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

June 2, 2014 (version 280)

J. Scott Armstrong
Professor of Marketing, 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
Consultant, University of Pennsylvania, Philadelphia, PA, U.S.A.,
Phone: 215-275-66550; ruidu@gse.upenn.edu

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

Andreas Graefe
Research Fellow, LMU Munich, Geschwister-Scholl-Platz 1, München, Germany,
Phone: +49 89 21800; a.graefe@lmu.de

Acknowledgments: Alexandra House helped to develop the software used in this study and was also involved in collecting data on the ratings of ads. Kay A. Armstrong, Heiner Evanschitzky, Rachel Kennedy, 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 provided peer review that led to substantial improvements. We take the responsibility for any remaining errors. We thank the Alex Panos Fund, which provided partial financial support for this project. Useful suggestions were received when this paper was presented at the 2011 International Symposium on Forecasting in Prague, the Center of Advanced Studies at LMU Munich in September 2013, the Business School at Vienna University in September 2013, and the 2014 Annual Conference of the International Communication Association. Laura Blagrave, Jennifer Kwok, and Lynn Selhat edited the paper.
Predictive Validity of Evidence-Based Persuasion Principles: An Application of the Index Method

Abstract. This study develops an index model for predicting the effectiveness of advertisements (measured as recall). The index incorporates 195 evidence-based persuasion principles that were derived from a prior review of empirical, mostly experimental, research. A rater assesses the application of each of the principles to an advertisement and the index model sums the points credited. The more relevant principles that are applied and the better the application of the principles, the higher the index score. To test the predictive validity of the index, self-trained novices rated 96 pairs of print advertisements for high-involvement utilitarian products for which recall data were available. The resulting index scores correctly predicted which ad had higher recall for 75% of the pairs. Predictions by experts using their unaided judgment were correct for 55%. Use of the persuasion principles index model thus reduced error by 43% compared to the method usually employed. The finding is consistent with previous research on the value of evidence-based findings, and on the value of the index method for modeling cumulative knowledge about situations involving many important variables. We regard this as a promising first step for the assessment of the Persuasion Principles Index.

Keywords: combining, index method, judgmental forecasting, print advertising, purchase intentions, recall.

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. We believe that we have made some progress on this problem.

In some fields, practitioners are expected to apply evidence-based knowledge. For example, in order to determine liability for harmful outcomes, courts examine whether medical doctors and engineers followed evidence-based procedures. We see no reason why scientific work on persuasion would differ from that in other fields. In this study, we tested the predictive validity of accumulated evidence-based knowledge on persuasion by using it to predict which advertisements will be more effective.
Importance of Experimental Evidence for Persuasion

We refer to advertising effectiveness as “persuasion” and use the term in its broadest sense. That is, persuasion includes all procedures, both direct and indirect, that lead people towards the acceptance of a message. In advertising, persuasion applies to all media whether still, motion, or sound. Persuasion, as used in this paper, pertains not only to advertising, but also to many aspects of life. As Adam Smith said in one of his lectures, “Everyone is practicing oratory on others thro the whole of his life.” McCloskey and Klamer (1995) estimated that one-quarter of American economy is persuasion.

What procedures can be used to persuade people? How can we show that the procedures are effective? The surest way to establish causal relationships is to analyze data from experimental comparisons of reasonable alternative hypotheses. The approach was well described by Chamberlin (1890, 1965). Experimentation is responsible for the dramatic advances in such fields as agriculture, engineering, and medicine (Kealey 1996; Gratzer 2008). We propose that experimental findings can help improve persuasion in advertising.

From Experiments to Principles

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

2. Difficult to understand.
4. Of uncertain applicability due to non-reporting of conditions (Armstrong et al., 2001).
5. Lacking in explicit advice on what to do and when.
6. Hard to remember when creating an ad.
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 help overcome some of the obstacles practitioners face in using experimental evidence to create persuasive advertisements, Armstrong (2010) summarized a century of experimental findings on persuasion effectiveness into a set of easily understood principles in his book, *Persuasive Advertising*. This search covered many fields in which persuasion is important including advertising, consumer behavior, language, law, marketing, mass
communications, politics, propaganda, psychology, and public opinion. The studies related to advertising included all media—including direct mail, magazines, Internet, TV, videos, billboards, posters, and radio. They were mostly, but not always, in English.

In all, roughly 2,400 papers and books were examined and relevant evidence was obtained from 687 of them. The 687 works themselves drew upon prior research such that the principles are based on more than 3,000 studies.

To help ensure that the Persuasive Advertising persuasion principles faithfully represented 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 and to the identification of additional evidence.

The Persuasive Advertising book describes a total of 195 persuasion principles. They provide advice on what to do in the form of condition-action statements. For example, one principle suggests: “Use two-sided arguments that refute strong opposing arguments.” While our knowledge about the principles improves over time, the principles do not seem to change. Also, with minor exceptions, the principles are invariant across cultures or languages. Studies done in different countries are consistent with one another; a finding that accords with industry experience.\(^1\)

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 is that the books tend to recommend actions without specifying conditions. And, the conditions for the persuasion principles are often not intuitive. For example, in what situations is it persuasive to use indirect conclusions? The principles in the Persuasive Advertising book include the primary conditions as follows: “If resistance is expected, use indirect conclusions when the arguments are strong and obvious.”

Conditions are often critical. For example, leading experts have often cautioned against the use of humor in advertisements and analyses of non-experimental data supported the experts. However, experiments found that humor is effective under some well-defined conditions, and harmful under other conditions (e.g., high-involvement products with strong arguments).

---

\(^{1}\) Personal correspondence with David Walker, Research Director for Ipsos ASI.
Some of the principles conflict with common sense, such as “Do not mix rational and emotional appeals in an ad.” Such principles are likely to be particularly effective at improving persuasiveness relative to current practice.

It is easy to find violations of the principles. 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 this principle reduce the satisfaction of not only the customers, but also the sellers’ agents (Armstrong 2010).

One-hundred-and-ninety-five is a large number of principles so, even with knowledge on persuasion summarized in the form of principles, it is difficult for people to apply this knowledge when creating an advertisement. To help overcome this obstacle, *Persuasive Advertising* provided checklists. Checklists can help people to apply all relevant knowledge. The key requirement is that the items on the checklist be logical and well supported. Think of skilled workers in engineering, manufacturing, aviation, aeronautics, and surgery. In such fields, checklists have been found to save lives (Gawande 2010; Hales and Provonost 2006). Furthermore, computer checklists have been found to be more effective than those on paper (Boorman 2001), so we have provided software for the application of the persuasion principles. The software is available at AdPrin.com.

**Assessing the Validity of the Principles**

Pioneering advertising practitioners, particularly Ogilvy (1983), distilled their experience and their knowledge of the research into advice on how to design effective advertisements. In fact, the initial list of persuasion principles for the *Persuasive Advertising* book drew heavily on Ogilvy (1983). Ogilvy’s book has withstood the test of time and continues to be among the best sellers in advertising. The writings of Ogilvy, along with those of leading ad men Claude Hopkins (1923), Rosser Reeves (1961), and Ken Roman et al. (2003) include many guidelines. The principles used in the research presented in this paper are, to a considerable extent, consistent with those ad men’s guidelines. This consistency between the ad men’s guidelines and the persuasion principles provides evidence of face validity.

Ninety-one percent of the principles are validated in the sense that they were based on experimental evidence. The other principles were included on the basis of logic and face validity. Some of them seem obvious, such as “Do not violate taste or standards.”

After the principles were developed, a colleague, Sandeep Patnaik, helped to test the concurrent validity of the principles. This 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 refer to as *WAPB*). These data include matched pairs of ads, along with their recall scores. Given the data were controlled for some key variables, such as product category, media placement, and size of ad, we refer to these as quasi-experimental data. A description of these data is provided in Armstrong (2010, pp. 300-301).

To test the concurrent validity, we listed the principles that were based on laboratory or field experiments and then tested them against the *WAPB* quasi-experimental data. This applied to 7 principles supported by field experiments, 26 by laboratory experiments, and 7 by meta-analyses of experimental findings. The directional effects of the quasi-experimental findings agreed with the experimental findings in all of these 40 comparisons (Armstrong and Patnaik 2009). This surprised us in that the *WAPB* findings were often based on small sample sizes. This testing did not lead to any substantive changes in 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 quasi-experimental results (for the details, see Armstrong 2010, p. 301).

The tests of face validity and concurrent validity help to assess whether the principles are useful for designing more effective ads. The current paper provides another test, this one based on *predictive validity*. We use the principles to predict whether ads that more effectively apply the principles are more effective. To do this, we needed to develop an instrument that can assess ads based on their use of principles. It was also necessary to develop a training program for people to properly rate the ads.

Our test of predictive validity provides evidence on whether the principles might be useful for pretesting advertisements. Finally, it provides for an assessment of the amount of error reduction that advertisers can expect to achieve from the application of principles.

**Making use of Evidence-based Principles**

Advertising researchers have long attempted to use many variables to assess the persuasive aspects of advertisements. Of particular note is Stewart and Furse’s (1986) analysis of before and after responses from thousands of viewers of 1,059 TV commercials for 356 brands from 63 firms in 12 product categories. They used regression analysis to assess the ratings of about 160 features of TV commercials on recall; each feature was correlated with comprehension and with persuasion. Their study inspired replications including Stewart and Koslow (1989), with an additional 1,017 commercials; Laskey, Fox, and Crask (1994), with an analysis of data on 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: few variables appeared to have substantive impact, and the directions of the effects were in many cases puzzling. Why was that?

One reason is that even sample sizes of over 1,000 commercials are inadequate for regression analysis when there are many predictor variables (Dana and Dawes 2004). More importantly, regression analysis of non-experimental data cannot estimate valid relationships from many variables no matter how large the sample size because the variables in non-experimental data correlate with one another. The practical limit of regression analysis is typically a handful of variables (Armstrong 2012). Another reason is that regression analysis has difficulty in controlling for the interaction caused by conditions. This was noted above, in the case where humor harms persuasion under one set of conditions while it improves persuasion under other conditions.

**Index method**

The index method provides a solution for dealing with problems that involve more than a handful of important causal variables. The general idea of this method is to rely on *a priori* analysis to develop forecasting models for problems that involve a large number of variables. That is, instead of estimating weights from given data, one uses prior knowledge to select relevant variables and assess their directional influence on the target criterion. Once the relevant variables are included and their directional impact on the criterion is specified, prior research showed that the magnitudes of effects are not very important (Dawes 1979, Dawes and Corrigan 1974).

Our 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 important variables, rate the extent to which each variable favors each alternative, and to then add the ratings to see which alternative is better.

Another early application of the index method involved calculating an index score based on whether prison inmates were rated favorably or unfavorably against a list of 25 factors influencing the chance of successful parole (Burgess 1939). That application of the index method recently made a comeback, with news articles reporting the use of several computer programs that calculate index scores based on up to 100 predictor variables derived
from criminology research. Predictors include whether the offender is married, the age of first arrest, the type of crime, and the last grade completed in school (Walker 2013).

More recently, research has tested the index method for forecasting U.S. presidential elections using biographical information about candidates (Armstrong and Graefe 2011) and using voter perceptions of each candidate’s ability to handle certain issues (Graefe and Armstrong 2013). The resulting index models provided forecasts that were competitive with those from established models using other methods and their use substantially improved the accuracy of the election forecasts (Graefe et al. 2014, figure 4). But their biggest advantage over traditional election forecasting models is that they provide advice on what actions to take (e.g., who to nominate and which issues to emphasize in a campaign).

The index method is ideal for comparing alternatives in situations where there are many important variables and a large body of knowledge about causal relationships. This is because it does not estimate the relationships from a given set of data, such as the attempts to use regression analysis on large samples of ads. Instead, the relationships are determined *a priori* based on well-supported theory and evidence gained from experiments.

Forecasting with the index method involves five steps (following Graefe and Armstrong, 2011):

1. Identify all variables that are important to the problem.
2. Specify the direction and magnitude (i.e., the weight) of each variable’s effect.
3. Use data about the situation of interest to determine the variable values.
4. Calculate the index score by applying the variable weights from step 2 to the values from step 3, and then sum the resulting values.
5. Use index scores to make the forecast.

**Creating the Persuasion Principles Index**

To develop the index method for predicting which advertisement will be more effective, we implemented the five steps described above to obtain ratings of how well an ad uses evidence-based persuasion principles. Table 1 describes the Persuasion Principles Index (PPI) procedure. The specific formulas are provided in the Research Repository at AdPrin.com.
Table 1: Persuasion Principles Index (PPI) Procedure

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Variables</td>
<td>All 195 principles published in Armstrong (2010) were considered causal variables of ad effectiveness. Raters use descriptions of the principles to decide whether a principle was relevant to an ad.</td>
</tr>
<tr>
<td>2. Direction and magnitude of influence</td>
<td>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 with larger effect sizes are weighted more heavily.</td>
</tr>
<tr>
<td>3. Rating of advertisement</td>
<td>(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 out of five raters were identical. If there are less than three identical ratings achieved, that principle was dropped from consensus.</td>
</tr>
<tr>
<td>4. Index score calculation</td>
<td>The Creativity Score measures the extent to which the ad used relevant principles. That is, the percentage of all relevant principles that were implemented well. The Weighted Mastery Score measures how effectively the relevant principles were implemented. The Persuasion Principles Index (PPI) is calculated as the unweighted average of the Creativity Score and Weighted Mastery score.</td>
</tr>
<tr>
<td>5. Forecast calculation</td>
<td>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.</td>
</tr>
</tbody>
</table>

We followed Franklin’s advice to use subjective weights for the variables. For example, principles that relate to strategy (e.g., identify benefits of the product being advertised) were weighted more heavily 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 the analysis; details are provided in the spreadsheets on the Research Repository at AdPrin.com. We made no attempt to search for optimum weights, nor was it possible to do so with our data.

We then converted the principles from the Persuasive Advertising book to questions for the raters. In doing so, we had to make decisions on how many conditions should be included in each rating question. In addition to the questions, we provided supplementary
information on each principle for the raters to view in case they should need help in
interpreting the question.

The questions were examined by the authors for clarity and reworded as required. One author coded many ads as part of the effort to improve the wording of the questions. To pre-test the instrument, we used Amazon Turk to select advertising novices to each rate 40 print ads. This led to many changes in the final procedures used. The results from the pre-tests are not included in the analyses presented in this paper.

**Data:** We used full-page print ads from *Which Ad Pulled Best (WAPB)*, editions four through nine, which were published from 1981 to 2002 (Burton 1981; Burton and Purvis 1986, 1991, 1993, 1996; Purvis and Burton 2002). These volumes 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).

The *WAPB* advertisements were for products sold by major U.S. firms. Gallup and Robinson provided day-after recall scores for all ads. This represents the percentage of respondents who can accurately describe an ad the day following exposure. Each *WAPB* volume includes 50 pairs of advertisements, except the ninth edition, which included only 40. Thus, the total number of pairs of *WAPB* ads considered for inclusion in this study was thus 290. Further description of the *WAPB* data is provided along with a description of the recall measure in Appendix B of Armstrong (2010).

We used only pairs of ads for the same brand and product. From these, we selected only ads for high-involvement utilitarian products because we expected the persuasion principles to be more useful for predicting the effectiveness of ads for such products. A large number of the persuasion principles are relevant for advertising such products. If the principles do not have predictive value in situations where they are expected to be most useful, there would be little value to them in situations where few principles apply, such as for vague corporate image ads.

The lead author, a research assistant, and Sandeep Patnaik of Gallup and Robinson each independently screened the ads. The sample was limited to those advertisements for which the three screeners agreed that the ad met our criteria. The selection process yielded a final sample of 96 pairs of ads.

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 would be that the relationship between the compliance with principles and the persuasiveness of the advertisements would be underestimated.

**Testing the Accuracy of the Persuasion Principles Index**

In this section we describe the selection of the raters, their training, the task, and the creation of the consensus for the PPI.

**Selection of Raters**

We wanted to develop procedures that could be used by clients, pretesting services, or advertising experts. We conducted our research primarily with novices. If those results show predictive validity, they would likely be a lower bound. In other words, the use by clients, experts, and especially pretesting services would be expected to be better. However, it was beyond our capabilities to address these issues in this study.

We recruited university students as raters for the task and paid them about $10 per hour. In addition, raters were hired from Amazon Mechanical Turk with $80 for the task (rating 20 pairs of ads) and a bonus awarded depending on the number of predictions they made that were correct. Seventeen raters were involved in total.

**Training the Raters**

All raters were first required to complete a self-training module on AdPrin.com during which 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.

**Rating task**

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

**Consensus ratings**

To improve reliability, we used five raters. An administrator summarized across raters by checking for principles on which three or more raters agreed. (The administrators had no knowledge of the recall scores.)
We developed a procedure for excluding unreliable raters, which examined how closely the raters’ ratings corresponded with the consensus rating on each principle. To do this, the software calculates a Rater’s Reliability Score. Raters who departed more than 10 percentage points from the average Reliability Score among the five raters were dropped and replaced by new raters.

**Results on the Predictive Validity of the Persuasion Principles Index**

On average, ratings by individual raters correctly predicted which ad in each pair had the highest recall score for 61.0% of the 96 pairs. Using the consensus of five raters (see Step 3b in Table 1) to calculate the index scores increased accuracy, as expected. The consensus PPI scores were correct for 74.5% of the 96 pairs. Thus, reliability was improved by combining across raters, and this substantially improved the predictive validity of the PPI.

As noted earlier, the WAPB report recall, whereas the PPI is designed to measure persuasion. The predictive validity is limited, then, by the 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 the product. This correlation implies that for binary data one would expect an upper limit on predictive validity in our study to be about 76%. In addition, the PPI is based on behavior, not intentions. Although there is a close relationship between intentions and behavior (Kim and Hunter 1993), especially for high-involvement products (Morwitz 2001), and such measures are unbiased (Wright and McRae 2008), this represents yet another problem in assessing the PPI. Best to test with behavioral data.

In addition to the problem with the criterion, there were challenges in selecting and training raters. We expect that a well-trained staff of raters would substantially improve the effectiveness of the PPI. Finally, prediction was difficult because the PPIs for many pairs of ads did not differ much.

**Accuracy Comparisons with Benchmark Forecasts**

We obtained forecasts from currently used methods for forecasting the effectiveness of advertisements in order to assess the value of the Persuasion Principles Index. Typically, advertisers test only a small proportion of advertisements. Perhaps the lack of testing is due to

---

2 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 of course. For example, if you used the one-tail binomial test, you would find that these results differ from chance at \( p < 10^{-6} \).
the confidence that they have in their ability to identify the most effective ads. If advertising practitioners are highly confident in their predictions, they might see no need to consider other approaches.

**Unaided judgment**

By unaided judgment, we mean that the subjects were given no instructions on how to make a prediction. Practitioners commonly predict the effectiveness of advertisements using their unaided judgment. To assess the value of expertise, we obtained unaided judgments from novices as well as from subjects with some experience in advertising. Finally, we used copy-testing. Copy testing is regarded by many as the most effective way to predict ad effectiveness.

**Subjects:** To obtain unaided judgmental predictions, we first sought advertising practitioners. Despite following many leads, we had limited success in gaining participation by practitioners. Those we contacted typically did not respond. Those who did respond typically said they were too busy or not interested. Thus, we can only provide weak tests on the value of expert judgment for choosing which ad will be more recalled.

We obtained predictions from seven practitioners 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 practitioners from Amazon Mechanical Turk. These participants claimed to have had at least one year of experience working in advertising.

Finally we recruited novices, including 113 unpaid volunteers who were mostly university students and 450 participants recruited through Mechanical Turk. The Turkers were paid $1 per batch of ads.

**Procedure:** We directed the expert and novice judges to an online questionnaire for one of the five batches of ads described above. The questionnaire asked, “Can you predict which advertisement had the better ‘day-after recall’? Think of recall as a measure of effectiveness.” In addition, the questionnaire asked, “How confident are you of your prediction?” and provided a scale from 50% to 100%, where 50% equals guessing. 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 instruments are provided on the Research Repository.)
Accuracy of unaided judgments

We had no expectations that the results would differ between unpaid and paid participants. Due to the similarity, the results were merged. A formal analysis of differences between the responses of volunteers and paid participants is available in the research repository at AdPrin.com.

Individual unaided judgments by novices were of some value for predicting the effectiveness of advertisements, with 54.1% of 10,809 individual judgments correctly identifying the ad with higher recall.

On average, the practitioners’ unaided judgments were correct for 55.4% of 2,764 predictions. The small effects on accuracy relative to the judgments of novices and guessing would be difficult to notice by unaided judgment. This is consistent with John Wanamaker’s observation that he did not know which half of his money spent for advertising was wasted.

Given the persuasive ads that were developed by industry leaders such as Bernbach and Ogilvy, we expect that some practitioners would do well. In other words, our test may underestimate the predictive skill of practitioners. For example, the 16 practitioners who were working at advertising agencies did better than the average expert in our sample, with 59.7% correct of the 320 predictions they made.

On the other hand, extensive prior research on the value of expert judgmental forecasts in complex uncertain situations (Armstrong 1980, Stewart 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 from novices (Armstrong 1991). In the light of the prior evidence, we suggest caution in interpreting our findings on practitioners’ forecasts.

Our practitioners were more confident about their predictions than were the novices: practitioners expected to achieve 85 percent correct predictions, whereas novices expected to achieve 78 percent correct predictions.

Combined judgments: It is often possible to improve the accuracy of individual forecasts by combining them, especially if the individual forecasts are based on different methods and different data (Graefc, Armstrong, Jones Jr., and Cuzán, 2014). However, we expected that the gains from combining would be small when only unaided judgments were combined and also given that the accuracy of individual judgments was poor.

To combine the forecasts, we identified the modal forecast; in other words, the ad in each pair that most of the judges expected to be more effective. Accurate predictions from combined judgment were those for which the modal forecast was the ad with the higher recall score. Ties were scored as half of an accurate prediction (i.e., 0.5).
Combining the 563 novices’ judgments increased accuracy from 54.1% to 59.0%. Combining the 144 practitioners’ judgments increased accuracy from 55.4% to 63.9%. Note however, that the use of such combined practitioners is apparently rare in advertising agencies. We expect, based on reading of secondary sources, such as the detailed observations on the behavior of agencies and clients, that these decisions are typically made in meetings (see, e.g., Armstrong 1996). In contrast to the benefits of simple equal weights combination across practitioners, group meetings are likely to harm predictions (Armstrong 2006). Leaders of creative agencies, such as David Ogilvy, George Lois, and Bill Bernbach were highly critical of meetings. Ogilvy said that, “Commercials should never be created in a committee.”

Copy Testing

While there are many types of copy testing, we provide only a single test here.

Subjects: We recruited 369 participants from Amazon Mechanical Turk to perform typical copy testing tasks. Each subject was paid $2 per batch, ten cents per pair of advertisements.

Procedure: The WAPB ads used in this research were published from 1981 to 2002. We were concerned that the 10 to 30-year age of the ads might influence our copy-testing participants’ perceptions and reactions. To address this problem, we gave the participants a role:

“We will show you 20 old black and white print advertisements. Please answer the 4 questions about each ad, imagining 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?

The intentions-to-purchase was calculated by averaging each participant’s likelihood, on a scale from 0 to 100, derived from each of the three questions. The intentions-to-purchase from each participant were then averaged for each ad. (Detailed information on the questionnaire is available on the Research Repository at AdPrin.com).
Three different procedures were implemented to obtain participants’ intentions-to-purchase:

Procedure 1: *Rate one ad from each pair.* We asked participants to rate either Ad A or Ad B.

Procedure 2: *Rate each pair twice.* We asked participants to rate the same pairs of ads twice, with the second rating conducted two weeks after the first one.

Procedure 3: *Rate two ads at different times.* We asked same participants to rate Ad A then, approximately two weeks later, Ad B.

To assess the accuracy of copy-testing forecasts, we determined whether the ad with the higher intention-to-purchase in each pair also had the higher recall score. Given that we had no prior expectations on the relative accuracy of the three procedures for obtaining intentions to purchase, we weighted each rating equally. Doing so yielded 5,285 predictions from 369 subjects.

**Comparative Results**

Across the three intentions elicitation procedures, 59.4% of the copy testing predictions were accurate. Thus, the copy-testing forecasts we obtained were 5 percentage points more accurate than the individual unaided judgments of novices and 4 percentage points more accurate than the individual practitioners’ judgments. The accuracy figures for the three procedures are presented in Table 2.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Number of participants</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate one ad from each pair</td>
<td>253</td>
<td>62.2</td>
</tr>
<tr>
<td>Rate each pair twice</td>
<td>33</td>
<td>50.6</td>
</tr>
<tr>
<td>Rate two ads at different times</td>
<td>83</td>
<td>58.2</td>
</tr>
<tr>
<td><strong>Sum/Average</strong></td>
<td><strong>369</strong></td>
<td><strong>59.4</strong></td>
</tr>
</tbody>
</table>

**Discussion**

Our objective for this study was to determine whether or not advertisements that conform 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. This finding
provides additional evidence to that provided by Armstrong (2010) and Armstrong and Patnaik (2009) on the validity of an evidence-based approach to persuasion. The accuracy of the forecasts we obtained from each method tested in this study is provided in Table 3.

Table 3: Accuracy of forecasts from index and benchmark methods

<table>
<thead>
<tr>
<th>Predictions</th>
<th>Percent correct</th>
<th>Percent error reduction v chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persuasion Principles Index (PPI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus of 5 raters per ad</td>
<td>96</td>
<td>74.5</td>
</tr>
<tr>
<td>Unaided novice judgments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual</td>
<td>10,809</td>
<td>54.1</td>
</tr>
<tr>
<td>Combined</td>
<td>96</td>
<td>59.0</td>
</tr>
<tr>
<td>Unaided expert judgments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual</td>
<td>2,764</td>
<td>55.4</td>
</tr>
<tr>
<td>Combined</td>
<td>96</td>
<td>63.9</td>
</tr>
<tr>
<td>Copy testing</td>
<td>5,285</td>
<td>59.4</td>
</tr>
</tbody>
</table>

The PPI forecasts were substantially more accurate than those from the currently used methods that we tested. Moreover, obtaining PPI forecasts is inexpensive. The costs for obtaining the ratings used to calculate the PPI scores are expected to be small relative to the costs of developing and placing advertisements. After one-hour of self-training, each rater took about 45 minutes to rate each pair of ads. While the time taken to obtain PPI forecasts exceeds the time taken to make unaided judgments (one minute per pair), and for copy testing (about two minutes per pair), the cost would nevertheless be trivial for advertising campaigns, especially for TV commercials.

While this study focuses on predictive validity, the PPI can also be used as a diagnostic tool. The ratings used to calculate the index scores identify violations of evidence-based principles detected by the raters that could be remedied by the advertiser. Advertisers could also endeavor to improve the application of principles identified as being currently used in the ads, and consider using principles that were not applied but were found to be relevant by the raters. For example, in the research described in this paper, a typical ad violated 2 principles, was only partly successful in applying 16 principles, and overlooked 25 relevant principles.

Though we tested the use of persuasion principles for print advertising, we expect that they could be profitably applied to online commercials, where they would be easy to test, and to TV commercials, for which the budgets are often enormous. 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).

Our findings are consistent with prior research and the effect sizes are large. This is heartening given that the results were obtained from raters who had little prior knowledge about the product and only a very short training period. In addition, the criterion (recall) that was available is not strongly related to sales or other desired behavior changes. We thus expect that the results underestimate the strength of the relationship between the PPI and advertising effectiveness.

However, this is only the first study on the predictive validity of the PPI. Replications and extensions are needed in order to further test the approach we have proposed. We did, however, have the benefit of an accidental replication. Due to a damaged hard drive, the ratings of the principles by nine raters were lost. Given the need for full disclosure, we decided to drop those ratings. We then recruited new raters. In addition, one of the authors reviewed all of the prior ratings and reanalyzed all of the ratings, correcting a number of small mistakes in doing so. The original PPI was correct for 76% of the predictions, whereas the PPI after replacing the “lost raters of the PPI” was correct for 74.5%.

Conclusions

This study provides yet another test of the validity of advertising principles. The principles, in the form of a Persuasion Principles Index (PPI), were used to predict the relative effectiveness of 96 matched pairs of print advertisements for high-involvement utilitarian products by leading advertisers.

The index method enabled us to assess the persuasiveness of advertisements by obtaining ratings of how well the ads complied with up to 195 persuasion principles. Advertisements that more closely followed the evidence-based principles were more effective than those that did not. The higher consensus index score correctly identified the more recalled ad for 74.5% of 96 pairs, compared with the method typically used by practitioners, unaided judgmental, which provided 55.4% correct predictions. This difference in accuracy represents a forecast error reduction of about 43%.

The Persuasion Principles Index (PPI) also provides information on how to improve advertisements by indicating which principles are violated, which could be applied more effectively, and which relevant principles are not used. We expect that Mr. Wannamaker would be pleased that we provided some insights into his concern over which half of his ad dollars were wasted.
The results provide further support for the conclusion that, as in other areas, applying knowledge in the form of evidence-based principles results in superior outcomes. The knowledge on advertising persuasion is now available in the convenient form of the PPI. The PPI can be rapidly and inexpensively applied to assessing and improving advertisements.

Conformance with evidence-based persuasion principles, as measured by the Persuasion Principles Index, provides yet another way to pre-test ads. The present study is only the first test of this approach. Extensions are needed to see if the findings hold up.

References


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


Nyilasy, G. & L. N. Reid (2009), Agency practitioners’ meta-theories of advertising,


Words: text only = 6,640; Total = 7,900