Viewpoint

How Reliable Are Neuromarketers' Measures of Advertising Effectiveness?

Data from Ongoing Research Holds No Common Truth among Vendors

DUANE VARAN
Murdoch University/
Audience Labs
D.Varan@murdoch.edu.au

ANNIE LANG
Indiana University/
The Media School
anlang@indiana.edu

PATRICK BARWISE
London Business School
pbarwise@london.edu

RENÉ WEBER
University of California,
Santa Barbara
renew@comm.ucsb.edu

STEVEN BELLMAN
Murdoch University/
Audience Labs
bellman@audiencelabs.
com

Buyers in search of new neuromarketing methods that potentially can predict advertising effectiveness face a daunting process. Vendors in this evolving industry offer a confusing range of often proprietary differences in methodology. The authors of the current article analyzed results from "Neuro 1"—the Advertising Research Foundation's first neuro-standards trial—and revealed that there is no common truth, no single scientific reality exposed as a result of these new methods. Addressing what they believe is a need for greater transparency—even after "Neuro 2"—which used publicly available methods, the authors demonstrated how a buyer can compare the validity of different vendors' measures.

INTRODUCTION

In two neuro-standards initiatives, the Advertising Research Foundation (ARF) collaborated with vendors and advertisers to explore the potential of using biometric and neuroscience-based methods for advertising research.

- In the ARF's first NeuroStandards Collaboration Project (Neuro 1), eight commercial vendors of biometric and/or neuroscience-based methods tested eight commercials from eight different sponsors. Neuro 1 tested measures available "in the market"—basically vendors' reports were reviewed by a panel of experts. These experts were independent of the ARF.

Those experts included three authors of this current paper (Lang, Barwise, and Weber). On the basis of the findings of this panel, the ARF's Neuro 1 report (Stipp and Woodard, 2011) concluded that the different vendors'

Management Slant

- Advertisers want measures of processing that go beyond mere exposure.
- Neuro vendors have cultivated an expectation that their measures are more reliable than traditional measures because they measure neurological and biological processes.
- The results of the current study question these strong claims and suggest methods that advertisers can use to choose their vendor carefully.
- Greater transparency about the constructs measured and methodologies used will advance the field of consumer neuroscience.

DOI: 10.2501/JAR-55-2-176-191

modified—and often proprietary—versions of these new neuro[1] methods reflected a lack of transparency about what was being measured and how.

- Neuro 2 was designed to test the utility of "the best" of these new methods for predicting actual sales tracked by scanner data. Neuro 2 used researchers from four universities to gather a single standard academic version of each neuro measure and implicit and explicit traditional measures from 300 participants (Vito, 2014). This design eliminated the transparency constraints associated with using actual vendors in Neuro 1.

When advertisers plan for the purchase of research from commercial vendors, most of them do not have the in-house capability to use the predictive measures identified by Neuro 2 (*e.g.*, fMRI). And, most vendors would argue, the results of Neuro 2 often do not apply to them, as they use different "proprietary" measures, not the "standard" ones tested in Neuro 2.

So even with the insight of the findings from Neuro 2, the data gathered from vendors in Neuro 1 remain relevant to advertisers, along with the following recommendations of the ARF's Neuro 1 report (Stipp and Woodard, 2011):

- Just because neuro measures are "new," there is no reason to ignore traditional research issues (*i.e.*, sample size, sample location, and sample composition) to ensure the reliability, interpretability, and statistical significance of these new measures.

- Potential buyers should establish
  - ✧ who actually is responsible for conducting the measurement;
  - ✧ the credentials of the people responsible for the measurement—specifically, their training;
  - ✧ the quality and reliability of the equipment being used; and
  - ✧ who interprets the data and whether those interpretations are based on evidence or hypothesis.

- Solely relying on a single vendor's interpretation for brand insights is risky because
  - ✧ the data are so unfamiliar and, therefore, merit more interpretation than traditional research because they are so unfamiliar;
  - ✧ a vendor's interpretation is just one among many that may prove useful to the buyer; and
  - ✧ a vendor often lacks a buyer's experience with the specific brand.
  - ✧ Some vendors have limited experience of marketing in general.

The Neuro 1 report did not disclose any advertisements' scores or even pictures of the vendors' results, respecting that "the advertisers [had] asked that specific findings about the commercials not be shared with other sponsors" (Stipp and Woodard, 2011, p. 20). This meant that a key recommendation of the report—that buyers ask vendors about the validity and reliability of their concepts and measures—naturally was daunting.

Many buyers would find it hard to question the quality, experience, and vendors of service providers because the concepts and measures are new and difficult. In fact, after the release of Neuro 1, many potential buyers may have concluded that neuro measures were not yet ready for use, "despite the fact that this project was clearly designed to provide information that would help apply neuroscience and biological methods to marketing communication issues more effectively" (Stipp and Woodard, 2011, p. 10).

The current study complements the Neuro 1 and Neuro 2 reports with examples of the kinds of data available from commercial neuro vendors that preserve the anonymity of vendors and sponsors.

The authors used these disguised data to illustrate how a buyer may be able make sense of vendors' offerings, before and after purchase. When choosing the most suitable vendor(s), a buyer needs to understand not only the predictive potential of specific measures (*e.g.*, as revealed by Neuro 2) but also the reliability and validity of the methods commercial vendors use to gather these measures and interpret their results.

The need to exercise due diligence and test a variety of neuro measures may seem to be at odds with the "hope and hype" surrounding them (Ariely and Berns, 2010). Some neuro vendors have marketed their research as offering "science-based insights into the unconscious in a way that suggests the absence of any uncertainty or element of interpretation" (Stipp and Woodard, 2011).

If the results really were error-free, however, there should be no differences between vendors' results if they measured the same variables with the same basic equipment. In fact, however, as Neuro 1 revealed, there are many differences in how different vendors produce their measures, just as there are differences between vendors in how they gather such traditional measures as recall and persuasion.

In the current study, the authors investigated how a buyer can compare the validity and reliability of different vendors' approaches despite the current lack of transparency around vendors' proprietary methods. The results led to new recommendations for both buyers and vendors of these measures and for future research.

[1] *In the remainder of this paper, the term* neuro *is used to refer to the full range of psychophysiological methods: eye tracking, biometrics such as heart rate and skin conductance, and measures of brain activity such as electroencephalograms (EEGs) and functional magnetic resonance imaging (fMRI).*

LITERATURE REVIEW

Reviews of neuro measures of advertising effectiveness have called for more validation work (Poels and Dewitte, 2006). The ARF's neuro-standards trials were, in part, an answer to this call. According to neuromarketers, "traditional market research methods—like consumer surveys and focus groups—are inherently inaccurate because the participants can never articulate the unconscious impressions that whet their appetites for certain products" (Singer, 2010, p. BU.4).

Neuro measures potentially offset many of the weaknesses associated with traditional measures, such as reliance on language and memory, temporal imprecision, and interruption of the experimental process (Ravaja, 2004). Even when subtle responses are available to conscious awareness, only the peaks or end points may be recalled (Fenwick and Rice, 1991; Fredrickson and Kahneman, 1993). A self-reported continuous response measure (CRM) such as dial turning (Biocca, David, and West, 1994; Boyd and Hughes, 1992)—available since the 1930s—counteracts these biases but at the cost of interfering with message processing (Potter and Bolls, 2012; Ravaja, 2004). A self-report CRM also suffers from brain–hand reaction-time lag (Young, 2002).

Although neuro measures for testing television commercials still are relatively new, they already have demonstrated significant long-term potential and already are providing valuable insights. One study has shown that brain scans produced by functional magnetic resonance imaging (fMRI) did a better job at predicting in-market performance than traditional survey measures (Falk, Berkman, and Lieberman, 2012). The promise of these measures is that they enhance traditional measures in two ways:

Neuro measures potentially offset many of the weaknesses associated with traditional measures.

- They provide complementary indicators of emotional and subconscious responses to advertisements, and

- many provide fine-grained continuous data during the viewing of a commercial. Unlike dial turning,
  - ✧ they involve no conscious effort by the subject;
  - ✧ they can sample data hundreds of times a second (Appel, Weinstein, and Weinstein, 1979); and
  - ✧ they can measure multiple variables simultaneously.

  For example, eye tracking of visual attention, in combination with other neuro measures, can identify the stimuli a viewer is responding to. This is important especially for advertisers looking to identify the key elements or key scenes in a commercial.

Despite these advantages, neuro measures are hard to interpret, especially with complex real-world stimuli like television commercials (Ravaja, 2004). Therefore, normal practice—and the practice recommended by the authors—is to use a combination of several complementary neuro and traditional methods.

Pros and Cons of Different Measures

There are three different basic types of neuro measures (Noble, 2013):

- Neurometrics are methods that directly measure brain activity. Electroencephalograms (EEGs) have been used since 1929 (Potter and Bolls, 2012); fMRI, detecting oxygenated blood in the brain, has been in use since 1992 (Weber, Mangus, and Huskey, 2015).

- Biometrics measure activity in the rest of the body, triggered when the brain notices novel, relevant, or motivating stimuli, such as the key moments in a commercial. Examples are sweating (skin conductance), heart rate, respiration, posture, and facial expression.

- Psychometrics also provide indirect measures of brain activity, using the "what-wires-together-fires-together" principle. These typically measure reaction times that reflect unconscious implicit associations and attitudes.

A major consideration for research buyers is the relative cost of these approaches: fMRI is extremely expensive per subject, so researchers tend to use small samples that may dangerously reduce statistical power (Button *et al.*, 2013). EEG is less expensive, has higher temporal resolution, and is portable (See below). Compared to EEG, both biometrics and psychometrics cost so little they can be deployed with larger scale. For example, webcam facial decoding and implicit memory word-completion tests can be included in online surveys.

The level of training required to collect the measures is an important consideration for buyers of neuroscience-based research. Specifically,

- psychometric tests require expertise to design and interpret (Krishnan and Chakravarti, 1999);
- collecting accurate biometric data also requires training (Potter and Bolls, 2012);
- signals originate from very precise locations that only highly trained operators can identify;

- even higher levels of training and expertise are required to collect reliable and valid neurometric data; and
- EEG signals are very weak and easily may be lost unless electrically quiet labs are used and the participant's skin surface is carefully prepared (Potter and Bolls, 2012).

Another dimension on which neuro measures differ is their temporal resolution—more specifically, how quickly they respond to an event in a commercial.

- EEG has high temporal resolution, detecting response times measured in hundredths of a second (Treleaven-Hassard *et al.*, 2010), but low spatial resolution. It is unclear where in the brain the signals come from (Polich and Criado, 2006).

- Conversely, fMRI has high spatial resolution but low temporal resolution—a delay of one to two seconds before blood flow begins and six seconds before it peaks (Weber *et al.*, 2015), which means that for some time it was difficult to ascribe fMRI activity to sequences of less than 15 seconds in duration. New technology does enable researchers to use sophisticated analytical techniques that allow a temporal resolution of a few seconds.

- Biometric measures also have time lags; time-series analysis shows that heart rate has a lag of up to five seconds (Wang, Lang, and Busemeyer, 2011).

All these lags need to be controlled for on an individual basis (*e.g.*, using time-series analysis) before responses can be time-locked to the frames of a commercial. Once the data are time-locked for speed of response, further time-series analyses may be able to detect "cognitive lags." These could range from rapid responses to aversive stimuli (like a snake) to very long and gradual responses (getting a subtle joke; Wang *et al.*, 2011).

Other differences between neuro measures include the different and highly artificial conditions under which they are collected—a concern especially acute for fMRI, which requires participants to watch commercials while lying down inside a noisy and claustrophobic giant magnet. Such conditions make participants less relaxed and raise questions about the ecological validity of their responses. In contrast, EEG and biometrics can be collected using portable, less obtrusive units allowing measurement in more natural conditions, although these measures may be compromised by the risk of environmental and movement noise (Potter and Bolls, 2012).

The most important consideration for a prospective buyer, however, is which commercially available neuro measures reliably and validly answer specific research questions. The authors of the current study believe the results of Neuro 1 can inform this decision.

NEURO 1 SUMMARY AND DATA ANALYSIS

Neuro 1 Objectives

The aim of Neuro 1 was to provide evidence-based guidance for advertisers to help them decide which kinds of neuromarketing research—as provided by a number of leading vendors—best addressed specific research questions. Eight vendors participated. Each was given the same eight 30-second commercials to assess (one from each of eight sponsors).

To reproduce what buyers could obtain from the market, each vendor used its standard combination of methods—subject to budget restrictions—and a general stipulation to sample U.S. consumers ages 18 to 49 years. Most conducted their research in November, 2010. Their reports were assessed by an international team of 12 subject-matter experts and six members of a senior review panel, which, as disclosed earlier in this study, included three of the authors of the current article. The ARF's report on the trial, including comments from the review panel and the vendors, was released in October, 2011 (Stipp and Woodard, 2011).

Because the eight vendors were asked to use their normal methods—including sample size and location—Neuro 1 was not designed to achieve directly comparable results. "As a result, differences in the findings produced by the vendors' studies cannot be interpreted as the result of different methods alone, as sample composition could have played a role" (Stipp and Woodard, 2011, p. 20).

From a scientific viewpoint, what vendors (working to a budget) offered had some limitations:

- Vendors typically provided little if any information on how they controlled for the lag issues discussed above.

- Their diagnoses and recommendations frequently depended on their subjective interpretation of the data.

- Sample sizes (and information about them) were limited. Six of the vendors used samples of between 25 and 140 participants. The other two gave no information about sample size. Such relatively small sample sizes could produce statistically significant results if a "within-subject" design were used.

- None used within-subject analysis in their standard practice, and none of them employed such a design to compare ads in Neuro 1. Only two of the vendors provided confidence intervals around their continuous data, and these indicated that data in only a couple of

seconds in a 30-second advertisement were statistically different from zero.

- Sample location also was an issue: Cross-cultural neuroscience has shown that it is wrong to assume that location does not matter for neuro measures because "a brain is a brain is a brain" (Han *et al.*, 2013). In fact, as these measures tap unconscious responses, location can matter more than for traditional measures.

  One of the eight advertisements provided for assessment in Neuro 1 featured the lesbian actress Ellen DeGeneres. It is very likely that this advertisement would have received a more favorable response in the San Francisco Bay area (the location used by one vendor) than in a more conservative city such as El Paso (the location used by another). As noted in the Neuro 1 report, when an advertisement includes "a spokesperson who might elicit different responses from different consumer segments," a larger, nationally representative sample may be needed (Stipp and Woodard, 2011, p. 22).

Neuro 1 compared what advertisers could buy from a range of leading vendors, each using a different combination of measures and working within a budget that limited sample size and sample locations. For the same budget, a technique like fMRI, which uses expensive, immovable labs and equipment, can sample fewer people in fewer locations than a more mobile technique like facial coding. Because of these differences between the vendors' measures, the results provided an excellent case study in the analysis of neuro-measure validity, specifically in the context of what is available in the market.

In contrast, Neuro 2 specifically was designed to compare representative versions of a number of different measures, including fMRI and skin conductance. Rather than comparing multiple measures from different commercial vendors, as Neuro 1 did—or choosing a single "best" version of each measure from the many available in the marketplace—Neuro 2 used four universities to gather a single "standard" academic version of each measure.

Neuro 1 Measures

Media researchers have used neuro measures to gauge two basic types of audience response: attention and emotion (Ravaja, 2004):

- Attention is the allocation of cognitive resources to process a stimulus measured by *decreases* in either heart rate (*e.g.*, Thorson and Lang, 1992) or alpha wave EEG activity (Appel *et al.*, 1979).

- For emotion, researchers measure two dimensions: valence and arousal (Ravaja, 2004):
  - ✧ Valence refers to the degree to which each of the body's two motivational systems is more strongly activated, the positive (appetitive) system or the negative (aversive) system. Valence can be measured by facial expression (Hazlett and Hazlett, 1999).
  - ✧ Arousal refers to the combined strength of activation of the two systems (Bradley and Lang, 2007) usually measured by skin conductance (Lang, Bolls, Potter, and Kawahara, 1999; Ravaja, 2004).

In Neuro 1, the eight vendors used five neuro measures (See Table 1):

- EEG;
- fMRI;
- facial coding—or second-by-second ratings of the amount of activity in "action units" of the face (*e.g.,* the brows and the cheeks)—to derive measures of positive and negative valence, and to identify

TABLE 1

Neuro Measures Used by the Eight Neuro 1 Vendors

Measure		Vendor #
EEG	Electroencephalogram (EEG) electrodes record very faint electrical signals on the surface of the scalp, often sampling faster than the speed of thought, 3/10ths of a second (Polich and Criado, 2006).	1, 2, 3, 8
fMRI	Functional magnetic resonance imaging (fMRI) detects greater blood flow associated with increased neural activity (brain cell metabolism) via the BOLD (blood-oxygen-level-dependence) signal (Weber, Mangus, and Huskey, 2015).	4
Facial Coding	Second-by-second ratings of the amount of activity in "action units" of the face (e.g., the brows and the cheeks) to derive measures of positive and negative emotion, and identify basic emotions such as happiness or fear (Ekman, Friesen, and Ancoli, 1980).	5
EMG	Facial electromyography (EMG) uses electrodes attached to the skin to detect invisible activity in facial expression muscles (Cacioppo, Petty, Losch, and Kim, 1986).	6
Biometrics	A suite of measures of the body's response to a stimulus, including skin conductance and heart rate (Potter and Bolls, 2012), motion (posture) changes (Dael, Mortillaro, and Scherer, 2012), and respiration (Ritz, Ayala, Rosemore, and Meuret, 2010).	3, 7

basic emotions such as happiness or fear (Ekman, Friesen, and Sncoli, 1980);
- EMG, specifically facial electromyography, which uses electrodes attached to the skin to detect invisible activity in facial expression muscles (Cacioppo, Petty, Losch, and Kim, 1986); and
- biometrics.

These measures were used in combination with traditional self-report measures and, in some cases, eye tracking. The vendors processed the continuous neuro data to derive measures with psychological labels such as "engagement" and "positive emotion" (See Table 2). For example, one vendor defined "engagement" very generally as "cortex electrical activity" whereas another defined it specifically as "an increasing willingness to pay."

These concepts corresponded to the three main responses measured by media psychologists—

- attention,
- positive emotion (valence), and
- arousal—

even though each vendor had its own, usually proprietary, combination of measures and data-reduction methods (the degree of transparency varied).

Despite vendors' claims to the unique nature of each analysis method, there were clear similarities across many of the methods and constructs that allowed vendors' measures to be compared, at least qualitatively. The authors of the current study concentrated on comparing measures of engagement and positive emotion, as more of the vendors provided measures of those two concepts. If these measures validly and reliably were tapping the same underlying brain or biological response, the authors argued, these measures should have aligned in the direction and intensity of those responses, if not the exact numerical quantity of response.

Assessing Content Validity among Neuro 1 Vendors

Before purchasing a neuro measure, a buyer can assess content validity by comparing how vendors describe the concepts they measure; however, as the authors of the current study found, definitions provided by vendors in Neuro 1 varied widely on concepts of engagement and positive emotion (See Table 2):

TABLE 2

Neuro 1 Vendors' Definitions of Two Shared Constructs: Engagement and Positive Emotion

Measure	Vendor's Definition	Measure	Vendor #
Engagement	"Neurons firing (electrophysiological measurements through EEG) in response to a specific stimulus. If neurons are not firing, the brain is not engaged and neurological functions, such as memory, attention, language processing and emotion are not taking place."	EEG	1
	"Physiologically the measurement of EE is cortex electrical activity. The result is a measurement of whether a consumer feels the ad conveys an emotionally self-relevant message to consumers."	EEG	3
	"A neural measure of an increasing willingness to pay."	fMRI	4
	"Percentage of participants who had at least one code-able emotional response during exposure to the stimulus and/or in response to questions."	Facial coding/ survey	5
	"Attention to something that emotionally impacts you."	Biometrics	7
	"A measure of personal relevance; is this subject matter relevant to me? Similar to what marketers would call salience."	EEG	8
Positive Emotion	"The Emotional Polarity Timeline identifies key moments of positive and negative emotions during the media."	EEG	1
	"Percentage of respondents who were predominately positive (more than 50% positive emotions)."	Facial coding	5
	"[EMG measure] of the smile muscle. The zygomatic muscle response can reflect spontaneous emotional expression, in response to humor for example. It can also indicate changes in mood states, including warm, positive feelings toward a character or a storyline, and/or the sense of resolution when a story reaches its climax."	EMG	6
	"Technically called 'motivational valence'—it equates broadly to a 'like' or 'dislike' response."	EEG	8

- Vendor 1 on "engagement": "Neurons firing [electrophysiological measurements through EEG] in response to a specific stimulus. If neurons are not firing, the brain is not engaged and neurological functions, such as memory, attention, language processing and emotion are not taking place."

  This definition suggests high face validity because it intuitively taps the concept of arousal. Brain activity, however, is not a continuous measure of memory recall while someone is watching an advertisement. Recall, by its very nature, is something that happens after exposure to a stimulus. During the viewing of the commercial, the most one can say is that certain brain activities may be predictive of subsequent recall.

- Vendor 8's definition of "positive emotion" (again based on EEG): "Motivational valence—(which) equates broadly to a 'like' or 'dislike' response"— also has high face validity. As in the case of Vendor 1, this is how the concept generally is understood: People are motivated to approach things they like and avoid things they dislike (See the generally accepted definition of valence under "Neuro 1 Measures" on page 180).

- Vendor 5's definition of "engagement" (based on facial coding), however, appears to have low face validity. The definition, "percentage of participants who had at least one code-able emotional response during exposure to the stimulus and/or in response to questions," combines arousal and valence and mixes a neuro measure with traditional survey measures. Vendor 5's measure is an aggregate measure, so statistical tests based on individual-level variance (*e.g.*, is engagement significantly different from zero?) would not be available for this measure.

Brain activity, however, is not a continuous measure of memory recall while someone is watching an advertisement.

- Vendor 4's definition of "engagement"—"A neural measure of an increasing willingness to pay"—lacks face validity because it seems more properly a definition of positive emotion, as willingness to pay is clearly a positive (appetitive) response (Bradley and Lang, 2007). Moreover, this definition suffers from the problem of "reverse inference" (Ariely and Berns, 2010). Neuroscientist Christopher Chabris gives an example of flawed reverse-inference reasoning, when discussing a study that measured activity in the amygdala, a part of the brain's emotion-processing system: "If it is true that scary things activate the amygdala, it does not follow that anything that activates the amygdala must be scary" (quoted in Felten, 2011, p. A17).

  Similarly, although there is published evidence for a correlation between activity in specific regions of the brain and actual purchase decisions (Knutson *et al.*, 2007; Weber *et al.*, 2015), it does not follow that activity in these regions must indicate an increase in willingness to pay. In defense of Vendor 4, however, this vendor did not rely on these previously published correlations to validate the claim that fMRI-detected activity indicates willingness to pay. Pre- and post-willingness to pay was measured by traditional self-reports, every time, to test the correlation asserted by the definition of the measure.

- Vendors 5 and 6 (on "positive emotion") used evidence from facial expression (EMG or human coding) to assess valence and arousal, respectively:
  - ✧ Vendor 5: "Percentage of respondents who were predominately positive (more than 50 percent positive emotions)."
  - ✧ Vendor 6: (EMG measure) of the smile muscle: "The zygomatic muscle response can reflect spontaneous emotional expression, in response to humor, for example. It can also indicate changes in mood states, including warm, positive feelings toward a character or a storyline, and/or the sense of resolution when a story reaches its climax."

  These measures have established construct validity for this purpose in a television-watching context (*e.g.*, Hazlett and Hazlett, 1999). Vendor 5's measure is again an aggregate measure, so it can't be used for most statistical tests.

- Vendor 7 measured biometric responses that, separately, had valid applications to television watching (Wang *et al.*, 2011). Some academic research, however, has suggested that these measures cannot be combined in one single measure as they do not tend to vary together (Lacey, 1967).

In short, deciphering vendors' definitions of the concepts they measure—and deciding whether these definitions have content validity—is difficult. Although the Neuro 1 report recommends that prospective buyers discuss concept definition and validity issues with vendors (Stipp and Woodard, 2011), the current authors further advise buyers to engage independent

consultants to help them to make sense of these discussions.

How to Assess Validity Empirically amid Chronic Inconsistency

If a buyer has purchased measures of the same concepts from different vendors—essentially using comparable tools to measure the same thing—in theory, that buyer should be able to compare results to see whether they agree.

The authors of the current study conducted that evaluation with the measures supplied by the eight vendors that participated in Neuro 1.

Convergent validity refers to reality checks against related variables measured at the same time, whereas predictive validity refers to correlations with measures made later—for instance, recall or sales. In general, best practice is to use multiple measures when assessing psychological constructs and to include experiential data related to the psychological construct of interest (Ariely and Berns, 2010).

- As the authors have noted, Vendor 4 followed this "best-practice" recommendation by correlating a neuro measure (fMRI) with a traditional measure of the same concept—willingness to pay—to validate the content definition of the neuro measure but, for some, the question remains: Why invest in an expensive neuro measure when a less expensive traditional measure is available? Although cost always is a consideration for brand managers, convergence with a traditional measure—willingness to pay or recall, to mention two—does not necessarily mean there is no need for the neuro measure. Indeed, the latter may predict other outcomes (*e.g.*, low-involvement persuasion without awareness) that traditional measures cannot detect (Heath, 2009).
- In Neuro 1, only one vendor (Vendor 8) used a single measure (EEG); one other company (Vendor 7) made no traditional measures. For these vendors, it was harder to assess the validity of their results.
- By contrast, like Vendor 4, Vendors 1, 3, 5, and 6 used combinations of neuro and traditional measures to assess the concepts of interest.

As discussed earlier, the main benefit of neuro measures is their potential to provide continuous measures of consumer response, allowing the identification of key seconds (or even key frames) of a commercial; but when the authors of the current study compared the vendors' continuous measures of "engagement" (See Figure 1) and measures of "positive emotion" (See Figure 2)—for each commercial—these measures showed a notable lack of consistency, with no seconds where all eight vendors were in agreement.

One vendor's high (engagement or positive emotion) response occurs during the same second as another vendor's low response or a third vendor's flat-line response. Only the inconsistency between Vendor 4's measure and other vendors' measures of engagement could be explained by a potentially inappropriate comparison of two different concepts: positive emotion and arousal (cf. Patnaik and Purvis, 2011). Another possible explanation is that the measures agreed but were misaligned with the commercial content because of differences in the way vendors controlled or did not control for lags in response (there would be no need to speculate if more details were available).

The current authors next aligned scores from the eight different Neuro 1 vendors averaged across time units that lasted more than one second: the individual

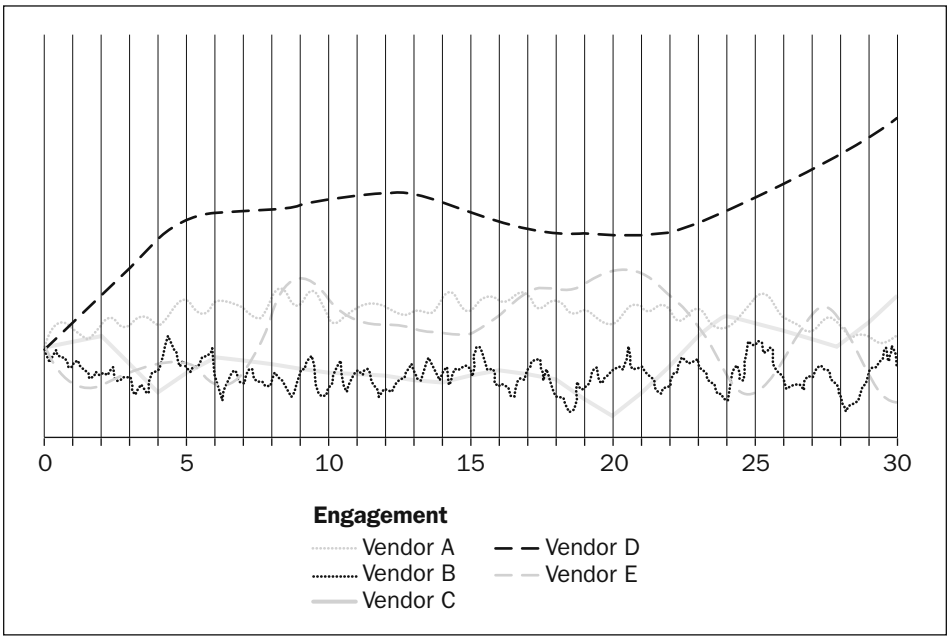

Figure 1 Continuous Response Measures of Engagement

Response measures from five of the eight Neuro 1 vendors, for a test commercial, vary widely.

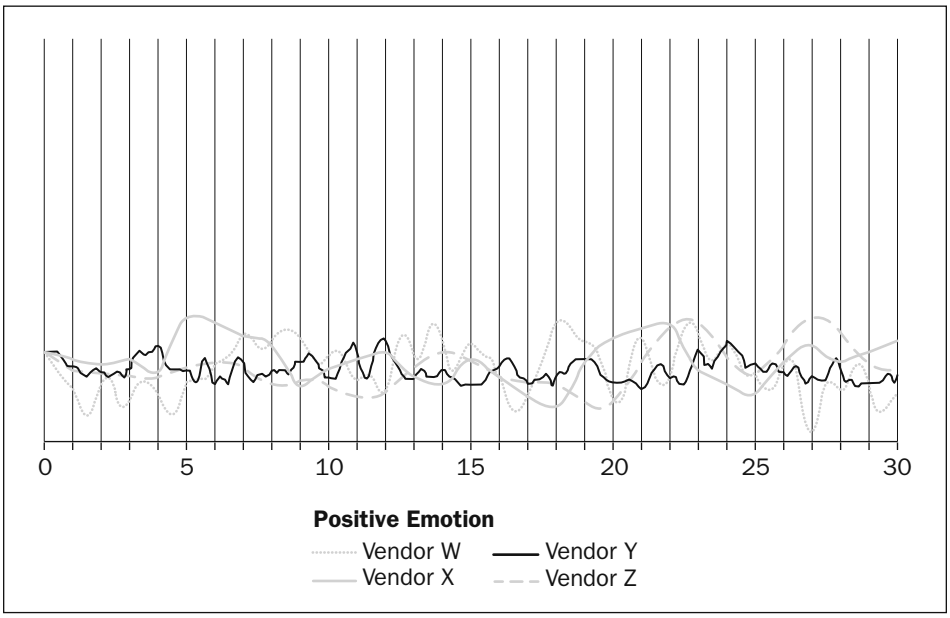

Figure 2 Continuous Response Measures of Positive Emotion

Response measures from four of the eight vendors, for a test commercial, vary widely.

scenes within a television commercial. These scene-level averages should have been more comparable than the second-by-second results, because averaging across multiple seconds largely should have eliminated the problems of lag issues and other differences across vendors, including different beginning and end points for scenes. Descriptions of the scenes in the vendors' reports (*e.g.*, "the germs scene") were used to match scenes across vendors. Once again, however, the vendors' average scores for individual scenes showed a notable lack of consistency (See Figure 3). This time, this inconsistency could be analyzed and was statistically significant ($p < 0.001$). For any individual scene,

- 2.2 of the 8 vendors, on average, thought the scene was a "good" point in the commercial;
- 1.0 thought the scene was a "bad" point; and
- 4.8 did not think the scene was worth commenting on.

Lag issues and interpretation issues may have precluded comparisons between the vendors using any unit of time less than the full 30 seconds of a commercial. With those limitations in mind, the authors compared how the vendors ranked the

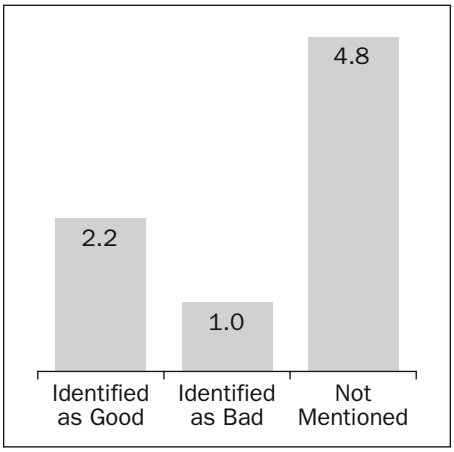

Figure 3 Average Number Of the Eight Vendors Commenting on a Key Scene Identified in Any of the Eight Test Commercials

TABLE 3

"Best" to "Worst" Vendor Rankings of the 8 Commercials A-H

		Vendor							
Measure		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Engagement	"Best"	D		D	C	C		F	
		B		G	A	A		D	
		C		C	G	F		B	
		H		E	E	H		C	
		G		H	D	D		H	
		F		F	F	G		A	
		A		A	B	B		E	
	"Worst"	E		B	H	E		G	
Positive Emotion	"Best"					E	D		
						F,D	H		
							E		
						B	B		
						H	A		
						G	F		
						C	G		
	"Worst"					A	C		

commercials from "best" to "worst" on three different measures:

- recall,
- positive emotion, and
- engagement (See Table 3).

Seven of the eight vendors provided overall assessments of the eight commercials (Vendor 8 provided only continuous measures). In this final comparison analysis, the lack of consistency once again was notable. The rank order correlations were large (r = 0.54) for the two positive emotion rankings but were small (r = 0.10) for the five engagement rankings (using the standard Cohen definitions of large, medium, and small correlations). Further, each vendor's overall evaluation of each commercial and recommendations on how to improve them depended on subjective judgments that, inevitably, introduced even more variation between the different vendors.

More Problems with Validity of Measures

The authors assessed the validity of the Neuro 1 vendors' measures more formally using the Multitrait-Multimethod (MTMM) correlation matrix (Campbell and Fiske, 1959; See Table 4). A MTMM matrix consists of measures of the same traits (constructs). As such, they should correlate more highly with one another than they do with measures of different traits made using the same method. For advertisers who have purchased multiple measures from different vendors, the MTMM correlation matrix offers a way to weigh results from a variety of research findings.

Correlations between measures of the same construct are called "validity correlations." Validity never can be higher than a measure's reliability, and considering that most neuro methods provide, at best, reliabilities between 0.6 and 0.7 (Vul, Harris, Winkelman, and Pashler, 2009), validities that exceed 0.6 are unlikely. Correlations or validities of 0.3 or above count as "acceptable"; and near or above 0.6 indicate "strong validity" (Vul *et al.*, 2009).

The authors assessed correlations between rankings of the eight commercials on measures of

- recall (A),
- positive emotion (B), and
- engagement (C; See Table 4).

There are three reasons for rankings rather than comparing raw scores:

- To ensure complete cooperation from participating companies, the ARF agreed with the Neuro 1 sponsors to withhold scores for individual advertisements (Stipp and Woodard, 2011, p. 20).

- The reports provided by the vendors did not include the raw data, so the authors of the current study could not convert the vendors' results into a common metric based on standard deviations (Patnaik and Purvis, 2011).

- Because of methodological differences across vendors (*e.g.*, how they dealt with lag issues), it was very difficult to align second-by-second—or even their scene-by-scene results. The only comparable time unit was the entire 30-second duration of each commercial (See Figures 1, 2, and 3).

As reported above, the strongest validity correlation that the authors found (r = 0.54) was between two measures of positive emotion (B1 and B2, Vendors 5 and

TABLE 4

Multitrait-Multimethod Matrix

Measure (Method)		A1	B1	C1	A2	B2	C2	C3	C4	C5
Unaided Recall (Vendor 1)	**A1**	—								
Positive Emotion (Vendor 6)	**B1**	−0.29	—							
Engagement (Vendor 1)	**C1**	*−0.07*	0.21	—						
Unaided Recall (Vendor 2)	**A2**	**0.27**	−0.37	0.02	—					
Positive Emotion (Vendor 5)	**B2**	−0.08	**0.54**	−0.04	0.00	—				
Engagement (Vendor 5)	**C2**	0.07	−0.45	**0.05**	−0.44	*−0.65*	—			
Engagement (Vendor 3)	**C3**	−0.67	0.05	**0.26**	−0.22	0.13	**−0.05**	—		
Engagement (Vendor 4)	**C4**	−0.52	−0.62	**−0.26**	−0.02	−0.51	**0.40**	**−0.33**	—	
Engagement (Vendor 7)	**C5**	0.26	0.17	**0.52**	−0.32	0.33	**0.29**	**−0.17**	**−0.40**	—

Validity correlations (same trait, different method) are shown in bold. Monomethod correlations (same method or vendor, different trait) are shown in italics.

6). However, there also was evidence of same-method bias for one of these measures (B2, Vendor 5), as it had an even stronger (absolute value) correlation with a measure of engagement by the same vendor (C2, $r = -0.65$).

The validity correlation for the two unaided recall measures was just under the 0.3 minimum (A1/A2, $r = 0.27$), because one of the two vendors measured recall in the United Kingdom, where the brands were unfamiliar. Again, the validity correlation was lower in absolute value than correlations with many other constructs in the same row and column ($r = -0.67$ to +0.26; See Table 4).

The validity correlations for the engagement construct ranged from "strongly negative" to "strongly positive" ($r = -0.40$ to +0.52), so their average correlation (after appropriate Fisher-transformation) was not much different from zero. The opposite signs may have been due to some measures having a negative correlation with arousal (alpha wave EEG) and others having a positive correlation (facial muscle activity). However, the average correlation between engagement measures increased only to $r = 0.29$ when the authors averaged the absolute values of the correlations.

An MTMM matrix also allows tests of convergent, discriminant, and predictive validity. For example, high-engagement commercials should be highly correlated with measures of (conscious) recall (Appel *et al.*, 1979; Young, 2002). The authors of the current study saw some strong correlations with recall, but they were all negative (See Table 4). Specifically, the negative correlations between the two recall measures and Vendor 3's EEG measure ($r = -0.67$ and $r = -0.22$) could be explained by the non-intuitive inverse relationship between attentive encoding and alpha brainwave activity, measured by EEG: High alpha wave activity is associated with sleep; low alpha activity (alpha blocking) is a signal of high attention (Pope, Bogart, and Bartolome, 1995).

Recall is not necessarily correlated with positive emotion (Zeitlin and Westwood, 1986) and, therefore, the authors of the current article looked for low correlations that would indicate discriminant validity (because recall and positive emotions are unrelated constructs) between measures of recall and positive emotion. And, in fact, the authors did observe a small negative average correlation between these measures ($r = -0.19$; See Table 4).

If Vendor 4's fMRI indication of increasing willingness to pay also were considered a measure of positive emotion rather than engagement, the negative correlation between this measure and recall ($r = -0.52$ and $r = -0.02$) suggested that consumers more likely would remember advertisements for brands they are less willing to buy.

Similarly, engagement should be uncorrelated with positive emotion, as people could experience high engagement/arousal with positive- or negative-emotion content.

Despite some very strong negative correlations ($r = -0.65$ to +0.22; See Table 4), the average correlation between measures of positive emotion and engagement also was small and negative ($r = -0.17$). Once again, though, for individual vendors' measures, strong correlations (*e.g.*, $r = -0.65$) with measures of a different construct suggested problems with the validity of these measures.

In summary, the current authors carried out an empirical comparison of the measures provided by the eight vendors that participated in Neuro 1 and found very little consistency—the opposite, they believe, of what would be expected if the vendors' neuro measures were more accurate and objective than traditional self-report measures.

Recall is not necessarily correlated with positive emotion.

DISCUSSION

Although neuromarketing methods are still at an early stage and not yet delivering their full promise (Ariely and Berns, 2010; Stipp and Woodard, 2011), advertisers should not abandon them. Advertisers want measures of processing that go beyond mere exposure; these new methods do provide continuous measures of instantaneous, subconscious emotional, and cognitive responses—free of the memory and social desirability biases that can be associated with self-reporting—and potentially more predictive of advertising effectiveness than traditional measures.

The ARF's neuro trials provided the first publicly available head-to-head comparisons of these new neuro measures. Neuro 1 was designed to evaluate the commercially available versions of these approaches, reproducing the real-world purchase decisions faced by prospective buyers of these new measures. Neuro 2, in turn, evaluated which pure approach (*e.g.*, fMRI versus EEG) was the most predictive of actual sales.

Chief among the recommendations of the Neuro 1 report (Stipp and Woodard, 2011) was that prospective buyers should discuss issues of reliability and validity with vendors. Using disguised data from Neuro 1 published here for the first time, the current study should further that goal.

The current authors first showed how buyers could compare the content validity of vendors' descriptions of the concepts their methods measure. The authors then showed how buyers could use correlational analysis to compare purchased measures of the same concept from

different vendors, and measures of multiple concepts from the same vendor.

The current study's comparison of results from Neuro 1 found, in fact, that there was no common truth among neuro measures of advertisement effectiveness. Specifically,

- the eight vendors disagreed about the definitions of similar concepts;
- there were low correlations between measures of the same concept made with the same method; and
- there was limited agreement on many aspects of the commercials, including their relative effectiveness.

These limitations, in fact, may result more from over-enhanced expectations than the real abilities of current neuro technology. In truth, traditional recall and persuasion measures can be equally unreliable (Lodish *et al.*, 1995). Many neuro vendors, however, have encouraged the belief that their measures are more reliable than traditional measures because they measure neurological and biological processes (Stipp and Woodard, 2011).

The results of the current study suggest that advertisers need to choose their vendor carefully, maintaining a healthy skepticism about the reliability and validity of vendors' methods. As noted earlier, this process may benefit from the contracting of independent third-party experts.

The ARF's Neuro 1 report noted there were disagreements between the vendors and highlighted the problems caused by sampling issues, especially when comparing reports from vendors in different locations (Stipp and Woodard, 2011). Sample location, however, is less of an explanation for differences between vendors when correlations are analyzed, as the authors of the current article believe they have demonstrated: Scores might differ between locations, but these findings should still be highly correlated if they measure the same concept (*e.g.*, attitude).

The inconsistency of information reflects the current stage of development of these methods.

The Neuro 1 data analyzed in this study were collected in 2010 and may be unrepresentative of the methods used by the eight vendors currently in 2015. The authors reviewed the vendors' websites to identify updates to their methodologies since Neuro 1 and did find some evidence of revised practices:

- In 2014, two of the EEG vendors were providing mobile EEG measurement, which theoretically would eliminate problems of sample location and allow in-store measurement at the point-of-purchase (POPAI, 2012).
- Head-mounted eye tracking allows vendors to pinpoint which screen is driving responses when television viewers are multitasking (Warc, 2014).
- Since 2010, many of these vendor websites have added pages devoted to "validation" and "due diligence" (Dooley, 2012) that address points and recommendations made by the Neuro 1 report.

The current study should not be seen as a verdict on any particular vendor or method or an argument that neuro measures are not valid, but the study does demonstrate that these measures do not reflect a common truth. The disagreements between vendors add a level of difficulty to a buyer's purchase decision in addition to the understandable difficulties of trying to understand this highly technical new area.

In the authors' view, the inconsistency of information reflects the current stage of development of these methods rather than any weakness on the part of the vendors who participated in the study. Indeed, as clients, the authors would be cautious about working with other vendors who were not prepared to submit their work to this type of rigorous peer review.

FUTURE RESEARCH

The groundbreaking results of the ARF's Neuro 1 and Neuro 2 collaborations suggest many directions for future research.

- The results of Neuro 1 suggest that, in the short term, more applied research is needed to test the validity of these approaches.
- Neuro 2 identified which neuro measures—collected using a "pure" academic approach—were the most predictive of actual in-market sales. Long-term theoretical research will be needed to understand why advertising "works" on these measures and how advertisers can optimize advertising effectiveness using these measures.

  Advertisers, however, may not be able to buy the "pure" measures endorsed by Neuro 2. Vendors supply modified, proprietary versions of these measures, or proprietary combinations of several measures. Sometimes, the proprietary combination includes traditional measures. And usually, the results, even when "pure" measures are used, are interpreted using the vendor's experience and judgment.

The authors believe that a short list of specific considerations that future research could address should include the following:

- Studies of the best ways to measure EEG responses to commercials: The current study found a low correlation between vendors using the same method (EEG) to measure the same construct (engagement). This low correlation must reflect differences between different vendors' construct definitions and methods—perhaps including measurement and sampling errors/differences.

  With this in mind, the authors believe that new studies need to explore the implications of averaging whole-brain activity versus comparing localized activity (*e.g.*, left versus right) and whether it is possible to reveal relatively noise-free responses without repeated exposure (*e.g.*, Silberstein and Nield, 2008).

- Studies of the best ways to measure facial expression: Although the authors of the current study found a relatively high correlation between facial-expression measures of positive emotion, future studies should investigate the trade-offs involved between using high participant-cost methods like fEMG and human coding versus less expensive computerized methods (Teixeira and Stipp, 2013).

- More research is needed about the concept validity of fMRI measures, with questions that include
  - ✧ which questions can be asked and answered with fMRI (Weber *et al.*, 2015);
  - ✧ which fMRI measures reflect "engagement" or "positive emotion";
  - ✧ how reliable and valid are fMRI measures of time periods shorter than 10 to 15 seconds;
  - ✧ on which brain regions and brain networks should fMRI measures focus? Different studies relate different regions and networks to advertising effectiveness, such as the nucleus accumbens (Knutson *et al.*, 2007), the medial prefrontal cortex (Falk *et al.*, 2012; Weber *et al.*, 2015), and the superior temporal sulcus (Bakalash and Riemer, 2013).

- As the brain is a network that encodes complex information in multiple areas, it is unlikely that there are "buy buttons" in the brain. It is more likely that measures of brain connectivity across cortical and sub-cortical networks reach the level of validity needed for solid predictions.

In short, the main task for future research in this area should be to increase the transparency and credibility of commercial versions of neuro measures so that buyers have fewer concerns about their reliability and validity than for traditional measures.

IMPLICATIONS

The current analysis of the Neuro 1 results also has a number of practical implications for buyers of neuro measures:

- Buyers should carefully discuss issues of reliability and validity with potential vendors before choosing their vendor(s) (Stipp and Woodard, 2011). Buyers can begin this discussion by comparing definitions of concepts across vendors. Buyers should ignore claims based on "neurobabble"—terminology such as "buy spots" (Lee, Broderick, and Chamberlain, 2007); "buy buttons" (Hubert and Kenning, 2008); and other "buy"-related areas that do not exist in the human brain. Likewise, they should eschew appealing pictures such as brain images and heat maps (McCabe and Castel, 2008) that often substitute for evidence. Given the complexity of the issues, the authors of the current study recommend that buyers engage independent third-party assistance during these discussions.

  Although a thorough, informed selection process is expensive, the authors believe it will be less costly in the long term than either ignoring the differences in neuro measures or blindly accepting vendors' claims.

- Buyers need to use the right tool for the right job. Although one resource could be Neuro 1's linking of research objectives with their potentially best traditional and neuro measures (Stipp and Woodard, 2011, p. 6), advertisers need to gain experience with a number of vendors to understand what the different measures mean and what their own objectives should be.

- After choosing one or more vendors, buyers should carry out a "post-research audit" to compare the new neuro measures with traditional measures or in-market results (Stipp and Woodard, 2011). Neuro measures need cross-validating with well-understood traditional measures and constructs, and their relationship with advertising effects needs to be better explained.

  Many of the vendors in Neuro 1 provided traditional measures as part of their service. Buyers, in fact, can use these measures—as did the authors of the current study—to carry out correlational checks of reliability and validity. Measures should be identified clearly and separated for these correlations to be meaningful, not intermingled to give traditional measures a "neuro halo" (*i.e.*, reporting overall advertising scores that combine self-report measures with neuro measures).

  In practice, these new neuro measures almost always will be used in conjunction with less expensive, better-established traditional methods. The

question, therefore, is: What is the incremental value of neuro measures? The main aim should be to use the combined insights from neuro measures and traditional research to suggest improvements in creative execution and advertising success, typically from minor edits.

Other possible applications include selecting segments/elements for reuse in shorter commercials and/or other communications (packaging, point-of-sale, other media) and selecting compatible television programs in terms of audience response rather than just demographics. Over time, many of the limitations of neuro measures gradually should reduce, increasing the number of ways in which buyers can use these methods cost-effectively.

For vendors to advance the field, there is a need for greater transparency, at least about the constructs measured and the basic methodology used. Key constructs, such as "engagement" (in its neuro context) should be standardized in a peer-review process (Rossiter, 2002) so that they become common currencies that can be compared across vendors and cross-validated in independent controlled studies (Stewart, 1984).

In fact, at the moment, many neuro vendors are competing as monopolists with unique proprietary or even patented methods. This may increase returns in the short term but risks a backlash against neuro measures.

The combination of exaggerated claims of efficacy, on the one hand, and secrecy surrounding the empirical evidence for neuro measures, on the other, contrasts with the transparency of neuro-economics research (Fisher, Chin, and Klitzman, 2010). Even in extraordinary cases where there is validity to stated capabilities, it is unlikely that vendors would be able to maintain their monopoly power for very long. These technologies are becoming more ubiquitous and less expensive to use, enabling other vendors and researchers to test vendors' claims independently. Soon, just like traditional vendors, neuro-measure vendors would have to make money from the quality of their research and not just from its scarcity or claimed unique potency.

CONCLUSION

The current study used data from the ARF's Neuro 1 trial to illustrate how buyers can investigate and discuss the reliability and validity issues associated with neuro measures of advertising effectiveness, as recommended by the ARF's Neuro 1 report (Stipp and Woodard, 2011). In the wake of Neuro 2, which identified standard neuro measures that are predictive of actual sales, the authors also suggested avenues for further research validating vendors' modifications of these standard measures.

Waves of interest in "pure" measures of advertising response have come and gone in the past, many times for the same reason: Though grand claims were made, they could not be replicated by other researchers (Stewart, 1984). To prevent this happening with this new wave of neuro measures, vendors will have to show that they have sufficient confidence in their measures that they are willing to let others test them independently. Neuro vendors should compete like opinion-poll vendors: on the quality of their data, not the uniqueness of their measures. JAR

Duane Varan is professor of audience research at Murdoch University in Perth, Australia and director of Murdoch's television-focused Audience Labs. He also is ceo of Austin, TX–based MediaScience, facilitator of the ESPN Lab, and he oversees "Beyond :30," a collaborative industry project exploring the changing media landscape. His work has been published in the *Journal of Advertising Research*, *Journal of Communication*, and *Journal of Economic Psychology*.

Annie Lang is distinguished professor of telecommunications and cognitive science at Indiana University. She studies mediated-message processing; in particular, the dynamics of how the human brain understands message structure and content. Her research has been published in the *Journal of Broadcasting & Electronic Media*, *Journal of Communication*, and *Media Psychology*, and she also edited the book, *Measuring Psychological Responses to Media Messages* (Routledge, 1994).

Patrick Barwise is emeritus professor of management and marketing at London Business School, a visiting senior fellow at the London School of Economics, and chairman of Which?, the UK's leading consumer organization. He joined LBS in 1976 after an early career at IBM and has published widely on management, marketing, and media. His book, *Simply Better: Winning and Keeping Customers by Delivering What Matters Most* (Harvard Business School, 2004) with Seán Meehan, won the American Marketing Association's 2005 Berry-MA Book Prize. Their follow-up book, *Beyond the Familiar: Long-Term Growth through Customer Focus and Innovation* (John Wiley & Sons), was published in 2011.

René Weber is professor at the University of California, Santa Barbara's (UCSB) communication department and director of UCSB's Media Neuroscience Lab. His research focuses on cognitive responses to mass communication and new-technology media messages, including video games. Weber's work, which applies both traditional social scientific, neuroscientific methodology, and functional magnetic resonance imaging has been published in major communication and neuroscience journals, including *Media Psychology* and *Human Brain Mapping*, and in three books.

Steven Bellman is associate professor and deputy director of Audience Labs at Murdoch University. His research on viewer responses to media content and advertising is funded by the "Beyond :30" project's sponsors, who include many of the world's leading television networks and advertisers. Bellman's work has been published in the *Journal of Marketing*, *Journal of the Academy of Marketing Science*, and *Management Science*, and he

is a co-author of *Marketing Communications* (Pearson-Prentice Hall, 2005).

REFERENCES

Appel, V., S. Weinstein, and C. Weinstein. "Brain Activity and Recall of TV Advertising." *Journal of Advertising Research* 19, 4 (1979): 7–15.

Ariely, D., and G. S. Berns. "Neuromarketing: The Hope and Hype of Neuroimaging in Business." *Nature Reviews Neuroscience* 11, 4 (2010): 284–292.

Bakalash, T., and H. Riemer. "Exploring Ad-Elicited Emotional Arousal and Memory for the Ad Using fMRI." *Journal of Advertising* 42, 4 (2013): 275–291.

Biocca, F., P. David, and M. West. "Continuous Response Measurement (CRM): A Computerized Tool for Research on the Cognitive Processing of Commercial Messages." In *Measuring Psychological Responses to Media*, A. Lang, ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.

Boyd, T. C., and G. D. Hughes. "Validating Realtime Response Measures." In *Advances in Consumer Research, vol. 19*, J. F. Sherry, Jr. and B. Sternthal, eds. Provo, UT: Association for Consumer Research, 1992.

Bradley, M. M., and P. J. Lang. "Emotion and Motivation." In *Handbook of Psychophysiology*, 3rd ed., J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, eds. New York, NY: Cambridge University Press, 2007.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14, 5 (2013): 365–376.

Cacioppo, J. T., R. E. Petty, M. E. Losch, and H. S. Kim. "Electromyographic Activity over Facial Muscle Regions Can Differentiate the Valence and Intensity of Affective Reactions." *Journal of Personality and Social Psychology* 50, 2 (1986): 260–268.

Campbell, D. T., and D. Fiske. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56, 2 (1959): 81–105.

Dael, N., M. Mortillaro, and K. R. Scherer. "Emotion Expression in Body Action and Posture." *Emotion* 12, 5 (2012): 1085–1101.

Dooley, R. (2012, August 30). "The Neuromarketing Challenge: First Response." Retrieved January 8, 2015, from http://www.neurosciencemarketing.com/blog/articles/challenge-innerscope.htm

Ekman, P., W. V. Friesen, and S. Ancoli. "Facial Signs of Emotional Experience." *Journal of Personality and Social Psychology* 39, 6 (1980): 1125–1134.

Falk, E., E. Berkman, and M. D. Lieberman. "From Neural Responses to Population Behavior: Neural Focus Group Predicts Population-Level Media Effects." *Psychological Science* 23, 5 (2012): 439–445.

Felten, E. "The Amygdala As Sales Tool." *Wall Street Journal*, November 30, 2011.

Fenwick, I., and M. D. Rice. "Reliability of Continuous Measurement Copy-Testing Methods." *Journal of Advertising Research* 31, 1 (1991): 23–29.

Fisher, C. E., L. Chin, and R. Klitzman. "Defining Neuromarketing: Practices and Professional Challenges." *Harvard Review of Psychiatry* 18, 4 (2010): 230–237.

Fredrickson, B. L., and D. Kahneman. "Duration Neglect in Retrospective Evaluations of Affective Episodes." *Journal of Personality and Social Psychology* 65, 1 (1993): 45–55.

Han, S., G. Northoff, K. Vogeley, B. E. Wexler, S. Kitayama, and M. E. W. Varnum. "A Cultural Neuroscience Approach to the Biosocial Nature of the Human Brain." *Annual Review of Psychology* 64 (2013): 335–359.

Hazlett, R. L., and S. Yassky Hazlett. "Emotional Response to Television Commercials: Facial EMG vs. Self Report." *Journal of Advertising Research* 39, 2 (1999): 7–23.

Heath, R. "Emotional Engagement: How Television Builds Big Brands at Low Attention." *Journal of Advertising Research* 49, 1 (2009): 62–73.

Hubert, M., and P. Kenning. "A Current Overview of Consumer Neuroscience." *Journal of Consumer Behaviour* 7, 4–5 (2008): 272–292.

Knutson, B., S. Rick, G. E. Wimmer, D. Prelec, and G. Loewenstein. "Neural Predictors of Purchases." *Neuron* 53, 1 (2007): 147–156.

Krishnan, H. S., and D. Chakravarti. "Memory Measures for Pretesting Advertisements: An Integrative Conceptual Framework and a Diagnostic Template." *Journal of Consumer Psychology* 8, 1 (1999): 1–37.

Lacey, J. I. "Somatic Response Patterning and Stress: Some Revisions of Activation Theory." In *Psychological Stress: Issues in Research*, M. H. Appley, and R. Trumbull, eds. New York, NY: Appleton-Century-Crofts, 1967.

Lang, A., P. Bolls, R. F. Potter, and K. Kawahara. "The Effects of Production Pacing and Arousing Content on the Information Processing of Television Messages." *Journal of Broadcasting & Electronic Media* 43, 4 (1999): 451–475.

Lee, N., A. J. Broderick, and L. Chamberlain. "What Is 'Neuromarketing'? A Discussion and Agenda for Future Research."

International Journal of Psychophysiology 63, 2 (2007): 199–204.

Lodish, L. M., M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens. "How T.V. Advertising Works: A Meta-Analysis of 389 Real World Split Cable T.V. Advertising Experiments." *Journal of Marketing Research* 32, 2 (1995): 125–139.

McCabe, D. P., and A. D. Castel. "Seeing is Believing. The Effect of Brain Images on Judgments of Scientific Reasoning." *Cognition* 107, 1 (2008): 343–352.

Noble, T. "Neuroscience in Practice: The Definitive Guide for Marketers." *Admap* 48, 3 (2013): 30–45.

Patnaik, S., and S. Purvis. (2011, March). "A Case Study Using the Heineken 'Weasel' Commercial: Comparative Analysis of Emotion Measures." Retrieved January 8, 2015, from http://www.quirks.com/articles/2011/20110307.aspx

Poels, K., and S. Dewitte. "How to Capture the Heart? Reviewing 20 Years of Emotion Measurement in Advertising." *Journal of Advertising Research* 46, 1 (2006): 18–37.

Polich, J., and J. R. Criado. "Neuropsychology and Neuropharmacology of P3a and P3b." *International Journal of Psychophysiology* 60, 2 (2006): 172–185.

POPAI [Point of Purchase Advertising International]. (2012). "2012 Shopper Engagement Study: Media Topline Report." Retrieved February 11, 2014, from http://www.popai.com/engage/docs/Media-Topline-Final.pdf

Pope, A. T., E. H. Bogart, and D. S. Bartolome. "Biocybernetic System Evaluates Indices of Operator Engagement in Automated Task." *Biological Psychology* 40, 1–2 (1995): 187–195.

Potter, R. F., and P. D. Bolls. *Psychophysiological Measurement and Meaning: Cognitive and Emotional Processing of Media*. New York, NY: Routledge, 2012.

Ravaja, N. "Contributions of Psychophysiology to Media Research: Review and Recommendations." *Media Psychology* 6, 2 (2004): 193–235.

Ritz, T., E. S. Ayala, J. Rosemore, and A. E. Meuret. "Breathing Dysregulation in Blood-Injury-Injection Phobia during Exposure." *Biological Psychology* 83, 1 (2010): 65–66.

Rossiter, J. R. "The C-OAR-SE Procedure for Scale Development in Marketing." *International Journal of Research in Marketing* 19, 4 (2002): 305–335.

Silberstein, R. B., and G. E. Nield. "Brain Activity Correlates of Consumer Brand Choice Shift Associated with Television Advertising." *International Journal of Advertising* 27, 3 (2008): 359–380.

Singer, N. "Making Ads That Whisper to the Brain." *The New York Times*, November 14, 2010.

Stewart, D. W. "Physiological Measurement of Advertising Effects." *Psychology & Marketing* 1, 1 (1984): 43–48.

Stipp, H., and R. P. Woodard. *Uncovering Emotion: Using Neuromarketing to Increase Ad Effectiveness*. New York: ARF, 2011.

Teixeira, T. S., and H. Stipp. "Optimizing the Amount of Entertainment in Advertising: What's So Funny about Tracking Reactions to Humor?" *Journal of Advertising Research* 53, 3 (2013): 286–296.

Thorson, E., and A. Lang. "The Effects of Television Videographics and Lecture Familiarity on Adult Cardiac Orienting Responses and Memory." *Communication Research* 19, 3 (1992): 346–369.

Treleaven-Hassard, S., J. Gold, S. Bellman, A. Schweda, J. Ciorciari, C. Critchley, and D. Varan. "Using the P3a to Gauge Automatic Attention to Interactive Television Advertising." *Journal of Economic Psychology* 31, 5 (2010): 777–784.

Vito, C. A. (2014, December 11). "Fox Researchers Use Brain Data to Predict Real-Life Success of TV Ads." Retrieved February 5, 2015, from http://www.fox.temple.edu/posts/2014/12/fox-researchers-use-brain-data-predict-real-life-success-tv-ads

Vul, E., C. Harris, P. Winkelman, and H. Pashler. "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition." *Perspectives on Psychological Science* 4, 3 (2009): 274–290.

Wang, Z., A. Lang, and J. R. Busemeyer. "Motivational Processing and Choice Behavior During Television Viewing: An Integrative Dynamic Approach." *Journal of Communication* 61, 1 (2011): 71–93.

Warc [World Advertising Research Council]. (2014, January 21). Dual screening increases TV engagement. Retrieved January 8, 2015, from http://www.warc.com/Content/News/N32477_Dual_screening_increases_TV_engagement.content?

Weber, R. J., M. Mangus, and R. Huskey "Brain Imaging in Communication Research: A Practical Guide to Understanding and Evaluating fMRI Studies." *Communication Methods and Measures* 9, 1 (2015): 5–29.

Young, C. E. "Brain Waves, Picture Sorts, and Branding Moments." *Journal of Advertising Research* 42, 4 (2002): 42–53.

Zeitlin, D. M., and R. A. Westwood. "Measuring Emotional Response." *Journal of Advertising Research* 26, 5 (1986): 34–44.