Extending validity testing of the Persuasion Principles Index

Nicole Hartnett
Ehrenberg-Bass Institute, University of South Australia, Adelaide, Australia
Luke Greenacre
Monash Business School, Monash University, Melbourne, Australia, and
Rachel Kennedy and Byron Sharp
Ehrenberg-Bass Institute, University of South Australia, Adelaide, Australia

Received 19 December 2018
Revised 6 September 2019
14 February 2020
22 May 2020
Accepted 31 May 2020

Abstract

Purpose – This study aims to independently test the predictive validity of the Persuasion Principles Index (PPI) for video advertisements for low-involvement products with a measure of in-market sales effectiveness. This study follows the inaugural test conducted by Armstrong *et al.* (2016) for print advertisements for high-involvement utilitarian products with a measure of advertising recall.

Design/methodology/approach – The method was in line with that developed by Armstrong *et al.* (2016) for rating advertisements and assessing the reliability of ratings. Consensus PPI scores were calculated for a data set of 242 matched pairs of television advertisements. For each pair, the authors determined whether the advertisement that better adhered to the persuasion principles performed better in-market.

Findings – Consensus PPI scores predicted the more sales effective television advertisement for 55% (confidence interval (CI) = 49%, 61%) of the 242 pairs. This result is no better than chance and much weaker than the result from the initial validation study, which found that the consensus PPI scores predicted the more recalled print advertisement for 74.5% (CI = 66%, 83%) of 96 pairs.

Research limitations/implications – This study replicated the application of the PPI as per Armstrong's guidelines and extended validity testing to a different set of advertising conditions. Findings indicate that better adherence to the persuasion principles produces only a weak, positive effect for predicting the performance of television advertisements for low-involvement products. A research agenda that flows from the results is discussed.

Practical implications – The authors suggest that the PPI in its present form is best used to predict advertising performance under conditions as per the inaugural validation test (Armstrong *et al.*, 2016).

Originality/value – Advertisers will require compelling evidence of the PPI's predictive accuracy to adopt the tool for pre-testing advertising. This study is the first independent test of the predictive validity of the PPI and its generalisability across advertising conditions. Another contribution of this study is the assessment of Armstrong's advice to remove unreliable ratings. The authors show that this procedure, surprisingly, does not improve the predictive accuracy of the PPI.

Keywords Television advertising, Creative, Index method, Prediction, Sales effectiveness

Paper type Research paper

Introduction

Advertising has long been the domain of the creative artists in the marketing industry. Now emphasis is shifting to some combination of art and science in the process of developing advertising. Although there is a substantial body of research into creative strategies and tactics that may increase (or decrease) the effectiveness of advertising, there have been few attempts to fashion these findings into useful guidelines to help advertisers to become more

The authors would like to thank and acknowledge Mars, Incorporated for providing the advertising data.

European Journal of Marketing
© Emerald Publishing Limited
0309-0566
DOI 10.1108/EJM-12-2018-0865

evidence-based with their creativity. One attempt to provide a systematic basis for creating and evaluating advertising is that of Armstrong (2010). He proposed a comprehensive set of 195 persuasion principles that could be used to design or choose more effective advertisements, drawing on evidence from thousands of academic and industry sources.

An index model for evaluating advertisements was later developed based on those principles (Armstrong *et al.*, 2016). The resulting Persuasion Principles Index (PPI) has attracted both praise and debate, with a near-unanimous call for more research into evidence-based advertising predictions, and the PPI in particular (Gendall, 2011; Green *et al.*, 2016; O'Keefe, 2016; Sharp and Hartnett, 2016; Woodside, 2016; Wright, 2016). Indexes provide a useful means to forecast outcomes affected by many causal variables (Armstrong and Green, 2018; Graefe and Armstrong, 2011). In the advertising context, indexes can be used to assess the potential relative effectiveness of different advertisements based on the creative strategies and tactics used by advertisers. Depending on their validity, indexes could help marketers to make better decisions about which advertisements they should air, and possibly inform the allocation of media expenditure across different advertisements.

Armstrong *et al.* (2016) conducted the first validation test of their index. Its predictive validity was assessed for advertisements that had a measure of day-after recall. Advertisements that better adhered to the persuasion principles were expected to achieve higher day-after recall. The data set consisted of 96 matched pairs of print advertisements for high-involvement utilitarian products (e.g. electronics and automotive). The prediction accuracy of the consensus PPI was 74.5% [1] (confidence interval (CI) = 66%, 83%). This result exceeded the predictions of alternative methods to assess advertising performance applied to the same 96 pairs, such as combined novice judgements (62%) (CI = 51%, 71%), expert judgements (64%) (CI = 55%, 73%) and a copy testing method with a measure of purchase intent (59%) (CI = 49%, 69%).

From this test, the PPI appears to show promise, though concerns were raised by commentators about some of the particulars of how it was tested. For example, the validity test focused solely on print advertisements exclusively for high-involvement utilitarian products (O'Keefe, 2016; Sharp and Hartnett, 2016), used a recall-based measure of advertising effectiveness rather than a behaviour-based measure (Sharp and Hartnett, 2016; Woodside, 2016) and did not evaluate the effectiveness of the rating reliability procedure used (O'Keefe, 2016). Further, the validation test made predictions for a sample of advertisement pairs that were a subset of a larger pool that had been used to develop some of the 195 persuasion principles (Sharp and Hartnett, 2016). A stronger validation test would be to use an altogether new sample of advertisements.

Responding to commentators, the creators of the PPI supported the view that:

> [. . .] replications and extensions using behavioral data and alternative implementations of the index method would help to better assess the effects of judging conformity with principles as a means of predicting relative advertising effectiveness (Green *et al.*, 2016, p. 317).

This study addresses these concerns. We applied the PPI to a new sample of television advertisements for low-involvement products (e.g. human and pet food) and compared predictions to a measure of sales performance. We further examined the impact of the rating reliability procedure on the PPI's predictions, to assess whether this procedure improves prediction accuracy.

Predicting the effectiveness of advertisements

Assessing the potential effectiveness of advertisements has been a challenge for advertisers since the birth of the industry. No two advertisements are the same by virtue of including complex combinations of information, emotional appeals, persuasion attempts and attention-getting tactics. There is no simple way to map the contribution of individual execution elements on to a selected measure of advertising effectiveness, of which there are many. Examining the combinations and potential non-linear interactions of these execution elements presents an even more complex problem. Nonetheless, we know that advertisements differ enormously in effectiveness (Hartnett *et al.*, 2016a; Jones, 1995; Wood, 2009) and advertisers want to know which advertisements are better and (ideally) why.

The most common tools that marketers use to assess advertisements *before* launch are managerial judgement and pre-testing. Managerial judgement attempts to draw upon experience gained from past advertising successes and failures. Decisions based on judgement have the benefit of speed. However, intuitive or unstructured judgements often prove to be poor predictors of marketing outcomes (Armstrong, 1991; Hartnett *et al.*, 2016b). Pre-testing can be used to augment or supersede judgement. Traditional pre-testing, which is still popular, typically exposes consumers to an early production of an advertisement in a controlled environment to gauge their responses. These responses can be recall (memory), message comprehension, persuasiveness or any number of other measures. Importantly, pre-testing contributes a data-driven approach to testing, which goes some way towards overcoming the flaws associated with judgement. Prediction may improve with pre-testing (Haley and Baldinger, 2000) but it is far from perfect, and evidence for its validity is limited and conflicting (Blair, 1987; Lodish *et al.*, 1995).

The index method provides another tool in the advertiser's toolbox. An index method takes a different approach from both managerial judgement and pre-testing by focusing on the creative strategies and tactics that form the advertisement. Compared with other methods, applying an index has the potential advantage of producing more diagnostic information about how an advertisement might be improved. The specific creative strategies and tactics captured by the index are drawn from the collective knowledge of what has been observed to work more often, subject to advertising conditions. Pre-testing most often compares to benchmarked performance, with a limited diagnosis for why an advertisement may or may not work in creative terms. In contrast, the index method can theoretically increase the persuasiveness of advertisements by helping advertisers to ideate using a wider variety of creative strategies and tactics, and to consistently consider the many complex inputs that advertising can draw upon (Armstrong, 2010, 2011).

Predictive validity of the Persuasion Principles Index

The PPI incorporates 158 persuasion principles that cover the broad areas of information, influence, emotion, mere exposure, overcoming resistance, acceptance, message and attention. Additionally, there are 24 principles specific to still media (e.g. print) and 13 principles specific to motion media (e.g. television). Most principles are supported by at least one source of empirical evidence. Initially some principles lacked extensive empirical support, and for this reason, the list of principles was evaluated using the *Which Ad Pulled Best* (*WAPB*) data (Armstrong and Patnaik, 2009). *WAPB* is quasi-experimental data that controls for some extraneous variables by analysing print advertisements in matched pairs that are similar with respect to the product, target market and media. In each pair, one advertisement performed better than the other. Performance is judged based on a measure of brand-prompted day-after recall collected by a market research company. In that research, 150 adults were interviewed about a magazine they had received the previous day containing the target advertisements. It was confirmed that adherence to any one of the 56 persuasion principles selected for this research phase had some positive impact on recall (Armstrong and Patnaik, 2009). This process helped to establish the final list of persuasion principles that comprise the PPI, and contributed evidence that Armstrong used to differentially weight principles when calculating a PPI score.

With the principles established and the weightings determined, the final index method was validated against a subset of the same *WAPB* data (Armstrong *et al.*, 2016). As previously noted, the validation produced successful predictions for 74.5% (CI = 66%, 83%) of the *WAPB* subset, which consisted of 96 pairs of advertisements for high-involvement utilitarian products. This specific subset was selected because the authors:

> [. . .] 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 (Armstrong *et al.*, 2016, p. 283).

Although the PPI performed well in this test, several commentaries have noted limitations, some of which are addressed in this paper. The most critical is the use of day-after recall as the measure of advertising effectiveness. Recall measures are imperfect proxies for purchase behaviour that arises in response to advertising exposure (Woodside, 2016). The use of recall does not capture *persuasion* as effectively as other behavioural measures do. In our study, we have used a population-level sales measure to validate the PPI, providing a stronger test of persuasion. Although it has been argued that sales measures can be excessively noisy (Wright, 2016), in our case, we used a measure of sales uplift benchmarked against a “no exposure” baseline of sales. Such an approach is well documented (Bellman *et al.*, 2017; Hartnett *et al.*, 2016a; Taylor *et al.*, 2013). Some commentaries suggested that tests conducted under different conditions (e.g. media, class of product and market) are necessary to assess how generalisable the PPI’s predictions really are (O’Keefe, 2016; Sharp and Hartnett, 2016). Print has long provided a fruitful avenue for copy creators to express their creativity, but video advertisements offer substantially more opportunity for creativity in execution. Video incorporates not only text and image copy, but also dynamic movement, audio and temporal elements. Consequently, video introduces many more execution elements to account for than print. Importantly, the PPI includes principles specific to motion media that were precluded from the initial validation study using print advertisements. These principles have thus not been validated at this point in time. There is also a need to validate the PPI for products that do not fit the classification of high-involvement purchase decisions. P&G, Nestlé and Unilever, which are sellers of low-involvement products such as shampoo, cereal and household cleaning products, are some of the world’s biggest advertisers, collectively spending more than $28bn in 2016 (Johnson, 2017). It would be useful to have evidence that the PPI works well across multiple classes of product. Some of the persuasion principles included in the PPI do specifically address low-involvement products (e.g. Principle 6.6.2 “Does the ad use celebrity endorsement to gain attention for a low-involvement product?”). In our study, we used television advertisements for low-involvement products to validate the PPI.

Data

A global consumer packaged goods (CPGs) manufacturer provided us with 312 television advertisements that had aired between 2000 and 2013. The advertisements spanned more than 60 competitive brands in 4 food categories sold across Europe and America. The brands varied in market share and the advertisements varied in length, with most running for 30 s (48%) or 20 s (33%).

Unlike the original validation study, which used day-after recall, this study used a measure of sales effectiveness to assess advertising performance. The main point of brand advertising is usually to drive sales; hence, sales effectiveness provides a strong external validity test of the persuasion principles and the resulting index. The sales effects of the advertisements were calculated from single-source data collected from large panels of households in the specified regions. The effects were measured based on purchases scanned in a four-week period, with advertising exposures logged via set-top boxes in the prior four weeks. The measure of short-term sales effectiveness is a proprietary index that compares brand purchases made by exposed and unexposed households, similar to Jones' (1995) *Short-Term Advertising Strength*, but uses layered contingency tables to account for other impacting and extraneous variables (e.g. the frequency of advertising exposure and promotional activity). The approach, therefore, estimates the probability of a brand's purchase being based on the consumer's exposure to the brand's advertising (vs not), such that the sales indexes isolate changes driven by the advertising execution alone. The sales indexes differed by category and country (i.e. some categories/countries were more responsive to advertising than others); hence, the raw index scores were converted to a three-level ordinal variable of *above-average*, *average* and *below-average* sales effectiveness. The determined cut-offs for *above-average* and *below-average* were ±5 index points from the average index score for the category/country, giving confidence that the relative sales effectiveness of the outer groupings was meaningfully different, as well as consistent across categories, countries and time periods.

In line with Armstrong *et al.* (2016), the purpose of the PPI is to compare the effectiveness of multiple advertisements, hence we created 242 pairs of advertisements from a pool of 312 advertisements to test the predictive validity of the PPI. The paired advertisements were for the same brand, product category and country. Advertisements were paired irrespective of length and about half of the pairs compared advertisements of the same length. It should be noted that whether the advertisements were of the same length or not did not practically (nor significantly) impact the PPI's prediction accuracy [χ^2 (1) = 0.772, p = 0.379]. Pairs covered combinations of all levels of sales effectiveness: *above-average with below-average* (106 pairs), *above-average with average* (68 pairs) and *average with below-average* (68 pairs). Sometimes one advertisement was present across multiple pairs. For example, if there were four advertisements for one brand from the same category/country, of which, one was an *above-average* performer and three were *below-average*, three pairs were created by matching the *above-average* advertisement with each of the *below-average* advertisements. Few pairs (3%) contained advertisements not present in any other pair.

Method

The method was in line with that developed by Armstrong *et al.* (2016). Raters assessed which persuasion principles applied to each advertisement in each pair. The PPI was calculated using those assessments, and the resulting predictions were compared with actual relative sales performance.

Selecting and training raters

A total of 26 raters were recruited, most of whom were university students. After an initial evaluation of the reliability of the ratings, another eight raters were recruited to replace unreliable ratings. Native language speakers were used to assess the pairs. The raters were remunerated for their time.

Raters undertook the training available on Armstrong's website (www.advertisingprinciples.com) and were supplied with a reference copy of *Persuasive Advertising* (Armstrong, 2010). Rating advertisements was a two-step process. Firstly, raters assessed whether a principle applied to Ad A and/or Ad B for a sample pair of print advertisements supplied by Armstrong. Secondly, raters assessed whether that principle was applied well, needed improvement, was violated or was not used by Ad A and Ad B. Discrepancies between what the rater produced and Armstrong's supplied solution for that pair were discussed with the principal researcher, who was well versed in the principles, to come to a mutual understanding of how to best rate each principle. Once raters had satisfactorily completed the training, they were allocated pairs of video advertisements in random order to rate over a number of sessions. Each session typically lasted 3–4 h to avoid rating fatigue. The raters rated both advertisements in a pair at the same time. Each pair was allocated to 5 independent raters, resulting in 1,210 ratings across the 242 pairs. The PPI scores were then calculated for all pairs based on a consensus rating.

Inter-rater reliability

Two reliability scores were generated for each rater for each pair:

(1) a rating reliability score (for each advertisement); and

(2) a relevance reliability score (for the pair), as per Armstrong's automated calculations.

The rating reliability score represents the number of principles for which a rater's rating agrees on whether each principle is applied well, needs improvement, is violated or is not used with at least two other raters (achieves consensus), divided by the number of principles for which a consensus rating is achieved across all five raters. The relevance reliability score uses the same process, but examines whether the principles are considered relevant to the pair or not. The rating reliability and relevance reliability scores are then averaged. The two reliability scores for the two advertisements are then averaged again, to give an overall reliability score for the pair. Finally, the reliability scores for each rater are averaged and the raters' differences from this average reliability score are used to assess their reliability.

Armstrong *et al.* (2016) recommend removing ratings when the reliability score differs from the average reliability score by more than 10 percentage points. Using this criterion, 139 of the 1,210 ratings had to be replaced with more reliable ratings, affecting 124 of the 242 pairs. The procedure was iterative, where we first replaced the *least* reliable rating (if indeed two ratings for a pair were more than 10 percentage points above and below the average reliability score), and reassessed the ratings relative to the new average reliability score. As will be discussed later, the impact of this rating reliability procedure on the predictive validity of the PPI is marginal in this case.

Calculating the Persuasion Principles Index scores

Index scores can be calculated for each rater's rating and for the combined consensus ratings, though the latter tends to be more reliable and predictive (Armstrong *et al.*, 2016). A series of sequential calculations generate the PPI scores, which are detailed in Table 1. The interpretation of the resulting scores is that the advertisement with the higher score within a pair is predicted to be more effective than the advertisement with the lower score.

Results

The consensus PPI scores predicted the more sales effective advertisement for 133 of 242 pairs, which is a prediction accuracy of 55% (CI = 49%, 61%). This result, statistically no better than a 50% chance, is much weaker than the result from the initial validation study, which reported a prediction accuracy of 74.5% (CI = 66%, 83%) (Armstrong *et al.*, 2016).

Some of the advertisement pairs had bigger or smaller differences in their relative sales performance. Presumably there should be a higher level of prediction accuracy for the more

Weighted score (per principle)	= ((2 × applied well) + (1 × needs improvement) – (2 × violated)) × (1 + important) × evidence × effect size
Maximum score (per principle)	= 2 × (applied well + needs improvement + violated) × (1 + important) × evidence × effect size
Strategy weighted score	= Sum of weighted scores for strategy principles/sum of maximum scores for strategy principles
Tactics weighted score	= Sum of weighted scores for tactics principles/sum of maximum scores for tactics principles
Weighted mastery score	= Strategy weighted score × 0.5 + tactics weighted score × 0.5
Strategy creativity score	= Sum of strategy principles applied well/sum of relevant strategy principles
Tactics creativity score	= Sum of tactics principles applied well/sum of relevant tactics principles
Creativity score	= Strategy creativity score × 0.5 + tactics creativity score × 0.5
PPI	= Weighted mastery score × 0.5 + creativity score × 0.5

Table 1. Intermediate steps in calculating the PPI

extreme pairs, which contrasted *above-average with below-average* sales performance, compared with the less extreme pairs, which contrasted *for above-average with average* and *average with below-average* sales performance. However, we found no significant difference in the prediction accuracy across these different pairings [2]. That said, the direction of the difference in prediction accuracy was in line with expectations. The more extreme pairs had a numerically greater prediction accuracy of 61% (CI = 52%, 71%), relative to *above-average with average* of 47% (CI = 35%, 59%) and *average with below-average* of 53% (CI = 41%, 65%). The prediction accuracy of 61% (CI = 52%, 71%) for the more extreme pairs is notable, as it predicted above statistical chance.

An alternative approach to assessing prediction accuracy is to examine the differences in the consensus PPI scores between Ad A and Ad B in the pairs. We expected larger absolute differences between scores when looking at the more extreme pairs compared with the less extreme pairs. However, there were no significant differences [$F(2,239) = 0.347$, $p = 0.707$]. The more extreme pairs of *above-average with below-average* had only a slightly greater average absolute difference of 6.3 index points compared with the less extreme pairs; 6.0 index points for *above-average with average* pairs and 5.7 index points for *average with below-average* pairs.

These results indicate that the PPI is sensitive to predicting extreme differences in sales performance. However, the lack of prediction accuracy for the less extreme pairs, coupled with the lack of difference in PPI scores between advertisements across the more extreme and less extreme pairs, suggests that the PPI is not so sensitive to real differences in sales performance across the whole distribution of advertisements.

Inter-rater reliability and accuracy of prediction

One concern raised about the procedure of replacing ratings is its potential to undermine the predictive validity of the PPI (O'Keefe, 2016). The naïve removal of ratings with reliability scores *more* than 10 percentage points *above* the average reliability score is tantamount to removing the best ratings.

To examine whether or not the procedure affected the results, analyses were run on the initial (unreliable) set of ratings before any had been replaced. The consensus PPI scores without the rating reliability procedure implemented predicted the more sales effective advertisement for 137 of 242 pairs, which is a prediction accuracy of 57% (CI = 50%, 63%). Although this small change would suggest that the rating reliability procedure had only a marginal effect on the PPI's prediction accuracy (a 2% decrease), such a blunt comparison hides a substantial change in the predictions of individual pairs.

Of the 124 pairs that had at least one rating replaced, the consensus prediction *reversed* for 36 pairs (29% [CI = 22%, 38%]). That is, the predicted more effective advertisement changed from Ad A to Ad B, or vice versa. It was only because the reversals counterbalanced each other that we came close to the prediction accuracy of 55% (CI = 49%, 61%) with the more reliable ratings. Therefore, the rating reliability procedure appears to have a substantive impact on calculating the PPI, even though in this circumstance, the overall prediction accuracy was not significantly affected.

Discussion and conclusion

This study sought to validate the PPI under conditions that differed from those in the inaugural validation study, using television advertising for low-involvement products with sales effectiveness as the performance measure. These conditions are relevant to many advertisers and offer a strong test of the PPI both practically and theoretically.

Although the initial validation study (Armstrong *et al.*, 2016) showed that the PPI had a very good prediction accuracy, our validation study presents a far weaker result with a prediction accuracy approaching random chance. Such a marginal result does not refute the potential value of the PPI but it does suggest a need to invest more in validating and developing it. Notably for our data set, the PPI was able to detect the most extreme differences in advertising sales effectiveness. This suggests that the PPI *could* prevent the poorest performers from being aired. The difficulties that the PPI has in predicting more moderate differences in performance is similarly present for more expensive and time-consuming commercial testing approaches (Lodish *et al.*, 1995).

The are several possible reasons for why our result is weaker than Armstrong *et al.* (2016). Most reasons arise from differences between the samples of advertisements used in the studies, as shown in Table 2. Any one, or a combination, of these differences may be a primary cause of the divergence in results.

The first potential reason for the weaker result is that the PPI may work best for print advertising. The principles are weighted towards experimental evidence from studies conducted with print advertisements and not video advertisements as we tested. It might be that the PPI is of limited usefulness in pre-testing anything except print (or static) advertisements. This concern highlights the potential danger in developing a generalised index from a limited set of advertisement types, with the development of a generalised index needing to draw on a large pool of diverse advertisements.

The second potential reason is that the PPI works best for advertisements for high-involvement products. The PPI is largely based on a high-involvement persuasion model of

Table 2. Differences between Armstrong *et al.* (2016) and the present study

Study details	Armstrong *et al.* (2016)	Present study
Advertising media	Print (full page)	Video (15–90 s)
Number of ad pairs	96	242
Product types	High involvement Utilitarian	Low involvement Staple and impulse CPG
Advertisement sources	*WAPB* data	Company data
Years of ads	1981–2003	2000–2012
Number of raters	13	26
Outcome measure	Recall	Sales effectiveness
Outcome measure timing	Day-after recall	Short-term effects (28 days)
Outcome measure sources	Gallup and Robinson	Company data
Prediction accuracy	74.5% (CI = 66%, 83%)	55% (CI = 49%, 61%)

advertising, and Armstrong *et al.* (2016, p. 283) did assert that the PPI would be more useful for high-involvement products "because consumers think more carefully about the offer". Our results are not inconsistent with the authors' suggestion that the persuasion principles are more useful for such products.

The third potential reason for the weaker result is that the original validation study was somewhat biased because it analysed a subset of the *WAPB* advertisements that were also used to develop and weight some of the persuasion principles that contribute to the PPI scores. The reuse of some of the advertisements meant that the test lacked external generalisability, as it was overly calibrated on the original advertisements. This effect gave rise to a form of confirmation bias, which does not extend to our data set of advertisements.

The fourth potential reason comes from the different advertising outcome measures used across studies. Although recall measures are a valuable intermediate outcome, the usefulness of the PPI is limited if it is unable to make predictions about the impact of advertising on sales. The PPI may be highly effective at predicting recall across a variety of advertising conditions, but ultimately advertisers are most interested in whether this type of tool can be used to predict sales driven by advertising. The evidence from this study suggests that the PPI in its present form may not be up to the task.

Further research could help to identify which of these reasons best explains the different results obtained. However, for now, we suggest that the PPI in its present form should be used by advertisers only under the conditions set out in the initial validation test; that is, with print advertisements for high-involvement utilitarian products. There may still be merit to the PPI, at least as a foundation on which to build, with new evidence and principles to extend its relevance. On developing the PPI further, our results suggest that there may be a need to have separate indexes or sub-indexes for different types of advertising that draw on different principles (e.g. for high- vs low-involvement products) or certain principles may be applied to both conditions but require different weightings.

We must ourselves acknowledge the problems that can arise in the process of matching advertisements. Quasi-experimentally matching the advertisements allowed us to control for *some* variables, specifically within the matching criteria (namely, product, target market and media). However, there were many number of unobserved variables that were not controlled for, such as seasonality, random world events and changes in business functions (e.g. supply chain factors). The only way to accommodate such factors is with an increasing number of replications that apply the PPI to an increasing number of advertisements.

One last potential weakness worth noting is that administering the PPI necessitates using human judgement to assess whether advertisements apply the persuasion principles and how well they do this. This then raises the potential concern of maybe we "did it wrong" in the present research; maybe our raters did not rate advertisements in the same way that Armstrong's raters did, despite following his publicly available training materials. The counter to this concern is that the PPI is well documented, and we made substantial effort to follow the documented procedures. Raters were trained and inter-rater reliability was assessed in this research. Moreover, the efforts made in this study would match those of any advertisers who implement the PPI. It could be argued, therefore, that if the way the PPI was implemented in this study was not sufficient, then most users would struggle to implement it, bringing its usefulness into question.

Seeking to improve the predictive accuracy of the PPI leads to the challenge of identifying whether new principles are needed and/or when principles need to be retired. One must also consider the weightings applied to those principles, which have been shown to impact on predictions (Green *et al.*, 2016). Achieving improvements by changing the composition and weighting of principles is likely to be a long process that requires more research on the contribution of individual principles to predictions (seemingly one at a time) to build evidence for the relative impact of each principle. This process requires the contributions of many researchers and many sets of data.

One of the benefits of an index method is its flexibility. Consequently, it will be relatively easy to build on Armstrong's (2010) solid foundation. New research can be incorporated at any time, which may improve the PPI's validity, as well as allay concerns that the principles emphasise some areas of the advertising literature over others (Gendall, 2011). Moreover, there is nothing to stop advertisers from adding their own unique principles relevant to their particular needs. The index method, in its flexibility, presents no limit or threat to creativity. New ideas and attempts at persuasion can be readily included in an index, creating further opportunities for testing alternative versions of the index. One of the advantages of the broader adoption of the PPI and associated validation studies is that it builds a library of data sets against which versions of the PPI can be evaluated.

An important contribution of our study was resolving the question about the procedure for dealing with unreliable ratings (O'Keefe, 2016). Reliability is a necessary but insufficient condition to establish the validity of the PPI. Armstrong's suggested procedure is to remove any *ratings* that diverge from the collective average reliability score by more than 10 percentage points. Our results show that following the suggested procedure did little to improve the overall predictive accuracy of the PPI, at least for our sample of advertisements. Traditional inter-rater reliability tests, particularly those in psychometrics, would often attribute low reliability not only to the specific rater's judgement, but to the possible subjectivity or ambiguity of an item (in this case, a persuasion principle*)*. Establishing a procedure for removing both unreliable raters and potentially unreliable principles needs to be considered.

Although these results are disappointing, the index approach remains supported in the forecasting literature. By sharing this research, we hope to inspire further attempts to develop and apply the index approach in advertising and marketing, to produce superior predictions and help build a body of evidence to drive improved marketing decision-making.

Notes

1. That is, the index predicted the better recalled advertisement for 71.5 of 96 pairs. In the case of ties (where index scores were identical for advertisements within a pair), this was considered "half right".
2. There was no significant difference in the prediction accuracy for pairs comparing *above-average with below-average* sales performance, *above-average with average* sales performance or *average with below-average* sales performance [χ^2 (2) = 3.559, p = 0.169].

References

Armstrong, J.S. (1991), "Prediction of consumer behavior by experts and novices", *Journal of Consumer Research*, Vol. 18 No. 2, pp. 251-256.

Armstrong, J.S. (2010), *Persuasive Advertising: Evidence-Based Principles*, Palgrave Macmillan, Basingstoke.

Armstrong, J.S. (2011), "Evidence-based advertising", *International Journal of Advertising*, Vol. 20 No. 5, pp. 743-767.

Armstrong, J.S. and Green, K.C. (2018), "Forecasting methods and principles: evidence-based checklists", *Journal of Global Scholars of Marketing Science*, Vol. 28 No. 2, pp. 103-159.

Armstrong, J.S. and Patnaik, S. (2009), "Using quasi-experimental data to develop empirical generalizations for persuasive print advertising", *Journal of Advertising Research*, Vol. 49 No. 2, pp. 170-175.

Armstrong, J.S., Du, R., Green, K.C. and Graefe, A. (2016), "Predictive validity of evidence-based persuasion principles: an application of the index method", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 276-293.

Bellman, S., Nenycz-Thiel, M., Kennedy, R., Larguinat, L., McColl, B. and Varan, D. (2017), "What makes a television commercial sell? Using biometrics to identify successful ads", *Journal of Advertising Research*, Vol. 57 No. 1, pp. 1-14.

Blair, M.H. (1987), "An empirical investigation of advertising wearin and wearout", *Journal of Advertising Research*, Vol. 40 No. 6, pp. 45-50.

Gendall, P. (2011), "A review of: persuasive advertising: evidence-based principles", *Marketing Bulletin*, Vol. 22.

Graefe, A. and Armstrong, J.S. (2011), "Conditions under which index models are useful", *Journal of Business Research*, Vol. 64 No. 7, pp. 693-695.

Green, K.C., Armstrong, J.S., Du, R. and Graefe, A. (2016), "Persuasion principles index: ready for pretesting advertisements", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 317-326.

Haley, R.I. and Baldinger, A.L. (2000), "The arf copy research validity project", *Journal of Advertising Research*, Vol. 40 No. 6, pp. 114-135.

Hartnett, N., Kennedy, R., Sharp, B. and Greenacre, L. (2016a), "Creative that sells: how advertising execution affects sales", *Journal of Advertising*, Vol. 45 No. 1, pp. 102-112.

Hartnett, N., Kennedy, R., Sharp, B. and Greenacre, L. (2016b), "Marketers' intuitions about the sales effectiveness of advertisements", *Journal of Marketing Behavior*, Vol. 2 Nos 2/3, pp. 177-194.

Johnson, B. (2017), "World's largest advertisers: spending is growing (and surging in China)", New York, NY, available at: https://adage.com/article/cmo-strategy/world-s-largest-advertisers-2017/311484/ (accessed October 2018).

Jones, J.P. (1995), *When Ads Work: New Proof That Advertising Triggers Sales*, Lexington Books, New York, NY.

Lodish, L.M., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B. and Stevens, M.E. (1995), "How TV advertising works: a meta-analysis of 389 real world split cable TV advertising experiments", *Journal of Marketing Research*, Vol. 32 No. 2, pp. 125-139.

O'Keefe, D.J. (2016), "Evidence-based advertising using persuasion principles: predictive validity and proof of concept", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 294-300.

Sharp, B. and Hartnett, N. (2016), "Generalisability of advertising persuasion principles", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 301-305.

Taylor, J., Kennedy, R., McDonald, C., Larguinat, L., El Ouarzazi, Y. and Haddad, N. (2013), "Is the multi-platform whole more powerful than its separate parts?", *Journal of Advertising Research*, Vol. 53 No. 2, pp. 200-211.

Wood, L. (2009), "Short-term effects of advertising: some well-established empirical law-like patterns", *Journal of Advertising Research*, Vol. 49 No. 2, pp. 186-192.

Woodside, A.G. (2016), "Predicting advertising execution effectiveness: scale development and validation", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 306-311.

Wright, M. (2016), "Predicting what? The strengths and limitations of a test of persuasive advertising principles", *European Journal of Marketing*, Vol. 50 Nos 1/2, pp. 312-316.

Corresponding author
Nicole Hartnett can be contacted at: Nicole.Hartnett@marketingscience.info