Releasing the death grip of null hypothesis testing

Releasing the death-grip of null hypothesis statistical testing (p < .05): Applying complexity theory and somewhat precise outcome testing (SPOT)

Arch G. Woodside

JOURNAL OF GLOBAL SCHOLARS OF MARKETING SCIENCE, 2017
VOL. 27, NO. 1, 1–15
http://dx.doi.org/10.1080/21639159.2016.1265323

Releasing the death-grip of null hypothesis statistical testing ($p < .05$): Applying complexity theory and somewhat precise outcome testing (SPOT)

Arch G. Woodside

School of Marketing, Curtin School of Business, Curtin University, Perth, WA, Australia

ABSTRACT

Even though several scholars describe the telling weaknesses in such procedures, the dominating logic in research in the management sub-disciplines continues to rely on symmetric modeling using continuous variables and null hypothesis statistical testing (NHST). Though the term of reference is new, somewhat precise outcome testing (SPOT) procedures are available now and, along with asymmetric modeling, enable researchers to better match data analytics with their theories than the current pervasive theory–analysis mismatch. The majority (70%+) of articles in the leading journals of general management, marketing, finance, and the additional management sub-disciplines are examples of the mismatch. The mismatch may be a principal cause for the scant impact of the majority of articles. Asymmetric modeling and SPOT rests on the principal tenets of complexity theory rather than overly shallow and simplistic symmetric modeling and reporting of NHST findings. Though relatively rare, examples of asymmetric modeling and SPOT are available now in the management literature. The current lack of instructor knowledge and student training in MBA and PhD programs of asymmetric modeling and SPOT are the likely principal reasons for this scarcity.

零假设统计检验（$p < .05$）将被取代：应用复杂理论精准结果测试（SPOT）

即使一些学者描述了一些程序中的弱点，管理子学科研究中的主导逻辑仍然依赖于使用连续变量和零假设统计检验（NHST）的对称模型。虽然参考信息是新的，但现在已经有一些精准的结果测试（SPOT）程序，连同非对称建模一起，和i研究人员能够哼好地匹配数据分析与其理论，而不是当前普遍的理论——分析不匹配。管理、营销、金融和其他管理学子学科中绝大多数（70%+）主要领导期刊中的文章都是不匹配的例子。不匹配可能是大多数文章的缺乏影响力的主要原因（精英排名并不是精英期刊排名）。不对称建模和精准的结果测试（SPOT）取决于复杂理论的主要理论，而不是过于简单的对称建模和零假设统计检验(NHST)的发现报告。尽管相对少见，不对称建模和精准结果测试的例子在管理文献中可以找到。目前教师知识缺乏以及博士和硕士课程建设的不全面是不对称建模和精准结果测试稀缺的主要原因。现在可用基于布尔代数的计算式软件，新的理论倡导不对称建模和精准结果检测取代零假设统计检验将在2020年到来。

ARTICLE HISTORY
Received 4 September 2016
Revised 9 September 2016
Accepted 12 September 2016

KEYWORDS
Asymmetric; continuous; discretize; NHST; SPOT; symmetric

关键词
不对称; 连续; 离散化; 零假设统计检验; 精准结果测试; 对称

CORRESPONDENCE TO Arch G. Woodside arch.woodside@curtin.edu.au

Introduction

Paraphrasing Gigerenzer (1991) and Kerlinger and Pedhazur (1973), tools-in-use influence theory construction as well as attempts to extract information from data. The easy-reading first edition of Kerlinger and Pedhazur (1973) emphasizes that multiple regression analysis (MRA) is a manner of thinking – theory construction – and not just a data analytic tool. Gigerenzer (1991) fully develops this thesis. Kerlinger and Pedhazur's (1973) book is highly useful for learning the nitty-gritty steps in today's dominant logic of symmetric theory construction and testing in behavioral and management research, that is, MRA. Studies by Gigerenzer (1991) and others (Gigerenzer & Brighton, 2009; McClelland, 1998; Ragin, 2008) are highly useful for learning the nitty-gritty steps in behavioral research that go beyond MRA to the algorithm-based asymmetric-theory construction and testing. Unlike symmetric modeling and analysis, such asymmetric modeling builds on a foundation of complexity theory, rather than linear theory, and Boolean algebra, rather than matrix algebra. While appearing widely in sociology and political science (see available articles and working papers at www.COMPASSS.org), in the teen years of the twenty-first century, research showing the advantages of asymmetric thinking, modeling, and analysis is only starting to appear in the management sub-discipline journals and books (e.g. Feurer, Baumbach, & Woodside, 2016; Fiss, 2011; Frösén, Luoma, Jaakkola, Tikkanen, & Aspara, 2016; Hsu, Woodside, & Marshall, 2013; Ordanini, Parasuraman, & Rubera, 2014; Woodside, 2016, 2017; Woodside, de Villiers, & Marshall, 1916).

Embracing the complexity turn

Taking the complexity turn in theory construction and analysis includes embracing a paradigm shift away from symmetric thinking and testing (e.g. analysis of variance, multiple regression analysis, and structural equation modeling) to thinking and testing asymmetrically via creating and testing algorithms. Symmetric research focuses on proposing and reporting findings via null hypothesis statistical testing (NHST) such as the relationship between an independent variable (X) and a dependent variable is linear or curvilinear, and significant statistically ($p < .01$) to permit rejection of the null hypothesis. Symmetric testing sometimes reports the "effect size of the relationship" being small, medium, or large, where X can be a simple variable (e.g. age, price, trust, size) or a regression model made up of several simple X's. While the relevant literature (Armstrong, 2012; Cohen, 1994; Hubbard, 2016; Gigerenzer & Brighton, 2009) thoroughly discredits the accuracy and usefulness of symmetric testing that relies on NHST, symmetric theory construction and NHST continue to be pervasive and the dominant logic in all of the sub-disciplines of management.

Telling weaknesses of symmetric modeling and testing in the management sub-disciplines

Here are some problems with relying on symmetric theory construction and testing. (1) In practice (reality) almost all relationships are significant statistically if the number of cases in a study is very large (e.g. $n \geq 1{,}000$). (2) Also an observed relationship can have a large effect size but contrarian cases are usually observable; for example, in a study data analysis might support a highly significant positive, linear relationship showing the probability of an observed correlation of $r = .50$, when the true correlation is zero is less than .001 ($p < .001$) and the "coefficient of determination" ($r^2 = .25$) indicates a reasonably large effect size estimate, but at the same time 6% or so of cases with low scores for X have high scores for Y and 4% or so of cases with high scores for Y have low scores for X. The usual practice in research is to ignore the presence of such contrarian cases. (3) While symmetric models are more accurate in fit validity than asymmetric models because they use more information on the associations of variations of the independent variables and the dependent variable, symmetric models are less accurate in prediction validation – predicting the scores for an outcome condition (here think: dependent variable) of cases in separate samples of cases (Gigerenzer & Brighton, 2009). Symmetric modeling via MRA is so powerful that fit validation of such models is usually high even if random numbers are used for the scores of the independent variables (Armstrong, 2012).

Almost all (95%+) studies published in the leading behavioral and management science journals ignore predictive validation even though examining predictive validation is crucial in estimating the usefulness of an empirical model. (4) Simple rudimentary screens such as taking the cases that manage to be in the top quintiles of three-plus independent variables as the cases that will have high scores in an outcome condition and rudimentary algorithm models (RAMs) in fuzzy-set qualitative comparative models (fsQCA) as asymmetric theories/tools outperform symmetric theory/tools in predictive validation (e.g. McClelland, 1998; Ordanini et al., 2014; Ragin, 1997). (5) The use of variable-based tools such as symmetric analysis is a mismatch in testing case-based identification models; most theory constructions in behavioral and management science are case-based proposals (Fiss, 2007, 2011).

Seeing both the trees and the forest: the value of information varies in different ranges of values in a distribution

Embracing the idea that not all variance informs equally is a foundational tenet in building and testing asymmetric models. Such an idea goes beyond transforming an original scale of values to work with scales having the same measurement properties or the use of logarithmic transformations. An elite university applying the following complex antecedent configuration as a decision rule is ignoring averages and standard deviations in two variables: accept students for admission in the university who are in the top quintile of their high school graduation class "AND" who are in the top quintile of the Scholastic Aptitude Test (2016). "AND" is the Boolean algebra logical condition that requires both simple conditions to be met. Ragin (2008) provides a full discussion on the thinking and analytic steps in "calibrating" a variable scale into a discretized calibrated scale ranging from a dichotomous membership (e.g. top 20% = 1.0, bottom 80% = .0) to a fuzzy-set membership (.0 to 10) scale. A fuzzy membership score attaches a truth value, not a probability, to a statement (e.g. the statement that a high school graduate is in the set of admissible students at a top-20 ranked university):

> A membership score of 1.0 indicates full membership in a set; scores close to 1.0 (e.g., 0.8 or 0.9) indicate strong but not quite full membership in a set; scores less than 0.5 but greater than 0.0 (e.g., 0.2 or 0.3) indicate that objects are more "out" than "in" a set, but still weak members of the set; a score of 0.0 indicates full non-membership in a set. The 0.5 score is also qualitatively anchored, for it indicates the point of maximum ambiguity (i.e., fuzziness) in the assessment of whether a case is more in or out of a set. (Ragin, 2008, p. 30)

The full membership set of financially wealthy Americans might be set to include American households having net wealth above $1 million in 2013 USD – a net worth for 10% of US households in 2013 (2013 Survey of Consumer Finances, 2016). Two cases (households) with a respective net worth of $2 million and $4 million USD would vary little in their membership scores for high net wealth – each would have a net worth score equal to 9.5 and 9.7, respectively. Note that the fuzzy-set calibrated score is for "high net worth" and not just "net worth." While value transforming to a distribution free scale (i.e. Z transformation) and logarithmic transformations are variable-based analytical operations, calibrated scoring is a case-based analytic operation.

Early attempts to solve the telling weaknesses in symmetric modeling and analysis

Bass, Tigert, and Lonsdale recommend eliminating variance with cross-tab cells

In the 1960s Bass, Tigert, and Lonsdale (1968) described the problem with the low explanatory power of symmetric tests. They offer a unique solution that was, and continues to be, ignored widely. While more compelling and useful solutions are available, the solution by Bass et al. (1968) illustrates the step of reducing variability in variables to increase explanatory power while maintaining use of the dominant paradigm of symmetric analysis. Bass et al. (1968) report that in their regression analysis of the data in their Table 7 (reproduced as Table 1 here) they assigned the mean score of each case to all households in the cell, thus eliminating the variation in beer purchases within each cell. For their regression analysis, Bass et al. weighted each cell mean by the sample size in each cell; the total number of cases in Table 1 and Bass et al.'s (1968) Table 7 equals 1400. Thus, for the first cell in Table 1, 124 cases are assigned the beer purchases equal to 10.01. Bass et al. (1968) eliminated all variations inside each cell to reduce noise of individual cases in their variable-based data analysis. This elimination of within-cell-variability procedure serves to more than triple the size of the explanatory power of regression models usually estimated using data from individual cases – from adjusted R^2's equal to .05 to .15 to adjusted R^2's equal to .35 to .50.

However, the Bass et al. (1968) procedure has severe shortcomings. The procedure Bass et al. (1968) use eliminates more than just noise in each cell. The procedure eliminates the possible extraction of additional information from each cell in the crosstab. In Table 1, most cells, including the two cells having the highest mean beer purchases, include non-beer purchase cases, that is, cases contrary to the main finding of high and low beer purchases occur in nearly every cell. For example, consider the cell having the highest mean beer purchases (12 oz. bottles per month) – bottom left corner in Table 1 – the cell with highest income and lowest education level. The cell mean equals 36.58; the standard error (se) equals 10.68 for the 7 cases in this cell. Consequently, the standard deviation (s) equals 28.26 ($s = se{*}n\wedge^{.05}$). The large value for s relative to the mean indicates that multiple cases with very low non-beer purchases occur in this cell. Rather than noise, a case identification hypotheses (CIH) perspective seeks to model both contrarian as well as supportive cases to the hypotheses. Who are the contrarian consumers to the finding of very high beer purchases among the cases who are very high in income and very low in education? Answering this question is possible but regression analysis is a poor tool for extracting such information.

Table 1. Cross-classification analysis of beer purchase of 1400 households, education by income.

Annual family income		Years of education 6	10	12	14	16
Under $3	Mean	10.01	6.53	6.18	12.27	15.21
	s.e.	1.64	2.38	1.74	7.78	13.17
	n	124	36	38	7	4
$3–4.999	Mean	27.74	20.27	11.70	17.40	1.79
	s.e.	3.93	4.03	2.98	9.51	1.65
	n	48	38	45	14	7
$5–7.999	Mean	25.23	26.03	22.63	24.27	16.80
	s.e.	2.62	2.45	1.85	3.58	3.85
	n	115	122	196	57	35
$8–9.999	Mean	27.72	24.21	32.14	21.78	23.23
	s.e.	6.47	4.51	3.07	3.92	3.78
	n	30	56	88	32	30
$10–14.999	Mean	34.24	24.05	21.54	30.63	24.18
	s.e.	6.47	4.51	3.07	3.92	3.78
	n	15	37	61	45	50
$15+	Mean	36.58	12.50	28.49	34.17	17.86
	s.e.	10.68	.00	6.93	8.52	3.80
	n	7	1	15	10	37

Notes: Data are for 1964; s.e. = standard error of the mean; *n* = number in sample.
Source: Data are from Table 7 in Bass et al. (1968, p. 270).

Figure 1 shows the findings of applying the findings in the Bass et al. (1968) data to asymmetric testing of one complex configuration of three antecedent conditions. The three conditions: gender: male; income: high; education: low, and beer consumption: high. The Boolean equation for this model appears as follows:

$$\text{gender} \bullet \text{income}\bullet \sim \text{education} \leq \text{beer\_consumption (Model)}$$

where the mid-level dot, "•" indicates the logical "AND" condition, male is a high score for gender (1) and female is dummy calibrated (0), and the sideways tilde sign, "~" indicates the negation of education (i.e. 1 minus the calibration score for education for each case).

The findings in Figure 1 include a high consistency index (C1) and moderate level of coverage (C2) for the model. Example computations of these two indexes appear in Ragin (2008). High consistency (C2 = .881 in Figure 1) indicates that most cases with high scores on the following statement have high scores on the outcome condition: male AND high income AND low education. Note in Figure 1 that close to half of the cases (i.e. dots) in the XY plot have high scores on the outcome (heavy beer consumption) – two to four alternative models are necessary to construct for useful complex configurations describing heavy beer consumption – an example here of the equifinality tenet. Figure 1's findings support the conclusion that the model is a somewhat, but not completely, precise outcome model; 2 of the 11 cases with high scores on the recipe model do not have high scores on the outcome model. A useful rule of thumb to adopt is that models should achieve a consistency index equal to, or above, .85 to conclude that the model is useful for predicting high scores in the outcome. Ragin (2008) recommends adopting a rule that consistency should be equal to or greater than .80. Of course, reporting models as useful should depend on achieving similar high consistency levels for the model in tests using separate sets of data – separate from the data used to test the model for fit validity. Also, such asymmetric testing should include constructing and reporting XY plots to show that high-consistency models are useful – with recipe scores of some range greater than .0 to .2 for example.

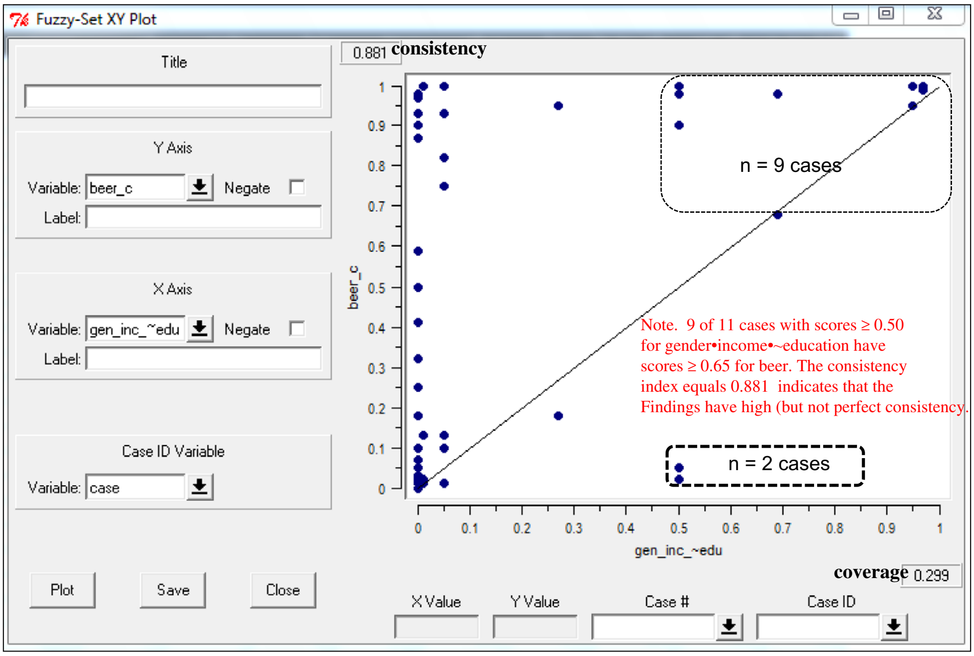

Figure 1. XY plot for complex antecedent model: gender•income•~education. Note: Model achieves high consistency: 9 of 11 cases with scores ≥ .50 for the model have high scores for calibrated beer consumption (consistency = .881).

McClelland's multiple quintile combination algorithms

McClelland (1998) offered a way forward in solving the dilemma of generalizing beyond the individual case while maintaining viewing the individual case in the output of a model – staying sufficiently close to see the trees while viewing the entire forest. In a human resources study to estimate future highly competent executives, McClelland (1998) ran a series of additive regression models using the standardized competencies to predict different dependent variables (e.g. annual bonus awards to executives and O (outstanding) versus T (typical) memberships were dependent variables). McClelland reports that the additive regression models did not yield stable results. In one regression analysis, the previously validated competencies yielded a multiple R of .52 ($p < .10$) with the bonus criterion [dependent variable]; however, in a second sample of 42 similar executives, the same regression formula failed to predict the same criterion (R = .24, not significant). In this latter sample, a regression model based on different competencies did predict the criterion (R = .58, $p < .01$), but this model yielded an R of only .08 when applied back to the original sample (McClelland, 1998, p. 334). For each sample of executives, McClelland first tested a model for its fit validity and then tested the accuracy of the model on an additional sample of executives for predictive validity. He also tested a model developed using data in a different sample of executives for fit validity and then tested the model for predictive validity on the first sample. This dual testing for fit and predictive validation on two samples is "cross-validation" of a model.

McClelland (1998) observed that many relationships between success (a unidirectional concern) and the frequencies of different competencies were not linear and these relationships are not described well by correlation coefficients (a symmetric – two-directional – test of association). Instead "tipping points" (Gladwell, 1996) describe these relationships well, that is, the identification of an outcome of interest associates consistently only with certain configurations of levels of two or more antecedent conditions. McClelland observed that most (55%) outstanding (O) executives displayed 8+ different competencies in behavioral event interviews (BEIs). He also observed that few (20%) typical (T) executives displayed 8+ different competencies in the BEIs. McClelland (1998, p. 334) recognized the necessity of conducting predictive validation using additional samples: "The critical question was whether the competency algorithm predicted actual performance in additional samples." Thus, using the 8+ competency algorithm he was able to correctly identify most top performers in three separate samples of top performers: 11 of 17, 9 of 11, and 6 of 7 outstanding executives. In additional samples, McClelland reports that most typical executives exhibited fewer than 8 competencies in the BEIs in each of 3 samples: 23 of 25, 7 of 8, and 10 of 11 typical executives had scores below 8 combined competencies in the BEIs. Thus, the specific algorithm – "if the executive exhibits 8+ competencies, predict the executive's performance to be outstanding" – was found to have high accuracy (not perfect) in two-directional tests in predicting Os and Ts in additional samples beyond the sample used to create the algorithm. The algorithm replaces the use of regression analysis and avoids the use of estimating weights for the individual 11 competencies.

McClelland (1998) reports that an additional examination of BEI data for different senior executives of a global firm resulted in his construction of the following somewhat more complex algorithm of BEI competencies. For identifying a senior executive as an O, the senior executive had to exhibit at least 1 of 3 individual-initiative competencies, at least 1 of 3 organizational-skill competencies, and a total of 6 out of 12 competencies that either most commonly differentiate significantly between O and T executives or are unique to the firm. This complex algorithm represents a unidirectional hypothesis of a case-identification theory. An executive who achieves the three parts of this algorithm will perform at the O level. This statement indicates sufficiency but not necessity; some other executives not achieving the three-part algorithm may perform at the O level and some may perform at the T level; thus, the claim is not made that achieving all parts of the algorithm is necessary for future O performance. The algorithm states that achieving all three parts of the algorithm is sufficient for identifying high O executives consistently. Given that some executives are likely to have high future performances who do not achieve all three parts of the algorithm, more than this algorithm alone is necessary for identifying these other high-performing executives. An algorithm is but one screening mechanism for identifying Os; other screening mechanisms may work equally well or better.

Decision-makers use screens pervasively in real-life contexts rather than using compensatory rules for identifying viable cases for solving choice problems. Configurations of screening rules are algorithms. Here is an example of one screening rule for selecting common stocks that will maintain or increase their present stock share price in future time periods: screen for stocks paying a dividend between $.05 and $.20 per share AND having a price/earnings ratio last year below 18 AND having a share price less than $30 AND have a "Buy" recommendation from more than two leading stock advisors AND have a price increase equal to or greater than 5% in the last four weeks. Notice this algorithm focuses on the identifying case outcomes and not whether or not any one independent variable associates with low and high variability in a dependent measure. Algorithm construction is one form of statistical sameness testing (SST). Constructing stock screens and other screens (i.e. recipes) represents fuzzy logic applications and "computing with words" (Zadeh, 1996).

The algorithm method that McClelland (1998) describes is a dramatic paradigm shift away from relying only on symmetric tests (e.g. regression analysis). While McClelland (1998) does not offer a general theory for this shift, his insights and data analytics are useful steps toward more formal theory construction using asymmetric reasoning. Rather than testing using NHST, McClelland (1998) tested using what Hubbard (2016) refers to as "statistic sameness testing" (SST) where the focus is on constructing models that identify a particular outcome consistently. The shift from NHST to SST thinking, theory construction, and extraction of information from data is an enormous help in stepping out of the quagmire of most behavioral and management science being ignored by everybody – including other researchers (see Pham, 2013 for a discussion of this quagmire).

Ragin's third step forward toward useful SPOT

Ragin (1997, 2008) provides the next step forward in case-based modeling without eliminating, but reducing, the variability relevant for useful analysis to predict both expected and contrarian relationships. Ragin (2008, p. 29) observes that one reason social scientists are reluctant to study social phenomena in terms of set relations is that they think that the study of set relations is restricted to nominal-scale measures. Not only are such scales considered "primitive," but interval and ratio scales that have been recoded to nominal scales (and thus "downgraded") are almost always suspect. Has a researcher selected cut-points in a biased way, to favor a particular conclusion? Fortunately, a well-developed mathematical system is available for addressing degree of membership in sets: fuzzy-set theory (Zadeh, 1965). Fuzzy sets are especially powerful because they allow researchers to calibrate partial membership in sets using values in the interval between .0 (non-membership) and 1.0 (full membership) without abandoning core set theoretic principles and operations (e.g. the subset relation). Set relations are central to social science theory, yet the assessment of set relations is outside the scope of conventional correlational methods (Ragin, 2008, p. 29).

In fuzzy-set analysis, all original values for variables are calibrated into membership scores ranging from .00 to 1.00 using a logarithmic function whereby a researcher specifies a "full-membership calibrated score" equal to .95 (or to .90 or another cut-off as a researcher might prefer), usually for all original values for a variable which are above the 95th or 90th percentile. The median value of an original scale is typically assigned a membership score equal to .50 that indicates the "maximum ambiguity score" and a score equal to .05 is assigned to the fifth or tenth percentile to indicate the calibrated full non-membership score in the condition. Computer software programs (e.g. fsQCA.com) are available for calibrating original values into calibrated membership scores. Alternative fuzzy-set models are possible to construct and test for consistency and coverage using these software programs. Models predicting high scores in an outcome condition or the negation of an outcome condition can be constructed and tested. Such a modelling approach predicts that the same outcome occurs consistently – all or nearly all cases with high scores in the model will have high scores in the outcome condition. The researcher can set consistency requirements for a model to equal .80, .85, .95, or higher to spell out requirements for SPOT. The findings for a model that achieves high consistency indicate that nearly all to all cases scoring high on the complex antecedent configuration (i.e. screen or recipe) also score high on the simple or complex outcome condition (e.g. 9 of every 10 cases high in the screen score high in the outcome condition). Most researchers using SST procedures (e.g. SPOT) accept a limited share of misidentification of cases not having high outcomes (e.g. conclude a model is useful if 9 of 10 cases identified by the model are identified accurately as having high scores in the outcome condition).

Embracing Somewhat Precise Outcome Testing (SPOT)

SPOT is an outcome of configurational modeling, that is, it involves constructing recipes of configurations of high and/or low membership scores of two or more ingredients that indicate a high score in an outcome condition. Configurational modeling recognizes that no one simple condition is usually sufficient or necessary for a high score in an outcome condition. SPOT follows from recognizing that a few complex conditions of ingredients are sufficient (though any one of these models is not necessary) for a high score in the outcome condition. Constructing complex conditional models (e.g. word models), each usually combining two to eight simple conditions that accurately predict the same outcome condition, is the objective of SPOT. Though still relatively rare, fuzzy-set modeling appears in the literature that predicts configurations of multiple outcome predictions (e.g. a combination of low market share and high profits as the outcome condition). Thus, applications of fuzzy logic embrace the complexity turn in case-based modeling. Adopting the complexity turn includes adopting the core principles of complexity theory and shifting beyond NHST to SPOT (cf. Woodside, 2013a). The following discussion describes these core principles.

Large effect size for a simple antecedent condition is neither sufficient nor necessary for indicating cases having high scores for an outcome condition. A specific level of a simple antecedent condition is insufficient for accurately predicting the level of an outcome condition. Thus, watching sports competitions on TV frequently by itself is not an accurate predictor of heavy beer purchases even if the correlation between the two behaviors is positive and statistically significant (e.g. $r = .53$, $p < .001$). Simple antecedent conditions associate consistently only rarely with a given level of an outcome condition (e.g. a sufficient symmetric association indicates a very high correlation).

A few recipes of complex antecedent conditions are sufficient for indicating high scores in an outcome condition (i.e. high Y cases). The performance of a specific level of an outcome condition (e.g. high firm success, heavy beer purchases, monthly holiday trip-taking) depends on specific recipes of antecedent conditions. Such complex antecedence conditions are configurations of simple conditions. Bass et al. (1968, p. 267) present recipes of heavy buyers for nine product categories in their Table 4. For example, persons "between 25 and 50, not college grad, TV more than 3.5 h" are heavy beer buyers; persons with "incomes 10,000 or over [high incomes for the year of the data collection], with high score or less ed." are heavy cream shampoo buyers.

Some antecedent conditions are necessary but insufficient for identifying high Y cases. Most simple conditions are neither necessary nor sufficient alone in predicting a level of an outcome condition. However, some conditions may be necessary but insufficient; for example, persons who are heavy beer consumers on a daily basis are all likely to be 25+ years of age. Some persons 25+ are heavy beer purchasers and some are not and all heavy beer purchasers are 25+ years of age.

A simple antecedent condition can associate with both the cases high and cases low in membership scores in the Y outcome condition. For many simple antecedent conditions, high levels of an antecedent condition associate with both a specific outcome condition and the negation of the same outcome condition – which depends on what additional ingredients occur in specific recipes of complex antecedent conditions associating with the specific outcome or the negation of this outcome. For example, frequent watching of sports competitions on TV may associate with heavy beer purchases as well as not buying beer at all.

The equifinality principle: different recipes of complex antecedent conditions are adequate for identifying cases having high scores in the outcome condition. This tenet is known as the equifinality principle – different routes (paths) are available that lead to the same outcome. For example, persons "between 25 and 50, not college grad, TV more than 3.5 h" are not all the people who are heavy beer purchasers.

The causal asymmetry principle: recipes associating with cases with a full non-membership score in an outcome condition are not mirror opposites of recipes associating with cases with a full-membership score in the same outcome condition. This tenet is known as the causal asymmetry principle. For example, Bass et al. (1968, p. 267) describe light beer purchasers to be "under 25 or over 50, college ed., nonprofessional, TV less than 2 h." Note that "nonprofessional" appears in this recipe but occupational category does not appear in the recipe for the heavy beer purchaser.

Resolving discretizing versus continuous variable controversies

The observations in the present editorial serve to support several conclusions and recommendations involving discretizing continuous variables. Except for naturally occurring dichotomous variables (e.g. gender), researchers should avoid dichotomizing continuous variables as Rucker, MacShane, and Preacher (2015) recommend. However, they are mistaken and offer bad advice in recommending preserving the continuous nature of the variable and analyzing the data via linear regression and in recommending that regression remain the normative procedure in research involving continuous variables. Discretizing to quintiles or deciles or calibrating from a variable to a fuzzy-set scale offers several beneficial outcomes and represents a fulcrum for a radical improvement in theory construction and testing. Such discretizing, calibrating, supports a major paradigm shift away from variable-based modeling to case-based modeling:

> We show that dichotomizing a continuous variable via the median split procedure or otherwise and analyzing the resulting data via ANOVA involves a large number of costs that can be avoided by preserving the continuous nature of the variable and analyzing the data via linear regression. As a consequence, we recommend that regression remain the normative procedure both when the statistical assumptions explored by Iacobucci et al. hold and more generally in research involving continuous variables. (Rucker, MacShane, & Preacher, 2015, p. 666)

Iacobucci, Posavac, Kardes, Schneider, and Popovich (2015a, 2015b) offer bad advice in their conclusion: "the bottom line is this: our research sets the record straight that median splits are perfectly acceptable to use when independent variables are uncorrelated" (Iacobucci et al., 2015b, p. 690). Most variables tend to be correlated significantly with large samples, but the additional reasons that the present editorial discusses are even more powerful reasons not to dichotomize continuous variable data. Using quintiles or fuzzy-set membership scores and asymmetric models rather than linear modeling outperforms regression analysis in predictive validation tests (using additional samples, see Gigerenzer & Brighton, 2009) and enables the researcher to avoid Armstrong's (1970) "Tom Swift" inventive linear data analysis that almost guarantees statistical significance and a paper's acceptance for publication:

> In one of my Tom Swift studies [Armstrong, 1970], Tom used standard procedures, starting with 31 observations and 30 potential variables. He used stepwise regression and only included variables where [Student's] t was greater than 2.0. Along the way, he dropped three outliers. The final regression had eight variables and an R^2 (adjusted for degrees of freedom) of 0.85. Not bad, considering that the data were from Rand's book of random numbers. (Armstrong, 2012)

Fitzsimons' (2008) brief editorial offers "a simple heuristic that makes us smart" (Gigerenzer, Todd, & the ABC Research Group, 1999). Fitzsimons (2008, p. 5) recommends, "Death to dichotomizing!" Embrace discretizing via the use of quintiles and simple algorithms (McClelland, 1998) and/or fuzzy-set calibrating of continuous variables and asymmetric analysis (Ragin, 2008) is a second simple heuristic that makes us smart.

If NHST is so bad, why does the procedure dominate as the method of science?

Gigerenzer (2004, 2008) provides a historical analysis about how relationship testing and NHST came to dominance in the behavioral science literature despite especially the fact that the procedure receives such severe criticism. So, the focus of this editorial does not include repeating his review. However, certainly the overwhelming focus on examining whether or not relationships are significant statistically in management, marketing, and psychological textbooks on research, the general lack of knowledge of how to examine case-based models for consistency (a form of SST, see Hubbard, 2016), and the near mindless use of stepwise regression analysis (Armstrong, 2012) are ingredients in the configuration of NHST high usage – a configuration that has both high consistency and high coverage.

For the related question of just how bad is NHST, the following summary by Gigerenzer (2004, pp. 591–592) of reviews of using NHST by several senior scholars in psychological research is informative:

> The seminal contributions by Frederick Bartlett, Wolfgang Köhler, and the Noble laureate I.P. Pavlov did not rely on p-values. Stanley S. Stevens, a founder of modern psychophysics, together with Edwin Boring, known as the "dean" of the history of psychology, blamed Fisher for a "meaningless ordeal of pedantic computations" (Stevens, 1960, p. 276). The clinical psychologist Paul Meehl (1978, p. 817) called routine null hypothesis testing "one of the worst things that ever happened in the history of psychology," and the behaviorist B.F. Skinner blamed Fisher and his followers for having "taught statistics in lieu of scientific method" (Skinner, 1972, p. 319). The mathematical psychologist R. Duncan Luce (1988, p. 582) called null hypothesis testing Herbert A. Simon (1992, p. 159) simply stated that for his research, the "familiar tests of statistical significance are inappropriate." It is telling that few researchers are aware that their own heroes rejected what they practice routinely. Awareness of the origins of the ritual and of its rejection could cause a virulent cognitive dissonance, in addition to dissonance with editors, reviewers, and dear colleagues. Suppression of conflicts and contradicting information is in the very nature of this social ritual.

The use of NHST is doubly bad. First the typical NHST reports stop at telling the statistical significance levels achieved by standardized partial regression coefficients (β's) in MRA models without testing these models for predictive validity using additional samples.

Second, they focus on reporting significance of β's of independent terms in regression models; the research never focuses on whether or not the model enables the researcher to identify cases of high interest or ask, by using additional samples of cases, if the model predicts high-scoring cases consistently for the outcome of interest. Answers to this question support the conclusion that MRA does less well than case-based asymmetric models in accurately predicting outcomes of interest (Gigerenzer & Brighton, 2009).

The missing link, somewhat sad to report, has been the availability of software to enable the near-mindless use of case-based model testing similar to the ease of using the available software for MRA. The ease-of-use is not the problem. The problem has been the lack of seeing both the necessity and the availability of easily testing case-based models for consistency (a form of SST rather than testing for statistical differences). The configuration of three events supports the claim that the management and marketing sub-disciplines are witnessing the early stages of a paradigm shift away from NHST and toward SPOT: the easy availability of free software (e.g. fsQCA.com) for case-based model testing, the work of a global organization providing training in such testing (COMPASSS.org), and the articles using case-based model testing in leading marketing journals (Fiss, 2011; Frösén et al., 2016; Ordanini et al., 2014; Woodside, 2013a).

For success, this paradigm shift will need to force marketing, finance, and management research textbook authors to present statistics as a toolbox rather than a single NHST hammer – in those textbooks in use in the first, and typically the only course, that undergraduate and MBA students will need to include SPOT in examining research methods and data analysis. Such methods are available widely and easy to use. Editors (e.g. Trafimow, 2014; Trafimow & Marks, 2015) having the courage to desk reject papers that report on NHST findings of statements of symmetric-only hypotheses that increases in X associate with increases in Y supports the growth of focusing research on modeling outcomes (e.g. SPOT). The editorial requirements by Trafimow and Marks (2015) will shock most researchers in the management sub-disciplines in 2017 and nearly all instructors of the research courses in finance, management, marketing, and psychology:

> The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP. With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers. Question 1. Will manuscripts with p-values be desk rejected automatically? Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about "significant" differences or lack thereof, and so on). Question 2. What about other types of inferential statistics such as confidence intervals or Bayesian methods? Answer to Question 2. Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP… Analogous to how the NHSTP fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Much like the transition years (1985–2005) of using attachments in mail via the Internet while still sending paper copies in envelopes via postal mail, the transition process from NHST to SPOT that is occurring now frequently includes findings from use of both tools in the same articles along with discussions that SPOT findings are particularly useful (e.g. Frösén et al., 2016; Ordanini et al., 2014; Woodside, Schpektor, & Xia, 2013). Based on time lapses of prior paradigm shifts (e.g. email from postal mail, internal combustion-driven vehicles from horses, electric from internal combustion-driven vehicles), the use of NHST in articles in scholarly journals in the management and psychology sub-disciplines might become a rarity only by 2030. The appearances of studies using SPOT in scholarly journals are necessary but insufficient for completing the paradigm shift away from NHST to SPOT. Without the now widely available use of case-based modeling of outcomes of interest for consistency, the paradigm shift would never happen. A missing link is now available and fully operational.

A hopeful close

Beyond exciting the reader with the idea of SPOT, the hope is that this editorial will propel you to read the literature that the editorial describes on how and why to discretize continuous data and construct and test complex antecedent conditions for predicting high scores in an outcome condition. Such model constructions occur naturally in real life in the form of firm stock ownership screening, selection of high school students for admission to universities and colleges, selecting possible romantic persons-of-interest to meet on dating websites, and screening for terrorists to arrest at airports. Formalizing case-based model construction and testing in business-to-consumer marketing and business-to-business marketing is possible and now available in the marketing literature (see Chang, Tseng, & Woodside, 2013; Cheng, Chang, & Li, 2013; Frösén et al., 2016; Feurer et al., 2016; Ordanini et al., 2014; Woodside & Baxter, 2013; Wu, Yeh, & Huan, 2014).

Table 2 is a helpful guide for understanding the terminology and procedures for doing SPOT. Table 2 shows and compares core concepts in asymmetric SPOT versus symmetric NHST modeling. The hope is that this editorial will aid and encourage your taking a complexity turn in your future research by reading Ragin (2008), Gigerenzer and Brighton (2009), and Woodside (2013a), learning how to use the complementary software available at fsQCA.com for SPOT, and beginning to construct and test simple discretized or fuzzy-set-based models. Closing with the wisdom in words by Gigerenzer et al. (1999), such models represent "simple heuristics that make us smart."

Table 2. Analogies and distinctions in using empirical positivism and fs/QCA.

Empirical positivism	fs/QCA
Independent variable	Antecedent condition
Dependent variable	Outcome condition
Interaction effect	Recipe
Objective 1: net effects of variables	Objective 1: multiple case ID recipe
Objective 2: high adjusted R^2	Highly consistent accuracy
● = multiply	● = and
+ = add	+ = or
Value	Membership score
Z transformation score	Calibrated membership score
Two-directional tests	One-directional tests
Analysis: regression; ANOVA	Analysis: algorithms

Acknowledgements

The author appreciates the review and comments of this paper by Carol M. Megehee, Coastal Carolina University, USA.

Disclosure statement

No potential conflict of interest was reported by the author.

References

2013 Survey of Consumer Finances. (2016). Retrieved January 4, 2017 from http://www.federalreserve.gov/econresdata/scf/scfindex.htm.

Armstrong, J. S. (1970). How to avoid exploratory research. *Journal of Advertising Research, 10*, 27–30.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting, 28*, 689–694.

Bass, F. M., Tigert, D. J., & Lonsdale, R. T. (1968). Market segmentation: Group versus individual behavior. *Journal of Marketing Research, 5*, 264–270.

Chang, C.-W., Tseng, T.-H., & Woodside, A. G. (2013). Configural algorithms of patient satisfaction, participation in diagnostics, and treatment decisions' influences on hospital loyalty. *Journal of Services Marketing, 27*, 91–103.

Cheng, C.-F., Chang, M.-L., & Li, C.-S. (2013). Configural paths to successful product innovation. *Journal of Business Research, 66*, 2561–2573.

Cohen, J. (1994). The earth is round (p <.05). *American Psychologist, 49*, 997–1003.

Feurer, S., Baumbach, E., & Woodside, A. G. (2016). Applying configurational theory to build a typology of ethnocentric consumers. *International Marketing Review, 33*, 351–375.

Fiss, P. C. (2007). A set-theoretic approach to organizational configurations. *Academy of Management Review, 32*, 1180–1198.

Fiss, P. C. (2011). Building better casual theories: A fuzzy set approach to typologies in organizational research. *Academy of Management Journal, 54*, 393–420.

Fitzsimons, G. J. (2008). Editorial: Death to dichotomizing. *Journal of Consumer Research, 35*, 5–8.

Frösén, J., Luoma, J., Jaakkola, M., Tikkanen, H., & Aspara, J. (2016). What counts versus what can be counted: The complex interplay of market orientation and marketing performance measurement. *Journal of Marketing, 80*, 60–78.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 254–267.

Gigerenzer, G. (2004). Mindless statistics. *Journal of socio-economics, 33*, 587–606.

Gigerenzer, G. (2008). *Ratinality for mortals*. Oxford: Oxford University Press.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristics: Why biased minds make better inferences. *Topics in Cognitive. Science, 1*, 107–143.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Gladwell, M. (1996). The tipping point. *New Yorker, 72*, 32–39.

Hsu, S. Y., Woodside, A. G., & Marshall, R. (2013). Critical tests of multiple theories of culture's consequences: Comparing the usefulness of models by Hofstede, Inglehart, and Baker, Schwartz, Steenkamp, as well as GDP and distance for explaining overseas tourism behavior. *Journal of Travel Research, 52*, 679–704.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. Thousand Oaks, CA: Sage.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015a). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology, 25*, 652–665.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, Matthew J., & Popovich, D. L. (2015b). The median split: Robust, refined, and revived. *Journal of Consumer Psychology, 25*, 690–704.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York, NY: Holt, Rinehart, & Winston.

Luce, R. D. (1988). The tools-to-theory hypothesis. Review of G. Gigerenzer and D.J. Murray, cognition as intuitive statistics. *Contemporary Psychology, 33*, 582–583.

McClelland, D. C. (1998). Identifying competencies with behavioral-event interviews. *Psychological Science, 9*, 331–339.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Ordanini, A., Parasuraman, A., & Rubera, G. (2014). When the recipe is more important than the ingredients: A qualitative comparative analysis (QCA) of service innovation configurations. *Journal of Service Research, 17*, 134–149.

Pham, M. T. (2013). The seven sins of consumer psychology. *Journal of Consumer Psychology, 23*, 411–423.

Ragin, C. C. (1997). Turning the tables: How case-oriented methods challenge variable oriented methods? *Comparative Social Research, 16*, 27–42.

Ragin, C. C. (2008). *Redesigning social inquire: Fuzzy sets and beyond*. Chicago, IL: The University of Chicago Press.

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology, 25*, 666–678.

Scholastic Aptitude Test. (2016). Retrieved September 9, 2016 from, https://collegereadiness.collegeboard.org/sat.

Skinner, B. F. (1972). *Cumulative record*. New York, NY: Appleton-Century-Crofts.

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science, 3*, 150–161.

Stevens, S. S. (1960). The predicament in design and significance. *Contemporary Psychology, 9*, 273–276.

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.

Woodside, A. G. (2013a). Moving beyond multiple regression analysis to algorithms: Calling for a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research, 66*, 463–472.

Woodside, A. G. (2013b). Proposing a new logic for data analysis in marketing and consumer behavior: Case study research of large-N survey data for estimating algorithms that accurately profile X (extremely high-use) consumers. *Journal of Global Scholars of Marketing Science, 22*, 277–289.

Woodside, A. G. (2016). *Bad to good: Achieving high quality and impact in your research*. Bingley: Emerald.

Woodside, A. G. (2017). *The complexity turn: Cultural, management, and marketing applications*. Cham: Springer.

Woodside, A. G., & Baxter, R. (2013). Achieving accuracy, generalization-to-contexts, and complexity in theories of business-to-business processes. *Industrial Marketing Management, 42*, 382–393.

Woodside, A. G., Schpektor, A., & Xia, X. (2013). Triple sense-making of findings from marketing experiments using the dominant variable based-logic, Case-Based Logic, and Isomorphic Modeling. *International Journal of Business and Economics, 12*, 131–153.

Woodside, A. G., de Villiers, R., & Marshall, R. (1916). *Incompetency and competency training*. Cham: Springer.

Wu, P.-L., Yeh, S.-S. Huan, & Woodside, A. G. (2014). Applying complexity theory to deepen service dominant logic: Configural analysis of customer experience-and-outcome assessments of professional services for personal transformations. *Journal of Business Research, 67*, 1647–1670.

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

Zadeh, L. (1996). Fuzzy logic: Computing with words. *IEEE Transactions on Fuzzy Systems, 4*, 103–111.