

Assessing the value of neurophysiological measurement for advertising pre-testing

Are Biometrics Better?

by

Haydn Joseph Northover

BSc (Psyc), BCom (Mktg), MCom (Mktg)

A dissertation submitted for the degree of Doctor of Philosophy

Ehrenberg-Bass Institute for Marketing Science
The University of South Australia

Supervisors:

Assoc. Prof Rachel Kennedy

Prof. Byron Sharp

Dr. Jane Leighton

June 2012

Assessing the value of neurophysiological measurement for advertising pre-testing

Are Biometrics Better?

Abstract

It is hoped that pretesting can achieve two things. First, that it can identify if an advertising execution is likely to be sales effective; this is an Evaluation task. Ideally it would also predict *how* sales effective the execution is (i.e. by how much), but this would be a difficult objective. The second task that pre-testing should achieve is to identify how the execution is working; this is a diagnostic task. This thesis set out to determine whether current biometric measures offer a better alternative to traditional, self-report measures for both tasks: Evaluative and Diagnostic.

An on-going pre-testing debate exists that has not been resolved despite years of study and numerous journal articles devoted to the task. Put simply, the pre-testing debate centers on whether pre-testing can predict advertising sales effectiveness. Pre-testing has been, and is still, dominated by self-report measures. These include recall, persuasion, likeability for instance; any measure that requires a respondent to recall from memory, or answer using a scale. The seminal ARF Copy Research Validation Project provided an answer on which metrics work better, yet did not end the debate.

In the background, our understanding of the brain has advanced along with developments in the techniques that measure it. While self-report measures were criticised for simply being proxies of the under-lying - and subtle – brain activity, there now exist measures that could more *directly* measure the brain activity. Notably, advances in technology have improved data collection for Electroencephalography (EEG), skin conductance and fMRI, as well as their subsequent interpretation. Despite this advancement, little controlled research in the marketing field has been conducted. This raises the question of how such tools perform compared to traditional, self-report advertising pre-tests.

The conjunction of these two developments – the continued pre-testing debate and advancement in brain science – forms the basis of this thesis. Specifically it tackles the questions: Are biometric / neurological measures the panacea the industry has been looking for? Do they tap the unconscious level of consideration inaccessible to self-report measures? If the answers to these questions are ‘Yes’, is there evidence they are also better predictors of sales?

The advancement of biometrics is not the only development in pre-testing. Alternative metrics are being introduced, particularly those relating to implicit memory; implicit memory is the dominant form of memory for common or routine decisions. It is therefore valuable to determine whether these additional measures are also beneficial to the evaluation of advertising.

A unique focus of this thesis is to use, as the dependent variable, sales data utilising a single-source approach. Such data is collected longitudinally, recording media usage and sales data, and enables quasi-experimental comparison of purchases made after exposure to the advertising versus purchases that aren’t preceded by exposures. As such the in-market sales effectiveness of a given piece of ad copy can be determined.

Five separate studies are reported:

1. A case study reporting an assessment of the same advertising executions by international pre-testing agencies using both traditional and biometric measures;
2. Measures of Likeability and EEG were assessed for Super Bowl ads over a three-year period. A portion of the Super Bowl ads were for up-coming movie releases allowing for an assessment between ad response and in-market behavioural response (measured as first week box-office returns);
3. Primary research taking a sample of confectionary executions (n=10), each with in-market sales performance data and tested using Skin Conductance and the self-report measures of Likeability and the Self-Assessment Manikin (SAM);
4. A pilot study assessing seven executions with known in-market sales performance using fMRI, implicit and explicit measures;
5. Three ad pairs with known (single-source) sales effectiveness were assessed using an EEG technique, specifically Steady State Topography, and separately

against a range of measures drawn from the literature and pilot study results. Specifically a Recognition Reaction Time measure that demonstrated a good correlation to sales effectiveness during the pilot study was the main evaluative measure used, along with traditional measures of recall, likeability and persuasion. Additionally a moment-by-moment diagnostic measure was also captured.

In total 182 ads were incorporated in this thesis, including 3 ad pairs with known in-market sales effectiveness for the same brand. Three biometric / neurophysiological techniques were employed (fMRI, EEG and Skin Conductance) including a timed brand recognition task. In addition to these, self-reported evaluative and diagnostic measures common to the ad research industry were included at all stages.

The findings demonstrate that on each occasion that a self-reported measure - akin to those used by large, multinational research organizations - was employed, its ability to predict effective advertising was minimal. For instance, no single measure, or combination of, correctly predicted 3 out of 3 ad pairs, while correlations to sales were always in the low range ($r = <0.2$). Conversely, where a biometric / neurophysiological measure was employed, its relationship to sales effective advertising was more predictive. For instance, skin conductance correlations to sales were above $r=0.5$. This relationship was not always consistent however, with further investigation warranted.

In summary:

STUDY 1: Examining the same ads when assessed by competing methodologies / suppliers (Ipsos, Emsense and Neuro-Insight) demonstrates that – for the small number of cases assessed – the same ad is not necessarily identified as superior, and nor do their moment-by-moment traces of ads necessarily correspond in any meaningful pattern. It is apparent that biometric and self-report measures of advertising do not highlight the same elements of that advertising that lead to its effectiveness.

STUDY 2: Likeability of Super Bowl ads – as measured by the *USA Today* ‘Ad Meter’ - demonstrated a small correlation to EEG Engagement for the same ads, as measured by SANDS Research; testing the same ads, biometrics and self-report measures in this context, aren’t measuring the same “thing”. The *USA Today* ‘Ad meter’ is the quasi

industry standard for measuring the “best” Super Bowl ads, yet this alternative EEG measure offered a different ranking of the “best” ads. In, and of itself, this is seemingly of trivial importance; yet achieving a low *USA Today* Likeability score has repercussions for the advertising agencies that produce them (some accounts have been terminated following a low likeability ranking). Hence it is important to link the scores for each ad, by each measure, to some form of dependent variable. This will provide an objective measure of effectiveness for each tool.

Using the first-week box-office returns, per theater, for each movie as a dependent variable, neither measure achieved a high positive correlation with box-office success (in fact likeability tends toward a negative relationship, $r = -.34$). For movies therefore, there appears to be little relationship between how much a trailer is ‘liked’, or how engaging it is (EEG) and subsequent box-office returns. *Post-hoc* exclusions of family films and franchises suggests there may be some relationship between EEG and box-office success ($r = .38$), which itself warrants further research.

STUDY 3: Average Arousal for individual ad executions, as measured by skin conductance, demonstrated a strong correlation to in-market sales performance of the same executions. The performance of skin conductance ($r = 0.52$) was higher than the self-report measures ($r = 0.04 - 0.1$).

The skin conductance result was only apparent for Australian respondents, yet not International respondents. The fact there exists such a discrepancy between the Australian and International respondents for the biometric measure yet not the self-report measure reinforces the view of many that the self-report technique is inadequate to tap the real measure of interest.

STUDY 4: There is wide variation in the predictive ability of pre-testing methods, with again the self-report measures failing to present a clear relationship to in-market effectiveness. This study highlights the potential efficacy of fMRI and reaction time.

STUDY 5: Peak memory encoding during branded moments – for the three ad pairs tested – did not accurately predict which was the winner amongst the 3 ad pairs. Memory encoding for the final branded moment also failed to predict the winner on 3 out of 3 occasions, only achieving directional and not significant results.

A faster reaction time to the brand indicated the ad winner significantly for 1 of 3 ad pairs, and overall, yet not significantly for the remaining two pairs. The Arousal (as measured by the Self-Assessment Manikin), significantly predicted the ad winner in 2 out of 3 occasions. A range of self-report evaluative measures was tested (recall, likeability, persuasion), with none demonstrating a predictive relationship, individually or when combined.

Comparison of the diagnostic measures, particularly neurological and self-report moment-by-moment traces, demonstrates that each measure is capturing something different i.e. the traces are dissimilar. Of note, the self-report trace displays a homogenous pattern amongst each of the ads tested (initially flat, lifting toward final branding) whereas the neurological moment-by-moment trace tends to react to individual elements within each ad, and not in a homogeneous pattern as described for the self-report trace.

Conclusion

Single studies are unable to build robust marketing knowledge. Only when studies are subsequently replicated can we start to be confident we are observing a real effect.

Findings that warrant further investigation include:

1. A **speeded recognition** task achieved a high correlation to sales (Study 4) and demonstrated promise identifying the ad pairs (Study 5);
2. The **Arousal** measure taken from the **Self-Assessment Manikin** achieved a high correlation to sales (Study 3) and was significantly predictive of the more sales effective ad from 2 of 3 ad pairs (Study 5);
3. Biometric **arousal** as measured by **skin conductance** achieved a high correlation to sales (Study 3) while arousal as measured by EEG and defined as **Emotional Intensity** was in line with the ad winner/loser pair (Study 5); and
4. **Traditional self-report measures** were consistently poorer at predicting subsequent behaviour compared to biometric measures (Study 2), achieved a

lower correlation to sales than biometrics (Study 3) and could not accurately predict the direction of the ad winner/loser pairs (Study 5).

As well as the answers to the questions that were posed, there were other lessons for those involved in advertising pre-testing. Memory encoding was only studied once (measured via EEG); the original expectation that peak memory encoding during any branded moment would predict the ad winner wasn't observed. However this area holds promise as memory encoding during the final branding moment was inline with the ad pairs, hence this technique warrants further consideration as an evaluative measure.

This thesis also suggests that no current pre-testing tool is reliable. In some circumstances I found pre-testing to be predictive of sales, yet this was not universal. There was no evidence that the commonly used self-report measures, when assessed against actual sales, were predictive. A number of factors will contribute to this lack of prediction – notable amongst them the unrepresentative “lab environment” of data collection. It is suggested future assessment of pre-testing validation needs to focus on sales effectiveness, ideally using single source, in-market sales performance data. This current thesis used such data and found traditional pre-testing measures wanting, with biometrics particularly the area of memory encoding and arousal demonstrating promising performance. However the evidence is not complete on this matter, hence caution is urged, while further research is clearly encouraged.