European Journal of Marketing

Evidence-based advertising using persuasion principles: Predictive validity and proof of concept

Daniel O'Keefe

Article information:

To cite this document:
Daniel O'Keefe , (2016),"Evidence-based advertising using persuasion principles", European Journal of Marketing, Vol. 50 Iss 1/2 pp. 294 - 300
Permanent link to this document:
http://dx.doi.org/10.1108/EJM-11-2015-0801

Downloaded on: 01 March 2016, At: 14:01 (PT)
References: this document contains references to 7 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 62 times since 2016*

Received 24 November 2015
Revised 24 November 2015
Accepted 26 November 2015

Evidence-based advertising using persuasion principles

Predictive validity and proof of concept

Daniel J. O'Keefe
Department of Communication Studies,
Northwestern University, Evanston, Illinois, USA

Abstract

Purpose – The purpose of this paper is to provide commentary on Armstrong, Du, Green and Graefe's (this issue) article.

Design/methodology/approach – The paper is based on reading and thinking about Armstrong *et al.*'s article.

Findings – One appealing way that advertising practice can be evidence-based is by applying dependable generalizations about what makes for effective ads. Armstrong *et al.*'s article offers data concerning the application of Armstrong's persuasive advertising: Evidence-Based Principles (2010) persuasion principles. The article does not provide convincing evidence for the predictive validity of the principles, but it does offer a clear proof-of-concept demonstration of the feasibility of principles-based advertising assessment.

Originality/value – The paper's value lies in its clarification of what claims Armstrong *et al.*'s data do and do not underwrite.

Keywords Advertising, Persuasion, Generalization, Evidence based

Paper type Research paper

J. Scott Armstrong has long been a champion of the idea that the design of effective advertising need not be a haphazard enterprise in which message designers hope to randomly stumble into some message that works (Armstrong, 2010). Instead, message designers can be guided by evidence-based principles (generalizations) concerning message effects. His work is, thus, part of a larger developing understanding of the value of systematic research for communication design (Edgar and Volkman, 2012; Johnson *et al.*, 2010).

It is understandable that some could have doubts about the feasibility of such undertakings. Armstrong (2010) offered 195 advertising principles that are complex and abstract. Using them to guide or assess message design choices might seem so challenging as to be (practically speaking) impossible.

In the article under discussion, Armstrong *et al.* (2016) address such doubts. They report a project in which Persuasion Principles Index (PPI) scores are used to predict which of the two ads in a pair had higher recall scores. The better-recalled ad (in a pair) was correctly identified in about 75 per cent of the cases using the PPI-based approach, a rather higher rate than obtained using other (currently used) methods.

I believe this article is best (most charitably) understood as a feasibility demonstration, rather than as a conventional research report intended to provide evidence of the predictive validity of the principles. The report can be read as something

European Journal of Marketing
Vol. 50 No. 1/2, 2016
pp. 294-300
© Emerald Group Publishing Limited
0309-0566
DOI 10.1108/EJM-11-2015-0801

like a "proof of concept" paper, one meant to show that a given approach has the potential for viable application. Specifically, the article can be seen as arguing that it is indeed possible to devise a workable procedure that will successfully predict the relative effectiveness of two ads (here concretized as recall).

Understood as a proof-of-concept project, however, the present report must be seen to support a correspondingly circumscribed set of claims. If one wanted to provide convincing evidence of the predictive validity of the 195 principles, then a different kind of project would have been needed. But if the point is to display the feasibility of principles-based advertising assessment, then the current procedures appear more sensible.

To sharpen this contrast, this commentary discusses several concerns that readers might have about the reported methods. The recurring theme in what follows is this: What might seem to be methodological weaknesses – when assessed against the aim of providing evidence of the predictive validity of the principles – are actually not worrisome, if one takes the article's purposes to be those of a proof-of-concept undertaking.

Potential methodological worries

Readers might naturally have concerns about five methodological aspects of the reported research: the advertisements that were analyzed, the selection of the principles, the selection of the raters, the relationship between the selection of principles and the selection of raters, and the role of the weights.

The advertisements analyzed

The article reports an analysis of 192 ads, but all of the ads were ads for "high-involvement utilitarian products". The rationale for this restriction was that the researchers "expected the persuasion principles to be more useful for such products", that is, more useful for predicting relative ad effectiveness for these kinds of products than for other kinds of products. After all, if the enterprise cannot be shown to be workable in circumstances where its effectiveness should be most easy to display, then its feasibility would indeed be questionable.

This restriction might be thought worrisome. After all, one cannot tell (from the present data) whether the reported procedures would work for other kinds of ads. That is, the restriction to certain kinds of ads means that the report cannot provide general evidence about the predictive validity of the principles.

But that concern is relevant only if the purpose of the report was to provide evidence of the predictive validity of the principles for ads generally. If the purpose is to provide proof-of-concept for the idea that a useful procedure is possible, then this methodological concern is irrelevant. In some ways, it does not matter exactly what sort of ads were analyzed, if the point is simply to show that it *is* possible to have *some* sort of procedure that works.

The selection of the principles

It is not clear how many advertising principles were used for the PPI ratings in this study. Armstrong's (2010) book identified 195 persuasion principles, but in the reported study, not all 192 ads were rated on each of the 195 principles. The set of principles was winnowed in two ways, one relevance-based and one agreement-based.

First, in some fashion, decisions were made about whether a given principle was relevant to an ad (with irrelevant principles being put aside): "Raters used descriptions of the principles to decide whether or not a principle was relevant to the ad being evaluated" (listed as Step 1 in Armstrong *et al.* 's Appendix). So when raters came to the task of rating how well a principle was applied in a given ad, such ratings were obtained only for "each principle that was assessed as relevant" (Step 3 in the Appendix). It is not entirely clear whether each rater was free to decide whether a given principle was relevant to a given ad, as opposed to some consensus being formed about the relevance of each principle to each ad. But in any case, some principles were discarded as not relevant.

Second, in some fashion, decisions were made about whether judgments about the application of a given principle exhibited sufficient agreement (among raters) to permit inclusion of that principle; principles not eliciting sufficient agreement were put aside. Armstrong *et al.* 's Appendix indicates that:

> […] 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.

But it is not clear whether the five raters had to produce at least three identical ratings for that principle for *every* ad for which the principle was deemed relevant. And it is not clear whether a dropped principle did not figure at all in the computation of consensus PPI scores for *any* of the ads, as opposed to (e.g.) not figuring in the computation of the consensus PPI score for those particular ads on which consensus was not achieved. But in any case, some principles were discarded as ones for which consensus could not be reached.

Given this culling of the principles, these data cannot be used to underwrite the predictive value of the 195 principles as a whole. After all, at least some of those 195 principles might not have figured in the computation of consensus PPI scores for any ad (because a principle might not have survived both the relevance-based winnowing and the agreement-based winnowing). And of those principles that did contribute to consensus PPI scores for at least one ad, some principles may have figured only very rarely across the set of ads, while others contributed much more often.

If the point of the report was to provide evidence for the predictive validity of the principles, then these procedures would naturally raise concerns. If that were the purpose, then it would be crucial to know exactly how the criteria for relevance-based and agreement-based winnowing worked, specifically which principles were tested and how extensively a given principle was tested – because without that information one could not see just what support the data offered concerning the various principles' predictive utility.

But if the report's purpose is something different, to provide proof-of-concept for the idea that a useful procedure is indeed possible, then these procedural concerns are irrelevant. In some ways, it does not matter how many principles were involved or how frequently any of them was used or exactly how they were winnowed – if the point is simply to show that it *is* possible to have *some* sort of procedure that works. The report need not be taken to claim that this particular procedure (for selecting or dropping principles) is inevitably the best or that others should follow this procedure. The claim can be just that a workable procedure is indeed possible.

The selection of the raters

The composition of the raters (who yielded the data for the consensus PPI scores for each ad) is not entirely clear from the report. In all, 17 raters were involved (13 university students and 4 raters from Mechanical Turk), but not all 17 raters rated all 96 ad pairs, which meant that different ads were rated by different sets of people. The reported consensus PPI ratings for a given ad were based on a set of five raters. It appears that five raters were initially chosen for a given ad, but then "raters who departed substantially from the consensus" were dropped and replaced by other raters. "Specifically, raters whose scores were more than 10 percentage points different from the average rater were dropped and replaced by new raters." This procedure for discarding raters differs from that of conventional inter-rater reliability assessment procedures [e.g. correlation coefficients or Krippendorff's alpha; Hayes and Krippendorff (2007)]. Consider, for example: A rater whose ratings were *consistently* 15 percentage points higher than that of the average rater would, by conventional standards, likely be deemed a perfectly good rater. But such a rater would apparently have been discarded by the present procedures.

If the point of the report was to provide evidence for the predictive validity of the principles, then these procedures would naturally raise concerns. For example, one would want to have a rationale for the choice of method for dropping and replacing raters (a rationale that justified using that procedure rather than well-established conventional methods).

But if the report's purpose is something different, to provide proof-of-concept for the idea that a useful procedure is indeed possible, then this procedural concern is irrelevant. In some ways, it does not matter exactly how raters were initially chosen or subsequently identified as unsuitable or precisely how replacement raters were chosen – if the point is simply to show that it *is* possible to have *some* sort of procedure that works. The report need not be taken to claim that this particular procedure (for selecting, dropping, and replacing raters) is inevitably the best or that others should follow this procedure. The claim can be just that a workable procedure is indeed possible.

The interplay of principle selection and rater selection

As the report indicates, there was both agreement-based winnowing of principles (where a principle was dropped if there was insufficient agreement among raters) and reliability-based replacement of raters (where raters were dropped and replaced if they exhibited unreliability with other raters). But it is not clear exactly how these fit together.

For instance, suppose the insufficient-agreement principles were dropped first, and then, inter-rater reliabilities were assessed on the remaining principles. If that was the procedure, then a reader might naturally wonder why there would be much rater unreliability, given that only principles with sufficient agreement were being analyzed. Alternatively, suppose unreliable raters were identified and replaced, and then, principles were dropped if ratings of that principle's application did not display sufficient agreement. If that was the procedure, then a reader might naturally wonder why there would be much disagreement if only reliable raters remained.

If the point of the report was to provide evidence for the predictive validity of the principles, then these unclarities would naturally raise concerns. One would want to have a much more straightforward accounting of exactly how the eventual composition of principles and raters was obtained.

But if the report's purpose is something different, to provide proof-of-concept for the idea that a useful procedure is indeed possible, then this concern is irrelevant. In some ways, it does not matter exactly how the selection of raters and the selection of principles were intertwined – if the point is simply to show that it *is* possible to have *some* sort of procedure that works. The report need not be taken to claim that its particular procedures are inevitably the best or that others should follow these procedures. The claim can be just that a workable procedure is indeed possible.

The role of the weights

In computing PPI scores, not all principles were given equal weight. Each principle was given an a priori weight, such that greater weight was given to strategic (as opposed to tactical) principles and to well-evidenced (as opposed to poorly evidenced) principles. But the contribution of these weights to predictive success is not made clear by the report. The description of the results indicates the degree of predictive success using the thusly weighted PPI scores, but not the predictive success using unweighted PPI scores.

If one purpose of the report were to show that using weighted PPI consensus scores was superior (in predicting relative recall) to using unweighted PPI consensus scores, then more detail about the study's results would be needed. In particular, one would want to see information reported about the accuracy of forecasts from unweighted PPI scores. One might also want to see information about the relative value of weightings based on the strategic-versus-tactical nature of the principle and weightings based on the strength of the evidence supporting the principle, for greater transparency about how the weighting procedure worked.

But if the report's purpose is something different, to provide proof-of-concept for the idea that a useful procedure is indeed possible, then such lack of detail would not be troubling. In some ways, it does not matter exactly what the weighting procedure was or whether use of the weights was essential to the observed improvement in predictive accuracy – if the point is simply to show that it *is* possible to have *some* sort of procedure that works. The report need not be taken to claim that its particular weighting procedure is inevitably the best or that others should follow this procedure or that using weights in this way is generally preferable as a predictive method. The claim can be just that a workable procedure is indeed possible.

Summary

It must be acknowledged that sometimes the manuscript seems to drift away from a proof-of-concept purpose. In places, the manuscript invites conclusions of a sort that might be underwritten by a more conventional research project, but which are not well justified by the project reported here. For example, the manuscript claims that "this study provides a test of the predictive validity of persuasion principles." But without telling readers how many principles were studied or which principles were studied or how extensive the evidence is for each studied principle, that claim is unlikely to be found compelling.

To express this point differently: Interpreted as a report of a conventional research project meant to provide evidence about the predictive validity of the principles, the present report does not make as convincing a case as one would like to see. Too much is left unclear about the methods and the results.

But understood as a proof-of-concept demonstration project, the present report is thoroughly successful (and hence that is the charitable reading). The article convincingly shows that it is indeed possible to train raters to use advertising principles to rate ads and that such ratings can yield improved prediction of relative ad recall. Any doubts one might have had about the feasibility of principles-based ad assessment procedures should be, if not quite laid to rest, at least considerably dampened by this report.

Evidence-based advertising practice without principles

To place Armstrong *et al.*'s (2016) work in a broader context, one might consider whether principles are really needed to reap the benefits of evidence-based advertising design. Evidence-based practice is certainly desirable, but advertising practice might be based on evidence in at least two rather different ways.

One is the approach under discussion in Armstrong *et al.*'s article: identify evidence-based principles (generalizations) and use those principles when designing advertisements. So if advertisements of Type A are on average more effective than advertisements of Type B (either in general or under specifiable conditions, for example, for specific types of products or recipients), then advertisers should design advertisements accordingly. This is evidence-based practice akin to its familiar form in biomedical contexts: if Drug A is on average more effective than Drug B (either in general or under specifiable conditions, for example, for particular types of patients), then medication should be prescribed accordingly.

This first kind of evidence-based advertising practice faces at least two substantial challenges. One is the identification of dependable principles [for some discussion of this issue, see O'Keefe (2015)]. A second is the challenge of using those principles to create concrete advertisements. Given some particular advertising problem (a given product, target market, medium and so on), it is not always easy to see how to translate a given generalization into specific ads [this challenge is emphasized by Jackson and Aakhus (2014)]. Neither one should be underestimated.

But in at least some advertising contexts, a second kind of evidence-based approach is possible. This is the sort familiarly known as A/B testing, in which, *in situ*, alternative ads can be compared for effectiveness. This can be especially attractive in certain online applications. For example, when advertisers see that Ad A is getting more clicks (or sales or whatever) than Ad B, Ad B can be dropped and Ad A used exclusively.

Notice that this second, "brute-force" message design procedure is distinctly evidence-based. It permits constant experimentation, allows for the more effective ad to be modified on a continuing basis and, thus, lets empirical results guide advertising design. And it sidesteps the two challenges of principles-based design: it does not need principles, and it does not face the problem of translating abstract principles into concrete messages.

Now sometimes, it will not be feasible for an advertiser to adopt this brute-force approach, and in such cases, one could justifiably recommend the use of principles-based design. And even in brute-force design, principles can be helpful both in designing the initial ads to be tested and in suggesting ongoing modifications. But these brute-force methods are very much evidence-based – even though the evidence invoked is not evidence that underwrites some general principle but rather evidence about the relative effectiveness of the specific ads under consideration.

To be sure, for one who is interested in how and why advertisements have their effects or for one interested in deriving broader message design principles, the sort of data collected in brute-force applications will naturally be less useful than the sort of data collected through systematic theoretically motivated research. But for one who is merely interested in which of two versions of an ad will be more effective – never mind why – brute-force data will be entirely sufficient.

Conclusion

Advertising practice might be evidence-based in various ways. The approach on offer here, principles-based advertising design, has undeniable appeal, but might be seen to face insurmountable practical difficulties. However, the present report demonstrates the feasibility of a principles-based approach for assessing ads.

References

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

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. 276-292.

Edgar, T. and Volkman, J.E. (2012), "Using communication theory for health promotion: practical guidance on message design and strategy", *Health Promotion Practice*, Vol. 13 No. 5, pp. 587-590.

Hayes, A.F. and Krippendorff, K. (2007), "Answering the call for a standard reliability measure for coding data", *Communication Methods and Measures*, Vol. 1 No. 1, pp. 77-89.

Jackson, S. and Aakhus, M. (2014), "Becoming more reflective about the role of design in communication", *Journal of Applied Communication Research*, Vol. 42 No. 2, pp. 125-134.

Johnson, B.T., Scott-Sheldon, L.A.J. and Carey, M.P. (2010), "Meta-synthesis of health behavior change meta-analyses", *American Journal of Public Health*, Vol. 100 No. 11, pp. 2193-2198.

O'Keefe, D.J. (2015), "Message generalizations that support evidence-based persuasive message design: specifying the evidentiary requirements", *Health Communication*, Vol. 30 No. 2, pp. 106-113.

About the author

Daniel J. O'Keefe is the Owen L. Coon Professor in the Department of Communication Studies at Northwestern University. His research focuses on persuasion and argumentation, with a special interest in research synthesis (e.g. meta-analysis). He is the author of *Persuasion: Theory and Research* (third ed., Sage Publications). Daniel O'Keefe can be contacted at: d-okeefe@northwestern.edu

This article has been cited by:

1. Kesten C. Green, J. Scott Armstrong, Rui Du, Andreas Graefe. 2016. Persuasion Principles Index: ready for pretesting advertisements. *European Journal of Marketing* **50**:1/2, 317-326. [Abstract] [Full Text] [PDF]