factor correlations of Model-C or Model-U to the values obtained from the Baseline Model to provide a statistical test of the biasing effects of the marker variable on substantive relationships. Table 5 shows no biasing effects of the marker variable on substantive relations. The mean amount of marker variance in each indicator was 1.71%.

To assess the effect of CMV on reliability estimates, the goal is to decompose a reliability measure into substantive and method variance components. This information is reflected in equation [1] and suggests that very little of the reliability associated with the latent variables in the model can be associated with CMV. Sensitivity analyses suggest that the manipulation of method factor loadings induced a relatively small corresponding change in the factor correlations, except for the correlations associated with larger confidence intervals and then were only associated with the final endogenous outcomes.

$$R_{\text{Total}}\ (\,.9598) = R_{\text{Substantive}}\,(.9338) + R_{\text{Method}}\,(.0260) \qquad [1]$$

Establishing Multi-Group Equivalence of Measurement

Measurement equivalence ensures that the measurement scales operate in the same way and that the underlying latent factor has the same structural property across the groups of interest. There are two types of equivalence tests: measurement invariance and structural invariance. *Measurement equivalence* refers to the extent to which measurement parameters of observed variables and their links to the unobserved (latent) variables are similar across groups, while *structural equivalence* focuses on the latent variables (factor variances) and their relations among the factors (factor covariances) (Bryne 2008). In a research inquiry in which a researcher systematically compares regression path coefficients between and among subpopulations, both sets of tests are required.

We focused on gender in assessing measurement equivalence between the two groups (n_{male} = 108, n_{female} = 103). Table 6 demonstrates evidence of sufficient invariance to move forward with substantive analyses. Our first step was omnibus test, which did not pass, but the subsequent measurement and structural invariance test results indicate that at least metric invariance is evident, which should suffice unless we are comparing group mean differences or group differences in factor inter-correlations.

Predictive Results

Predictive analyses were run using MPlus 7. Figure 1 and Table 7 present the results of predictive analyses of the theoretical model using both Maximum Likelihood (ML) and Bayesian estimation algorithms. ML estimation is the default for many SEM programs and is based on choosing the value of the parameter which maximizes the associated likelihood function. However, Muthen & Asparouhov (2012) assert that using ML and likelihood-ratio χ^2 testing involve unnecessarily strict models to represent hypotheses from substantive theory, often resulting in rejection of models and model modifications that may capitalize on chance. In the Bayesian implementation inherent in MPlus, however, the estimation procedure produces an analysis that better reflects substantive theories by incorporating prior information and posterior predictive checking which is known to be less sensitive than likelihood-ration ratio χ^2 testing to ignorable degrees of model misspecification. Bayesian estimation uses prior beliefs about the likely value of a parameter (Muthén & Asparouhov 2010). In the current study, informative priors are differentiated from non-informative priors by applying a small-variance prior of 0.01 that will allow the cross-loading variation lying between -.2 and +.2. The reason to choose a Bayesian analysis in addition to ML estimation is that instead of relying on point estimates and asymptotically-justified confidence bounds and test statistics, the Bayesian approach bases inferences on exact prior distributions for the parameters and latent variables estimated by Markov Chain Monte Carlo (Dunson et al. 2005). In other words, it answers a different underlying question and relies on more available information.

Model fit statistics for Bayesian estimation are shown with parameter estimates in Figure 1 and include a 95% confidence interval for the difference between the observed and the replicated chi-square value, posterior predictive *p*-value (PPP), and deviation information criterion (DIC). A PPP value around 0.5 and 95% confidence interval including zero close to the middle of the confidence interval indicates a good fit (Muthén & Asparouhov 2010). The results indicate that the Bayesian analysis with informative priors results in improved PPP's of 0.142 with 95% confidence interval including zero, indicating that the observed-data statistic does not differ much than what would have been generated by the model while the Bayesian analysis with non-informative priors results in PPP's of 0.06 which shares the ML likelihood-ratio χ^2 results. Note that the p values associated with Bayesian estimate is interpreted differently than the classical p values in that it represents the proportion of the posterior distribution that is below zero with a positive estimate while it represents the proportion of the posterior distribution that is above zero with a negative estimate. Thus, it explains the probability that the estimate is the likely effect. For example, the p value of .002 for the path from Subjective Disconfirmation to Satisfaction indicates that there is only 2% probability that the estimate is not the effect we had expected. In other words, this effect is not likely to be negative estimate. While not reported in Figure 1, the 95% credible interval (akin to a confidence interval) for this estimate is (0.116, 1.314) can be correctly interpreted as the interval that contains the population parameter with 95% probability. Note that the interpretation of DIC is similar to that of akaike information criterion (AIC) of maximum likelihood estimation (i.e., small values indicate a better fit).

Readers will also note that p_{rep} values are presented in Table 7 per the proposed framework. Killeen (2005) provides a table for interpreting the p_{rep} scores. In the case of the current research, all p_{rep} scores can be interpreted as confidently replicable except for the relationships associated with consumer expectations. The key take-away from Figure 1 and Table 7 is, except for the relationship between Expectations and Satisfaction for the Bayesian estimation without prior information, the NHST ML estimation model and Bayesian models show consistency of results. We therefore are relatively confident in our interpretation of the relationships implicit in the expectancy disconfirmation paradigm of consumer satisfaction, except for the influence of consumer expectations as a unique exogenous influence.

The theoretical model is supported by the data based on the overall model fit indices reported in Figure 1. Thus, we can interpret the typical NHST hypotheses in the assessed model. However, as previously discussed, the effect sizes of interest herein that augment the predictive analyses involve the correlations expressed in Table 8. These results suggest that the current study generally found stronger inter-correlations among the concepts in the expectancy disconfirmation model of satisfaction than typically expected. This further suggests that the expectancy disconfirmation theoretical model is particularly effective in explaining young consumers' satisfaction and intentions based on the Wal-Mart shopping experience.

DISCUSSION/CONCLUSIONS

We hope that our proposed framework leads to greater dialogue among marketing researchers as to the utility from augmenting traditional NHST practices. First, the proposed framework arguably improves every stage of the research process, beginning with earliest planning stages. Intuitively, a stronger framing of research questions should yield better answers through marketing research. Further, the proposed transition involves relatively unobtrusive changes in marketing researcher standard practices. Greater attention to the framing of research questions as well as methods of inference articulated herein can only benefit the research process and subsequent research outcomes.

Second, incorporating estimation-based inquiries based on effect sizes leads to stronger evidence underlying conclusions from research. Moving from phenomenon existence (the domain of NHST) to questions of *How Much?* adds to the information content underlying inference. Such information will often be particularly interesting to marketing practitioners.

Third, the proposed framework adds additional information by incorporating measures of precision in analyses. Cumming (2012) asserts that the addition of confidence intervals affords a measure of precision in planning by first selecting a target confidence interval width and then using analysis to determine what sample size is likely to produce confidence interval width no larger than that target. In other words, precision can be used to replace/augment traditional power considerations, which are founded on NHST principles.

Fourth, an emphasis on estimation will allow for new insights not currently available through a reliance on NHST methods. For example, we can begin to evaluate whether observed effect sizes vary across research settings (e.g., "Do observed effect sizes vary depending on the form of expectation in disconfirmation or the criterion satisfaction variable?"). In the predictive study we find that inter-factor correlations as effect sizes are noticeably larger than expected based on extant meta-analyses. It remains unclear whether this is sample specific or represents a trend. Another example would be the consideration of effect size trends across time.

Fifth, an emphasis on both NHST and effect sizes potentially opens up new techniques to solve marketing research riddles. For example, Monte Carlo studies represent real opportunities to better understand marketing phenomena. The availability of comprehensive meta-analytic results allows for the development of data sets with known distributions reflective of observed effect sizes across studies. Thus, observed effect-size distributions can be created and then modified to investigate issues of interest to marketing researchers such as the influence of varying levels of measurement error, or various sample sizes on overall model fit indices. An emphasis on meta-analysis in marketing science research may also bring progress toward greater comfort with replication practices in marketing research.

Sixth, a transition to greater use of Bayesian estimation can provide a number of practical advantages that can advance marketing research. Above all, it allows intuitive interpretation of findings. While frequentist approaches such as *p* value and confidence interval ambiguously answer the proposed questions (i.e., an effect exist or not or 95 out of 100 confidence intervals include the true population value - in fact, this approach does not answer whether the research did find an effect or included the true population value), Bayesian approach produces straight summary of the results based on probabilities. Again, *p* value tells whether a parameter value is positive or negative and credible interval indicates the interval that contains the population parameter with 95% probability. Next, while the use of informative prior is somewhat limited, it can be helpful in explaining the data with no relevant information being omitted. In summary, Bayesian estimation can help marketing researchers and/or practitioners make better decision by eliminating some uncertainty associated with NHST via more intuitive and richer information.

REFERENCES

Anderson, R. E. (1973). Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of Marketing Research*, 13 (February), 38-44.

Aquinis, H, Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011b). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management* 37 (1), 5-38.

American Psychological Association (2010). *Publication manual of the American Psychological Association, 6th Edition*. Washington, DC, Author Unknown.

Arora, N., Allenby. G., M., & Ginter, J. L. (1998). A hierarchical Bayes model of primary and secondary demand. *Marketing Science*, 17 (1), 29-44.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397-438.

Bagozzi, R. P. & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, 40 (8), 8-34.

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

Bakker, M,, van Dijk, A. & Wircherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7 (6), 543-554.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37, 325-335.

Borenstein, M,, Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, United Kingdom: John Wiley & Sons.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measurement instrument: A walk through the process. *Psicothema*, 20 (4),:872-882.

Boulding, W., Kalra, A,, Staelin, R., & Zeithaml, V. A. (1993). A dynamic process model of service quality: From expectations to behavioral intentions. *Journal of Marketing Research*, 30 (February), 7-27.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: The Guilford Press.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

Cummings, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge, Taylor & Francis Group.

Davis, D. F., Golicic, S. L., Boerstler, C. N., Choi, S., & Oh, H. (2013). Does marketing research suffer from methods myopia? *Journal of Business Research*, 66, 1245-1250.

Dunson, D. B., Palomo, J., & Bollen, K. (2005). Bayesian structural equation modeling. *Statistical and Applied Mathematical Sciences Institute*, Research Triangle Park, NC 27709-4006, Technical Report #200504, http://www.samsi.info/communications/bayesian-structural-equation-modeling, [online].

Easley, R. W., Madden, C. S., & Dunn, M. G. (1995). Conducting marketing science: The role of replication in the research process. *Journal of Business Research*, 48, 83-92.

Easley, R. W. & Madden, C. S. (2013). Replication revisited: Introduction to the special issue section in business research. *Journal of Business Research*, 66, 1375-1376.

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, 5, 75-98.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, Published online 11 September 2011, Accessed 4/4/2013, http://mres.gmu.edu/pmwiki/uploads/Main/Fanelli2011.pdf, [online].

Forstmeier, W. & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecological Sociobiology*, 65, 47-55.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1 (4), 379-390.

Gelman, A. (2013). P-values and statistical practice. *Epidemiology*, 24:69-72.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996) Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology*, 33, 175-183.

Hubbard, R. & Vetter, D. E. (1996). An empirical comparison of published replication research in accounting, finance, management, and marketing. *Journal of Business Research*, 35 (2), 153-164.

Hubbard, R. & Lindsay, R. M. (2013). The significant difference paradigm promotes bad science. *Journal of Business Research*, 66, 1393-1397.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, 17 (2), 139-153.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19 (5), 640-648.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives of Psychological Science*, 7 (6), 645-654.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. United Kingdom: John Wiley & Sons, Ltd.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345-353.

Kim, Y. K. & Smith, A. K. (2005). Crime and punishment: Examining customers' responses to service organizations' penalties. *Journal of Service Research*, 8 (2), 162-180.

Kline, R. B. (2004). What's wrong with statistical tests – and where do we go from here? in *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC, US: American Psychological Association.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6, 299-312.

LeCroy, C. W. & Krysik, .J (2007). Research note: Understanding and interpreting effect size measures. *Social Work Research*, 31 (4), 243-8.

Lenk, P. J. & Rao, A. G. (1990). New models from old: Forecasting product adoption by hierarchical Bayes procedures. *Marketing Science*, 9 (1), 42-53.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Massey, Lindsey L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, 34, 171-187.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35 (2), 293-334.

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication studies. *Journal of Advertising*, 24 (4), 77-87.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2009). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, 29, 322-346.

Muliak, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing, in *What if there were no significance tests?* Harlow, L. L., Muliak, S. A., and Steiger, J. H. (Eds.). Hillsdale, NJ: Erlbaum, 65-116.

Muthen, B. & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 17 (3), 313-335.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5 (2), 241-301.

Neath, A. A. (2010). Statistical inference, statistics education, and the fallacy of the transposed conditional. *Section on Statistical Education*, http://www.statlit.org/pdf/2010NeathASA.pdf, accessed 4/4/2013, [online].

Nord, W. R. (2012). On doing the wrong things for the wrong reasons: Two misguided organization studies practices. *Journal of Management Inquiry*, 21, 443-447.

Oliver, R. L. (1997). Whence Consumer Loyalty? *Journal of Marketing*, 63, 33-44.

Oliver, R. L. (2009). *Satisfaction: A behavioral perspective on the consumer, 2nd Edition*. London, England: M. E. Sharpe.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64, (1):12-40.

Park, H. & Kim, S. (2013). A Bayesian network approach to examining key success factors of mobile games. *Journal of Business Research*, 66, 1353-1359.

Pauker, S. P. & Pauker, S. G. (1979). The amniocentesis decision: An explicit guide for parents, in *Risk, communication, and decision making in genetic counseling*. Epstein, C. J., Curry, C. J. R., Packman S, Sherman SN, Hall BD (eds), Vol. 15, New York: The National Foundation, 289-324.

Perugini, M. & Bagozzi, R. P. (2001). The role of desires and anticipated emotions in goal-directed behaviours: Broadening and deepening the theory of planned behavior. *British Journal of Social Psychology*, 40, 79-98.

Phillips, J. & Noble, S. M. (2007). Simply captivating. *Journal of Advertising*, 36 (1), 81.

Podsakoff, P. M., MacKenzie, S. B., & 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.

Pritchard, M. P., Havitz, M. E., & Howard, D. R. (1999). Analyzing the commitment-loyalty link in service contexts. *Journal of the Academy of Marketing Science*, 27 (3), 333-348.

Rossi, P. E. & Allenby, G. M. (2003) Bayesian statistics and marketing. *Marketing Science*, 22 (3), 304-328.

Rouder, J. N. & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877-903.

Sawyer, A. G. & Ball, D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, 18 (August), 275-90.

Sawyer, A. G. & Peter, J. P. (1983). The significance of statistical significance tests in marketing research.*Journal of Marketing Research*, Vol. XX (May 1983), 122-33.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13 (2), 90-100.

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in *What if there were no significance tests?* Harlow, L.L., Mulaik, S. A. & Steiger, J. H., eds., Hillsdale, NJ: Erlbaum, 37-64.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366.

Steenkamp, J. E. M. & van Trijp, H. C. M. (1991). The use of LISREL in validating marketing constructs. *International Journal of Research in Marketing*, 8, 283-299.

Szymanski, D. M. & Henard, D. H. (2001). Customer satisfaction: A meta-analysis of the empirical evidence. *Journal of the Academy of Marketing Science*, 29 (1), 16-35.

Terui, N., Chun, S. H. T. & Ogawa, K. (2011). Hierarchical Bayes modeling of the customer satisfaction index. *Service Science*, 3 (2), 127-140.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44 (5), 423-432.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3 (1), 4-70.

Wagenmakers, E. (2007). Theoretical and review articles: A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, 14 (5), 779-804.

Wagenmakers, E, Grunwald, P. (2006). Commentary: A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*, 17, 641-642.

Wagenmakers, E., Wetzels, R., Boorsboom, D., & van de Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100 (3), 426-432..

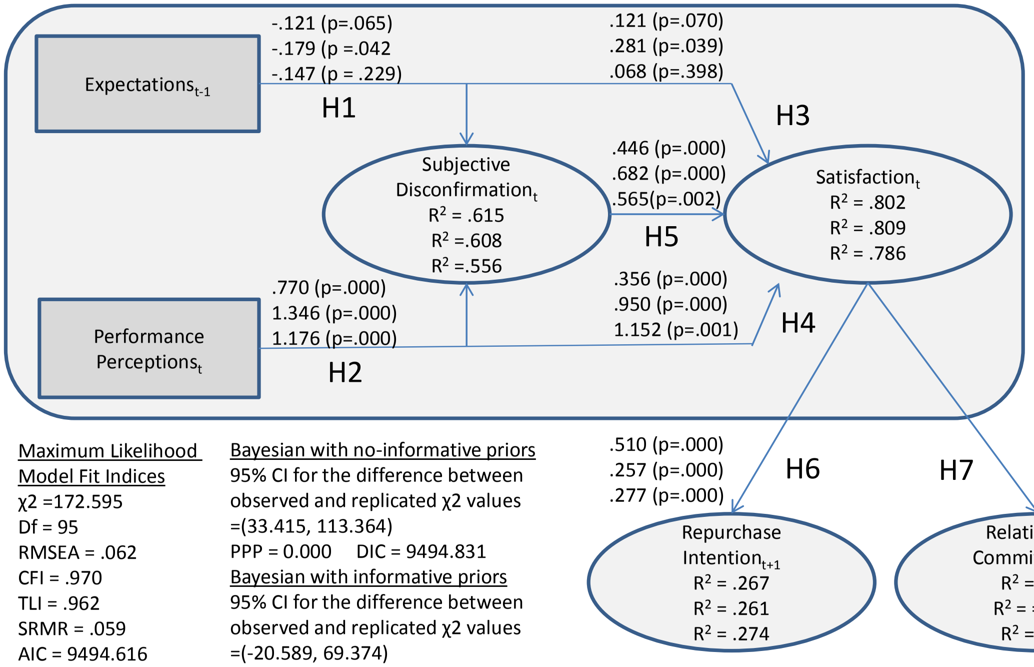
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. .J, & Wagenmakers, E. (2011) Statistical evidence in experimental psychology. *Perspectives on Psychological Science*, 6 (3), 291-298.

Williams, L. J., Hartman, N., & Cavazotte, F. (2010) Method variance and marker variables: A review and comprehensive CFA marker technique. *Organizational Research Methods*, 13 (3), 477-514.

APPENDIX

Figure 1

The Expectancy-Disconfirmation Model of Satisfaction & Predictive Model Results



All reported paths are unstandardized. Path order from top to bottom includes maximum likelihood, Bayesian with no information priors, and Bayesian with information priors.

Maximum Likelihood Model Fit Indices
χ2 =172.595
Df = 95
RMSEA = .062
CFI = .970
TLI = .962
SRMR = .059
AIC = 9494.616

Bayesian with no-informative priors
95% CI for the difference between observed and replicated χ2 values =(33.415, 113.364)
PPP = 0.000 DIC = 9494.831

Bayesian with informative priors
95% CI for the difference between observed and replicated χ2 values =(-20.589, 69.374)
PPP = 0.142 DIC = 9474.020

Table 1
Criticisms of NHST

General Issue	Specific Problem/Criticism	Comment	Source
Misunderstanding of "Statistical Significance"	Viewing the p-value as the probability that the results occurred because of sampling error or chance.	There is no way in classical statistical significance testing to determine whether the null hypothesis is true or the probability that it is true. The value of p is the probability of obtaining a value of a test statistic (e.g., D) as large as the one obtained – conditional on the null hypothesis being true – p(D|Ho).	Berkson (1942); Carver (1978); Pauker and Pauker (1979); Sawyer and Peter (1983); Falk and Greenbaum 1995; Nickerson (2000); Levine et al. (2008);
	Viewing the p-value as the probability that the results will be replicated in the future	Nothing in classical statistical significance testing says anything about the probability that the same results will occur in future studies.	Sawyer and Peter (1983); Nickerson (2000); Levine et al. (2008);
	Viewing the p-value as the probability that the alternative hypothesis is true (i.e., the proposed theory is true).	Rejection of the null hypothesis at a predetermined p-level supports the inference that sampling error is an unlikely explanation of results but it gives no direct evidence that the alternative hypothesis is true. One is, in effect, accepting the null hypothesis as true when one takes the failure of p to reach a conventional level of significance as evidence that prior to experimental treatment an experimental and control group were equivalent with respect to some measure of interest. A basic limitation of NHST is that it does not allow a researcher to gather evidence in favor or the null (Wetzels et al. 2011; Rouder and Morey 2012). This is an example of the fallacy of the transposed conditional. Neath (2010) , described as argument that P(B/A) is equal to P(A/B). This fallacy is why a decision in favor of Ha is not necessarily enough to provide evidence that Ha is true and Ho is false.	Sawyer and Peter (1983); Nickerson (2000); Levine et al. (2008); Wetzels et al. (2011); Rouder and Morey (2012); Wagenmakers (2007); Neath (2010); Wahenmakers et al. (2011);
	Confusion about the role of sample size and the level of statistical significance	There is no inherent bias against statistically significant results obtained from properly selected small samples. Moreover, because effect size is a measure of the strength of the relationship and large effects are more likely to be replicated than small ones, researchers should have more confidence in the study with the smaller sample.	Sawyer and Brown (1981); Sawyer and Peter (1983); Levine et al. (2008)
	Belief that a small p-value means a treatment effect of large magnitude; Belief that statistical significance means theoretical or practical significance	When one concludes on the basis of a statistical test that the difference between two means is statistically significant, one is saying only that a difference of the observed magnitude is unlikely to be obtained between two samples drawn at random from the same population.	Nickerson (2000);
Potential Researcher Influence on Evidence of Statistical Significance	Sensitivity of NHST to Sample Size	Whether or not one assumes that the null hypothesis is always or almost always false, when it is false the probability that a statistical significance test will lead to rejection increases with sample size. This criticism has been the focus of some of the strongest criticisms to NHST because it means that conclusions drawn from experiments often depend on decisions experimenters have made regarding how many participants to run. In addition, it is possible for studies with very large samples to demonstrate statistical significance for differences that are too small to be of any theoretical or practical interest. Specifically, the null hypothesis is never exactly true and will therefore always be rejected as the number of observations grows larger. Kline (2004) further argues that the nil hypotheses is almost always false.	Nickerson (2000); Kline (2004); Wagenmakers (2007); Levine et al. (2008); Hubbard & Lindsey (2013);
	Potential inflated effect sizes due to selection thresholds and suboptimal power.	Ioannidis (2008) argues that the effect sizes of newly discovered true (non-null) associations are inherently inflated on average. One reason is that the combination of requiring an association to pass a certain threshold of statistical significance based on a study with suboptimal power.	Ioannidis (2008)
	Misunderstanding the Subjectivity of Statistical Tests	Because researchers make many subjective decisions that greatly influence the probability of rejecting the null hypothesis, it is misleading to consider the process of statistical significance as objective solely because of the objectivity of the objectivity associated with the mathematics.	Sawyer and Peter (1983); Wagenmakers (2007); Ioannidis (2008);
	Flexibility in data collection, analysis, and reporting dramatically	There appears to be a measure of inflation in false positive (or failure to reject the null hypothesis) in psychological research published based on NHST. In many cases a researcher is more	Simmons et al. (2011); Forstmeier and Schielzeth (2011);

	increases actual false-positive rates.	likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. This is because the literature suggests that when researchers are faced with ambiguous analytic decisions, they tend to conclude (with convincing self-justification) that th4e appropriate decisions are those that result in statistical significance. Also, Forstmeier and Schielzeth (2011) explicitly criticize the often seen practice of fitting generalized linear models with multiple predictors by simplifying models through the deletion of non-significant items (i.e., an issue of model selection). The result is the "winners curse" wherein false positive results arise primarily through an overestimation among significant predictors that often cannot be reproduced in follow-up studies.	Rouder and Morey (2012);
	Mismatching statistical analyses and design	Kline (2004) argues that statistical analyses and design are frequently mismatched, and that statistical assumptions underlying NHST methods are infrequently verified.	Kline (2004)
General Issues Related to The Validity of a Reliance on Evidence Derived from Statistical Significance	Misconceptions related to Type I errors	Belief that alpha is the probability that if one rejects the null hypothesis one has made a type I error; Belief that the value at which alpha is set for a given experiment is the probability that a Type I error will be made in interpreting the results of that experiment; Belief that the value at which alpha is set is the probability of Type I error across a large set of experiments in which alpha is set at that level;	Nickerson (2000)
	Statistics can be a poor indicator of individual's behaviors.	The value of statistical significance tests lies in their ability to focus on aggregate central tendencies, thereby reflecting little in the way of specific individual's behaviors.	Wagenmakers (2007);
	The NHST p-value does not measure/quantify statistical evidence.	In order for the p-value to qualify as a measure of statistical evidence, a minimum requirement is that identical p-values convey identical levels of evidence, irrespective of sample size. However, p-values overestimate the evidence against the Ho. In addition, the point or nil-null is almost always false (Levine et al. 2008).	Wagenmakers (2007); Levine et al. (2008);
	NHST lacks logical validity	*Modus tollens* is a valid argument form in logic, also known as "denying the consequent". Statistical tests are typically patterned after modus tollens (Nickerson 2000). However, the logic of NHST is difficult to reconcile with the principles of modus tollens.	Nickerson (2000); Levine et al. (2008);
	Noninformativeness of NHST Outcomes	NHST does not provide measures of the size of an effect or the strength of a relationship between an IV and a DV. They only give evidence of whether or not a statistically significant effect has been obtained, and if so, what direction.	Nickerson (2000)
Potential Negative Influence on the Practice of Marketing Science	NHST provides no incentives to researchers to develop precise hypotheses.	NHST simply requires one to test an unspecified hypothesis (H1) against "chance" (Ho).	Nickerson (2000)
	There exists a significant publication bias toward statistically significant results; Belief that failure to reject the null hypothesis is evidence of a failed experiment	Results from statistical significance tests are perceived to be valuable when they support the favored hypothesis but are commonly discounted when they support the null. Nickerson (2000) argues that this publication bias creates a biased literature because: (1) Type 1 errors are likely to go undetected; (2) Type II errors are frequently made; and (3) the differences that are reported in the literature may be larger on average than the population effects they represent. In addition, review articles of topics will only represent statistically significant results.	Sawyer and Peter (1983); Nickerson (2000); Fanelli (2012); Ioannadis (2012); Bakker et al. (2012);
	An absence of avenues to replicate results can undermine overall confidence in the scientific practice.	Ioannidis (2012) argues that the ability to self-correct is a hallmark of science, but does not always happen to scientific evidence by default. The absence of unbiased and efficient replication mechanisms makes it difficult to maintain high levels of scientific credibility.	Hubbard & Lindsey (2013); Ioannidis (2012); Bakker et al. (2012);

Table 2
Recommendations to Overcome Criticisms of NHST

General Issue	Specific Remedy	Comment	Source
Misunderstanding of "Statistical Significance"	Consider alternatives to NHST-based p-value arbitrated research questions.	Increase the use of non-nil null hypotheses, or as something other than one of zero difference/effect/correlation, or as a nil hypothesis. Replace the phrase "test of significance" with "test against the null hypothesis." This avoids the confusion of statistical significance with "amount" of significance. Increase specificity about alternatives to the null hypothesis. The expression of specific alternative hypotheses moves in the direction of Bayesian analyses. Nickerson (2000) asserts that what constitutes appropriate statistical treatment and justified interpretation of the outcome depends to no small degree on the experimenter's pre-experiment intentions and expectations.	Nickerson (2000); Sawyer and Peter (1983); Nickerson (2000)
	Consider alternatives to NHST-based p-value arbitrated answers to research questions.	Report results by emphasizing effect size and substantive significance instead of merely focusing on the p-value associated with the resulting test statistics. Provide some indication of effect size either along with or in place of results of statistical significance tests. Nickerson (2000) also notes that statistical significance is a function of two factors – effect size and sampling error. Therefore, he suggests reporting these as well. Utilize ranges of effect size (e.g., confidence intervals) as opposed to point estimates (e.g., Ho). Avoid the use of point null hypotheses, instead using a range of values that will be considered effectively null. At a minimum, the reporting of directional hypotheses are recommended. Report power analyses associated with reported results. Only if the power of a test is high should one conclude that the null hypothesis is true (or approximately true. Even then, researchers should be sensitive to the fact that even a test of great power does not prove that it is true. Also, care must be taken to ensure that observed power is not so great that effects too small to be of interest will prove statistically significant. Bakker and Wicherts (2011) argue that in a sample of psych l papers with NHST only 11% referred to power as a rationale for the choice of sample size or design. Calculate and report the p_{rep} value in addition to the traditional p value. The p_{rep} value represents the probability of replicating an effect (Killeen 2005). The use of this statistic has been officially encouraged by Psychological Science (Wagenmakers 2007). Provide some indication of variability or precision of measurement. Include measures of measurement error, descriptive variability (e.g., standard deviation), and standard error of the mean. Report confidence intervals around point estimates as these are more informative than significance tests, and provide both an estimate of effect size and an indication of uncertainty related to accuracy.	Hubbard & Lindsey (2013); Sawyer and Peter (1983); Nickerson (2000); Wetzels et al. (2011); Sawyer and Brown (1981); Bakker and Wicherts (2011); Cumming (2012); Killeen (2005); Wagenmakers (2007);
	Increase the use of Bayesian analytic approaches whenever possible.	The Bayesian approach directly compares the null and alternative hypotheses and allows one to consider more fully the possibility that the null hypothesis is true. This method also permits evidence to strengthen either the null hypothesis or its alternative. Wagenmakers (2007) argues for Bayesian hypothesis testing using the BIC approximation. The most common problem with Bayesian analyses can be difficulty in specifying prior probabilities.	Sawyer and Peter (1983); Wagenmakers (2007); Wagenmeyer and Grunwald 2006; Nickerson (2000); Wetzels et al. (2011); Shiffrin et al. (2008);
Potential Researcher Influence on Evidence of Statistical Significance	Consider additional steps to reduce potential researcher bias.	Consider alternatives to NHST-based methods in deriving results. For example, consider the use of three-outcome tests, or parameter-estimation and model-fitting techniques. Employs more use of meta-analyses as summaries of extant literature. Consider greater use of strong inference methods. Strong inference is a model of scientific inquiry that emphasizes the need for alternative hypotheses, rather than a single hypothesis in order to avoid confirmation bias. Use more than one statistical analysis technique whenever possible. Compelling results should yield similar conclusions, irrespective of the statistical paradigm that is used to analyze the data. Publish the results of all experiments and analyses, whether or not they attain statistical significance. Marketers should report all observed p-values, including those that are non-significant. This makes the reporting of results more of an objective process, facilitates the aggregation of results across studies via meta-analyses,	Sawyer and Peter (1983); Wagemaker et al. (2011): Nickerson (2000); Bakker et al. (2012);
	Increase the use	The Bayesian approach directly compares the null and alternative hypotheses	Sawyer and Peter

	of Bayesian analytic approaches whenever possible.	and allows one to consider more fully the possibility that the null hypothesis is true. This method also permits evidence to strengthen either the null hypothesis or its alternative. Wagenmakers (2007) argues for Bayesian hypothesis testing using the BIC approximation. The most common problem with Bayesian analyses can be difficulty in specifying prior probabilities.	(1983); Wagenmakers (2007); Wagenmeyer and Grunwald (2006); Nickerson (2000); Wetzels et al. (2011);
	Increase the practice of encouraging replications.	Replications will allow for the establishment of predictive values for non-nil hypotheses based on previous results. Replication practices will also facilitate the identification of priors for use in Bayesian analyses. See Schmidt (2009) for a classification scheme based on a functional approach to replication.	Hubbard & Lindsey (2013); Nickerson (2000); Schmidt (2009); Ioannidis (2008, 2012);
	Consider greater use of published guidelines	Guidelines for authors of confirmatory studies (Wagenmaker 2011) 1. Fishing expeditions should be prevented by selecting participants and items before the confirmatory study takes place; no further selection or subset testing should take place once the confirmatory experiment has started. 2. Data should only be transformed if it has been decided beforehand. 3. In simple examples, such as when the dependent variable is success rate or mean response time, an appropriate analysis should be decided before the data have been collected. Author guidelines for the problem of false-positive publications. (Simmons et al. (2011) 1. Authors must decide the rule for terminating data collection before the data collection begins and report this rule in the article. 2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data justification. 3. Authors must list all variables collected in a study. 4. Authors must report all experimental conditions, including failed manipulations. 5. Also report what the statistical results are if eliminated observations are included. 6. If an analysis includes a covariate, authors must report the statistical result of the analysis without the covariate.	Wagenmakers et al. (2011); Simmons et al. (2011)
General Issues Related to The Validity of a Reliance on Evidence Derived from Statistical Significance	Increase the use of Bayesian analytic approaches whenever possible.	The Bayesian approach directly compares the null and alternative hypotheses and allows one to consider more fully the possibility that the null hypothesis is true. This method also permits evidence to strengthen either the null hypothesis or its alternative. Wagenmakers (2007) argues for Bayesian hypothesis testing using the BIC approximation. The most common problem with Bayesian analyses can be difficulty in specifying prior probabilities.	Sawyer and Peter (1983); Wagenmakers (2007); Wagenmeyer and Grunwald 2006; Nickerson (2000); Wetzels et al. (2011);
	Increase practice of encouraging replications.	Replications will allow for the establishment of predictive values for non-nil hypotheses based on previous results. Replication practices will also facilitate the identification of priors for use in Bayesian analyses. See Schmidt (2009) for a classification scheme based on a functional approach to replication.	Nickerson (2000); Schmidt (2009); Ioannidis (2008, 2012);
Potential Negative Influence on the Practice of Marketing Science	Increase the use of Bayesian analytic approaches whenever possible in addition to NHST-based methods.	The Bayesian approach directly compares the null and alternative hypotheses and allows one to consider more fully the possibility that the null hypothesis is true. This method also permits evidence to strengthen either the null hypothesis or its alternative. Wagenmakers (2007) argues for Bayesian hypothesis testing using the BIC approximation. The most common problem with Bayesian analyses can be difficulty in specifying prior probabilities.	Sawyer and Peter (1983); Wagenmakers (2007); Wagenmeyer and Grunwald (2006); Nickerson (2000); Wetzels et al. (2011);
	Consider greater use of published guidelines	Reviewer guidelines for the problem of false-positive publications. 1. Reviewers should ensure that authors follow commensurable guidelines. 2. Reviews should be more tolerant of imperfections in results. 3. Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions. 4. If justification of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.	Simmons et al. (2011)

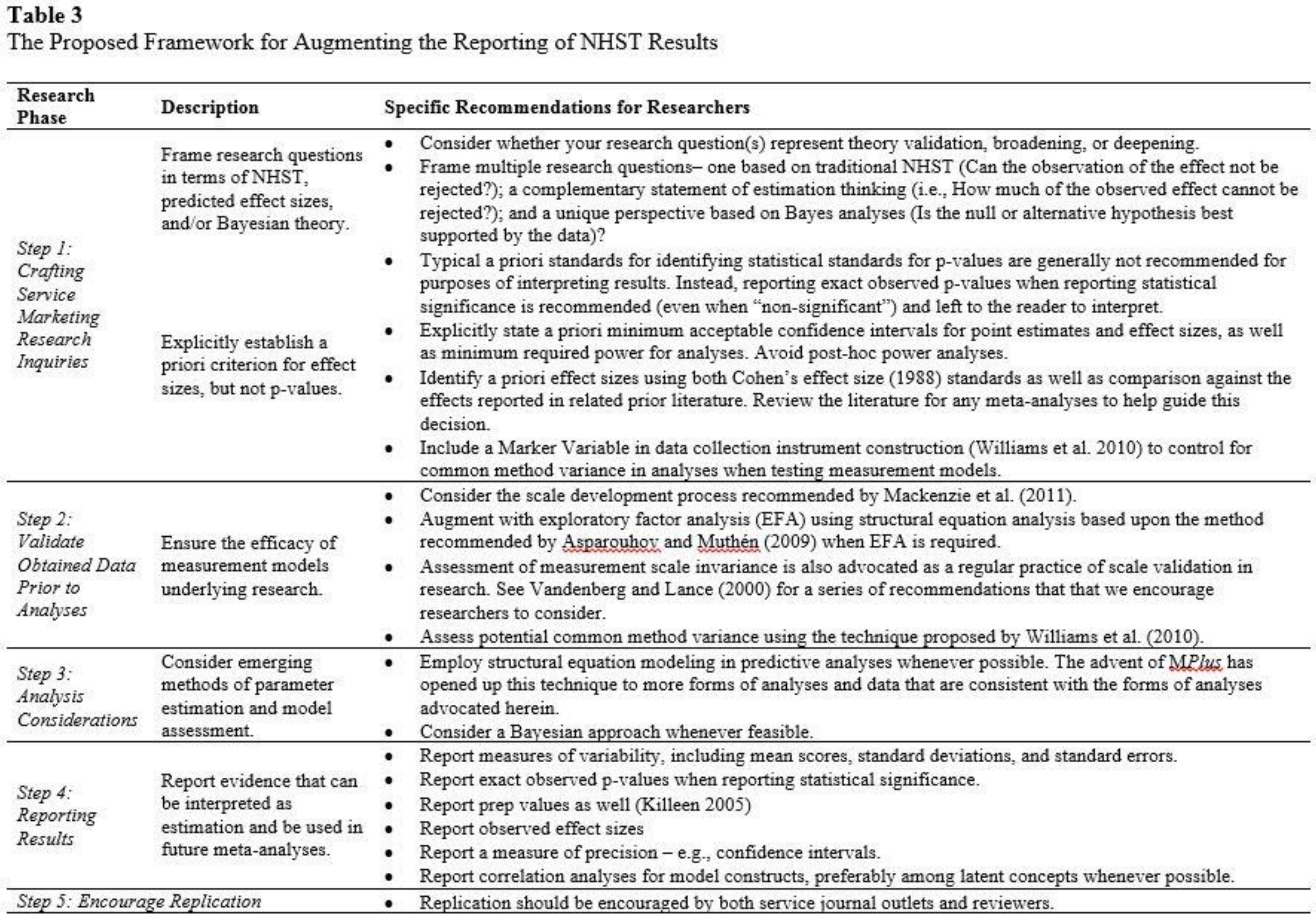


Table 3
The Proposed Framework for Augmenting the Reporting of NHST Results

Research Phase	**Description**	**Specific Recommendations for Researchers**
Step 1: Crafting Service Marketing Research Inquiries	Frame research questions in terms of NHST, predicted effect sizes, and/or Bayesian theory. Explicitly establish a priori criterion for effect sizes, but not p-values.	• Consider whether your research question(s) represent theory validation, broadening, or deepening. • Frame multiple research questions– one based on traditional NHST (Can the observation of the effect not be rejected?); a complementary statement of estimation thinking (i.e., How much of the observed effect cannot be rejected?); and a unique perspective based on Bayes analyses (Is the null or alternative hypothesis best supported by the data)? • Typical a priori standards for identifying statistical standards for p-values are generally not recommended for purposes of interpreting results. Instead, reporting exact observed p-values when reporting statistical significance is recommended (even when "non-significant") and left to the reader to interpret. • Explicitly state a priori minimum acceptable confidence intervals for point estimates and effect sizes, as well as minimum required power for analyses. Avoid post-hoc power analyses. • Identify a priori effect sizes using both Cohen's effect size (1988) standards as well as comparison against the effects reported in related prior literature. Review the literature for any meta-analyses to help guide this decision. • Include a Marker Variable in data collection instrument construction (Williams et al. 2010) to control for common method variance in analyses when testing measurement models.
Step 2: Validate Obtained Data Prior to Analyses	Ensure the efficacy of measurement models underlying research.	• Consider the scale development process recommended by Mackenzie et al. (2011). • Augment with exploratory factor analysis (EFA) using structural equation analysis based upon the method recommended by Asparouhov and Muthén (2009) when EFA is required. • Assessment of measurement scale invariance is also advocated as a regular practice of scale validation in research. See Vandenberg and Lance (2000) for a series of recommendations that that we encourage researchers to consider. • Assess potential common method variance using the technique proposed by Williams et al. (2010).
Step 3: Analysis Considerations	Consider emerging methods of parameter estimation and model assessment.	• Employ structural equation modeling in predictive analyses whenever possible. The advent of *MPlus* has opened up this technique to more forms of analyses and data that are consistent with the forms of analyses advocated herein. • Consider a Bayesian approach whenever feasible.
Step 4: Reporting Results	Report evidence that can be interpreted as estimation and be used in future meta-analyses.	• Report measures of variability, including mean scores, standard deviations, and standard errors. • Report exact observed p-values when reporting statistical significance. • Report prep values as well (Killeen 2005) • Report observed effect sizes • Report a measure of precision – e.g., confidence intervals. • Report correlation analyses for model constructs, preferably among latent concepts whenever possible.
Step 5: Encourage Replication		• Replication should be encouraged by both service journal outlets and reviewers.

Table 4
Effect Sizes and Associated Confidence Intervals

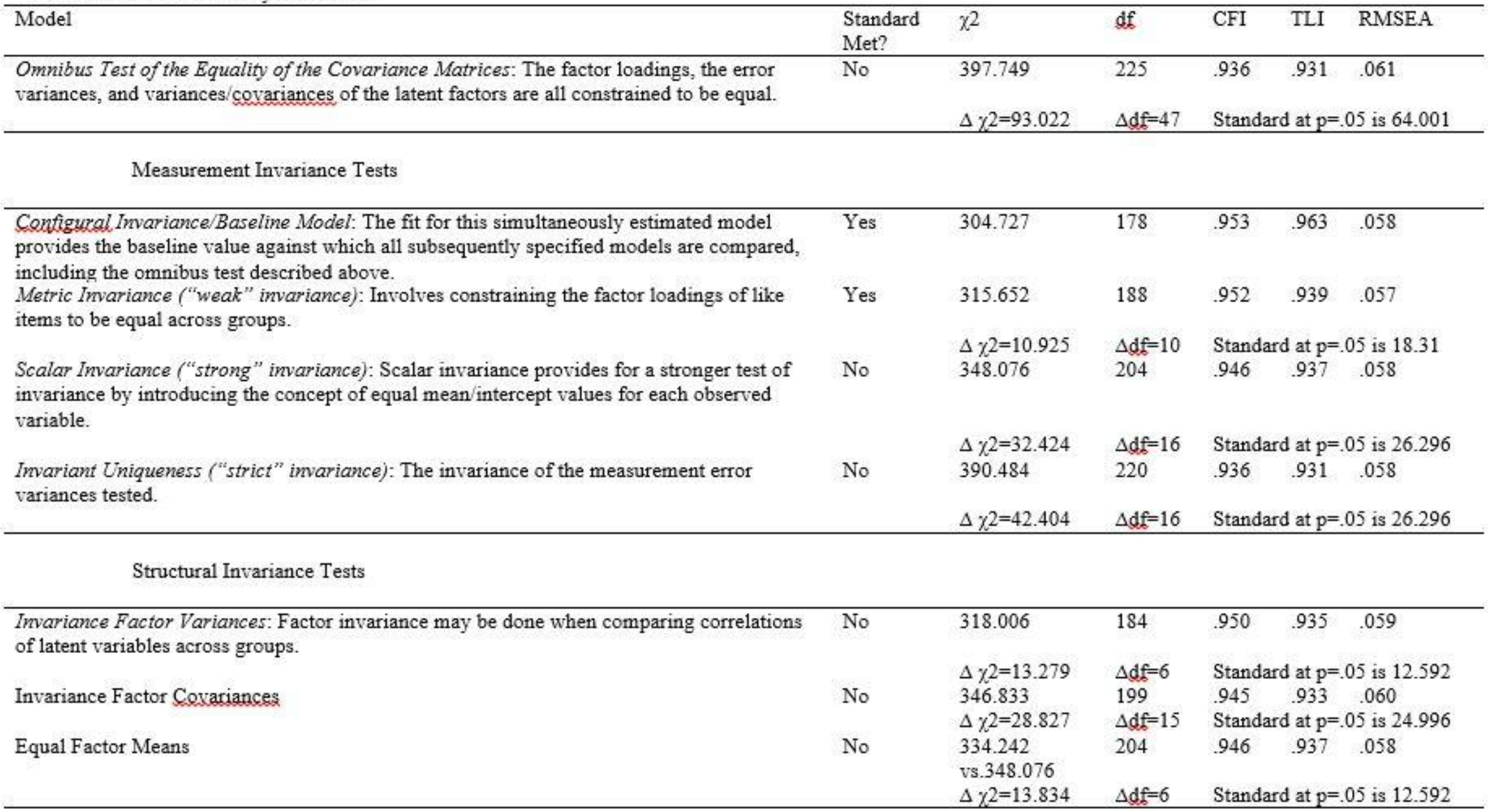
Associated Constructs	Number of Studies	Number of Correlations	Cumulative N	Effect Size	95% Confidence Interval
Disconfirmation ↔ Satisfaction	30	137	37,879	.460	Not reported
	20	36	14,251	.445	.320-.554
	50	177	52,130	.453	.439-.468
	4 Missing	36	14,251	.386	.250-.507
	50	177	52,130	.424	**.349-.494**
Performance ↔ Satisfaction	21	159	88,959	.340	Not reported
	15	22	9,666	.615	.527-.690
	36	181	98,625	.489	.178-.712
	0 Missing	22	9,666	.615	.527-.690
	36	181	98,625	.489	**.178-.712**
Expectation ↔ Satisfaction	8	17	5,927	.270	Not reported
	14	21	10,401	.279	.170-.381
	22	38	16,328	.276	.262-.290
	1 Missing	21	10,401	.292	.183 -.395
	22	38	16,328	.282	**.261-.304**
Performance ↔ Disconfirmation	7	23	3,435	.490	Not reported
	11	18	6,806	.492	.269-.665
	18	41	10,241	.491	.476-.506
	6 Missing	18	6,806	.299	.047-.516
	18	41	10,241	.367	**.350-.383**
Expectation ↔ Disconfirmation	7	23	4.445	.020	Not Reported
	7	11	4,937	.024	-.205-.250
	14	34	9,382	.022	.002-.042
	0 Missing	11	4,937	.024	-.205-.250
	14	34	9,382	.022	**.002-.042**

Note: Please note that the order of presentation within cells from top to bottom are the results of Szymanski and Henard (2001), the results of the current research meta-analysis (post-2000), the combined meta-analyses not adjusted for publication bias, the effect size adjusted by the trim-and-fill method, and the final combined effect size based on the values corrected for potential publication bias. Bolded ranges in the final column represent the anticipated confidence intervals explicated in the research hypotheses identified in Table 6.

Table 5
Chi-Square, Goodness-of-Fit Values, and Model Comparison Tests for Williams et al. (2010) Marker Variable Analyses

Model	χ^2	df	RMSEA	CFI	TLI
CFA	300.354	156	.066	.952	.942
Baseline	315.560	169	.064	.952	.946
Method-C	305.954	168	.062	.955	.949
Method-U	283.887	153	.064	.957	.946
Method-R	284.475	168	.057	.962	.957
Chi-Square Model Comparison Tests					
	$\Delta\chi^2$	Δdf	Chi-Square Critical Value: .05		
Baseline vs. Method-C	9.606	1	3.84		
Method-C vs. Method-U	22.067	15	25.00		
Method-C vs. Method-R	21.479	15	25.00		

Table 6
Measurement Invariance by GENDER



Model	Standard Met?	χ2	df	CFI	TLI	RMSEA
Omnibus Test of the Equality of the Covariance Matrices: The factor loadings, the error variances, and variances/covariances of the latent factors are all constrained to be equal.	No	397.749	225	.936	.931	.061
		Δ χ2=93.022	Δdf=47	Standard at p=.05 is 64.001		
Measurement Invariance Tests						
Configural Invariance/Baseline Model: The fit for this simultaneously estimated model provides the baseline value against which all subsequently specified models are compared, including the omnibus test described above.	Yes	304.727	178	.953	.963	.058
Metric Invariance ("weak" invariance): Involves constraining the factor loadings of like items to be equal across groups.	Yes	315.652	188	.952	.939	.057
		Δ χ2=10.925	Δdf=10	Standard at p=.05 is 18.31		
Scalar Invariance ("strong" invariance): Scalar invariance provides for a stronger test of invariance by introducing the concept of equal mean/intercept values for each observed variable.	No	348.076	204	.946	.937	.058
		Δ χ2=32.424	Δdf=16	Standard at p=.05 is 26.296		
Invariant Uniqueness ("strict" invariance): The invariance of the measurement error variances tested.	No	390.484	220	.936	.931	.058
		Δ χ2=42.404	Δdf=16	Standard at p=.05 is 26.296		
Structural Invariance Tests						
Invariance Factor Variances: Factor invariance may be done when comparing correlations of latent variables across groups.	No	318.006	184	.950	.935	.059
		Δ χ2=13.279	Δdf=6	Standard at p=.05 is 12.592		
Invariance Factor Covariances	No	346.833	199	.945	.933	.060
		Δ χ2=28.827	Δdf=15	Standard at p=.05 is 24.996		
Equal Factor Means	No	334.242 vs.348.076	204	.946	.937	.058
		Δ χ2=13.834	Δdf=6	Standard at p=.05 is 12.592		

Table 7
Study Research Hypotheses

Hypothesis	Observed Result (Unstandardized B)	Conclusion

H1 NHST: Subjective disconfirmation is positively related to consumer expectations.
H1 Effect Size: Subjective disconfirmation is positively related (i.e., correlated) with consumer expectations between the range $.002 \le \rho \le .042$.
	B = -.121 (p=.062, p_{rep} = .8616)
	95% CI = -.2482-.00064
	B = -.121 (p=.042)
	95% CI = -.382-.018
	B = -.147 (p=.229)
	95% CI = -.499-.276
		Expectations appear positively related to consumers' subjective disconfirmation judgments, although expectations fail to directly predict such judgments.

H2 NHST: Subjective disconfirmation is positively related to consumer perceptions of performance.
H2 Effect Size: Subjective disconfirmation is positively related (i.e., correlated) with consumer perceptions of performance between the range $.350 \le \rho \le .383$.
	B = .770 (p=.000, p_{rep} = .9856)
	95% CI = .6563-.8837
	B = 1.346 (p=.000)
	95% CI = 1.052-1.679
	B = 1.176 (p=.000)
	95% CI = .782-1.604
		Perceptions of performance appear to be both positively related to consumers' subjective disconfirmation judgments, and to directly predict such judgments.

H3 NHST: Satisfaction is positively related to consumer expectations.
H3 Effect Size: Satisfaction is positively related (i.e., correlated) with consumer expectations between the range $.261 \le \rho \le .304$.
	B = .121 (p=.07, p_{rep} = .8517)
	95% CI = -.0103-.2523
	B = .281 (p=.039)
	95% CI = -.035-.631
	B = 0.068 (p=.398)
	95% CI = -.421-.559
		Expectations appear positively related to consumers' satisfaction judgments; although expectations fail to directly predict such judgments.

H4 NHST: Satisfaction is positively related to consumer perceptions of performance.
H4 Effect Size: Satisfaction is positively related (i.e., correlated) with consumer perceptions of performance between the range $.178 \le \rho \le .712$.
	B = .356 (p=.001, p_{rep} = .9856)
	95% CI = .1874-5246
	B = .950 (p=.000)
	95% CI = .478-1.556
	B = 1.152 (p=.001)
	95% CI = .332-.1.955
		Perceptions of performance appear to be both positively related to consumers' satisfaction judgments, and to directly predict such judgments.

H5 NHST: Satisfaction is positively related to subjective disconfirmation.
H5 Effect Size: Satisfaction is positively related (i.e., correlated) with subjective disconfirmation between the range $.349 \le \rho \le .494$.
	B = .446 (p=.001, p_{rep} = ..9856)
	95% CI = .2814-6106
	B = .682 (p=.000)
	95% CI = .387-1.082
	B = .565 (p=.002)
	95% CI = .116-1.314
		Consumers' subjective disconfirmation judgments appear to be both positively related to consumers' satisfaction judgments, and to directly predict such judgments.

H6 NHST: Consumer repurchase intentions are positively related to satisfaction.	B = .510 (p=.001, p_{rep} = .9856)
	95% CI = .361-.659
	B = .257 (p=.000)
	95% CI = .151-.370
	B = .277 (p=.000)
	95% CI = .112-.483
		Consumers' intentions appear to be both positively related to consumers' satisfaction judgments, and to directly predict such judgments.

H7 NHST: Consumer relationship commitment is positively related to satisfaction.	B = .579 (p=.001, p_{rep} = .9856)
	95% CI = .4202-.7378
	B = .279 (p=.000)
	95% CI = .172-.410
	B = .515 (p=.000)
	95% CI = .273-.669	

Note: The order of presentation in the cell titled Observed Result (Unstandardized B) is Maximum Likelihood estimation followed by Bayesian estimation with no information priors, and finally Bayesian estimation with information priors.

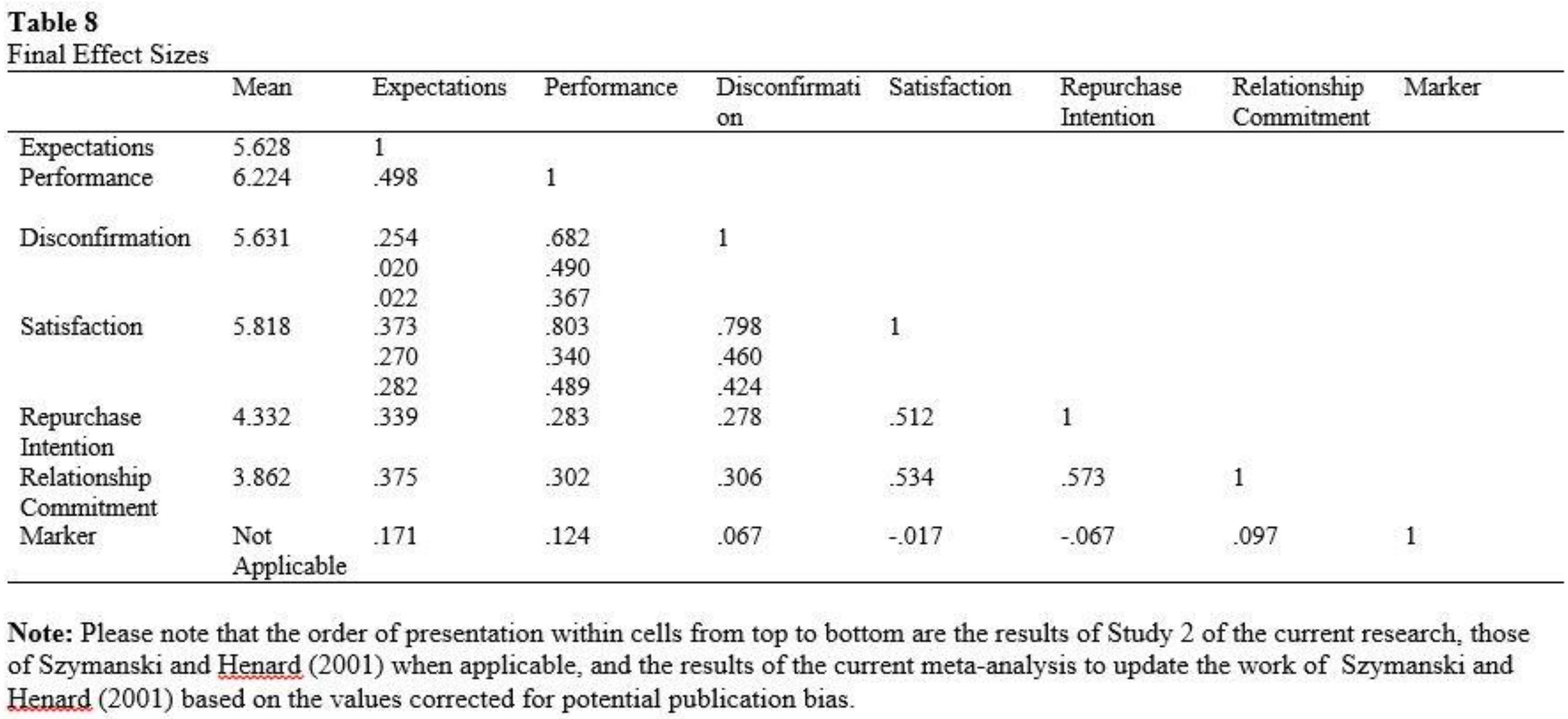


Table 8
Final Effect Sizes

	Mean	Expectations	Performance	Disconfirmation	Satisfaction	Repurchase Intention	Relationship Commitment	Marker
Expectations	5.628	1						
Performance	6.224	.498	1					
Disconfirmation	5.631	.254 .020 .022	.682 .490 .367	1				
Satisfaction	5.818	.373 .270 .282	.803 .340 .489	.798 .460 .424	1			
Repurchase Intention	4.332	.339	.283	.278	.512	1		
Relationship Commitment	3.862	.375	.302	.306	.534	.573	1	
Marker	Not Applicable	.171	.124	.067	-.017	-.067	.097	1

Note: Please note that the order of presentation within cells from top to bottom are the results of Study 2 of the current research, those of Szymanski and Henard (2001) when applicable, and the results of the current meta-analysis to update the work of Szymanski and Henard (2001) based on the values corrected for potential publication bias.

Study Measures

Concept	Question	Source	Reliability	Validity
Expectation$_{t-1}$	What is your prediction of the overall level of service that Wal-Mart will actually provide you? (1=Poor Service/8=Excellent Service) What is your prediction of the overall level of service you would consider to be reasonable, or that Wal-Mart *should* provide? (1=Poor Service/8=Excellent Service) DROPPED Overall, I would characterize my overall level of expectation of service from Wal-Mart as: (1=Very Low/8=Very High)	Parasuraman, Zeithaml, and Berry (1988) Boulding et al. 1993 Oliver 1997	.846	.734
Performance$_t$	The overall level of service that Wal-Mart actually provided to you was: (1=Poor Service/8=Excellent Service) The overall level of service that Wal-Mart should have provided to you was: (1=Poor Service/8=Excellent Service) Overall, I would characterize my overall evaluation of service from Wal-Mart as: (1=Very Low/8=Very High)	Parasuraman, Zeithaml, and Berry (1988) Boulding et al. 1993 Oliver 1997	.811	.619
Subjective Disconfirmation$_t$	The overall level of service that Wal-Mart actually provided to you was: (1=Much Worse Than I Expected/8=Much Better Than I Expected) The overall level of service that Wal-Mart should have provided to you was: (1=Much Worse Than I Expected/8=Much Better Than I Expected) DROPPED Overall, I would characterize my overall evaluation of service from Wal-Mart as: (1=Much Worse Than I Expected/8=Much Better Than I Expected)	Parasuraman, Zeithaml, and Berry (1988) Boulding et al. 1993 Oliver 1997	.939	.886
Satisfaction$_t$ (Three Parcels) 8-Point Likert Items	Sat 1 This is one of the best service encounters I could have experienced. This experience is exactly what I needed. I truly enjoyed this purchase experience. Sat 3 I feel bad about buying from Wal-Mart (-). I am NOT happy that I purchased from Wal-Mart (-). Sat 4 I was treated fairly when I went to Wal-Mart During my experience with Wal-Mart, I received the output that I needed relative to my input.	Oliver 1997 Created	.811	.376
Behavioral Intention$_t$ 8-Point Likert Scales	I intend to make gift purchases from Wal-Mart in the future. I plan to make additional gift purchases from Wal-Mart in the future. I intend to keep using Wal-Mart for my retail gift-giving needs even after the experiment.	Created	.871	.695
Relationship Commitment$_t$	I will actively seek to maintain my relationship with Wal-Mart as a retailer after the experiment. I am committed to a marketing relationship with Wal-Mart in the future. I will mention Wal-Mart to others quite frequently in the future.	Created	.791	.564
Marker (Resistance to Change)	My preference to use my current favorite retailer would not willingly change. It would be difficult to change my beliefs about my current favorite retailer. Even if others recommended another retailer to me, I would not change my preference for my current favorite retailer. To change my preference for my current favorite retailer would require major rethinking.	Pritchard et al. (1999)	.881	.649