Observation-Oriented Modeling: Going Beyond “Is It All a Matter of Chance”?

Educational and Psychological
Measurement

2017, Vol. 77(5) 855–867

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416667985

journals.sagepub.com/home/epm

James W. Grice¹, Maria Yopez¹,
Nicole L. Wilson², and Yuichi Shoda²

Abstract

An alternative to null hypothesis significance testing is presented and discussed. This approach, referred to as observation-oriented modeling, is centered on model building in an effort to explicate the structures and processes believed to generate a set of observations. In terms of analysis, this novel approach complements traditional methods based on means, variances, and covariances with methods of pattern detection and analysis. Using data from a previously published study by Shoda et al., the basic tenets and methods of observation-oriented modeling are demonstrated and compared with traditional methods, particularly with regard to null hypothesis significance testing.

Keywords

observation-oriented modeling, integrated model, inference to best explanation, null hypothesis significance testing

Introduction

Trafimow and Marks' (2015) recent ban on null hypothesis significance testing (NHST) in *Basic and Applied Social Psychology* has yet again brought this controversial procedure to the forefront. It was approximately 16 years ago that the American Psychological Association created a task force to discuss both the merits

¹Oklahoma State University, Stillwater, OK, USA

²University of Washington, Seattle, WA, USA

Corresponding Author:

James W. Grice, Department of Psychology, Oklahoma State University, 116 North Murray, Stillwater, OK 74078, USA.

Email: james.grice@okstate.edu

and weaknesses of the NHST procedure. Their conclusions were published in 1999 (Wilkinson, 1999), and at that time the task force did not call for an outright ban on p values. Instead, a more balanced approach toward data analysis was recommended, which included complementing NHST with effect sizes, confidence intervals, and graphs of the data. Serious criticisms of the NHST procedure have nonetheless continued to appear in the literature, with scholars arguing that its weaknesses far outweigh any benefits it might bring to scientific research (see Gigerenzer, 2004; Ioannidis, 2005; Lambdin, 2011; Wadman, 2013; Ziliak & McCloskey, 2008).

Regardless of the final, long-term outcome of the NHST debate, the information obtained from a traditional p value (e.g., in the context of an ANOVA or regression) is meager. In fact, in the most positive light imaginable the NHST procedure is one tool among many that could ostensibly be used as an aid for the detection of phenomena. Distinct from empirical observations (data), “phenomena are relatively stable, recurrent, general features of the world” (Haig, 2014, p. 33), and it is these phenomena scientists seek to detect and ultimately explain through their research. For example, consider the Flynn effect (Flynn, 2009). Over many decades, James Flynn noted increases in average scores on standardized tests of intelligence in a number of societies around the world. These observed increases were believed by some psychologists (although not without controversy) to point to a genuine phenomenon, namely, a general increase in human intelligence. According to Haig (2014), NHST could be used as an aid for detecting such phenomena in empirical data. However, in agreement with Hubbard (2015), he strongly prefers a different tool; specifically, the “significant sameness” approach that emphasizes confidence intervals over p values. Meta-analyses, which aggregate results from numerous studies, can also be regarded as important tools for phenomenon detection; and Bayesian statistics can fulfill this role in research as well by providing posterior probabilities for different hypotheses that can then be evaluated.

The NHST procedure is therefore one of several tools that could be used for phenomenon detection. What has become abundantly clear from the NHST debate, however, is that these tools can be extremely limiting if they are overemphasized or utilized in a simplistic, ritualized manner (see Gigerenzer, 2004; Gigerenzer & Marewski, 2015). First, they can result in a narrow view of science as an endeavor primarily concerned with the estimation of population parameters. The central goal of NHST and the so-called new statistics (Cumming, 2012) of confidence intervals and effect sizes is, after all, to provide estimates of population means, variances, covariances, or other parameters. Second, the tools of phenomenon detection listed above often depend on assumptions that are either dubious or rarely met in practice. These assumptions underlie the models used (e.g., linearity, interval, or ratio scaled measurement) as well as the computation or derivation of the probability values resulting from the analyses (e.g., assumptions of random sampling, distributional assumptions). Third, by focusing on population parameters, investigators can lose sight of the individuals in their studies, which then leads very often to confusion regarding the meaning of their results, for example, when between-person aggregate

statistics are interpreted at the level of the individual (see Lamiell, 2013). Finally, using a probability value and an arbitrary cut-point (e.g., .05) to determine whether or not results are worthy of interpretation not only distracts researchers from attending to the magnitudes of their observed results (effect sizes) but also distracts them from reasoning causally about the phenomenon they have detected or are seeking to detect.

Perhaps no other article in recent years has demonstrated these limiting influences of NHST and other tools of phenomenon detection than Bem's (2011) article on psi phenomena. In nine studies, he used a variety of methods to examine participants' prescient responses to random stimuli. Using NHST as his primary tool, he estimated the population parameter of correct, prescient responses to be greater than what would be expected by chance alone ($d = .22$, $p < .001$; Bem, Utts, & Johnson, 2011). By focusing on group averages and parameter estimation, the study was mute on the results at the level of the individuals, some of whom may have exceeded chance on a consistent basis and some of whom may actually have performed worse than chance on a consistent basis. Surely such differences would be important in studying psi phenomena? Moreover, given the nature of his studies randomization played a key role, not just in terms of the validity of the p values from NHST (which rely on the assumption of random selection or random assignment) but also in terms of the experimental methods. Yet as argued by Alcock (2011), it appears nonrandom bias may have entered into the studies in subtle ways that comprised their validity. Last, critics argued that Bem should have used Bayesian statistics (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011); and if he had done so, his data would have been interpreted as providing weak or nonexistent evidence of psi phenomena. Even the Bayesian analysis, however, does not go much beyond phenomenon detection, as Bem's entire approach precluded any serious effort to explain exactly why a standardized effect of .22 would be expected after averaging mountains of data across nine studies. Reasoning causally about psi phenomena at the level of the individuals in the study (where the causes must impart their effects) was precluded by the methodology and analysis procedures. Consequently, while overemphasizing probability in an effort to determine if Bem has in fact detected some phenomenon, the discussion has not risen above the question, "Is it all a matter of chance?"

Observation-oriented modeling was developed to help psychologists and other scientists go beyond these limited goals. In a practical way, observation-oriented modeling can be regarded as a set of analysis tools that are not centered on traditional aggregate statistics such as means, medians, variances, and correlations. Instead, like exploratory data analysis (EDA; Behrens & Yu, 2003; Tukey, 1977), these tools rely primarily on the visual examination of data to detect and explain dominant patterns within a set of observations. Using the OOM (observation-oriented modeling) software, for instance, investigators have shown how traditional t tests and chi-square goodness-of-fit tests can be replaced by simple and compelling pattern matching techniques (Grice, 2015; Grice, Barrett, Schlimgen, & Abramson, 2012), and how repeated-measures ANOVA can be replaced by a straightforward ordinal pattern

analysis technique (Grice, Craig, & Abramson, 2015). The observation-oriented modeling methods have moreover been shown to provide a number of practical benefits compared with traditional statistics, including transparency of results, immunity to outliers, relatively assumption-free analyses, and clear and readily interpretable “effect sizes.” In these ways, observation-oriented modeling can be considered as a novel tool for phenomenon detection.

More important, however, observation-oriented modeling requires researchers to seek explanatory inferences through model building rather than inferences to population parameters through NHST and Bayesian statistics. This model building is facilitated by moving beyond Hume’s truncated view of causation to Aristotle’s richer view of causality and moving from aggregates to individuals. These shifts in viewpoint, and several others (see Grice, 2014), essentially provide the framework for thinking formally about the underlying structures and processes of the phenomena. In other words, integrated into the OOM software is an approach toward conceptualizing and analyzing data that seeks to move beyond phenomenon detection to causal explanation. In this way, a largely statistical way of thinking gives way to an openness in research practice (see Trafimow, 2014) and to a more rigorous commitment to causes and their effects.

In summary, observation-oriented modeling is like EDA because it relies primarily on techniques of visual examination to detect dominant patterns within a set of observations. Going beyond EDA, however, observation-oriented modeling encourages researchers to explain—on an *a priori* or *post hoc* basis—the phenomena underlying their data, thus promoting model building and development. It also synchronizes visual examination of the data with transparent analyses that are person-centered, intuitive, and compelling. These tenets and features of observation-oriented modeling have been described and elaborated elsewhere, along with demonstrations of the OOM software, using data from previously published studies (Grice, 2011, 2014; Grice et al., 2012; Grice et al., 2015). Because this illustrative approach has been effective in highlighting the differences between traditional statistical methods and observation-oriented modeling, we similarly chose to reconceptualize and reanalyze data from a previously published study.

Example Study and Observation-Oriented Modeling Analysis

In a recent study, Wilson (2008) asked students to complete daily diary ratings regarding stressful events in their lives. The participants rated how stressful they considered each daily event to be using an 11-point (0 to 10) rating scale. They also rated each event using a list of 24 features of “internal situations” (e.g., thoughts and feelings being experienced in a given situation, such as feeling defeated, feeling anxious, fear of failure). Again, these ratings were made on 11-point scales. Drawing on the cognitive affective processing system (CAPS) theory (Mischel & Shoda, 1995), Shoda, Wilson, Chen, Gilmore, and Smith (2013) analyzed data from Wilson’s (2008) study to identify the features of internal situations stress.

Central to the CAPS theory is how each person construes the potentially stressful situation. Imagine two participants. The first person experiences situations (e.g., arguments) as stressful when they invoke feelings of incompetence and frustration, while the second person may find situations stressful when they invoke feelings of betrayal and discouragement. A consistent relationship between situation features and perceived levels of stress should then emerge, but they would differ with regard to the content of the specific situational triggers that bring about stress. Shoda et al. (2013) referred to these stable individual differences as “behavioral signatures,” and the goal of their statistical analyses was to identify these signatures (p. 557). More specifically, in Shoda et al. (2013), the analyses entailed individually predicting the ratings of stress from the situation feature ratings for each participant in the study using hierarchical linear modeling. The situation features yielding relatively high regression slopes that were distinctive for each person were considered to be potential triggers, or possibly consequences, of stress for a given participant.

Approaching Shoda et al.’s hypotheses and data from the vantage point of observation-oriented modeling, we begin with the construction of an integrated model. Figure 1 shows the cognitive and affective components hypothesized to underlie the observed responses for the hypothetical person described above who felt incompetence and frustration during arguments. As can be seen, the person is considered to have objectively argued with a coworker. Because CAPS emphasizes the subjective experience of the person, the model shows that the person construes (or judges) the interaction to have been negative. This judgment is represented by the pentagon surrounding the participant and coworker stick figures separated by a negative sign. The CAPS approach also emphasizes “if . . . then . . .” relationships among components in the model. With regard to the judged negative interaction, the participant simultaneously experiences negative feelings of incompetence and frustration that are represented as diamonds in the model. These feelings are linked to the negative judgment with double-headed arrows labeled as “Fo” to represent formal causes. Formal causes (see Grice, 2011, 2014; Rychlak, 1988) are generally concerned with pattern, shape, and structure, which may include logical structure. In this part of the model, the “if . . . then . . .” relationships between the negative judgment and negative feelings are implicative; specifically, if the interaction was judged as negative, then feelings of anxiety and inferiority were also experienced. The diamonds, which represent “feelings,” are enclosed in pentagons to represent the fact that the participant is asked to make a judgment about the feelings using an 11-point scale.

The simultaneous negative judgment and feelings then operate as an efficient cause of the participant predicating “I am stressed,” as denoted by the arrow labeled “Ef” in Figure 1. The predication is represented as a circle in the model that encloses the stick figure of the participant, and efficient causes are concerned with changes in structures and processes that occur over time (see Grice, 2011, 2014; Rychlak, 1988). The “if . . . then . . .” relation between the conjoint negative judgment/negative feelings and the predication of being a stressed person is thus considered to be an efficient cause, which is why Shoda et al. (2013) regarded stress as an outcome to be

Figure 1. Integrated model for Shoda et al.'s study.

predicted by the emotional signatures in their statistical analyses. The predication is moreover enclosed in a pentagon to again indicate that the participant must judge stressfulness using a 0 to 10 rating scale.

The integrated model represents the causal structures and processes presumed to underlie the observed ratings. Consistent with Figure 1, Shoda et al. considered the stress ratings to be the predicted effects in their hierarchical linear modeling analyses while the ratings of different feelings were treated as the predictors. Analysis methods in observation-oriented modeling revolve round the evaluation of patterns of raw observations rather than the examination of aggregate statistics such as means, variances, correlations, and regression weights. The simple “eye test” or more severe “inter-ocular traumatic test” (Edwards, Lindman, & Savage, 1963) therefore plays a central role in the analysis. Figure 2, for example, shows responses for Participant 9 from Shoda et al.’s study. As can be seen, this person kept 53 days of diary data. Features of internal situations (e.g., feeling anxious) and levels of stress were rated on 11-point scales (0 to 10), and the top panel of Figure 2 shows clearly that relatively high ratings of anxiety were paired with relatively high ratings of stress across the 53 days. Anxiety for Participant 9 therefore could be a trigger for, or efficient cause of, stress (or vice versa, or a reflection of a common, third variable(s)). The bottom panel shows the patterns of ratings for feelings of inferiority and for stress, and as can be seen, the two patterns do not match at all. Feelings of inferiority do not appear to be related to stress. Moreover, Participant 9 rarely indicated having any feelings of inferiority for a majority of the 53 days, which is important for the way

Figure 2. Anxiety, inferiority, and stress ratings for Participant 9.

these data are being analyzed, because the goal is to match the relative shapes of the patterns. The integrated model in Figure 1 is not sophisticated enough to generate exact, testable predictions regarding the magnitudes of the situation and stress ratings. The less precise goal here is to expect patterns in the situation and stress ratings to coincide across the 53 days. Figure 3, for instance, shows ratings of feeling overwhelmed and stress for Participant 5 who kept 63 days of diary data. The two patterns match in terms of their overall shape even though the stress ratings are almost always higher on the 11-point scale than the feeling overwhelmed ratings. For this person the feeling of being overwhelmed would therefore be considered as a potential trigger for stress. For comparison purposes, the bottom panel of Figure 3 shows ratings for feeling nervous, which is not related to stress for Participant 5.

The patterns in Figures 2 and 3 can be supplemented with a summary statistic referred to as the percent correct classification (PCC) index. Given that relative changes in ratings across the days were of primary interest, the PCC index is computed as a simple percentage of matches between the signs of differences in ratings. On the first day, for example, Participant 9 rated anxiety and stress as 8 on the scale (see Figure 1). On the second day, both ratings dropped, but not to the same value (3 for feeling anxious and 1 for stress) on the scale; then on the third day the ratings increased, but again not to an equal value. The relative differences on the scale (positive or negative) can be compared for each pair of these three days: Day 1 versus Day 2, Day 1 versus Day 3, and Day 2 versus Day 3. For all three pairwise comparisons, the signs of the difference scores matched, indicating perfect (100%) agreement. Extending this procedure across all 53 days, 916 of the 1,326 comparisons

Figure 3. Overwhelmed, nervousness, and stress ratings for Participant 5.

were tallied as matches (or correct classifications), yielding a PCC index equal to 69.08%. In other words, for all possible pairs of days the relative difference between the anxiety and stress ratings were frequently equal in sign, indicating consistent increases or decreases in both. It should be briefly noted that the total number of comparisons was computed as ${}_{52}C_2$ rather than ${}_{53}C_2$ because an anxiety rating was missing for the 43rd day. With a maximum of 100%, the PCC value of 69.08% is impressive, and as shown in Table 1, anxiety yielded the highest PCC index of the 24 “feelings” for Participant 9. The median PCC index was equal to 54.20%, and the lowest value was noted for feeling inferior (33.03%), the ratings for which are shown in the bottom panel of Figure 2.

Table 1 also reports what are referred to as *chance* values, or *c* values, for Participant 9. These values can be examined to shore up or clarify the visual examination of the observations (as shown in Figure 1) as well as the PCC indices. The *c* value is a probability computed on an entirely post hoc basis from a distribution-free randomization test (see Winch & Campbell, 1969). Specifically, the effect (stress) observations for Participant 9 are randomly shuffled across the 53 days and the PCC index is computed. This process is repeated for 1,000 trials (as determined by the investigator) and the resulting PCC values recorded in a frequency histogram. The number of PCC values from randomized versions of the actual observations that equal or exceed the PCC value from the original data is then determined and converted to a proportion, the *c* value. As a strict probability, low *c* values indicate more unusual, or rarely obtainable PCC indices from random shuffles of the actual observations. The *c* value for anxiety for Participant 9 (PCC = 69.08%) was less than .001, supporting the distinctiveness of the pattern in Figure 2 and the PCC index compared with randomized versions of the same data. Table 1 also shows, however,

Table 1. PCC Indices and *c* Values for Comparisons Between Ratings of Stress and 24 Rated Feelings (Participant 9).

Feeling/fear	PCC	<i>c</i> Value
Feeling anxious	69.08	<.001
Feeling incompetent	50.60	<.001
Feeling exhausted	54.75	<.001
Feeling behind	53.65	.002
Feeling defeated	46.12	.008
Feeling discouraged	48.47	.028
Feeling excluded	42.28	<.001
Feeling helpless	60.86	<.001
Feeling inferior	33.03	.062
Feeling nervous	65.23	<.001
Feeling overwhelmed	60.78	<.001
Feeling confused	49.26	.013
Feeling rushed	65.18	<.001
Feeling frustrated	59.29	<.001
Feeling irritated	55.92	<.001
Feeling self-doubt	51.29	.004
Feeling uncertain	47.37	.075
Feeling demand from others	64.33	<.001
Feeling time is wasted	42.67	.134
Feeling betrayed	40.06	.002
Feeling expectations were violated	49.26	.006
Fear of failure	59.58	<.001
Fear of letting others down	62.43	<.001
Fear of being viewed by others as incompetent	60.03	<.001

that even low PCC values were unusual. As with traditional *p* values, randomization tests will almost always yield small probabilities (*c* values in OOM) when the sample sizes are large, as with the 1,326 comparisons for the anxiety and stress ratings. Thus, while they provide assurance that the results are not easily obtained by randomizing the data, the *c* values should not be used as the basis for determining which effects are of theoretical and practical significance. Instead, the eye test and PCC indices are primarily relied on to interpret the feelings and stress rating patterns like those shown in Figure 2.

Analyses for each of the participants in Shoda et al.'s study would thus proceed by examining the patterns of feeling and stress ratings like those shown above for Participant 9. The PCC indices would provide summary statistics indicating how well the patterns were matched with regard to the signs of differences in magnitudes across the daily ratings, and they could be used to identify the strongest and weakest efficient causes of stress. The *c* values could also potentially be used if they varied highly across the 24 feelings for each participant. Solely for the sake of summarizing the observation-oriented findings for all 13 participants, however, we considered patterns with PCC indices of at least 60% (arbitrarily chosen) to indicate strong

connections between the feeling and stress ratings. These patterns were also visually examined to ensure strong correspondence between ratings. Using these criteria, a convincing trigger was not observed for 5 of the 13 participants. At least one convincing trigger was observed for the remaining eight participants, and the ratings for one participant indicated that 19 of the 24 situation features were clearly related to stress! The two most common efficient causes of stress, discovered for four participants, were feelings of discouragement and feelings of being overwhelmed. Feelings of failure, exhaustion, being behind, frustration, and anxiety were each found to be potential efficient causes of stress for three persons in the sample of 13 persons.

Discussion

Without computing a single mean, standard deviation, variance, or correlation, not to mention a regression coefficient or standard error in a complex regression analysis, the triggers for stress were clearly identified for 8 of the 13 people in Shoda et al.'s study. Moreover, the NHST paradigm with a null hypothesis entailing assumptions about a parameter (e.g., mean, regression weight, correlation) for some arbitrarily defined population was not used. Instead, the analysis proceeded primarily through visual examination of the observations themselves in an effort to detect a theoretically meaningful pattern. The PCC index, a summary statistic, was used as an aid in evaluating the patterns within each person's diary ratings. Lastly, the c value, a probability statistic, was also considered in an effort to shore up the interpretations of the patterns and PCC indices. As expected, given the large number of observations per participant, the computed c values were often very small even when the PCC indices did not appear to be particularly large. Thus, while the c values were helpful in identifying results not easily obtained from randomized versions of the data, they were not relied on as a basis for determining the theoretical and practical significance of these results.

Unlike the widespread and unwarranted use of the p value in NHST as a criterion for "significance," the probability statistic in observation-oriented modeling is used only as supplementary evidence and only when it is unclear that the results are not easily obtained through random pairings of the observations. The c values from the OOM software are derived from randomization tests which, as noted by Howell (2015), are free from assumptions underlying traditional parametric statistical analyses (e.g., normal population distributions, equal population variances). Also distinct from NHST, the goal of observation-oriented modeling is to draw an *inference to best explanation* (Haig, 2005, 2014), which is embodied in the integrated model in Figure 1 of this article. The ultimate goal is to construct a causal model, showing in an explicit way the structures and processes underlying the observations. Causes and their effects are therefore central to how data are approached and understood, which stands in contrast to considering the data primarily through a conceptual framework comprising stochastic processes and random variables. Ideally, the integrated model

would be created prior to designing the study, as it would then provide more exact predictions regarding the patterns within the observations.

For Shoda et al.'s study, the patterns of observations like those shown in Figures 2 and 3 could be meaningfully examined using the eye test and PCC indices. The PCC index is like the effect size one finds in the so-called "new statistics" that have been added to the NHST approach (Cumming, 2012). With a range of 0% to 100%, the PCC index is easily understood, but we cannot emphasize enough that in observation-oriented modeling the PCC index is never to be regarded as a stand-alone statistic. It can only be interpreted unambiguously in the context of the graphed pattern of data, and in this article, it was used as a rough guide to help identify visually compelling graphs. The "eye test" is therefore of utmost importance because it focuses the investigator's attention on the actual observations as they are organized within the framework of the units of observation. The units of observation themselves are determined by the researcher on the basis of an integrated model or on the basis of a more general theory or set of assumptions.

Shoda et al. used the CAPS model (Mischel & Shoda, 1995) to reason that stress vulnerability signatures exist and also vary across individuals. The integrated model in Figure 1 provides a more specific explication of the structures and processes comprising the diary ratings, and it does not assume that the variables are measured on an interval scale, consistent with the argument. "There is no evidence that the attributes that psychometricians aspire to measure (such as abilities, attitudes and personality traits) are quantitative" (Michell, 2011, p. 245). The OOM software can be used regardless, because observation-oriented modeling does not rely on the assumption of continuous quantities. Moreover, the "if . . . then . . ." relationships posited by the CAPS model were represented as formal and efficient causes in Figure 1. Are the causes in fact accurately identified in the integrated model, and do they jibe with Mischel and Shoda's (1995) understanding of "if . . . then . . ." relations? As explained by Rychlak (1988), Aristotle's four causes open up new frontiers for psychological science because they permit researchers to think in more complex ways about natural systems. Grice (2011) describes how the four causes can be seen in explanatory models from other sciences, including the model of the atom and biochemical models of cellular function. Bringing the four causes into the CAPS model could therefore expand its breadth as well as improve its validity in explaining the human psyche and behavior. Nonetheless, given they are correlational, the current data cannot distinguish among the posited formal and efficient causes in Figure 1. As noted above, efficient causes are concerned with changes in structures and processes that occur over time (see Grice, 2011, 2014; Rychlak, 1988). Ideally, then, observations should be ordered in time, with the hypothesized causes preceding the effects. In this way, the integrated model can be regarded as critical for designing future studies to accurately test the proposed causes.

In summary, drawing on the CAPS model, Shoda et al. posited the existence of stress vulnerability signatures that could be observed to vary across individuals. The overarching goal of their research was to determine how these signatures could be detected so that, in a clinical setting, individualized treatment plans could be

developed to reduce client stress. Recognizing that the focus of Shoda et al.'s study was the patterns observed within each individual rather than across individuals, and working within the framework of observation-oriented modeling, we created an integrated model of the causes and effects. Using the OOM software, we then examined the data at the appropriate level of analysis (i.e., the level of the persons) using simple, yet compelling, graphing techniques. A summary statistic, the PCC index, was also used to help identify the triggers of stress for each person in the sample. A distribution-free probability statistic was lastly considered to provide assurance that the results could not easily be obtained from randomized versions of the same data. However, for evaluating the theoretical and practical significance of the results, the eye test and PCC indices were primarily relied on. What we accomplished, then, is consistent with the Gestalt shift occurring in personality science, moving from the traditional focus on averages across individuals to patterns of variations observed *within* each person.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Alcock, J. (2011, March). Back from the future: Parapsychology and the Bem affair. *Skeptical Inquirer*. Retrieved from http://www.csicop.org/specialarticles/show/back_from_the_future
- Behrens, J. T., & Yu, C.-H. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology* (Vol. 2, pp. 33-64). New York, NY: Wiley.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407-425.
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101, 716-719.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Flynn, J. R. (2009). *What is intelligence: Beyond the Flynn effect*. Cambridge, England: Cambridge University Press.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587-606.
- Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41, 421-440.
- Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. New York, NY: Academic Press.

- Grice, J. W. (2014). Observation oriented modeling: Preparing students for the research in the 21st century. *Innovative Teaching*, 3, 3. doi:10.2466/05.08.IT.3.3
- Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology*, 6, 1007. doi:10.3389/fpsyg.2015.01007
- Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences*, 2(1), 1-22.
- Grice, J. W., Craig, D. A., & Abramson, C. I. (2015, September 8). A simple and transparent alternative to repeated measures ANOVA. *Sage Open*. doi:10.1177/2158244015604192
- Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, 10, 371-388.
- Haig, B. D. (2014). *Investigating the psychological world*. Cambridge: MIT press.
- Howell, D. (2015, June 28). *Randomization test using R*. Retrieved from <https://www.uvm.edu/~dhowell/StatPages/R/RandomizationTestsWithR/RandomizationTestsR.html>
- Hubbard, R. T. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. Thousand Oaks, CA: Sage.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Lambdin, C. (2011) Significance tests as sorcery: Science is empirical—Significance tests are not. *Theory & Psychology*, 22(1), 67-90.
- Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology*, 31, 65-67.
- Michell, J. (2011). Qualitative research meets the ghost of Pythagoras. *Theory & Psychology*, 21, 241-259.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246-268.
- Rychlak, J. (1988). *The psychology of rigorous humanism* (2nd ed.). New York: New York University Press.
- Shoda, Y., Wilson, N. L., Chen, J., Gilmore, A. K., & Smith, R. E. (2013). Cognitive-affective processing system analysis of intra-individual dynamics in collaborative therapeutic assessment: Translating basic theory and research into clinical applications. *Journal of Personality*, 81, 554-568.
- Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, 36, 1-2.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1-2.
- Tukey, J. W. (1977). *Exploratory data analysis*. New York, NY: Pearson.
- Wadman, M. (2013). NIH mulls rules for validating key results. *Nature*, 500, 14-16.
- Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100, 426-432.
- Wilkinson, L. (1999). Task force on statistical inference. Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, 54, 594-604.
- Wilson, N. L. (2008). *Identifying the features of stressful situations* (Unpublished doctoral dissertation). University of Washington, Seattle.
- Winch, R., & Campbell, D. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, 4, 140-143.
- Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice and lives*. Ann Arbor: University of Michigan Press.