IN THIS ISSUE:

From theory to common practice: consumer neuroscience goes mainstream

Using single-source data to measure advertising effectiveness

Using crowdsourcing and image processing to automate data collection
The world of measurement is changing.

Thanks to recent advances in data collection, transfer, storage and analysis, there’s never been more data available to research organizations. But ‘Big Data’ does not guarantee good data, and robust research methodologies are more important than ever.

Measurement Science is at the heart of what we do. Behind every piece of data at Nielsen, behind every insight, there’s a world of scientific methods and techniques in constant development. And we're constantly cooperating on ground-breaking initiatives with other scientists and thought-leaders in the industry. All of this work happens under the hood, but it’s not any less important. In fact, it's absolutely fundamental in ensuring that the data our clients receive from us is of the utmost quality.

These developments are very exciting to us, and we created the Nielsen Journal of Measurement to share them with you.

WELCOME TO THE NIELSEN JOURNAL OF MEASUREMENT

SAUL ROSENBERG

The Nielsen Journal of Measurement will explore the following topic areas in 2016:

**BIG DATA** - Articles in this topic area will explore ways in which Big Data may be used to improve research methods and further our understanding of consumer behavior.

**SURVEYS** - Surveys are everywhere these days, but unfortunately science is often an afterthought. Articles in this area highlight how survey research continues to evolve to answer today's demands.

**NEUROSCIENCE** - We now have reliable tools to monitor a consumer's neurological and emotional response to a marketing stimulus. Articles in this area keep you abreast of new developments in this rapidly evolving field.

**ANALYTICS** - Analytics are part of every business decision today, and data science is a rich field of exploration and development. Articles in this area showcase new data analysis techniques for measurement.

**PANELS** - Panels are the backbone of syndicated measurement solutions around the world today. Articles in this area pertain to all aspects of panel design, management and performance monitoring.

**TECHNOLOGY** - New technology is created every day, and some of it is so groundbreaking that it can fundamentally transform our behavior. Articles in this area explore the measurement implications of those new technologies.
Welcome to the 2nd issue of the Nielsen Journal of Measurement!

In this issue, we’re delighted to feature three papers that outline important advances in measurement capabilities made possible by new technology. The first paper, “From theory to common practice: consumer neuroscience goes mainstream,” discusses the tools of consumer neuroscience today, the theory they’re drawing from, and the consumer behaviors and market dynamics that those techniques can help unveil, particularly in the field of video advertising. This paper provides insights from some of the top scientists in the field and outlines recent breakthroughs we’ve been able to achieve by combining multiple tools on the same case studies.

The second paper, “Using single-source data to measure advertising effectiveness,” bridges the traditional data gap between ad exposure and sales through the use of single-source data. New technology developments come in the shape of loyalty-card purchase data and advanced algorithms to process and calibrate big data. The process described in this paper is helping write a new chapter in the book on advertising effectiveness. Developed over many years and in close cooperation with top industry players, it has already been the basis of thousands of successful case studies.

The third paper, “Using crowdsourcing and image processing to automate data collection,” examines how we’re piloting the use of smartphone cameras, optical character recognition algorithms and modern crowdsourcing techniques to capture grocery receipts directly from consumers in a number of markets around the world where point-of-sale data isn’t universally available. To get an accurate picture of all retail establishments in those markets, we must complement that point-of-sale data with something else, and that something else has so far come from manual data collection efforts. The paper describes a very promising proof-of-concept initiative in the U.K. that is paving the way towards automation.

Finally, we’ve included four shorter pieces in this issue to give you a preview of some exciting new work we’re engaged in to address the ever-changing media landscape. It’s no big secret that we’re not watching television content today the same way we did only a few years ago. The old broadcasting mold isn’t broken, but it’s seriously being challenged by new digital systems. While these new programming and delivery options are empowering for viewers, they’re more complex to measure for research companies. In these four short pieces—we call them ‘snapshots’—we review novel engineering solutions (both hardware and software) to capture TV ratings data in that new environment, and we examine some of the changes in viewer behavior that those solutions are already allowing us to uncover. We’re even exploring new ways to predict ratings in the future. Enjoy this new issue of the journal!

JEROME SAMSON, MANAGING EDITOR
IN THIS ISSUE

SNAPSHOTS

In each issue of the Journal, we start with a few snapshots to introduce current measurement topics in a summary format. We expect to develop many of these snapshots into full-length articles in future issues of the Journal.

1. SMALLER CHEAPER STRONGER: THE NANO METER .......................................................... 6
2. KIT AND CABOODLE: A SOFTWARE DEVELOPMENT KIT TO MEASURE ALL DIGITAL TV IMPRESSIONS .......................................................... 8
3. NOTT ALONE: IS OTT MAKING IT COOL AGAIN TO WATCH TV TOGETHER? .......................................................... 10
4. USING MACHINE LEARNING TO PREDICT FUTURE TV RATINGS IN AN EVOLVING MEDIA LANDSCAPE ........................................................................ 12

FEATURED PAPERS

Full-length articles that illustrate how Nielsen is thinking about some of the most important measurement challenges and opportunities in the industry today.

1. FROM THEORY TO COMMON PRACTICE: CONSUMER NEUROSCIENCE GOES MAINSTREAM ........................................................................... 15
2. USING SINGLE-SOURCE DATA TO MEASURE ADVERTISING EFFECTIVENESS .................................................................................... 24
3. USING CROWDSOURCING AND IMAGE PROCESSING TO AUTOMATE DATA COLLECTION ........................................................................ 32
Television is a very technical environment. To make sense of what content is being viewed, where, when, how and by whom, researchers around the world have long relied on sophisticated engineering solutions, and, for the most part, these solutions have answered the call beautifully. Today, however, the world of television is exploding with new viewing options—there’s not just more content, but more ways to watch it on one’s own schedule, and to have it curated to one’s personal tastes. Today’s connected consumer can truly choose to watch anytime, anywhere, and on the device of their choosing. With the flexibility of live, on demand or even binge viewing, consumers are now much more in control of their media consumption than ever before, and the technology to measure their behavior needs to evolve to keep pace.

Enter the Nano meter. It’s smaller, cheaper and faster to install in panelist homes and easier to maintain than existing metering devices. It’s also built from the ground up to be a wireless system that consumes very little energy and interfaces with a powerful cloud infrastructure. Those technical benefits are extremely important, not only because of the cost-savings involved, but because they’re expected to improve panel cooperation, boost data quality, and make it cost-effective to increase the size of television panels where and when needed—a crucial piece of the solution to address media fragmentation in the marketplace today and into the future.

In order to understand the extent of those improvements, let’s review first how television measurement is done today at Nielsen for traditional television, and in an advanced market like the U.S.

Traditional TV—also referred to as linear TV—is delivered to audiences (national or local) as a ‘linear’ experience: your favorite TV show gets delivered to you at a prescribed date and time, on a prescribed channel, and carries commercials from advertisers who have bought time during that broadcast hoping you’ll be receptive to their messages. Everyone who watches that show sees the same ads. To enable audience measurement in that type of environment, content is encoded at the source and decoded in a panel of 40,000 households that are carefully selected to represent the total viewing population.

The encoding process relies on hardware and software encoders that are installed at various distribution points of the media broadcast chain, such as broadcast networks, national cable networks and local TV stations, live or on demand. Nielsen has approximately 3,500 encoders in operation around the country. Those devices insert an inaudible audio ‘watermark’ into the content to help identify the originator and distributor, as well as the precise time of the broadcast.

Once they’ve agreed to participate in a panel, Nielsen households are visited by a field technician who installs a number of electronic devices to monitor every TV in the house. A ‘set meter’ is attached to each TV set to capture tuning information (what program is being watched), and a ‘people meter’ is attached to each TV set to capture viewing information (who is currently watching that program). During installation, the audio from the TV and other media devices attached to it (set-top box, DVD player, Roku, Apple TV, game console, etc.) is routed to the set meter so that we may determine the source of the content (which device is actively feeding content to the TV) and extract the Nielsen watermark from it. On the rare occasions when the watermark cannot be decoded, a backup system is in place to compute an audio ‘fingerprint’ of the signal on the fly (a unique identifier of what’s being played), and compare it to a battery of fingerprints captured at an independent media monitoring site nearby. Nielsen has nearly 1,000 such monitoring sites in operation around the country.
This innovative engineering architecture has earned Nielsen an Engineering Emmy Award, and has been the foundation of the gold-standard TV ratings business in the U.S. for many years now, but it’s still fundamentally a wired system. It has many separate parts that need to be installed, configured, tested and connected together, and that complexity has an impact on operational and maintenance costs. Current meters cost upwards of $500 and take an average of 60 minutes to install. Some installations are so difficult that they require multiple technician visits and time-consuming work to conceal wires and equipment. Some households find all of this to be too disruptive and when that happens, a new household with the same characteristics has to be recruited to fill the void, which adds dramatically to the cost of the system.

The Nano meter is the product of a careful analysis of all aspects of the current system: form design, pain points during installation and maintenance, communication, hardware and software architecture. By leveraging low energy computer engines spurred by Internet of Things (IoT) developments and integrated components, the Nano combines all measurement functions in a single compact box with a sleek modern design. In the Nano, the people meter and the set meter are integrated into a single device that can be powered right from a Smart TV’s USB port. Using Bluetooth and WiFi technology, it can communicate with other elements around the house—such as wearables, smartphones, or even a new breed of streaming meters developed by Nielsen to capture over-the-top (OTT) and broadband content delivery. It comes with remote management capabilities to help reduce future field technician visits. It provides the same industry-proven content identification techniques (watermarking and fingerprinting) in a tighter package, and has the capability to communicate with the back office via a dedicated cellular modem or the panelist’s own broadband connection—opening the door to real-time data collection.

The Nano meter is built for the future, but its benefits aren’t limited to advanced viewing environments or complex installation configurations: at $85 per unit and with an install time around 15 minutes per site, it’s expected to have a major impact on panel economics and cooperation rate (and therefore, ultimately, on data quality) once it rolls out to U.S. panels in 2017. We will review this engineering breakthrough in more detail in a full-length paper, and examine in particular how it performs against important industry benchmarks. We will also explore how the engineering team at Nielsen is developing innovative solutions to measure the world of digital television and dynamic ads.

It’s an exciting time to be an engineer in media research!
Video streaming technologies are changing how we watch television today. Think about how, not that long ago, we were still arranging our lives around TV schedules—not just for sports and news, but for other programs as well. We would fix dinner, get the laundry out of the way, and sit down for our weekly dose of Thursday night comedy.

Today, at a minimum, we have the option to catch up on the shows we’ve missed, but our liberation from the TV guide goes well beyond DVRs: we can now pay just for the channels and the shows we want to watch, and watch them when we want; we can subscribe to massive libraries of licensed and original content and watch entire series all at once; even the commercials we see are starting to be catered to our personal tastes; and, of course, while we can tunnel all of that exciting content to our connected television sets, we can also watch it on the go on our laptops, tablets and smartphones.

These advances are based on digital technologies, and while they’re empowering for the consumer, they can be a challenge for media researchers. There are many more devices to measure and, therefore, more engineering solutions that need to be developed. But most fundamentally, it’s the proliferation of channels and on-demand content that’s creating the highest hurdle, because it’s stretching the capabilities of the panel-based measurement solutions that have been the backbone of media research all these years. Simply put, even with a panel of 40,000 households and some 100,000 people (which is the current size of the national television panel that Nielsen operates in the U.S.), there are television shows today whose audience is too small to show up reliably in the ratings data.

One part of the solution is to increase the size of the panels we use in the industry for television measurement. Nielsen has major engineering developments underway that will make that possible. But there’s no question that the solution also needs to involve engineering developments that can capture and make sense of TV usage outside those panels—something we refer to as ‘census data collection.’

How do we accomplish this? Nielsen engineers have created a software library—called the software development kit, or SDK—to install on all content owner apps, aggregator apps and browser pages that render television content. It’s not a physical meter, but rather a collection of software plug-ins that can be embedded in existing applications and turned on to capture impressions. If the consumer is watching content on a laptop, then the SDK is embedded in the browser. If the consumer is watching on an iPhone, then the SDK is embedded as an iOS app. If the consumer is watching on a Roku box, then the SDK is embedded in the Roku media player. In order to minimize the number of independent SDK implementations, we have also partnered with Adobe to provide an implementation option of their SDK that has our measurement built in. This is typically offered to mutual clients of Nielsen and Adobe.

The SDK needs to be flexible enough to recognize a variety of unique content identifiers. In the case where the content originates from linear TV, it already contains a Nielsen watermark—an inaudible code that’s inserted in the audio signal by one of thousands of encoders deployed at content distribution facilities around the country. That watermark is the foundation of the current TV ratings system, but it can be

SNAPSHOT #2

KIT AND CABOODLE: A SOFTWARE DEVELOPMENT KIT TO MEASURE ALL DIGITAL TV IMPRESSIONS

BY ARUN RAMASWAMY, Chief Engineer, Nielsen

See “Smaller Cheaper Stronger: the Nano Meter” in this issue
used as well if that same content is distributed and consumed as digital content. It just needs to be transcoded first because audio is difficult to get to on digital platforms, for security reasons. To address this issue, Nielsen created software that is embedded in most leading transcoders (those appliances that create the streaming content) to extract the watermark from the audio and insert it as metadata in the digital stream. This metadata tag, called ID3, is supported on most leading streaming formats and is much easier to extract from the streaming content.

In the case where the content is natively digital and doesn’t have a Nielsen watermark to begin with, we rely on metadata from the providers’ content management systems (CMS) to identify what’s being played: name of program, episode, genre, etc. We may also apply our own tags if a particular provider’s CMS is difficult to access. Tagging digital content with CMS tags is an important function not just for programs, but also for ads. Dynamic ad exchanges and programmatic technologies are revolutionizing the media industry. The ads a consumer sees with their television content might be the same as those associated with that content in the linear TV experience, but they might also change after a period of time or be an entirely different set of ads from the outset. Without a separate tag to identify those ads, we wouldn’t know what ads people were exposed to.

Once the identification mark is captured for content and ads, whether in the form of an ID3 tag or a CMS tag, it’s aggregated with other marks and transmitted to back office facilities for final processing and analysis. But there’s one piece that’s still missing before digital content and ad ratings may be computed: demographics.

Only a small portion of the census impressions collected by the SDK correspond to Nielsen panel members. We have detailed personal data for panel members (demographics, lifestyle data, technology ownership, etc.), but not for the millions of consumers sending us data from their laptops, smartphones and other connected devices. Through data partnerships with carefully selected third-party providers and using procedures that safeguard consumers’ privacy, we’re able to receive aggregated demographic data that is calibrated against our panel data. This allows us to put a face on all the faceless impressions, and finally combine those digital ratings with linear ratings to paint a complete picture of a show’s performance—its ‘total audience.’

We don’t live in a world anymore where one technology, or one delivery mechanism, completely dominates the entertainment industry. The options have multiplied, the viewer is in charge now, and that means that media researchers need to be quick to adapt to emerging technologies. Audience measurement has never been more complicated, but with the right architecture in place and a culture of engineering innovation, the industry is up to the challenge.
Co-viewing in television is the process of watching content alongside other people, typically members of the same household. Entertainment products are often consumed collectively, and television is no exception. In fact, watching television has traditionally been considered a social activity. But the digital age is starting to erode that premise: With more television content being watched every day on laptops, smartphones and tablets, it seems that watching TV is slowly becoming an individual pursuit.

Are over-the-top (OTT) devices capable of reversing that tide? OTT devices are now in 20% of U.S. households. They’re typically connected to a big screen TV in the home, and make it possible for consumers to watch TV content via dedicated apps from major TV networks and streaming services. They come with all the convenience and flexibility we’ve come to expect from digital entertainment (enormous video libraries, on-demand viewing, unlimited viewing for a fixed monthly fee) but on a big screen and in the comfort of our living rooms. Is OTT making it cool again to watch TV together?

The short answer is: Undeniably, yes, but rates vary by age, daypart, and other factors.

Studying co-viewing for OTT devices is important for a number of reasons: Program developers need to understand who exactly is watching their shows and if watching on an OTT device is likely to affect certain demographic groups more heavily than others; advertisers need to understand how co-viewing on an OTT device might affect how their ads are perceived; and sociologists are eager to understand shared viewing patterns and some of the new dynamics driving our social interactions.

In early 2015, Nielsen partnered with Roku, a leading OTT device provider, to deliver the first ever audience measurement service on TV-connected devices. To facilitate census-based measurement, Nielsen embedded a piece of software (called a software development kit, or SDK) directly into the apps of the OTT provider to track ad impressions. Since that data came from devices and not panel homes, we didn’t know who watched the content. To solve this problem, we implemented two crucial steps: First, we used a third-party data provider to identify the household and person-level characteristics (e.g., income, age, gender) associated with the OTT device and calibrated that data against our National People Meter (NPM) panel; then, we developed a model to predict which specific household members viewed each ad impression, based on historical NPM TV data stemming from television sets that were connected to an OTT device.

The launch of this OTT measurement service was a breakthrough in our understanding of OTT usage, and it continues to grow in terms of clients (supporting an increasing number of publishers and advertisers) and data volume.

\*See “Kit and Caboodle: a Software Development Kit to Measure all Digital TV Impressions” in this issue
(capturing millions of impressions every day). In 2016, we embarked on a co-viewing study using data collected from that OTT measurement service. This study involved analyzing a large volume of campaign data across a variety of sources: 18 million ad impressions from 15 advertising campaigns across programs representing more than two dozen genres.

We found that the overall co-viewing rate for OTT was 34%—lower than what it is for traditional broadcast TV (43%) but much higher than TV co-viewing on mobile devices (14%). We were also able to determine that OTT co-viewing was a non-random phenomenon—it varied based on age, for example. Kids (2-12 year olds) co-view the most: seven out of 10 of this age group co-view with at least one other person in their home. Among teens (ages 13-17), females were more likely to co-view than males (63% vs. 54%). For all other age groups, however, males and females co-view at a similar rate. We also found that OTT co-viewing is much more prevalent in primetime (44%) than during daytime (25%). We will expand on those findings in an upcoming full paper.

The initial results are consistent with what we know of co-viewing in traditional TV, but there are significant differences along demographic and technology lines, as we start to expand measurement across different OTT providers. As OTT penetration keeps rising, Nielsen is committed to including OTT devices in its digital ad ratings and continuing to design innovative measurement techniques that can keep pace with the consumer marketplace. We believe this is an excellent example of how panels and census-based data can be brought together to better understand modern viewing trends.
SNAPSHOT #4

USING MACHINE LEARNING TO PREDICT FUTURE TV RATINGS IN AN EVOLVING MEDIA LANDSCAPE

BY JINGSONG CUI, VP, Data Science
SCOTT SEREDAY, Manager, Data Science, Nielsen

Media companies and advertisers rely on TV ratings every day to measure the success of TV shows, verify that their audience size and composition are delivering against media-buy targets, and make-good in case the numbers come up short. From that point of view, TV ratings are metrics that measure the past, or at best the present, of TV viewing.

But media companies are also using ratings to predict the future. Ratings set expectations and affect programming decisions from one season to the next. They also help set advertising rates well in advance of when a campaign might actually air. In the U.S., for instance, TV networks sell the majority of their ad inventory for the season at the “upfront,” an event they organize only once a year (between March and May). This means that the rate for the ads you’re seeing on TV today might have been negotiated more than a year ago.

In order to predict what a show’s rating might be in three, six or 12 months, researchers are using forecasting models. Many of those models have been used for years with little or no modification. They’ve been successful at predicting ratings and have done a great job of supporting the exchange of billions of advertising dollars each year. But fast changes in the TV ecosystem are making it increasingly difficult to develop reliable models.

Consider the list of recent technology innovations in the media industry: Viewers are increasingly using their laptops, tablets and smartphones to watch content; streaming services like Netflix and Amazon Prime have reached mass-adoptions; new TV-connected devices are reshaping the big-screen experience. People are time-shifting, streaming and binge watching—they’re more in control of the media they consume than they’ve ever been. Their behavior is not only more complex, but more unpredictable as well.

At Nielsen, we have access to many data resources that measure how people consume media. Before adding digital TV data into the mix (as input as well as output of our forecasting models), we wanted to examine whether it was possible to first improve how we predicted ratings for traditional TV, using traditional TV data as our only source. Thanks to the Nielsen National People Meter, we have high-quality data that goes back many years, with consistent methodology and a robust panel of nationally-representative viewers.

We tapped into this rich data at a very detailed level to create new predictive models: Variables like historical Live+7 ratings (i.e., ratings that include live audiences as well as viewers up to seven days after the initial broadcast), C3 ratings (commercial ratings that include playback up to three days afterwards), HUT (the percentage of households using television at any point in time), reach, household ratings, demographic ratings, day of week, hour of day, and the identity of the network are some of the key pieces of information we used as input variables; and we capitalized on advanced machine learning and statistical algorithms (like ridge regression, random forest and gradient boosting) to identify relevant data relationships.
Working in cooperation with a client, we conducted a number of proof-of-concept studies to test and validate the models we created. We designed our models to predict future ratings at a granular level (hour-blocks for small demographic groups, like males ages 2-5 or females 65+), but we also rolled up those figures to the network level. In order to understand how our models performed against reality, we used a holdout period of two quarters to compare our forecasts as well as our client’s internal forecasts to true ratings data. For example, we accurately predicted an average Live+7 rating of 1.94 for persons 30-34 on Network A between 9 p.m. and 10 p.m. on Tuesdays during second-quarter 2015, based solely on historical data up to the first quarter of 2014. Predictions were very accurate at the network level, where we had a 99% R-squared (percentage of variance explained), but they were more difficult at the more granular hour-block daypart level, or for some of the smaller demographic groups. Even at the hour-block level though, our model’s R-squared still topped 95% and significantly outperformed a model that our client had been relying on up to that point. Across more than 2,000 day-time projections, our forecasts were 41% more accurate for R-squared and 16% more accurate for weighted absolute percentage error (WAPE)—two key measures in forecasting accuracy.

We’ll share more details about those proof-of-concept models and the tests we conducted in an upcoming paper. The key takeaway of this project is that we were able to convert big and noisy behavioral data into predictive modeling features, and do so in a very efficient (and automated) manner. But every decimal point of a rating point has enormous financial implications, and we need to keep pushing the envelope by adding new input variables (such as ad spend or program-specific data), building ways to quickly adapt to changes in programming packages and channel lineups, testing new forms of regression and classification algorithms, or even combining multiple promising models into one.

While this project focused on traditional TV, it’s interesting to note that the impact of digital data is reflected in changes in TV ratings in the historical data—and thus in our predictions as well. But this is an indirect measurement of a cumulative effect, and no substitute for a model that would focus specifically on over-the-top viewing, for instance, or viewing on a smartphone app. In addition to the next steps outlined above, the use of digital data will be an important element to improve our forecasts in the future.

In the end, we also need to recognize that each client has intimate knowledge of its programs, as well as a strong intuition about how those programs might be received in the future. That “human element” should not be ignored when we put together predictive models, and can be especially valuable when reacting to significant and unforeseen changes in the marketplace. A system that integrates rich data, powerful machine learning algorithms and domain expertise can achieve better results than either could accomplish on their own.
On a typical day, the average consumer may be exposed to thousands of commercial messages, yet many of them won’t succeed in breaking through the clutter, nor have any discernible impact on that individual’s attitudes or behavior.

What separates effective advertising from that which fails to deliver a return on investment? Early thinking was that advertising worked by communicating relevant facts about the benefits of a product or service. Most believed that a rational consumer could then evaluate the value proposition of competing offerings in order to arrive at some welfare-maximizing conscious choice relative to his or her well-defined needs and desires. Large-scale studies of advertising effectiveness now have provided a robust body of evidence strongly suggesting that this is not the case.

What is becoming clear from more recent research is that effective advertising succeeds in eliciting an emotional response from consumers. For example, in a 2009 study reported in the Journal of Advertising Research, Binet and Field reviewed more than 800 ad campaigns in the U.K., each with clearly stated business objectives and “hard” business outcomes (e.g., sales, market share, price sensitivity, profit). To their surprise, the data clearly suggested that the more that emotions were at the center of the campaign, the bigger the business impact, and that the most positive outcomes for...
advertising campaigns were often those that included little or no rational content at all. More recent converging evidence for this view has been reported by the consumer packaged goods industry, suggesting that advertising with emotional content is on average nine times as effective at driving sales than non-emotional advertising.

Measuring emotion in advertising is challenging, but it’s become imperative for brand managers to find tools that dig beneath consumers’ conscious responses to advertising, and instead measure their emotional responses. Thankfully, advances in neuroscience now are making this a reality. But as with every new tool, it’s only possible to interpret the results properly when a solid theoretical foundation is in place. Before presenting the tools of consumer neuroscience and their practical applications, let’s briefly review that theory.

THEORETICAL BACKGROUND

Prominent advertising and branding thought leaders, like Robert Heath, have promulgated a theory of “low attention” processing in advertising effectiveness. In this view, emotional advertising works at a non-conscious level, strengthening implicit associations to a brand that can influence subsequent purchase decisions. In contrast, advertising that requires a high level of attention and conscious thought can actually undermine brand relationships. Similarly, Byron Sharp, director of the Ehrenberg-Bass Institute for Marketing Science, argues in his book How Brands Grow (2010) that successful brands are those that enjoy a high level of “mental availability,” and that advertising can nurture such ease of access by reinforcing simple, consistent, distinctive, and easy to remember brand cues that elicit an automatic, instinctual emotional response.

The findings of superior effectiveness of emotional advertising are consistent with a broad body of academic research related to the role of emotions in decision-making—work that has emerged in the fields of psychology, behavioral economics, and neuroscience in recent decades. For example, beginning in the 1970s, the psychologists Amos Tversky and Daniel Kahneman initiated a program of investigation that convincingly questioned the notion that humans are rational decision makers. Their research found that framing a scenario in terms that differed in affective or emotional intensity had a significant impact on decision making, despite the equivalent outcomes of offered choices. In Thinking Fast and Slow, the Nobel laureate Kahneman builds on this early research and a wide variety of other related work to argue that human decision-making is the outcome of two relatively independent cognitive systems: a “slow” system that is more accessible to consciousness and is deliberate, effortful, and rational versus a “fast” system that is less accessible to consciousness and is more intuitive, automatic, dependent on associative memory, and emotional. He reviews a variety of evidence to suggest that influences from the less accessible “fast” emotional system can have a disproportionate impact on the casual decision-making that is characteristic of everyday judgments and choices.

Growing evidence from neurobiology also provides convergent support for the notion that emotional responses play important roles in routine decision-making. In this view, emotional responses are automatic, coordinated, brain and body reactions to events in the mental environment. The emotional system is always on and always working. While in general the environmental event stimulus could be internal (e.g., a thought or memory) or external (i.e., as experienced by one of our five senses), marketers and market researchers are mostly concerned with responses to external stimuli (e.g., brands, products and their related consumer touchpoints). Emotional responses work by tagging sensory information for relevance, signaling importance and directing attention, memory and decision-making resources in the brain that ultimately impact some future behavior. When an individual is confronted with a stimulus that elicits an emotional response, information about that response is manifested in the body and is stored as a “somatic marker” in the prefrontal cortex (and several other parts) of the brain. This view is expressed, for example, by neurologist Antonio Damasio in Descartes Error: Emotion, Reason, and the Human Brain. The prefrontal cortex is one of the most highly evolved parts of our brains, and it has many connections with areas of the brain involved with emotions and valuation of information. When an individual is later confronted with a future similar experience, relevant prior emotion-based somatic markers are accessed from memory centers and provide non-conscious feedback to help inform decisions.

In the context of consumer behavior, such somatic markers may become integrated with the constellation of associations in the brain that constitute the individual’s “brand knowledge.” In marketing, an emerging goal of communications is thus to better leverage such emotional responses to create and amplify “meaning” and trigger emotional “approach” (as opposed to “ignore” or “avoid”) motivations in consumers. Successful advertising is that

*Source: WARC event report, “Procter & Gamble Validates Emotional Marketing,” by Steven Whiteside, March 2015*
which can lead consumers toward lasting connections through memory associations with positive sentiment, and ultimately an endorsement or purchase of a brand, product or service.

This evolving understanding of the psychological and neurological bases of consumer decision-making presents a measurement conundrum. Traditional approaches to consumer insights have in part relied upon measurement of overt consumer purchase behavior in panels or at a point of sale, and in part on evaluating consumers’ self-reports about potential motivational drivers of such behavior through focus groups and surveys. However, limits to the reliability of self-report measures of behavioral drivers have long been a concern, as they elicit only consciously accessible information from consumers. Obviously, when asked, people can usually provide “explanations” for their behavior. But the accuracy of such reports is compromised inasmuch as people lack direct introspective access to many of their internal cognitive and emotional processes.

A large body of research has demonstrated a lack of reportable subjective awareness of mental drivers of judgments and decisions. This limitation of self-report methods would appear particularly acute when trying to understand emotional factors in decision making, since much of the influence of emotional response occurs during non-conscious processing. But if self-report has serious limitations as an approach to understanding consumer behavior, what are the alternatives?

THE EMERGENCE OF CONSUMER NEUROSCIENCE AND ITS METHODS

Advances in basic and applied research on human behavior and neural information processing are not just providing new insights into the drivers of consumer behavior. They also are yielding new neuroscience-based tools that can help design more powerful communication approaches. Over the last decade, a growing consensus has emerged that measurements of brain and body physiology as well as rapid implicit response methods are now able to provide valuable information for understanding consumer behavior. Innovative creative teams and brand marketers have been leveraging such advances to improve their odds of winning in an increasingly complex and cluttered media environment. It has been our experience that at this point, most of the top consumer-facing marketers in the world have at least experimented with such methods, and several key leaders have begun to employ them on a routine basis.

The growing use of such methods in the commercial sphere has evolved into the market research sub-discipline of consumer neuroscience, an approach that is yielding new insights for evaluating and optimizing the effectiveness of marketing communications. Neuroscience-based approaches to consumer insights hold promise to help fill the knowledge gaps left by more traditional market research methods. By measuring relatively automatic behaviors and capturing physiological responses, the limitations and biases encountered in self-report surveys and focus groups are partly circumvented. As a result, adoption is growing quickly and becoming more mainstream. One reflection of such growth is the fact that, over the past several years, marketing departments at prominent business schools have begun to add scholars with expertise in consumer neuroscience to their marketing department faculty, and to add related courses to their graduate and undergraduate curriculums. In parallel, content related to consumer neuroscience has also been playing an increasingly prominent role in the conferences of long-standing market research industry organizations such as the Advertising Research Foundation (ARF) and ESOMAR. The field has even spawned its own nascent professional society, the Neuromarketing Science & Business Association, which already has over 1,600 members from more than 90 countries.

Consumer neuroscience, both as a commercial market research activity and a field of academic research, employs a wide range of measurement methodologies that originate in traditional experimental psychology and biomedical research settings. Here, we briefly describe some of those tools. We have categorized them based on what they are measuring, and how directly (or indirectly) they are indexing brain activity:

1. Tools that measure observable behavior

A first set of tools measure various aspects of observable behavior rather than a physiological response. Because it is relatively simple and inexpensive to implement them, they are among the more commonly used techniques in the commercial world. They are included here as consumer neuroscience techniques because, unlike traditional articulated survey responses, these types of behaviors are rapid and can be immediately influenced by factors outside of conscious control. Examples include:
Implicit Response Testing. This technique can be used to try to understand information—specifically semantic associations or “feeling states”—that individuals are unable or unwilling to verbalize. Consumers are provided a simple stimulus such as a word or picture to react to, and precise measurement of the timing of the responses (in the form of a rapid key press or other simple finger movement) can show non-conscious associations with brands and products. This technique has been widely used for understanding branding, brand and message positioning, ad messaging and responses to packaging, but is limited by the need to take measurement after exposure to the test stimulus.

Facial Coding. With advances in camera technology and computer vision methods, this methodology has recently been automated with near real-time software measuring the emotional facial expressions of consumers as they experience marketing content. Involuntary facial movements occur when a person experiences emotional states, and facial expressions are frequently viewed as near universal across cultures. The technique can be a useful diagnostic tool to understand whether a stimulus has elicited a specific facial expression (e.g., a positive smile or a negative frown) and is increasingly used as an aid to evaluate ad effectiveness. It must be kept in mind, however, that facial expressions evolved to communicate our feeling states in a social context and therefore occur at relatively low levels in the context of the passive media upon which the vast majority of marketing communications occur (i.e., television, internet, out of home signage).

Eye Tracking. Infrared cameras are used in eye tracking to monitor the direction of a consumer’s gaze and eye movements and pinpoint where they are looking—whether on screen, on a store shelf or elsewhere. Screen-based studies often employ fixed-position cameras for eye tracking whereas ambulatory studies (e.g., in-store shopper studies) and studies of mobile devices usually employ head-mounted eye-tracking equipment. While eye tracking in isolation can provide specific feedback about whether consumers are noticing specific elements of advertising creative, packaging, or product placement in the way marketers intended, the approach often is used in combination with other technologies. The technique provides no information about the nature of the brain’s emotional or memory response to the object of gaze.

2. Tools that measure some aspect of autonomic nervous system activity

Other common methods in consumer neuroscience use a variety of techniques to measure aspects of autonomic nervous system (ANS) activity. The ANS serves to regulate peripheral body organs in a largely automatic and non-conscious fashion, and thus includes control of heart rate, respiration, sweating and salivation, digestion, and a variety of other functions. Such measures are often collectively referred to as “biometrics,” and they have been employed for decades in psychophysiology laboratories to observe changes in individuals in response to stimuli that are emotionally arousing, or tasks that require variations in mental effort.

Pupillometry. One long-standing biometric technique is to measure momentary changes in the diameter of the pupil of the eye in response to some brand communication, often performed in conjunction with eye-tracking studies. Pupil diameter is under ANS control, and small dilations in pupil diameter occur in response to presentations of increasing cognitive load or emotionally arousing stimuli. One challenge in pupillometry studies is that it is often difficult to differentiate what is causal, as the size of pupil dilation in response to changes in cognitive or emotional inputs is small relative to changes associated with variations in light intensity.

Heart Rate. An additional commonly employed biometric technique is the measurement of changes in the frequency and variability of heart beats. Heart rate can provide a robust measure of variations in arousal due to exposure to a stimulus of interest, or engagement in some task of interest. Inasmuch as variations in heart rate (and respiration) can be strongly affected by physical exertion, relying on it as a measure of emotional arousal in isolation can be difficult. However, with the trend toward wearable devices, heart rate is increasingly collected outside of controlled lab spaces and used in environments such as in-store and in-home studies.

Skin Conductance. Probably the most commonly employed biometric technique in consumer neuroscience is the measurement of electrodermal changes, or “galvanic skin response” (GSR). Skin conductance is a measure of electrical conduction from the ANS—typically measured on the palms or fingertips due to high concentration of special perspiration cells. When experiencing a fight-or-flight response, stress, emotional engagement, or other factors, skin conductance increases.
from phasic changes in specialized glands and can be measured as an indication of arousal. As for heart rate, the trend toward wearable devices is making it possible today to collect skin conductance outside of controlled lab spaces.

3. Tools that measure changes in brain physiology

The most detailed and informative pictures of the consumers’ internal brain response to marketing materials are those provided by measurements of the central nervous system (CNS), comprised of the brain and spinal cord. And, in particular, by methods that directly index changes in brain physiology. Although a variety of methods have been developed for measuring CNS activity, two in particular have played an important role in the field of consumer neuroscience:

- **Functional Magnetic Resonance Imaging (fMRI).** Magnetic resonance imaging is a technique from neuroradiology that uses strong magnetic fields produced by large magnetoms, radio waves, and variable field gradients to non-invasively create highly detailed images of internal anatomy. Initially introduced in the 1990s, this technique is used to identify the level of oxygenation of blood in different regions as it changes over time. Active areas of the brain require more oxygen for metabolic purposes; by measuring localized changes in the oxygen level of the blood, inferences can be made about the relative functional activity of specific areas of the brain. By registering such activation maps with maps or atlases of structural anatomy, it is possible to make inferences of relative activation of specific brain structures. Studies using fMRI methods have been used extensively in academic consumer neuroscience to understand mechanisms of memory, valuation, reward, and decision-making, among other interesting topics. Several factors have limited more routine commercial application of the technology, including the size and complexity of the equipment installations, the per-participant operational expense of the equipment, and the time needed to execute fMRI experiments and analyses. Furthermore, participants must lie very still on a horizontal gurney while their head is surrounded by the magnetom, which can cause claustrophobia and limit the types of behaviors participants may engage in.

At a more basic level, measuring brain blood oxygenation levels is an indirect measure of brain activity. Brain cells or neurons communicate through electrochemical signaling that occurs on a millisecond timescale. However, with fMRI, several seconds may elapse before enough brain activity occurs to induce an observable change in regional brain blood oxygenation. Most studies that have utilized fMRI to study materials such as television commercials produce a single brain map to characterize the ad as a whole. This, in turn, makes it difficult to gauge scene-level impacts over the course of the ad. Despite excellent spatial resolution, fMRI provides poor temporal resolution, making it difficult to pinpoint which second of an advertisement renders a specific response.

- **Electroencephalography (EEG).** An alternative approach to assessing variations in CNS activity is the measurement of EEG (or “brain waves”), a technique that has now been widely adopted for routine commercial use. Typically observed at a distance from the brain by placing sensors on the scalp, EEG measures the mass effect of rhythmic current flowing between brain cells. EEG is typically sampled at a rate of hundreds of times per second. Abnormalities in the expected patterns of EEG rhythms are the clinical “gold standard” for diagnosing epilepsy, and changes in EEG patterns are commonly used for identifying levels of alertness during surgery and for understanding stages of sleep. Task-related changes in the EEG occur on a sub-second time scale, and thus brain responses to marketing materials can be captured with a high degree of temporal precision. However, EEG is not a true 3D imaging technique, and, as a result, pinpointing the sources of the signals to particular brain structures is generally not possible when the signal is recorded from the scalp.

While the field of consumer neuroscience is relatively new, studies of how EEG changes in response to mental stimulation have been conducted for over 90 years, and measurements of EEG in the context of advertising research have been conducted for more than 30 years. Most studies that use EEG methods to analyze advertising content use one or more spectral features of the ongoing EEG to determine whether that content requires effortful attention, is emotionally engaging, and is memorable. The corresponding metrics are benchmarked against an extensive library of basic and applied research literature.

For example, the shifting of attention in response to environmental cues, or the deliberate focusing of
attention in response to manipulations of task demands, are reliably associated with a reduction in the amplitude of certain types of EEG oscillations. Conversely, increases in the amplitude of those same oscillations are often seen during lapses of attention and boredom. There is also an extensive literature using EEG metrics to index emotional motivation or engagement. For the past 20 years, we have known that asymmetries of EEG indices of activation of the left versus right prefrontal cortices are associated with emotional experience. In particular, EEG indices of relatively greater left-hemisphere activation are associated with approach motivation (or being drawn towards a stimulus), whereas relatively greater right-hemisphere activation may be associated with avoidance motivation (that is, the withdrawal of engagement with a stimulus). Accordingly, frontal EEG asymmetry metrics of approach motivation have been shown to increase with arousing static images and engaging video advertising content. Such measures have also been used to predict virtual purchase decisions in laboratory settings.

The benefits of EEG are not limited to emotions: With respect to memory activation, changes in parameters of EEG during the viewing of television commercials have been shown to be correlated with increased likelihood of post-viewing recognition or recall of individual scenes, or brand and product information. Activating memory during marketing communications enables learning by creating and reinforcing connections to existing brand or product representations, so that the information can be utilized in future interactions, and it drives behavior such as future purchasing decisions. With measurements that can be made at a sub-second level, these dimensions of brain response can provide highly insightful scene-level diagnostics to help guide creative development.

The technical challenges and equipment expenses associated with EEG are relatively modest compared to those for fMRI. However, significant care and expertise are still required to draw meaningful insights. Foremost, EEG signals can be readily contaminated by non-brain and non-physiological artifacts from a variety of sources. To obtain the most reliable results, most routine testing must be conducted in highly-controlled laboratory environments, and sophisticated signal-processing methods must be employed to eliminate residual contaminants from recordings. Moreover, to make meaningful inferences from the EEG, the signals must generally be recorded from sensor arrays that cover the entire scalp, and applications are limited to static imagery (e.g., print ads or packaging) or short-form video (e.g., video advertising). Measurements obtained with highly-reduced sets of electrodes, or with equipment that is not medical-grade, or that are performed by personnel without significant training in the associated methodologies, have a high likelihood of being unreliable.

Broadly, the technologies used in consumer neuroscience can help explain and interpret brand performance more effectively than ever before. To date, these tools have perhaps been most extensively applied to the evaluation and optimization of early and late stage advertising. They have been widely used to inform package design and shelf assortments, as well as other aspects of the retail experience: point-of-sale materials, design of merchandising displays and aisles, pricing, and online experiences. In our own laboratories, neuroscience-based tools frequently have been used to evaluate product experiences, new product concepts and designs, and brand and product positioning. Finally, neuroscience-based tools are now starting to be used more frequently to explore the impact of full-length content experiences—whether in video format, print, or as web-based entertainment.

FROM LABORATORY OBSERVATIONS TO MARKETPLACE DYNAMICS

How do measurements made with such tools relate to broader-scale, real-world behavior? Can findings from small sample studies, typically performed in rarefied laboratory environments, predict how consumers might perform in the marketplace?

There are a growing number of studies in which small, laboratory-based samples have been used to predict population-level activity in the market. For example, using fMRI to measure fluctuations in regional cerebral blood oxygenation, Falk and colleagues (then at UCLA, now at the University of Pennsylvania) found that call-center volume for different public health direct-response advertising campaigns could be predicted by measures of brain reactivity in a small “neural focus group” of participants exposed to public service advertisements (PSAs). Interestingly, while measures of changes in activity from different brain regions—including the ventromedial prefrontal cortex—were found to be significant predictors of call center volume, the participants’ subjective ratings of the relative persuasiveness of different pieces of copy were not significantly related to the outcome.
University reported results from a study where they examined the degree to which laboratory studies of neural measures were associated with below average sales lift. In addition to advertising effectiveness, researchers have also conducted laboratory measurements of changes in emotional response and memory activation in response to exposure to television commercials. In particular, we’ve been able to model the sales lift associated with hundreds of television advertisements for which EEG-based measures of ad performance had been obtained in the laboratory, and found a close correlation between the two. That is, other factors being equal, creative executions that are significantly more neurologically engaging tend to yield above-average in-market sales lift for their brand. Specifically, creative executions that were above-average in the degree to which they engaged viewers were responsible for approximately 25% greater sales lift, whereas ads that scored below average on the EEG measures were associated with below average sales lift.  

In one of the largest studies of this nature completed to date, researchers at Temple University collaborated with the ARF in a study sponsored by large advertisers and media companies to investigate the relationship between a wide variety of consumer neuroscience technologies and in-market sales lift, as estimated by market mix modeling. In that study, an area of the brain called the ventral striatum (typically associated with emotional or behavioral reward) was the strongest predictor of real-world, market-level response to the advertising tested. The team at Temple University also partnered with a team of researchers led by one of the authors (Marcy, then at Innerscope Research) on a Super Bowl study that combined biometric responses with fMRI results. Results showed that ads that had very high levels of emotional response as measured by the biometrics also showed increased activity in the ventral striatum, as well as in other important emotional and memory centers—such as the prefrontal cortex, amygdala and hippocampus. 

At Nielsen, in the context of collaborative client work, we have also conducted laboratory measurements of changes in emotional response and memory activation in response to exposure to television commercials. In particular, we’ve been able to model the sales lift associated with hundreds of television advertisements for which EEG-based measures of ad performance had been obtained in the laboratory, and found a close correlation between the two. That is, other factors being equal, creative executions that are significantly more neurologically engaging tend to yield above-average in-market sales lift for their brand. Specifically, creative executions that were above-average in the degree to which they engaged viewers were responsible for approximately 25% greater sales lift, whereas ads that scored below average on the EEG measures were associated with below average sales lift.  

In addition to advertising effectiveness, researchers have examined the degree to which laboratory studies of neural responses to entertainment content can generalize to population-level behavior. For example, several years ago, a group of researchers led by Gregory Berns at Emory University reported results from a study where they measured the fMRI response of a small group of participants who were scanned while they listened to 15-second clips of music. Ratings of subjective liking of those clips were not found to be linked to the sales potential of the music. However, neural measurements from regions of the brain associated with reward processing were found to be significantly correlated with the subsequent overall cultural appeal and commercial success of those songs (as measured by Nielsen Soundscan). 

In another study recently reported in Nature Communications, a group of researchers using EEG-based methods examined similarities between lab participants who were asked to watch an episode of primetime television programming. They reported that between-subject similarities in EEG patterns were significant predictors of both variations in population-level program viewership (as measured by fluctuations in Nielsen TV ratings data), as well as variations in population-level Twitter activity related to that program (as measured by Nielsen Social). Our own independent research has demonstrated similar predictions of Twitter activity from EEG-based measures of high-engagement segments of television programming across a range of programming genres. 

These are just some of the recent industry initiatives that provide strong support for the notion that laboratory measurements of consumers’ responses obtained using neuroscience-based techniques can be robust predictors of population-level marketplace dynamics. 

INTEGRATION OF METHODS

Fueled by early success stories, the discipline of consumer neuroscience developed rapidly over the past decade. As exciting new techniques were being developed or adapted from medical research to the commercial sphere, many neuroscientists and other research practitioners saw a chance to put them to use to answer a wide variety of marketing questions. In their enthusiasm, they sometimes failed to recognize that each of the various techniques in consumer neuroscience research has limitations as well as strengths, and that the most appropriate tool to use will be in large part defined by the nature of the problem to be solved. 

By applying these tools too broadly, or by relying on one technology to answer too many questions, researchers...
sometimes ended up oversimplifying problems and overpromising solutions. That is, much like the old parable about a group of blind men and an elephant, the different measurement methods commonly used in consumer neuroscience research can each provide useful insights into how audiences respond to marketing messages. But no one approach provides a complete picture of the nature of the beast.

In order to circumvent the limitations imposed by individual measurement methods, we have been experimenting with a more holistic approach—one that combines a variety of methods to evaluate marketing communications. Our initial focus for this integrated approach has been on video advertising. Video advertising remains a powerful way for marketers to reach large audiences and to drive ROI. While the landscape for video advertising has changed dramatically in recent years (especially with digital providers creating avenues beyond traditional television), one truism remains: Executed well, video advertising is one of the most trusted advertising formats around, in both traditional and digital media; Executed poorly, video ads just contribute to a cluttered media landscape.

In an effort to improve consumer insights for video advertising and help advertisers increase ROI, Nielsen has been developing breakthrough methods for combining neurometric, biometric, eye-tracking, facial-coding, and self-reports into one comprehensive assessment tool. This integrated measurement approach, called Video Ad Explorer, recently has been implemented in Nielsen’s global network of neuroscience laboratories, and is becoming the “new normal” for applications like video advertising. The quality of the diagnostics is beyond anything we had been able to achieve before, and the initial response from clients from a variety of advertising verticals has been very encouraging.

To illustrate this integrative approach, let’s outline how we applied it to a particular public service advertisement. This was part of a study conducted recently in collaboration with the Ad Council. For this project, we applied a full suite of diagnostic tools to past advertisements from the Ad Council’s “Fatherhood Involvement” campaign, a long-standing and very successful campaign that seeks to inspire and support men in their commitment to responsible fatherhood.

The example case is from a 30-second PSA entitled “Cheerleader.” This was a humorous ad showing a father running through his grade-school daughter’s cheerleading routine to help her to practice. While participants (all fathers) viewed the ad, a variety of measurements were made, including measurements of central (whole-head EEG recordings) and autonomic (GSR and heartrate) nervous system responsivity. Measurements of overt behavior including eye tracking and facial coding were also recorded, and participants also provided self-reports about the ad.

The PSA begins with an older woman sitting by herself in an apartment, scowling as she hears loud noises from outside. The camera pans outside to reveal a man enthusiastically engaged in a cheerleader song-and-dance, and then pans out more to reveal his small daughter following her father’s example. After the midpoint, a voiceover intones “the smallest moments can have the biggest impact on a child’s life.” The pair repeats the cheer and the voiceover segues to a call-to-action that invites the viewer to call a number or visit a website for parenting tips and other resources.

While subjective reports, in the form of a questionnaire, indicated that viewers enjoyed the ad, the neuroscience measures of viewer response painted a more complex picture of engagement (see Fig. 1). The biometric engagement trace revealed peaks of autonomic arousal during the reveal of the cheerleading father and daughter, and for the closing

---

2The Ad Council is a US-based non-profit organization that marshals resources and volunteer talent from advertising and media to create and deliver PSA campaigns
3The ad can be seen here: [https://www.youtube.com/watch?v=hTl3V0xV2U](https://www.youtube.com/watch?v=hTl3V0xV2U)
sequence. The EEG engagement trace (computed from EEG-based measures of emotional motivation, memorability, and attentiveness) showed a much more complex series of peaks while the viewers built a mental model of the ad narrative in response to momentary scene-level changes. Facial coding further suggested that the participants first mirrored negative emotion from the woman in the apartment, were then surprised when the father/cheerleader was revealed, and were amused/happy during the closing sequence. Eye tracking (not included in the figure above) provided further diagnostics by revealing competition between the actors and other key information (such as the dial-in number) during the closing sequence—suggesting a clear opportunity for optimization.

Differences between methods were of particular interest. The biometric results identified that there were two non-conscious focal points in the ad. The faster changes in the EEG provided a more granular basis to recommend scene-level adjustments. And the directionality of emotional expression provided by the facial-coding results allowed us to improve our interpretation of the EEG and biometric responses. However, it is important to also note that we recorded responses in the EEG trace in the absence of systematic changes in facial expressions, suggesting that facial coding in isolation would have missed significant neural changes. The level of details and insights gained from this multi-layer analysis made a big difference for the client and the agency that made the ad.

Collectively, what is emerging from such integrated results is the fact that they provide different and unique information about the patterns of consumer response associated with viewing video content. Taken together, the techniques of consumer neuroscience are helping overcome the shortcomings of earlier frameworks, and offer a remarkably accurate and detailed “read” of whatever combination of features a marketer wants to test.

FROM THE HALLS OF ACADEMIA TO THE CMO’S DESKTOP

It’s a real breakthrough: From the halls of academia to the CMO’s desktop, neuroscience-derived measurements are providing new insights into the how and why of effective marketing communications. And from video ads to in-store displays, from product packaging to emerging forms of marketing outreach, consumer neuroscience’s much-improved diagnostics capabilities are rapidly making it an essential part of the creative process.

Does it mean that we have all the answers now? Of course not. Human beings are complex. We don’t necessarily respond to the most obvious stimulus in an advertising message. We sometimes avoid the things we like, and seek out the things we don’t like. The way we consume content is changing all the time, and competition for the hearts and minds of consumers is stronger than ever. We watch more and more content on small-screen mobile platforms, and we multi-task more and more. Our brain states in these situations are unlikely to be the same as when we’re watching content on a big screen, or sitting down by ourselves in the comfort of our living rooms. Neuroscience tools need to be fine-tuned to capture consumers’ reactions when they’re on-the-go or being distracted. And more and more advertising campaigns are multi-platform efforts where it can be difficult to tease out the effect that each platform contributes to the overall impact of the campaign.

But progress is being made. In the coming years, a more comprehensive theoretical understanding of media consumption will emerge, and we will roll out new measurement tools and techniques to address new marketing challenges. Consumer neuroscience is not just a powerful research discipline today: it will play an integral part in defining the measurement solutions of the future.
How do we know if advertising works? Is there a way to directly measure the in-store sales results of an advertising campaign? Scholars and marketers have grappled with measuring advertising effectiveness for decades. The long-standing joke is that, as John Wanamaker is said to have remarked more than a century ago, “half the money we spend on advertising is wasted—we just don’t know which half.” Thankfully, we’re long past that stage in our understanding of how advertising works, but there’s no question that measuring advertising effectiveness and optimizing performance remain a challenge for advertisers and media companies alike.

In recent years, massive shopper databases have allowed the industry to develop exciting new research methods. In particular, researchers at Nielsen Catalina Solutions (NCS) have introduced a breakthrough methodology based on single-source datasets to measure the direct effect that an ad campaign has on sales. These ‘sales effect studies’ offer a powerful solution to measure changes in purchase behavior among households that saw an ad campaign, and compare it to the behavior of similar households that were not exposed to it. This paper dives into the details of how those studies are put together.
THE BENEFITS OF SALES EFFECT STUDIES

Whereas traditional ad effectiveness studies relied on separate panels and datasets to collect ad exposure data and sales data, sales effect studies are based on a single-source dataset—that is, ad exposure and sales data are captured for the same households. Before single-source, researchers could only attempt to draw conclusions through correlations between disparate data sources, and changes in spending for a product could only be evaluated at an aggregate market level. Such an analysis was often complicated by external factors, and it was difficult to identify what particular creative execution or placement was driving sales—if at all. The advent of single-source data has made it possible to control for external factors, and attribute changes in purchase behavior much more accurately to advertising seen by the household.

These new studies provide value to a broad spectrum of companies. For advertisers, they are a critical piece of a holistic approach to planning, activating and measuring campaigns. With the insights gained from evaluating a campaign, they can allocate budget and impressions to the most effective media and improve returns on future advertising spend. For media companies, on the other hand, sales effect studies demonstrate accountability to advertisers and help make a case for increased spend in the future.

Advertising campaigns are run with a number of objectives in mind. At the most basic level, they aim to increase sales of a brand—both in the short term and the long term. However, more detailed objectives vary widely. Ad campaigns may seek to convince buyers to try out a new product, or bring lapsed buyers back to an established brand. They could be trying to shift share from competing brands, or increase spend among existing brand buyers. Single-source data makes it possible to quantify changes in purchase behavior in enough detail not only to calculate incremental spend, but to answer many other questions as well: Did a large number of new households try the product? Which competitors was share taken from? Which customer segments are most responsive to advertising, and are my existing customers buying my product more regularly? All of these insights can be used to uncover the brand’s real strengths and plan more effective campaigns in the future.

NCS uses sales effect studies to evaluate campaigns on a number of channels, including digital, mobile, TV, radio, magazines and direct marketing using a client’s proprietary data. Our single-source dataset contains over two years of purchase data, spanning more than 44,000 brands across hundreds of categories. The purchase data of the single-source dataset comes from the store loyalty cards of 90 million unduplicated households, a subset of Catalina’s data warehouse. Catalina’s retailers include major grocery stores and drugstore chains, and the data is commingled for a holistic view of spending for each household. Product categories represented in the data include food, beverage, baby products, pet food, and general merchandise, just to name a few. To date, studies have been delivered for thousands of campaigns, resulting in a series of advertising benchmarks that allow advertisers to compare campaign performance to industry standards.

Let’s examine how those studies are conducted.

A SALES EFFECT STUDY TYPICALLY INVOLVES THE FOLLOWING STEPS:

- TAGGING THE CAMPAIGN AND SELECTING HOUSEHOLDS
- APPLYING ALL-OUTLET ADJUSTMENTS TO THE PURCHASE DATA
- MATCHING TEST AND CONTROL HOUSEHOLDS
- MEASURING HOUSEHOLD-LEVEL INCREMENTAL SALES
- IDENTIFYING THE DRIVERS OF INCREMENTAL SALES
- CALCULATING REACH AND RETURN

TAGGING THE CAMPAIGN AND SELECTING HOUSEHOLDS

When a sales effect measurement is ordered, tags are embedded in all creative units that are part of the campaign, in order to identify not only which households are exposed, but also which specific version of the creative execution
they were exposed to. This is considered the “media exposure” element of the single-source dataset. The media types supported for tagging include addressable TV, digital, and mobile. Radio campaigns are tracked with codes, and magazine campaign exposures can be determined based on the households that subscribe to the corresponding publisher as well as secondary (or pass-along) readership. Linear TV data is collected from Nielsen panels as well as set-top boxes.

Online publisher partners, such as Yahoo or AOL, use first-party cookies to link exposures to a unique household. This is called a “direct match,” and tagging is not required in those situations because the publisher already has all the necessary linking information. With third-party cookies, on the other hand, a third-party company is required to link exposure to households and overlay demographic information. This is a technique that’s typically used for programmatic ads, and referred to as an “indirect match” or “cookie match.” In either case, no personally identifiable information is attached to any household. To protect the privacy of all households represented in the data, each household is simply represented by an anonymized ID number.

To be included in a sales effect study, households must meet a pre-period “purchase static” to ensure that they are actively using the loyalty cards that form the basis of the purchase data used in the study (or, for Homescan households, that they are actively scanning their purchases). An additional static is applied for the post-period. The second static maintains, as closely as possible, the same ratio as the pre-period static. For example, if a study has a pre-period of 4 quarters and a post-period of 2 quarters, and the pre-period static requires that a household purchase a particular product in two of those four quarters in order to qualify (e.g., a ratio of 1/2), then that household would need to register a purchase in at least one of the two post-period quarters (e.g., the same ratio of 1/2) in order to be counted.

The purpose of the statics is to retain the greatest possible percentage of dollars spent while excluding households where large amounts of data are missing. They ensure that households with no spend on a particular product or category truly didn’t purchase it (in statistics, those are referred to as “true negatives”). Statics are necessary to balance the need to protect data quality with the need to retain enough households to offer a robust basis for analysis. They’re critically important and an ongoing topic of investigation for the research community. We’re currently using different statics for different product departments (or product categories), as shown in the following table:

<table>
<thead>
<tr>
<th>STATIC</th>
<th>DEPARTMENTS INCLUDED</th>
<th>MINIMUM PURCHASE THRESHOLD</th>
<th>MINIMUM PURCHASE FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOOD</td>
<td>BAKERY, DAIRY, DELI, FROZEN FOODS, GROCERY, MEAT, PRODUCE</td>
<td>$30</td>
<td>3/4 QUARTERS</td>
</tr>
<tr>
<td>HEALTH CARE</td>
<td>HEALTH CARE</td>
<td>$5</td>
<td>2/4 QUARTERS</td>
</tr>
<tr>
<td>PERSONAL &amp; BEAUTY</td>
<td>BEAUTY CARE, PERSONAL CARE</td>
<td>$5</td>
<td>2/4 QUARTERS</td>
</tr>
<tr>
<td>NON-FOOD</td>
<td>GENERAL MERCHANDISE, HOUSEHOLD CARE, PET CARE</td>
<td>$10</td>
<td>3/4 QUARTERS</td>
</tr>
</tbody>
</table>

¹NCS is committed to protecting the privacy of all households represented in its data. A dedicated privacy council is constantly reviewing and revising NCS protocols to ensure that they meet industry standards. As the possibilities of single-source data are developed further, maintaining security will remain a central concern.
Sample sizes can vary widely, depending on the length of the campaign and the size of the product category. For a Homescan study, a typical sample size is a few thousand households. Digital studies typically produce the largest test groups. In one extreme case, a multi-month campaign for the orange juice category had a sample size of approximately two million households. The only sample size constraint that must be met is a minimum buyer count for the item or items being studied. Without a robust buyer count, the difference in purchasing between test and control groups would not be statistically significant. For frequent shopper studies, the buyer count is generally larger, but here too it depends on the product category. The instant breakfast study had a bit more than 4,000 buyers. In contrast, the household item study had over 100,000 buyers.

How do we identify the products that are purchased by the households in the study? All items and categories in sales effect studies are defined using classifications already existing in NCS data, or using custom UPC lists provided by clients. Precise product definitions allow a client to assess a campaign’s impact on the purchase of a particular subset of the brand, and category definitions make it possible to calculate how much share the brand has captured from competitors within a particular grouping of products. UPC lists that are both comprehensive and up-to-date are critical to get accurate results.

APPLYING ALL-OUTLET ADJUSTMENTS TO THE PURCHASE DATA

Catalina’s retail partners include major drugstore chains and grocery stores. These retailers add up to a very large frequent shopper dataset (FSD), but it’s not an exhaustive list of all outlets where a consumer might purchase goods and services. Thankfully, both the Nielsen Homescan panel and Nielsen’s Retail Measurement Services (RMS)\(^1\) can help. Intentionally selected to be nationally representative in terms of key demographics, the Homescan panel provides a comprehensive look into household purchasing. It can be used as the sole source of purchase data for certain sales effect studies, but we’re also using it to calibrate purchase data from the FSD and estimate where else consumers might have purchased the products in the study. This process, called all-outlet adjustment (AOA), leverages the overlap between Homescan and Catalina’s FSD. It works by estimating the proportion of Homescan buyers who shop at FSD retailers, calculating how much they spend inside and outside FSD retailers, and scaling dollar differences accordingly. We use the RMS data in a similar way to calibrate sales projections. Via the AOA process, we can account for sales beyond FSD outlets, and correctly identify households that did or did not make any purchase anywhere within the product category during the time period covered by the study.

A recent study for an instant breakfast brand was run using FSD data and showed that households in the test group spent an average of $2.04 on the product in the post-period. However, with the all-outlet adjustment, that average dropped to $0.40. This reflects the fact that the penetration of the instant breakfast category is limited. Few households purchase products like powdered mixes or bottled ready-to-drink shakes. Additionally, households that purchase those products do so primarily at grocery stores, which are well-represented in frequent shopper data. As a result, we need limited adjustment to account for purchases of the brand at retailers not captured by FSD loyalty cards.

In contrast, another recent study looked at a common household product. From the raw FSD data, the test group spent an average of $3.23 on the brand being studied. But after the all-outlet adjustment, the average surged to $8.20. This increase reflects both the fact that the category is purchased by a large proportion of households (meaning that no downward adjustment is needed), and the fact that the product is commonly purchased at retailers other than the ones in the frequent shopper data. In both cases, using the adjusted value to calculate the incremental spending driven by the campaign results in a more accurate assessment of its true impact.

The all-outlet adjustment has also a substantial effect on product penetration. For the household product, frequent shopper data alone indicated that it had a penetration of approximately 23% among the test group. However, after the all-outlet adjustment, the penetration jumped to nearly 43%. This adjustment captures the large number of consumers who purchase the product exclusively at non-FSD retailers. On the other hand, the breakfast brand had a penetration of about 17% within households included in the study (which were required to purchase in the instant breakfast category in order to qualify), but an all-outlet penetration of just over 2%. This decrease reflects the low penetration of the category as a whole.

\(^1\)The Nielsen RMS data captures sales and price data from virtually every major retail chain at the point of sale and at the UPC level
MATCHING TEST AND CONTROL HOUSEHOLDS

Once the list of exposed households that will form the test group is created, each of the households on that list is matched with a household in the NCS data pool that was not exposed to the ad campaign (a “non-exposed household”) and most closely resembles it. The purpose of matching is to be able to attribute any increase in spending to advertising rather than to an existing preference for the product being studied. To that end, the variables used to determine test-control household pairs are customized for each project to pertain to the particular product category.

We use hundreds of matching variables. Those include: past spending on the advertised item being studied, competitor brands, and category as a whole; purchase amount at different retailers; recency of purchase; dollar purchases with a discount; and demographic information. Match variables include not only aggregated variables that capture total spend for the pre-period and post-period, but also monthly spend variables to ensure that households are at a similar stage in their buying cycle. For example, if a household consistently purchased a product for the first six months of the pre-period but did not make any purchases in the final six months, it would not necessarily match well with a household that did make recent purchases, even if the total dollar amount they spent over the course of the year was the same. Controlling for the stage in the purchase cycle helps separate out the impact of advertising from purchasing by a household that had simply run out of the product and would have purchased it at that point in time anyway.

Purchase-based variables are weighed more heavily than demographic variables. Specifically, the variables that receive the greatest weight in the match relate to buying patterns during the year prior to the ad campaign. These include dollars spent on the item, the number of units purchased, the number of purchase occasions, and average dollars spent per trip. These variables help establish that the test-control pairs resemble one another in their buying patterns of the brand being studied. A household that bought a large quantity of a product on one occasion might have the same overall spending as another that bought smaller quantities on multiple separate shopping trips, but matching on occasions and dollars per trip helps ensure that the paired households exhibit a similar consumption behavior. That way, an extra purchase occasion in response to the ad campaign can be separated out from a regular frequent purchase.

By comparison, demographic variables such as household income, presence of children, and the age of the head of household are weighed very lightly because these characteristics are not highly predictive of future purchase behavior. To the extent that they are—for example, parents of young children are much more likely to buy baby products—this information is already largely contained in the purchase data. However, information about the demographics of households that saw a campaign is provided in the form of indices that show whether a certain group is overrepresented among those exposed.

Control households may be paired with more than one test household because linear television campaigns have such high reach that there are not enough control households for a one-to-one match. However, research has shown that as long as resampling is limited to prevent any single control household from having too much influence, it does not adversely affect the quality of the results.

To guarantee that the incremental sales impact seen in the post-period is truly attributable to advertising exposure, the total difference in certain key variables between the test and control groups is less than 1% in the year prior to the start of the campaign. Key variables include total purchase, price, and purchase on deal. That 1% difference does not necessarily apply to individual household pairs. For example,
MEASURING HOUSEHOLD-LEVEL INCREMENTAL SALES

Once test-control pairs are created, the final step is to compare the aggregated purchases of the two groups in the post-period, thus quantifying average incremental sales per household (see our sidebar: metrics and calculations). Decomposition of the incremental sales can also be calculated by aggregating the differences for sub-groups in order to evaluate if a particular segment of consumers responded particularly well. This information helps the advertiser decide if there is a better way to allocate impressions for future campaigns.

In the household product study, the test group spent an average of $8.20 on the brand throughout the four-month period being studied, after all-outlet adjustment. The control group spent an average of $7.99. The results suggest that being exposed to the campaign led to an incremental spend of $0.21 per household, a 3% increase. Statistical significance was also calculated based on the distribution of the purchase data and the sample size. A significance level of 99% confirms that the results were conclusive and not a result of sampling variation. The robust sample size of many studies makes a significance level over 90% common even for studies where purchases increase by only one or two percent.

For this particular campaign, the client ran two different versions of the creative, and the exposure data was used to split the test group into those who saw one version or the other and an overlap group that saw both. Purchase data for each of the groups showed that one of the creative executions drove the entirety of the incremental revenue. While the results of the decompositions are not always statistically significant, they can highlight the greater success of one portion of the campaign compared to others. They can be computed for a variety of custom segments based on demographics, ad exposure, or purchase history.

A commonly used decomposition is one where we split the exposed group into subgroups based both on their past category purchases and on their loyalty to the brand being studied. For example, households that were already heavy buyers of the category but were only moderately loyal to the brand might have responded well to the campaign and become more loyal to the brand. This type of insight can help brand managers identify more specialized consumer segments and other diagnostics to improve future campaigns.

For parent brands with multiple sub-brands, the incremental revenue can be split up by sub-brand to determine which ones saw the greatest gain. In the household brand study, one particular sub-brand of the parent brand accounted for less than 20% of its sales in the pre-period, but represented over 40% of incremental sales in the post-period. Campaigns do not impact all products within a brand family equally, making the sub-brand contribution a valuable insight when planning what products to emphasize in future creative executions.

Finally, using the purchase data, the brand being studied can be compared to competitors in terms of category share requirements. This makes it possible for the advertiser to see which brands, if any, it’s taking share from. The competitor products can be defined based on a custom UPC list or based on product designations within NCS data, allowing for flexibility when determining how purchasing of a specific competitive set has changed.

A distinction must be made between the results seen in a sales effect study and overall market performance. The incremental sales are reflective of sales among the test households only, so spending among exposed households should not be expected to align with spending among all households. Single-source data makes it possible to separate out exposed households, distinguishing their purchase behavior from the purchase data of all buyers during that time span. This type of analysis isolates the impact of the ad campaign among those exposed to the advertising, and shouldn’t be confused with measures of sales in the total market.
IDENTIFYING THE DRIVERS OF INCREMENTAL SALES

The behavioral changes that drive incremental revenue are not always the same. For instance, an advertiser may want to know whether buyers are new to the brand or whether they had purchased it in the past already. It’s possible to answer that question using purchase data from the pre-period. Households that didn’t purchase in the pre-period but did in the post-period are classified as trial buyers, while households that purchased in the pre-period and again in the post-period are classified as repeat buyers. These totals are compared for the test and control groups. In the household product study, the brand was well-established and already had a high penetration. Over 40% of households in both the test and control groups were repeat buyers. But the percentage was higher within the test group, suggesting that the campaign successfully led more buyers of the brand to stay loyal to it.

Differences among households that made a purchase are also analyzed by calculating the dollar buying rate. This number represents the average household spend during the post-period, excluding those that didn’t buy the brand being studied. In the case of the household brand, the average dollar buying rate was $0.14 higher for the test group than for the control group. This suggests that the campaign not only drove more households to make a purchase, but also drove those households to spend more. However, increased penetration was a larger contributor to its success.

The dollar buying rate can be decomposed even further, into the number of purchase occasions and the average dollar amount spent per shopping trip. The average number of purchases made by the test and control groups for the household brand was almost identical. This suggests that the campaign did not drive purchases on a greater number of separate occasions. Incremental sales came instead from consumers choosing to buy more of the brand at a time.

The purchase drivers were significantly different for the breakfast brand. The campaign drove an overall lift of 28%, a number reflective of a smaller brand with room to grow its sales. Of the total lift, 10% came from increased penetration and 10% came from a higher number of purchase occasions. This campaign ran for less than a month, but it successfully persuaded consumers to choose the brand more frequently, at the expense of its competitors.

CALCULATING REACH AND RETURN

The approach to estimating campaign reach varies depending on the data available from each media platform and the particular project’s specifications. In general, the percentage of households from the NCS dataset that have been exposed is multiplied by a universe number to project to the general population. For example, in digital campaigns using third-party cookie matching, the number of unique households reached is divided by the total number of active online households in the NCS dataset to obtain a reach percentage, which is multiplied by the total number of online households—estimated at 96.5 million as of 2016—to get a final count. The reach percentage for linear TV studies comes from Nielsen Ad Intel and Nielsen’s national TV panel. For some media types, such as magazines, the publisher provides an estimated reach.

For cross-media studies, reach can be estimated for the entire audience that saw the campaign on a certain platform. Alternatively, it can be estimated for the audience that only saw it on that platform, with an overlap group capturing households that saw it in more than one place. Estimation of the overlap is made possible by exposure data from a large number of partners, and research is ongoing to improve estimates not only between partners on the same platform (such as multiple digital publishers) but also between platforms, such as linear TV and digital.

The total amount of incremental sales attributable to the campaign is calculated by multiplying the per-household incremental sales by the reach number. This calculation is based on the assumption that incremental sales are consistent between households in the sample and exposed households not represented by NCS purchase data. The samples on which the studies are based are robust enough that this assumption can be justified.

With total incremental sales calculated, other key indicators such as return on advertising spend (ROAS) can be calculated and compared to past campaigns and to benchmarks for the specific product category. With the insights provided in these studies, advertisers can determine effective ways to improve future campaigns. Single-source data takes the guesswork out of determining which aspects of a campaign are driving results.
CONCLUSION

Sales effect studies are a powerful tool available to researchers to measure the direct effect that an ad campaign has on in-store sales. One avenue for future research is to enhance the list of variables used in the matching process. While the purchase and demographic variables currently in use are robust and relevant to product buying patterns, there might be others that could help improve the attribution of incremental sales to the ad campaign. The relative weights of the variables could also be refined, along with the way the values are standardized.

This research is opening new doors to understand advertising effectiveness. For instance, with buyer basket analysis, we're starting to understand how ad exposure might affect cross-buying behavior—that is, consumers' propensity to purchase other products alongside the item being studied, during the same shopping trip. We're also starting to understand the many ways in which exposure on multiple platforms might combine to affect purchasing. The opportunities provided by single-source data are virtually limitless and have only begun to be explored.  

METRICS AND CALCULATIONS

Average Incremental Sales (can be in dollars, volume, units):
(Test Group Spend – Control Group Spend)/Total Analysis Households

Total Incremental Sales:
Average Incremental Sales X Total Households Reached

Penetration (percentage of households that made a purchase):
Number of Buyers/Total Analysis Households

Buy Rate (average household spend on the brand during the post-period):
Average Incremental Sales/Penetration

Purchase Frequency (average number of purchases made during the post-period by households that bought the brand being studied – this excludes non-buyers)

Purchase Amount (average amount spent on each purchase occasion):
Buy rate/Purchase Frequency

ROAS (return on ad spend):
Total Incremental Sales/Spending on Campaign

DPM (for digital studies, return per 1,000 impressions):
Total Incremental Sales/Impressions *1000

Corresponding metric for TV studies is DPP, or return per GRP

SALES IMPACT CALCULATION

Penetration x Purchase Amount x Purchase Frequency

How many Brand Buying HHs
How much of the products did HHs buy per occasion during the test period?
How often did HHs buy Brand products during the test period?

Buying Rate

How much are they buying?

Total Sales

HH = Households
INTRODUCTION

To measure the consumer packaged goods retail trade, marketing research companies such as Nielsen typically collect data directly from retailers who provide electronic point-of-sale (POS) information from their check-out scanning systems. This is by far the most accurate data available, but its collection is dependent upon retailer cooperation. If some of the retailers in the sample design do not cooperate, there can be some level of bias in the reported data. We can eliminate that retailer cooperation dependency by collecting data directly from individuals in population-projectable consumer panels, such as Nielsen’s consumer panel services (CPS), but the size of panels in certain regions of the world makes it difficult sometimes to report data at the granularity clients need to track performance. Increasing sample size is a solution, but it’s not always possible due to the costs of managed panels and the difficulty of recruiting reliable panelists.

With the growing worldwide adoption of new technologies like mobile smartphones, crowdsourcing, and virtual payments, new opportunities now exist to collect purchase information directly from consumers, in large numbers, and do so in a way that is both economical and less burdensome.

*To operate its flagship consumer panel services, Nielsen provides each panelist with a portable scanning device and requires that they scan the bar code of each and every product they purchase*
KEY MARKET CATALYSTS FOR AUTOMATION

According to Gartner, the smartphone market today has reached 90% penetration in the mature markets of North America, Western Europe, Japan and parts of Asia/Pacific. More than 1.4 billion smartphone units were sold worldwide in 2015. It’s become second nature for people to use apps on their mobile phones, and to use their mobile phones to share content over the internet. With this rapid adoption of smartphones by consumers comes opportunity for data collection. The quality of built-in cameras in the new generation of smartphones has improved to the point where images taken with those phones have a good enough resolution to serve as input data for automated processing. Asking people to take pictures of their store receipt is the basis for the project we’re describing in this paper.

People are not as reluctant to participate in cooperative projects as they might have been in the past. In fact, the rise of modern technologies and social media is helping shine a bright new light on the world of crowdsourcing. Without the technological barriers and social stigma, the notion of recruiting people to take pictures of their store receipts with their phones is not far-fetched anymore. With proper engagement and motivation, we can use the crowd very efficiently. Today’s online volunteers are very comfortable with digital rewards. No need to deal with physical goods or currency anymore: We can reward participants with “virtual coins,” and use the internet to transfer and manage those rewards.

To test these assumptions, we ran a proof-of-concept study in the United Kingdom. The U.K. was an ideal test market for two reasons: First, the overall smartphone penetration in the U.K. at the start of 2016 was around 68% (close to 91% for people under 35 years old, and 60% for people over 35). High enough for reasonable representation, and with improvements for better representation of the whole population expected in the near future. Second, we estimate that the combined market share of the two big non-cooperating discounters in the U.K. (Aldi and Lidl) is over 10%, so there is real market demand to provide more precise insights about these chains.

TRYING OUT THE NEW PROCESS IN THE U.K.

We kicked off our proof-of-concept project in January 2016, and by the end of August, 8,000 users had signed up and uploaded at least one receipt from their camera phone. So far, we have collected more than 400,000 images. In the first stage of the project, we focused on process and the best ways to engage with participants. We started with a small-scale initiative with approximately 800 users contributing pictures of their receipts back in January, and planned for a gradual increase to 4,000 active users by the end of the year.

Through random web-based invitations, we ask users to install our app on their smartphone. During the sign-up process, we collect some basic demographic information: their age, household size and postal code, as well as their email address. This information is used later in the process to measure and control for bias. The app is really simple: Users take a photo of any purchase receipt they get when they buy any product in a store. The photo is taken from within the application (that is, not with the phone’s standalone camera app), so that the app may immediately assess whether the quality of the image is up to our standards, and alert the user to retake the shot if needed. A typical receipt includes the description of all purchased items, the date and time of the purchase, the street address of the shop, as well as the amount paid. Each image is sent to the cloud for us to access, download and process the information further.

The user is rewarded with virtual coins for uploading even a single receipt. Additional coins can be earned if the app is used daily, or if the user invites others to become participants. Our objective isn’t to recruit a managed panel with this project, hence the way we recruit participants (using referrals, for instance) and the structure of the reward system itself are not subjected to the same scrutiny. We do want the rewards to be motivational, but of small value, so that they don’t unduly influence the shopping behavior of our participants. The virtual coins that people earn while participating in the study can be redeemed to enter a sweepstake, or saved up and eventually cashed out. There’s no minimum that needs to be achieved before earning the right to take part in a sweepstake, so a user may have just signed up, uploaded one receipt, and already earned the right to bid on a reward. The more coins users bid on a reward, the better their odds of winning.

---

*http://www.gartner.com/newsroom/id/3339019
*http://www.gartner.com/newsroom/id/3215217
*Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people (like an online community) rather than from employees or suppliers
*Pew Research Center - http://pewrsr.ch/1RX3Iqq
While we instruct users to take pictures of all their receipts, we cannot be certain that they’re complying. Perhaps they’re only sending us a few when convenient, or occasionally submitting receipts from other shoppers (e.g., other household members or even neighbors) in order to earn more rewards. To detect these situations, we have developed statistical algorithms to determine probabilities that users might under- or over-deliver. To maximize the usefulness of the receipts we receive, we’re looking at them as being representative of the universe of all shopping trips in the country to a specific chain. This is less subject to bias, but it also means that we need to establish a good independent estimate for the total number of shopping trips to each chain during the reported period.

INITIAL PARTICIPATION BUILD-UP

This is a proof-of-concept, and it was important for us to focus on the mechanics and logistics of the project rather than make sure that the participants were absolutely representative of the population. The initial phase allowed us to build quality checks to detect users whose activity was not compliant with our request (e.g., sending an image of a product instead of a receipt, or sending an obsolete receipt, or a receipt already uploaded by someone else) and implement ways to warn those end users. Users who repeatedly broke the rules were suspended, and not allowed to redeem any reward. We also learned that users stayed active on average only one month, so we started to experiment with ways that we could extend the duration of their commitment. Early data results have confirmed our initial hypotheses about some bias areas: We have low-representation among the oldest population group (they are the least technology-savvy of all population groups), and the average basket size is slightly less than what we originally estimated. We expect to be able to counter-balance these biases by using our managed household panel data for calibration.

For efficiency, we process all receipts via a sophisticated optical character reading (OCR) solution. OCR comes with great automation benefits, but it’s also an enormous technological challenge in our particular situation. One complication is dealing with a large volume of images with noisy background, and developing an algorithm to remove that noise is not a trivial endeavor. Another difficulty is to properly transform each receipt image into a set of database entries. The algorithm needs to be capable of finding the relevant information on each receipt; interpret each data point accurately regardless of its position on the receipt; and determine the correct meaning of the product descriptions found on that receipt.

IMAGE CAPTURE AND ITS CHALLENGES

It’s critical that the quality of each image satisfies a minimum-level requirement, and machines have yet to catch up with the human eye’s capability of deciphering text that might be blurry, printed on crinkled paper (e.g., warped, wrinkled, partly folded), in uneven lighting, or with faded or missing characters. Additionally, receipts might show up placed at an angle on the picture, or with objects in the background, or they could even be of an object that is not a receipt to begin with. Think about online CAPTCHA systems: To prove you’re not a bot, those systems present you with strings of blurry, distorted characters that are typically easy for a human being to make sense of, but nearly impossible to decipher for a bot relying on OCR algorithms. We want to prevent a situation where the store receipts we receive are as unreadable as those CAPTCHA string challenges.

For OCR to be successful with pictures of paper receipts, uploaded images need to be de-skewed, with as little background noise as possible, and the print needs to be clear. Image quality at the source is an all-important determinant for the success of OCR processes. We have a user guide to assist users in placing the camera in an optimal position, and to instruct them to take pictures of their receipts (even long, multi-part receipts) with as little background noise as possible. As long as users follow these instructions, there’s no need for them to perform more complex operations—such as cropping the image or any other special editing.

Taking the OCR out of the phone and putting it in the cloud was an important design decision. The Nielsen app can ascertain picture quality directly on the phone and alert the user to retake the picture if needed, but the more complicated image processing steps are performed out of the phone. There are many different types of smartphones in circulation, and only high-end mobile phones have enough processing power to perform sophisticated image processing and OCR with any reasonable degree of success. We decided to develop our own app so that our success rate wouldn’t be a function

---

4For an illustration of how big data can effectively be calibrated using panel data, see “The value of panels in modeling big data” by Paul Donato, Nielsen Journal of Measurement, Vol 1, Issue 1 (July 2016)
of the type of smartphone device used by participants. Taking
the OCR out of the phone also allowed us to try new detection
algorithms with much more flexibility than if we had to
upgrade the app on everyone’s phone every time we wanted
to change something.

PREPARING THE IMAGES FOR OCR

Once an image reaches the Nielsen server, the machine
work begins. The first checkpoint is that of image quality.
The image has already passed a first test on the phone,
but we’re able to apply more advanced algorithms on the
server to detect blurriness and determine if a receipt is
good enough for OCR. If it passes that test, our next step
is to identify the store chain via a combination of logo and
character recognition techniques. We built a dictionary of the
most frequent logos and their variants so that the tool can
match the image to the ones in the dictionary. A number of
innovative patent-pending techniques are at play to extract
relevant regions of interest on the image, detect a logo and
identify the source of the receipt.

If the image cannot pass those tests, it’s siphoned off
to a manual intervention stage where someone takes
the necessary steps—like manually cropping the region
of interest before sending the image to the de-skewing
algorithm, or visually identifying the chain before sending
the receipt back for further processing. When we started the
project in the U.K., half of the receipts at that stage required
some level of manual intervention, but the experience built
over the course of the last few months has allowed us to
bring that figure down to 25%, with plenty of room for further
improvements.

NIelsen’S UNIQUE OCR SOLUTION

Nielsen’s patent-pending innovation consists of wrapping
the OCR engine with a manual correction function that uses
machine learning to teach the function itself to autocorrect
going forward. Effectively, what this means is that manual
corrections on a small percentage of images end up having a
much larger impact, as learnings are gained by the machine
and applied to future batches. Early tests indicate a 90-95%
recognition rate based on this machine learning mechanism.

A recent run showed an impact rate of 22 times the size of
the manually corrected batch: that is, for every receipt that
underwent manual correction, 22 similar receipts benefited
from the learning and were corrected automatically.

Downstream, the next step is to make sense of the text that
has come through. We’re using a specialized algorithm to
classify the recognized text into relevant data types for the
business (e.g., product name, quantity, price, etc.). The
information needs to correspond to products that have
already been inventoried by the system—or if the product
is not recognized, invite a manual intervention to add the
product to the database.

A number of things could still go wrong in that late stage.
For instance, it may be difficult to determine whether a new
product is indeed a new product, or an existing product with
an alternate spelling. We may also find a discrepancy between
the sum of values of all products on the receipt, and the
total sum at the bottom of the receipt. As we are still in the
proof-of-concept stage, the data dictionaries are still a work
in progress. We’ve found that many products appear only on
one receipt, and the large number of such products poses a
special challenge to the team responsible for cross-coding
these descriptions. The maintenance of the product and store
dictionaries is one of the largest cost components in the
whole production process. But it’s well worth the investment:
Our team recently discovered that many receipts from one of
the retailers in the study weren’t getting processed properly
because the chain had just switched to a new logo. Once that
correction was made in the visual library, the recognition rate
for that retailer jumped by 35%.

PUTTING IT ALL TOGETHER

There are many different pieces to this process: collecting
the images, performing image quality checks, identifying
the retail chain, extracting the relevant regions-of-interest,
performing the OCR itself, and finally classifying the related
text—calling upon manual intervention along the way, as
needed. In order to keep track of everything and follow every
image along its life cycle, we built an end-to-end web facility
that all team members can use to measure progress.
AUTOMATIC RECOGNITION OF CROWDSOURCED PAPER RECEIPTS
SUCCESS RATE FOR TEST RETAILER - NIELSEN PROOF OF CONCEPT (U.K. 2016)

The illustration above provides a glimpse of the steps involved and their current success rates for one of the chains in the study. Out of every 100 images captured and uploaded by participants, 85 currently pass our initial quality checks. The 15 images that don’t pass this early check are either incomplete, unreadable, duplicates or are not receipts to begin with. Of the 85 that pass through, 17 correspond to long receipts that we don’t currently handle automatically, although that’s an active area of development. The 68 images that correspond to single receipts are then prepared for OCR processing, and 75% of those (i.e., 51 images) make it through to the actual OCR stage. Of those, 48 are successfully processed and only three need to be checked.

These early results are very encouraging, but there’s plenty of room for improvement. Our success rate keeps climbing from month to month as discoveries are made regarding the nature of the receipt images flooding in, and as our software stack becomes more sophisticated to address new scenarios.

CONCLUSION

The proof-of-concept study currently underway in the U.K. is an outstanding opportunity for Nielsen to test the viability of a new paradigm in data collection. It doesn’t replace point-of-sale scanner data, of course, but it provides an effective method to fill in the gaps when the point-of-sale data in a particular market doesn’t quite cover the universe of retail outlets. For our CPS managed panels, the study opens the door to potentially transition collection from Nielsen proprietary handheld scanners to mobile apps that lessen the requirement of scanning each item purchased to simply capturing an image of the overall receipt. Receipts come in many forms, but our algorithms are getting smarter all the time at handling special cases. While our software development effort is far from finished, we already have enough confidence in the quality of the automated outcome to start working on potential rollout plans and expanding this work to other countries.

The illustration above provides a glimpse of the steps involved and their current success rates for one of the chains in the study. Out of every 100 images captured and uploaded by participants, 85 currently pass our initial quality checks. The 15 images that don’t pass this early check are either incomplete, unreadable, duplicates or are not receipts to begin with. Of the 85 that pass through, 17 correspond to long receipts that we don’t currently handle automatically, although that’s an active area of development. The 68 images that correspond to single receipts are then prepared for OCR processing, and 75% of those (i.e., 51 images) make it through to the actual OCR stage. Of those, 48 are successfully processed and only three need to be checked.

These early results are very encouraging, but there’s plenty of room for improvement. Our success rate keeps climbing from month to month as discoveries are made regarding the nature of the receipt images flooding in, and as our software stack becomes more sophisticated to address new scenarios.

CONCLUSION

The proof-of-concept study currently underway in the U.K. is an outstanding opportunity for Nielsen to test the viability of a new paradigm in data collection. It doesn’t replace point-of-sale scanner data, of course, but it provides an effective method to fill in the gaps when the point-of-sale data in a particular market doesn’t quite cover the universe of retail outlets. For our CPS managed panels, the study opens the door to potentially transition collection from Nielsen proprietary handheld scanners to mobile apps that lessen the requirement of scanning each item purchased to simply capturing an image of the overall receipt. Receipts come in many forms, but our algorithms are getting smarter all the time at handling special cases. While our software development effort is far from finished, we already have enough confidence in the quality of the automated outcome to start working on potential rollout plans and expanding this work to other countries.