Predictive Validity of Evidence-Based Persuasion Principles:

An Application of the Index Method

January 23, 2015 (Version 328R)

This paper was published in the *European Journal of Marketing*. 50 (2016), 276-293.
(followed by Commentaries, pp. 294-316).

J. Scott Armstrong
The Wharton School, U. of Pennsylvania, Philadelphia, PA, U.S.A. and Ehrenberg-Bass Institute, Adelaide, Australia; Phone: 610-622-6480; armstrong@wharton.upenn.edu

Rui Du
University of Pennsylvania, Philadelphia, PA, U.S.A., Phone: 215-275-66550; ruidu@gse.upenn.edu

Kesten C. Green
University of South Australia Business School and Ehrenberg-Bass Institute, Adelaide, Australia; Phone: +61 8 830 29097; kesten.green@unisa.edu.au

Andreas Graefe
LMU Munich, Geschwister-Scholl-Platz 1, München, Germany; Phone: +49 89 21809466; a.graefe@lmu.de

Acknowledgments: Kay A. Armstrong, Heiner Evanschitzky, Rachel Kennedy, Nick Lee, Shengdong Lin, Leonard Lodish, Jörg Matthes, Barbara Phillips, Sandeep Patnaik, Rik Pieters, Denise Rousseau, Martin Schreier, Byron Sharp, Dave Walker, Malcolm Wright, and Mark Wu, along with two anonymous reviewers, provided peer review that led to substantial improvements. Alexandra House helped to develop the software used in this study and was also involved in collecting data and rating ads. We received advice on the design of the copy testing procedures from Dave Walker, Sandeep Patnaik, and Don Esselmont. We thank the Alex Panos Fund and the Ehrenberg-Bass Institute for partial financial support for this project. Useful suggestions were received when earlier versions of this paper were presented at the 2011 International Symposium on Forecasting in Prague, the Center for Advanced Studies at LMU Munich in September 2013, the Business School at Vienna University in September 2013, the 2014 Annual Conference of the International Communication Association, and the 2014 International Symposium on Forecasting in Rotterdam. Laura

Blagrave, Hester Green, Emma Hong, Jennifer Kwok, and Lynn Selhat edited the paper. We take responsibility for any remaining errors.

Predictive Validity of Evidence-Based Persuasion Principles: An Application of the Index Method

Abstract

Purpose: To test whether a structured application of persuasion principles might help improve advertising decisions. Evidence-based principles are already used to improve decisions in other complex situations, such as those faced in engineering and medicine.

Design/Methodology/Approach: Scores were calculated from the ratings of 17 self-trained novices who rated 96 matched pairs of print advertisements for adherence to evidence-based persuasion principles. Predictions from traditional methods—10,809 unaided judgments from novices and 2,764 judgments from people with some expertise in advertising, and 5,285 copy-testing predictions—provided benchmarks.

Findings: The higher adherence to principles consensus score correctly predicted the more effective ad for 75% of the pairs. Copy testing was correct for 57%, and expert judgment was correct for 55%. Guessing would provide 50% correct predictions.

Research limitations/implications: Ads for high-involvement utilitarian products were tested on the assumption that persuasion principles would be most effective for such products. The measure of effectiveness that was available—day-after-recall—is a proxy for persuasion or behavioral measures.

Practical implications: Pretesting ads by assessing adherence to evidence-based persuasion principles in a structured way helps in deciding which ads would be best to run. Such a procedure also identifies how to make an ad more effective.

Originality: This is the first study in marketing, and in advertising specifically, to test the predictive validity of evidence-based principles. In addition, the study provides the first test of the predictive validity of the index method for a marketing problem.

Keywords: advertising, combining predictions, copy testing, expertise, intentions, judgmental forecasting,

In the late-1800s, department store owner John Wanamaker was reputed to have said, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." More than a century later, advertising experts still have difficulty predicting which advertisement will be more effective. This paper describes efforts to develop a better method for evaluating the effectiveness of ads.

Practice in diverse fields has improved thanks to the application of the scientific method. Chamberlin (1890, 1965) observed that some fields advance rapidly, while others do not. He concluded that rapid progress occurred when researchers employed experiments to test multiple reasonable hypotheses. Kealey's (1996) review of scientific research supports Chamberlin's conclusion. For example, agriculture showed little progress for centuries. That changed in the early 1700s, when wealthy British landowners began to experiment with alternative approaches. The resulting agricultural revolution led to enormous gains in productivity.

Medicine provides another example of the application of the multiple reasonable hypotheses approach to complex problems. Diseases are so complex that doctors are unable to learn from experience about the best treatments for a patient. Researchers conducted useful experiments, but practitioners paid little attention. Starting around 1940, however, doctors began to make use of experimental findings that were published in scientific journals (Gratzer 2006). Over time, it became increasingly likely that if a doctor failed to follow the evidence-based medical practices he would be sued. Large gains in life spans resulted.

For another example, engineers are expected to apply evidence-based knowledge. If a building, bridge, or mine collapses, courts examine whether the engineers followed evidence-based procedures.

In this study, we tested the predictive validity of evidence-based knowledge on persuasion by using that knowledge to predict which advertisements will be more effective. We hypothesized that those advertisements that adhere closely to evidence-based persuasion principles would be more effective than those that do not. Following the method of multiple reasonable hypotheses, we obtained benchmark predictions from reasonable alternative methods.

Experimental Evidence on Persuasion

We refer to advertising effectiveness as "persuasion," and use the term in its broadest sense to include all influences—both direct and indirect—that lead people towards action. Persuasion principles apply to all media whether still, motion, or sound.

Researchers in persuasion, advertising, and related fields have published a large body of experimental evidence on persuasion over the past century. Advertising practitioners, however, rarely draw on that evidence because relevant studies are typically:

1. Hard to find and obtain (Armstrong 2011).
2. Difficult to understand.
3. Unreliable, due to lack of replication (Hubbard and Vetter 1996).
4. Of uncertain applicability, due to non-reporting of conditions (Armstrong, Brodie and Parsons 2001).
5. Lacking in explicit advice on what to do and when.
6. Hard to remember.
7. Ignored by practitioners in the belief that they have learned what works best from their experience (Helgesen 1994; Nyilasy and Reid 2009).
8. Ignored by practitioners in the belief that the best advertising is unconventional and "breaks the rules" (Nyilasy and Reid 2009).

To overcome some of the obstacles practitioners face in using experimental evidence to create persuasive advertisements, Armstrong (2010) summarized a century of experimental findings as a set of operational principles, or condition-action statements. His search for evidence on persuasion covered many fields, including advertising, consumer behavior, law, marketing, mass communications, politics, propaganda, psychology, and public opinion. Studies that related specifically to advertising encompassed all media, including direct mail, magazines, Internet, TV, videos, billboards, posters, and radio. And the studies employed a variety of criteria, including sales, intentions to act, behavioral changes, and attitude changes.

In deriving the principles, roughly 2,400 papers and books were examined, and relevant evidence was obtained from 687 of them that, in turn, drew upon more than 3,000 studies.

The formulation of principles was guided by the generalizations of experts in persuasion. While the experts' generalizations are useful in general, they do not apply under all conditions. Principles go beyond generalizations, in that they are conditional. For example, Aristotle's generalization to use two-sided arguments becomes a principle—in the sense used in the paper—with the addition of the condition that it applies when one "refutes strong opposing arguments."

Experimental research has led to knowledge about conditions that are often not intuitively obvious. For example, the principles "if resistance is expected, use indirect conclusions when the arguments are strong and obvious" and "do not mix rational and

emotional appeals in an ad." Adhering to such counterintuitive principles is likely to be particularly effective at improving persuasiveness relative to current practice.

Knowing the conditions that will apply in the situation, and their effects, is often critical for the correct application of a principle. For example, leading experts have often cautioned against the use of humor in advertisements, and analyses of non-experimental data supported this opinion. However, experiments found that humor is effective under some well-defined conditions, but is harmful under other conditions—e.g., for high-involvement products with strong arguments.

In addition to identifying conditions, experimental research in various fields led to the identification of many new principles. Leading examples include the work of Festinger, Reicken and Schacter's (1956) on cognitive dissonance, and Cialdini's (1984) work on the principles of influence. Other behavioral researchers, too, have discovered principles that are contrary to the conventional wisdom.

It is easy to find situations in which the principles are violated. For example, one of the most frequently violated principles is "Do not invite customers to evaluate their satisfaction while using a product (or service)." Violations of that principle reduce the satisfaction of not only the customers, but also of the service providers.

The review by Armstrong (2010) led to the development of 195 persuasion principles. While knowledge about the principles improves over time, the underlying principles appear to be unchanging. Also, with minor exceptions, the principles are the same across cultures and languages.

It is difficult to find the persuasion principles in advertising books. An audit of a convenience sample of three practitioner handbooks and nine popular university advertising textbooks found none of the 195 principles (Armstrong 2011). The primary reason for the absence of principles is that the books seldom specify conditions under which their advice is persuasive.

Prior Assessments of the Persuasion Principles' Validity

Face validity

Pioneering advertising practitioners distilled their experience and their knowledge of the research into advice on how to design persuasive advertisements. The initial list of principles for this project drew heavily on Ogilvy (1983), a book that is still useful and is among the best sellers in advertising. The writings of Ogilvy, along with books by eight other leading advertisers including, for example, Hopkins (1923), Reeves (1961), and Roman, Maas

and Nisenholtz (2003) include many generalizations. The persuasion principles used in the research presented in this paper are, to a considerable extent, consistent with those generalizations. For example the generalization, "long copy sells" was found to require only one minor condition to transform it into a principle. The consistency between the experts' generalizations and the persuasion principles provides evidence for the *face validity* of the principles. Armstrong (2010) describes in full how persuasion principles were derived from the experts' persuasion generalizations.

To help ensure that the persuasion principles faithfully represent the research findings, efforts were made to contact all researchers whose contributions were used to develop the principles. The great majority of researchers who could be contacted replied. Their corrections and suggestions led to many useful changes in the wording of the principles and to the inclusion of additional evidence (Armstrong 2010). The process of checking and correcting the representation of research also constitutes evidence for the face validity of the principles.

Concurrent validity

Ninety-one percent of the principles were validated in that each was based on experimental evidence. The remaining nine percent of the principles were included on the basis of logic, such as "Do not violate taste or standards." A summary of the evidence is available on the AdPrin.com site under the heading "Strength of Evidence on Principles." Given that (a) the great majority of the principles are based on experimental evidence, (b) the experimental evidence for most of the principles is based on more than one study, and (c) the effect size estimates for many principles are large, the principles have concurrent validity.

After the principles were developed, a colleague, Sandeep Patnaik, helped the first author of this article to further assess the concurrent validity of the principles. The assessment involved testing principles one by one against the print ads that had been published in a series of books known as *Which Ad Pulled Best* (hereafter referred to as *WAPB*). These data include matched pairs of ads, along with their recall scores. A description of these data is provided in Armstrong (2010, pp. 300–301).

Armstrong and Patnaik (2009) found that the directional effects in the quasi–experimental *WAPB* data pairs were consistent with the principles that are supported by experimental evidence for all of the 40 principles for which comparisons could be made. Specifically, the *WAPB* data were consistent with 7 principles supported by field experiments, 26 principles supported by laboratory experiments, and 7 principles supported by meta-

analyses of experimental findings. These agreements were surprising, given that there were few relevant *WAPB* pairs for many of these principles.

The concurrent validity testing against *WAPB* data did not lead to substantive changes in any of the principles, although one minor principle was dropped because it was based only on the opinions of advertising experts and it was not supported by the data (see Armstrong 2010, p. 301 for details). The finding of concurrent validity strengthens the case for using the principles.

In contrast, consider the poor agreement between findings from experimental and non-experimental data. Armstrong and Patnaik (2009) examined 24 principles for which both types of findings were available. The directions of the effects were different for 8 of the principles. The finding suggests that one should be skeptical about the generalizability of non-experimental findings. Unfortunately, 25 of the 195 principles lacked any experimental evidence, and 40 were based on only one experiment. Thus, many of the principles lack evidence on concurrent validity.

A Procedure to Predict Ad Effectiveness Using Evidence-based Principles

Evidence on the validity of the persuasion principles from experiments does not address the issues of (1) whether practitioners can make effective use of adherence to the principles, or (2) whether any gains in predictive accuracy would be substantial, or (3) whether any gains would come at a reasonable cost.

To the extent that adherence to the principles has predictive validity, it would provide a useful way to pretest advertisements in order to improve them or to select those ads that are most effective. The primary purpose of this paper is to assess the predictive accuracy of evidence-based persuasion principles

Advertising researchers have previously attempted to assess the effects of various features of ads by using regression analysis. Of particular note is Stewart and Furse's (1986) analysis of before and after responses from thousands of viewers of 1,059 TV commercials encompassing 356 brands from 63 firms in twelve product categories. Their regression analyses assessed the relationships between roughly 160 features of TV commercials and recall, comprehension and persuasion. Their study inspired replications including Stewart and Koslow (1989), with an additional 1,017 commercials; Laskey, Fox, and Crask (1994), with 1,100 thirty-second commercials for fast-moving food and household items; Stanton and Burke's (1998) analysis of 601 commercials; and Phillips and Stanton's (2004) analysis of 5,000 commercials. The findings of these studies were disappointing in that few variables

appeared to have substantive effects and the directions of the effects often seemed to be inconsistent with rational expectations.

The findings of the regression studies using non-experimental data are, however, not so puzzling for those who are familiar with the literature on the limitations of regression analysis. Even sample sizes of 1,000 or more are inadequate when there are many predictor variables. Regression analysis of non-experimental data cannot estimate valid relationships from many variables no matter how large the sample size, because the causal variables in non-experimental data correlate with one another, some important variable cannot be included, and some important variable do not vary. The practical limit of regression analysis is typically a handful of variables (See Armstrong, 2012, for a discussion on this issue.) Thus, the problem of how to predict the effects of the 195 principles on effectiveness cannot be solved by regression analysis.

Index method

To address the problem of forecasting with many causal variables identified from much prior knowledge, we turned to the index method. Instead of estimating the importance of variables from a given data set, the index method uses prior knowledge to select variables and to determine the magnitude and direction of weights.

Inspiration for the index method came from Benjamin Franklin. Franklin's friend and fellow scientist, Joseph Priestley, was considering a new job, and asked Franklin for advice. On September 19, 1772, Franklin wrote a letter in reply, in which he described his "method of deciding doubtful matters" (Sparks 1844, p. 20). Franklin's advice was to list all variables known to be important, rate the extent to which each variable favors each alternative, and to then add the ratings to see which alternative is better.

An early formal application of the index method involved calculating index scores for prison inmates based on whether they rated favorably or unfavorably against a list of 25 factors believed to influence the chance of successful parole (Burgess 1936). The application of the index method to that problem recently made a comeback, with news articles reporting the use of computer programs that calculate index scores based on up to 100 predictor variables derived from criminology research. Predictors include such variables as whether the offender is married, the age of first arrest, the type of crime, and the last grade completed in school (Walker 2013).

Recent research tested the index method for predicting U.S. presidential elections by using the index method with biographical information about candidates (Armstrong and Graefe 2011) and another with voter perceptions of each candidate's ability to handle

important issues (Graefe and Armstrong 2013). The index scores provided predictions that were competitive with those from established methods, including regression analysis. But their biggest advantage over traditional selecting candidates and deciding which issues to emphasize in a campaign.

Creating a Persuasion Principles Index

We converted the principles into questions for the raters. To make the task simple for the raters, only the most important conditions were included in the questions. The raters were, however, free to follow links to supplementary information on each principle.

The questions were checked for clarity and reworded as required. One author coded many ads as part of the effort to improve the wording of the questions. We also pretested questions by asking research assistants to each rate 40 print ads. That process led to changes in the wording. No results from these pre-tests were included in the analyses presented in this paper.

We applied the five steps described in Graefe and Armstrong (2011) to develop an index model with causal variables corresponding to the persuasion principles (see the Appendix for more details):

1. Identify all variables (principles) that are important to the problem.
2. Specify the direction and magnitude for each variable's effect (the weight on each principle).
3. Determine the values for each variable (the rating on the use of each principle).
4. Calculate the index score by applying the weights from step 2 to the values from step 3, and then sum (across principles). We refer to the index score as the Persuasion Principles Index (PPI).
5. Use the index scores to make the predictions.

To improve reliability, the procedure we developed facilitates combining the ratings of several raters to achieve a consensus rating for each principle. The procedure follows Franklin's advice to use subjective weights for the variables. Principles that relate to strategy—e.g., identify benefits of the product being advertised—are given more weight than those based on tactics—e.g., how to punctuate a headline. Also, principles supported by much evidence were weighted more heavily than those supported by little evidence. The weights were all specified *prior* to doing the analysis. We made no attempt to search for optimum weights, nor was it possible to do so with our data. A copy of the software used for this study is provided in the Research Repository at AdPrin.com.

In sum, then, the structured procedure that we developed for predicting ad effectiveness from evidence-based persuasion principles is based on evidence-based forecasting principles. In particular, the procedure follows three forecasting guidelines described by Armstrong, Green, and Graefe (2015): (1) use prior information to select variables and determine effect sizes, (2) use all available information, and (3) combine judgments.

Testing the Predictive Validity of a Persuasion Principles Index

In this section we describe the data, the selection and training of raters, the task, and the creation of consensus ratings of adherence to persuasion principles in order to derive an index score for each ad.

Data

For our test of predictive validity, we used full-page U.S. print ads from *Which Ad Pulled Best (WAPB)* editions four through nine, that were published from 1981 to 2002 (Burton 1981; Burton and Purvis 1986, 1991, 1993, 1996; Purvis and Burton 2003). These books have been used in advertising courses for more than three decades. The *WAPB* ads have also been used in prior research studies (e.g., McQuarrie and Phillips 2008; McMackin and Slovic 2000; Tom and Eves 1999). Further description of the ads is provided in Appendix B of Armstrong (2010).

Gallup and Robinson provided day-after recall scores for all ads. The scores are the percentage of respondents who accurately described an ad the day following their exposure to it. (A description of the recall measure is provided in Appendix B of Armstrong, 2010.) Our test's ability to assess the extent of predictive validity is limited, then, by the less-than-perfect correlation between recall and behavior. Zinkhan and Gelb (1986) found a positive relationship ($r = .52$) between recall of ads and people's intentions to buy a product. This correlation implies that for binary data one would expect an upper limit on accuracy in this study to be about 76%.

We used pairs of ads for the same product and brand. From those, we selected ads for high-involvement utilitarian products. We expected the principles to be more useful for such products because consumers think more carefully about the offer, and they are likely to find it easy to evaluate the reasons why a given utilitarian product might solve their problem. Using these criteria, the lead author of this article, a research assistant, and Sandeep Patnaik of

Gallup and Robinson each independently screened the ads. The final sample was 96 pairs of ads agreed upon by all three screeners.

We regard these *WAPB* data as quasi-experimental because each pair is identical with respect to the target market, product, brand, size of ad, and media placement. The timing of the ad placements was approximately the same, although some placements were separated by as much as a year.

The *WAPB* data are not ideal. The net effect of the shortcomings of these data is that the relationship between compliance with persuasion principles and the effectiveness of an advertisement is likely to be *under*estimated in our test.

Method

Selection and Training of Raters

We aimed to develop procedures that could be used by all practitioners who are concerned with persuasion. They include those in advertising agencies, corporations, sole proprietors, pretesting services, government agencies, and so on. To enable this, the persuasion principles were stated in ways such that intelligent people would be able to make useful judgments. In other words, the procedure does not require experts to do the ratings.

We recruited 13 university students for the rating task and paid them $10 per hour. We also hired four raters from Amazon Mechanical Turk for $80 each for the task of rating a batch of about 20 pairs of ads, plus a bonus payment based on the number of correct predictions derived from their ratings.

All raters were first required to complete the self-training module provided on AdPrin.com. As part of their training, they received feedback based on the consensus ratings provided by two of the authors and another expert on the rating system. The training session took about an hour.

Recruiting and training raters via Mechanical Turk proved to be substantially faster and less expensive than hiring university students. In addition, the quality of the Turkers appeared to be on a par with that of the students.

Rating Task

The participants rated both ads in each matched pair at the same time. To make the task manageable, we organized the 96 pairs into batches of 18 to 20 pairs of ads. The task was nevertheless a sizable one that, including the training, took about 16 hours per rater to complete. The batches of ads that were used are available in the Research Repository on AdPrin.com.

Consensus Ratings

We used five raters for each pair of ads. An administrator, who had no knowledge of the recall data, copied the ratings from each individual rater into a summary spreadsheet that in turn generated consensus ratings from agreement between the ratings of three or more raters. The administration task was divided between two administrators.

We calculated each rater's reliability score. These scores were used to reject raters who departed substantially from the consensus. Specifically, raters whose scores were more than 10 percentage points different from the average rater were dropped and replaced by new raters. Details are provided on the Research Repository on adprin.com.

Results on the predictive validity of adhering to persuasion principles

The ratings by individual raters correctly predicted which ad in each pair had the higher recall for 61.0% of the 96 pairs. Consensus PPI scores were correct for 74.5% of the 96 pairs.[1] As we expected, a consensus approach to combining across raters improved reliability and, consequently, predictive validity.

Accuracy of Benchmark Predictions

The purpose of this article, as stated above, is to assess the predictive validity of a structured approach to measuring adherence to evidence-based persuasion principles. Following the method of multiple reasonable hypotheses, we obtained benchmark predictions from two pre-testing methods: unaided judgment and copy testing.

Unaided judgment

Practitioners commonly predict the effectiveness of advertisements using their unaided judgment, such as by thinking about whether potential customers will like the ad. We obtained unaided judgments from novices as well as from people with some experience in advertising.

[1] We do not provide tests of statistical significance because they are detrimental to the effective use of findings (e.g., see Armstrong, 2007; Ziliak and McCloskey, 2008). Decisions should be based on costs and benefits. Readers are free to ignore our recommendation. For example, if you used the one-tail binomial test, you would find that these results differ from chance at $p < 10^{-6}$.

Method

To obtain unaided judgmental predictions, we first sought help from advertising practitioners. Despite following many leads, we had little success in gaining participation by practitioners. Few of the practitioners we contacted responded. Most of those who did respond informed us they were too busy or not interested. We did obtain predictions from 16 practitioners—seven recruited via personal contacts with people at two U.S. advertising agencies, and nine recruited from a Microsoft advertising department in China. In addition, we recruited 128 participants from Amazon Mechanical Turk who claimed to have had at least one year of experience working in advertising.

Novice participants included 113 unpaid volunteers, mostly university students and recent graduates. We recruited a further 450 novice participants through Mechanical Turk, and paid them $1 per batch of ads.

We directed the expert and novice judges to one of the five batches of ads described above. They were asked, "Can you predict which advertisement had the better 'day-after recall'? Think of recall as a measure of effectiveness." In addition, they were asked, "How confident are you of your prediction?" and were provided with a scale from 50% to 100%, where 50% equals guessing, against which to answer. The online questionnaire automatically recorded the time that the participants spent judging each pair of ads. The median time spent by judges was about one minute per pair of advertisements. The survey instruments are provided in the Research Repository at AdPrin.com.

Findings

We had no expectation that the judgments of unpaid and paid participants would differ. The judgments turned out to be similar, so we merged the results.

Individual unaided judgments by novices were of some value for predicting the effectiveness of advertisements: 54.1% of 10,809 novice judgments correctly identified the more-recalled ad. The experts' judgments were correct for 55.4% of 2,764 predictions. The experts were more confident about their predictions than were the novices. They expected 85% of their predictions to be correct, whereas novices expected 78% percent of theirs to be correct.

Given that industry leaders—especially David Ogilvy, who was an advocate of research—anticipated some of the principles, we expected that some practitioners would do well. In other words, our test was likely to underestimate the predictive skill of leading practitioners. The results suggest that this might be the case. For example, the 16 practitioners

with advertising roles did better than the average expert in our sample, achieving 59.7% correct out of 320 predictions.

On the other hand, extensive prior research on the value of expert judgmental predictions in complex uncertain situations (Armstrong 1980; Stewart 2005; Tetlock 2005) found that there is a modest threshold level of expertise beyond which further expertise does not lead to better predictions. Moreover, in the domain of consumer behavior, a study found that practitioners' predictions were not more accurate than those of novices (Armstrong 1991).

Formal combining of judgments often improves accuracy relative to individual predictions, especially if the individual predictions are based on different knowledge and information (Graefe, Armstrong, Jones Jr., and Cuzán 2014). However, we expected that the gains from combining would be small when unaided judgments were combined given that the accuracy of the individual judgments was poor.

To combine the judgmental predictions, we identified the modal prediction; that is, the ad in each pair that most of the judges predicted would be more effective. Ties were scored as half of an accurate prediction (i.e., 0.5).

Combining the 563 novices' judgments increased the accuracy of the predictions from 54.1% to 59.0%. Combining the 144 experts' judgments increased accuracy from 55.4% to 63.9%. Formal procedures for combining the independent judgments of practitioners are apparently not common in advertising agencies. Based on secondary sources, including detailed observations on the behavior of agencies and clients, judgments on ads are typically made in meetings (see, e.g., Armstrong 1996). Unlike combinations of independent judgments, predictions from group meetings are likely to be less accurate than those of the individuals (Armstrong 2006). Leaders of creative agencies—including David Ogilvy, George Lois, and Bill Bernbach—were highly critical of meetings.

Copy Testing

Copy testing is currently the primary evidence-based approach to testing advertisements. There are many types of copy testing. We conducted a single test that used three ways of deriving predictions of which ad would be more effective in the form of intentions-to-purchase.

Method

Participants for the copy-testing task were recruited from Amazon Mechanical Turk. Each participant was paid $2 per batch—i.e., ten cents per pair of advertisements.

Because the *WAPB* ads used in this research study were published from 1981 to 2002, we were concerned that the age of the ads might influence the copy-testing participants' reactions to them. To address this problem, we asked the participants to adopt a role to "imagine that you were in the market for this kind of product at the time the advertisement was run. Specifically, imagine that the item being advertised is an example of a product that you, a family member, or an acquaintance would like to buy within 12 months."

For each ad, the participants were asked:

Q1: How likely would you be to seek further information about this brand of <type of product in the ad> after seeing this ad?

Q2: If you wanted to compare different brands of <type of product in the ad>, how likely is it that you would include this brand in your comparison?

Q3: How likely would you be to purchase this brand of <type of product in the ad> within 12 months of seeing this ad?

To obtain participants' intentions-to-purchase, participants rated: either ad A *or* ad B; the same pairs of ads twice, with the second rating conducted two weeks after the first one; or ad A then, two weeks later, ad B. Intentions-to-purchase were calculated as an average of each participant's responses, on a 0-to-100 scale, to each of the three questions, Q1 to Q3 above.

Findings

To assess the accuracy of copy-testing predictions, we examined whether the ad in each pair with the higher intention-to-purchase also had the higher recall score. Average intentions-to-purchase scores from the first procedure provided accurate predictions of which ad had the higher recall for 62.2% of pairs. For the second procedure, it was 50.6%, and for the third, 58.2%.

We had no prior expectations on the relative accuracy of the three procedures for obtaining intentions to purchase, and so we weighted each procedure equally. The combined prediction was correct for 57.0% of the 5,285 predictions from 369 subjects. Additional details are provided in the AdPrin.com Research Repository.

Discussion

Our objective for this study was to determine whether or not advertisements that adhere more closely to evidence-based persuasion principles are more effective. Given that 74.5% of the consensus predictions from the Persuasion Principles Index were correct

compared to the 50% that could be expected from guessing, the answer is yes. The finding is consistent with evidence on the face validity of the principles provided in Armstrong (2010) and the concurrent validity tests in Armstrong and Patnaik (2009). Given that we had only a proxy for effectiveness, such that the upper limit for effectiveness was estimated to be 76%, we were surprised by the accuracy.

The accuracy of each method tested in this study is summarized in the Table. Asterisks designate the benchmark methods that are often used in practice.

Table: Accuracy of Predictions from Index and Benchmark Methods

	Predictions	Percent correct
Persuasion Principles Index (PPI)		
Individual	480	61.0
Consensus of 5 raters per ad	96	74.5
Unaided novice judgments		
Individual *	10,809	54.1
Combined	96	59.0
Unaided expert judgments		
Individual *	2,764	55.4
Combined	96	63.9
Copy testing *	5,285	57.0

Predictions from novices who, after one hour of training, used an index method procedure to assess adherence to principles were, at 61% correct, 7 percentage points more accurate than the predictions of novices who used their unaided judgment. The accuracy of the individual index method predictions is substantially greater than the 55.4% achieved by judges with at least some experience in advertising. The individual index method predictions are also more accurate than predictions from copy testing at 57.0% correct.

The improvement in the reliability of ratings that was achieved by using the consensus of five raters led to substantial increases in accuracy. At 74.5% correct, index method predictions based on consensus ratings were 13.5 percentage points more accurate than those based on individual ratings.

We expect that further gains with the use of the Persuasion Principles Index might be obtained by selecting raters who demonstrate that they are good at the rating task, by providing them with additional training, and by using raters who have experience with the evidence-based principles rating system.

We are not aware that advertising agencies use structured combinations of many experts' predictions. Moreover, employing many experts on such a task does not seem practical. Thus, the most effective of the currently used methods was, as expected, copy testing. The accuracy of copy-test predictions was, at 57% correct, substantially lower than the 74.5% correct from consensus assessments of adherence to principles obtained using the Persuasion Principles Index.

The cost of achieving the gains in accuracy is modest. After about one hour of self-training, each rater took about 45 minutes to rate each pair of ads. While that is greater than the time taken to make unaided judgments—at one minute per pair—and copy testing—at two minutes per pair—the additional cost in financial terms would be trivial relative to the potential benefits from running more effective ads, especially for TV commercials. In addition, the pre-testing can be done early by using storyboards or rough mock-ups of commercials.

Though we tested the use of persuasion principles for print advertising, we expect that they could be applied to online and TV commercials. The persuasion principles are also likely to be useful for other communications, such as political campaigns and management presentations, as described in Armstrong (2010, Appendices G and H). Indeed, the principles are expected to apply to all efforts to persuade. McCloskey and Klamer (1995) estimated that one-quarter of the American economy is persuasion. In addition, Adam Smith said in one of his Lectures on Jurisprudence, "And in this manner every one is practising oratory on others thro the whole of his life" (Smith, 1975).

Our findings are consistent with prior research, and the effect sizes are large. The results are encouraging given that they were obtained from raters who had only a short training period for rating adherence to persuasion principles. In addition, the criterion (recall) is not strongly related to persuasion and behavior. We expect, therefore, that our results underestimate the strength of the relationship between adherence to persuasion principles and advertising effectiveness for high-involvement utilitarian products and services. Given the shortcomings in the experience of the participants and the data on effectiveness, we were surprised to see that the 75% of correct predictions.

Moreover, our results may underestimate the gains in accuracy from using the persuasion principles relative to the accuracy of unaided judgment. The unaided judges were specifically asked to assess relative recall for a pair of ads. For example, an ad for a soft drink containing an image of a friendly polar bear would be memorable, but might not be persuasive.

This is the first study on the predictive validity of rating the effectiveness of ads by assessing their adherence to persuasion principles. Replications and extensions are needed in order to further test the expected error reductions. We did, however, have the benefit of an accidental replication. The ratings of the principles by nine raters were lost due to a damaged computer hard drive. Given the need for full disclosure, we decided to drop the ratings from the lost data from our analysis, and to recruit replacement raters. The original PPI consensus ratings were correct for 76.0% of the predictions, whereas with the replacement raters, the PPI predictions were correct for 74.5% of the predictions.

We are interested in generalizable research findings, and so we have tested the value of a structured application of evidence-based principles. We expect that further research will lead to improvements in the current principles and to the identification of additional principles. We also expect that improvements can be made to the index method that we developed for rating adherence to principles.

The PPI software that we developed and used in our research is offered as part of full disclosure. There is no copyright or patent. The program is written in Excel so that advertisers and consulting firms can follow what has been done and make changes as they see fit. In making changes, our advice from forecasting research is that expert knowledge should only be used as inputs to the method, and not to revise the predictions.

While this study is concerned primarily with predictive validity, the ratings of adherence to principles can also be used to improve ads. For example, in the tests described in this paper, a typical ad violated two principles, was only partly successful in applying 16 principles, and overlooked 25 relevant principles. We consider that the information on how to improve ads is a key benefit of pretesting ads by assessing adherence to persuasion principles.

Conclusions

This study provides a test of the predictive validity of persuasion principles. Adherence to the principles was used to predict the most effective ad in each of 96 matched pairs of print advertisements for high-involvement utilitarian products by leading advertisers. Adherence to principles was assessed using the index method, in the form of predictions from a Persuasion Principles Index (PPI).

Advertisements that more closely followed the evidence-based principles were more effective than those that did not. The PPI scores correctly identified the more-recalled ad for 74.5% of 96 pairs. Our findings provide further support for the conclusion that, as in other

fields, applying knowledge in the form of evidence-based principles using a structured method results in better predictions.

Consistent with prior research on situations with many important variables and good prior knowledge, predictions from an index method were substantially more accurate than those from unaided judgment, the method typically used by advertising practitioners. Unaided judgment provided 55.4% correct predictions. Compared to unaided judgment, then, PPI predictions reduced error by about 43%.

Also consistent with prior research, combining unaided judgmental predictions improved accuracy. It improved the accuracy of novices' judgmental predictions from 54.1% for individuals to 59.0% for the combined predictions. For experts, combining improved accuracy from 55.4 to 63.9 percent.

By assessing adherence to evidence-based persuasion principles, one can choose the more effective ad. Moreover, one can improve ads by using more of the relevant principles, improving the application of relevant principles, and avoiding violations of principles.

We expect that Mr. Wannamaker would be pleased with the progress that has been made in addressing his concern over which half of his advertising dollars were wasted. Regrettably, it is too late to help him. In our defense, the solution required using generalizations developed by leading thinkers and advertising practitioners, findings from thousands of researchers over the past century, and Benjamin Franklin's advice on using the index method. They are the heroes of this effort to benefit from a scientific approach to advertising.

References

Armstrong, J. S. (1980), "The Seer-Sucker theory: The value of experts in forecasting", *Technology Review*, 82(7), pp. 16–24.

Armstrong, J. S. (1991), "Prediction of consumer behavior by experts and novices", *Journal of Consumer Research*, 18(2), pp. 251–256.

Armstrong, J. S. (1996), "How should firms select advertising agencies? A review of Randall Rothenberg's *Where the Suckers Moon*", *Journal of Marketing*, 60(3), pp. 131–134.

Armstrong, J. S. (2006**), "**How to make better forecasts and decisions: Avoid face-to-face meetings"**,** *Foresight: The International Journal of Applied Forecasting*, 5(Fall), pp. 3–15.

Armstrong, J. S. (2007), "Significance tests harm progress in forecasting", *International Journal of Forecasting*, 23(2), pp. 321–327.

Armstrong, J. S. (2010), *Persuasive Advertising: Evidence-based Principles*, Palgrave MacMillan, Hampshire. U.K.

Armstrong, J. S. (2011), "Evidence-based advertising: An application to persuasion", *International Journal of Advertising*, 30(5), pp. 743–767.

Armstrong, J. S. (2012), "Illusions in regression analysis"**,** *International Journal of Forecasting*, 28(3), pp. 689–694.

Armstrong, J. S., Brodie, R. J., & Parsons, A. G. (2001), "Hypotheses in marketing_science: literature review and publication audit", *Marketing Letters*, 12(2), pp. 171–187.

Armstrong, J. S., & Graefe, A. (2011), "Predicting elections from biographical information about candidates: A test of the index method", *Journal of Business Research*, 64(7), pp. 699–706.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). "Golden rule of forecasting: Be conservative", *Journal of Business Research (forthcoming)*, available at: www.goldenruleofforecasting.com.

Armstrong, J. S., & Patnaik, S. (2009), "Using quasi-experimental data to develop empirical generalizations for persuasive advertising", *Journal of Advertising Research*, 49(2), pp. 170–175.

Burgess, E. W. (1936). „Protecting the public by parole and by parole prediction", *Journal of Criminal Law and Criminology*, 27, pp. 491–502.

Burton, P. W. (1981), *Which Ad Pulled Best?,* 4th ed., Crain Books, Chicago IL.

Burton, P. W., & Purvis, S. C. (1986–1996), *Which Ad Pulled Best?, 50 Case Histories on How to Write and Design Ads That Work*, 5th–8th eds., NTC Business Books, Lincolnwood, IL.

Chamberlin, T. C. (1890, 1965), "The method of multiple working hypotheses", *Science*, 148, pp. 754–759. (Reprint of an 1890 paper).

Cialdini, R. B. (2010), *Influence: Science and Practice,* Allyn and Bacon, Boston, MA..

Festinger, Leon, Henry W. Reicken & Stanley Schacter (1956), *When Prophecy Fails*, University of Minnesota Press, Minneapolis.

Graefe, A., & Armstrong, J. S. (2011), "Conditions under which index models are useful: Reply to bio-index commentaries", *Journal of Business Research*, 64, pp. 693–695.

Graefe, A., & Armstrong, J. S. (2013), "Forecasting elections from voters' perceptions of candidates' ability to handle issues", *Journal of Behavioral Decision Making*, 26, pp. 295–303.

Graefe, A., Armstrong, J. S., Jones Jr., R. J., & Cuzán, A. G. (2014), "Combining forecasts: An application to elections", *International Journal of Forecasting*, 30(1), 43–54.

Gratzer, D. (2006), *The Cure: How Capitalism Can Save American Health Care*, Encounter Books, New York.

Helgesen, T. (1994), "Advertising awards and agency performance criteria", *Journal of Advertising Research*, 34(4), pp. 43–53.

Hopkins, C. C. (1923), *Scientific Advertising*, Cool Publications Limited, Bramhall, U.K.

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

Kealey, T. (1996), *The Economic Laws of Scientific Research*, Palgrave Macmillan, Bramhall, U.K.

Laskey, H. A., Fox, R. J. & Crask, M. R. (1994), "Investigating the impact of executional style on television commercial effectiveness", *Journal of Advertising Research*, 34(6), pp. 9–16.

McCloskey, D., & Klamer, A. (1995), "One quarter of GDP is persuasion", *American Economic Review*, 85(2), pp. 191–195.

McMackin, J., & Slovic, P. (2000), "When does explicit justification impair decision making?", *Applied Cognitive Psychology,* 14(6), pp. 527–541.

McQuarrie, E. F., & Phillips, B. J. (2008), "It's not your father's magazine ad: Magnitude and direction of recent changes in advertising style", *Journal of Advertising*, 37(3), pp. 95–106.

Nyilasy, G., & Reid, L. N. (2009), "Agency practitioners' meta-theories of advertising", *International Journal of Advertising*, 28(4), pp. 639–668.

Ogilvy, D. (1983), *Ogilvy on Advertising*, Crown, New York.

Phillips, D. M., & Stanton, J. L. (2004), "Age-related differences in advertising: Recall and persuasion", *Journal of Targeting, Measurement, and Analysis for Marketing*, 13, pp. 7–20.

Purvis, S. C., & Burton, P. W. (2003), *Which Ad Pulled Best?: 40 Case Histories on How to Write and Design Ads That Work*, 9th ed., McGraw-Hill/Irwin, Boston.

Reeves, R. (1961), *Reality in Advertising*, Alfred A. Knopf, New York.

Roman, K., Maas, J., & Nisenholtz, M. (2003), *: What works, what Doesn't – and why,* 3rd ed., Kogan Page Limited, London.

Smith, A. (1976). Meek, R. L., Raphael, D. D., Stein P. G., eds., *Lectures on Jurisprudence*: *Report of 1762–3*, page 56 of section vi. Oxford University Press, Oxford.

Sparks, J. (1844). *The Works of Benjamin Franklin*, Vol. 8, Charles Tappan Publisher, Boston.

Stanton, J. L., & Burke, J. (1998), "Comparative effectiveness of executional elements in TV advertising: 15- versus 30-second Commercials", *Journal of Advertising Research*, 38(6), pp. 7–13.

Stewart, D. W., & Furse, D. H. (1986), *Effective Television Advertising: A Study of 1000 Commercials*, Lexington Books, Lexington, MA.

Stewart, D. W., & Koslow, S. (1989), "Executional factors and advertising effectiveness: A replication", *Journal of Advertising*, 18(3), pp. 21–32.

Tetlock, P. C. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton.

Tom, G., & Eves, A. (1999), "The use of rhetorical devices in advertising," *Journal of Advertising Research*, 39(4), pp. 1–5.

Walker, J. (2013), "State parole boards use software to decide which inmates to release; Programs look at prisoners' biographies for patterns that predict future crime", available at: http://online.wsj.com/news/articles/SB10001424052702304626104579121251595240852 (accessed October 11, 2013).

Ziliak, S. T., & McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press, Ann Arbor.

Zinkhan, G. M., & Gelb, B. D. (1986), "What Starch scores predict", *Journal of Advertising Research*, 26(4), pp. 45–50.

Total words = 8,160

Text = 6,600

Appendix: Persuasion Principles Index (PPI) Details

Step	Description	
1. Variables	All 195 principles published in Armstrong (2010) were considered as causal variables of ad effectiveness. Raters used descriptions of the principles to decide whether or not a principle was relevant to the ad being evaluated.	
2.Direction and magnitude of influence	Each principle in Armstrong (2010) is formulated in such a way that compliance has a positive influence on ad effectiveness. Principles supported by more evidence and those expected to have larger effect sizes are weighted more heavily.	
3.Rating of advertisement	(a) *Individual ratings:* For each principle that was assessed as relevant, raters rated how well the principle was applied in the ad using the scale: applied well = +2; needs improvement = +1; not used = 0; violated = -2.	(b) *Consensus ratings:* Ratings from five raters were used to calculate consensus ratings on how well a principle was applied. A consensus was achieved when the ratings of three or more (out of five) raters were identical. When there were fewer than three identical ratings for a principle, that principle was dropped from the PPI
4. Index score calculation	First, the *Creativity Score* was calculated as the percentage of all relevant principles that were implemented well. Second, the *Weighted Mastery Score* assesses how effectively the relevant principles were implemented, relative to the ideal of all used principles having been well applied. Finally, the PPI was calculated as the unweighted average of the *Creativity* Score and *Weighted Mastery* score.	
5. Prediction	An ad with a higher PPI score implements principles better than one with a low PPI score and is, therefore, predicted to be more effective.	