Counting What Will Count: How to Empirically Select Leading Performance Indicator[1)]

Koen Pauwels, Amit Joshi

Abstract

Facing information overload in today's complex environments, managers look to a concise set of marketing metrics to provide direction for marketing decision making. While there have been several papers dealing with the theoretical aspects of dashboard creation, no research creates and tests a dashboard using scientific techniques. This study develops and demonstrates an empirical approach to dashboard metric selection. In a fast moving consumer goods category, this research selects leading indicators for national-brand and store-brand sales and revenue premium performance from 99 brand-specific and relative-to-competition variables including price, brand equity, usage occasions, and multiple measures of awareness, trial/usage, purchase intent, and liking/satisfaction. Plotting impact size and wear-in time reveals that different kinds of variables predict sales at distinct lead times, which implies that managerial action may be taken to turn the metrics around before performance itself declines.

Keywords : Metrics, Leading Indicators, Marketing Dashboards

Koen Pauwels | Associate Professor, Business Administration, Ozyegin University (koen.h.pauwels@ozyegin.edu.tr)

Amit Joshi | Associate Professor, Department of Marketing, College of Business Administration, University of Central Florida(ajoshi@bus.ucf.edu), Corresponding author.

---

1) The Authors Are Grateful To Marije Teerling For Research Support And To Marnik Dekimpe, Shuba Srinivasan, The Marketing Science Institute, And Participants At The 2006 And 2007 Marketing Science Conferences And At The Tuck School Of Business For Insightful Comments.

I. Introduction

With the call for marketing accountability increasing, managers regularly turn to marketing metrics to inform them about the direction company performance is heading (Rust et al. 2004; Webster et al. 2005). At the same time, continued improvements in data collection, storage and analytics generate a wealth of potentially useful metrics. The fragmentation of media, proliferation of marketing channels as well as product lines, and mass customization complicate the marketing landscape (Hyde et al. 2004). As a result, large marketing departments nowadays track tens and sometimes hundreds of metrics of various buyer readiness stages as "leading indicators" of market performance(Ambler 2003).

Management's inability to effectively process all this information(Krauss 2005), let alone integrate it for decision making and CEO status reports, points to the need to identify and summarize key factors. Indeed, a survey of over 1000 C-level managers reveals that only 17% of marketing executives have a comprehensive system to measure marketing performance, and that they outperformed others in revenue growth, market share and profitability(CMO council 2004). The Marketing Science Institute(2006) acknowledges the problem of information overload by including among its 2006-2008 research priorities "separating signal from noise in detecting emerging external trends," and "the role of marketing dashboards".

A marketing dashboard is "a relatively small collection of interconnected key performance metrics and underlying performance drivers that reflects both short and long-term interests to be viewed in common throughout the organization"(Pauwels et al. 2009, p.177). As many as 40% of U.S. and UK companies report substantial efforts in developing and using such dashboards(Clark et al. 2006; Reibstein et al. 2005). Clark et al.(2006) find that using dashboards improves company performance. Indeed, "properly created dashboards provide the mechanism to drive effective management and resource allocation decisions"(Wind 2005, p.870).

Managers are interested not just in theory-based identification of candidate metrics, but also in empirical techniques for efficiently populating the dashboard, establishing relations between metrics and providing what-if analysis(Stewart 2008; Marketing NPV 2005). Reibstein

et al.(2005) term (1) identification of metrics as the first stage in the dashboard development process, which next progresses through, (2) populating the dashboard with data, (3) establishing relationships between the dashboard items, (4) forecasting and "what-if" analysis, and (5) connecting to financial consequences. Current practice typically does not go beyond Step 2, prompting Pauwels et al.(2009) to call for academic research on selecting dashboard metrics that are leading indicators of performance and on quantifying their contribution to performance prediction.

This study addresses this gap in the literature. Using a unique dataset that contains 99 regularly-measured metrics for a frequently purchase national brand and its competitors, the study creates multiple dashboards using various econometric techniques and then selects the most predictive set of metrics based on out-of sample forecasting fit criteria. Moreover, this research relates the variables included in the selected dashboard with short term and long term strategic outcomes for the brand. Thus, the research empirically completes all stages of dashboard creation, being the first study to do so.

The rest of this paper is organized as follows. Section 2 is a discussion of past theory that helps us in deciding what metrics ought to enter a dashboard. Section 3 briefly discusses the method for dashboard creation and the data used for an empirical demonstration. Section 4 creates multiple dashboards for a national frequently purchased snack brand as well as a store brand with a discussion of the obtained dashboard. Section 5 concludes with limitations and suggestions for future research.

Ⅱ. Research Background

Which metrics should a marketing dashboard track? Marketing theory and construct development provides a rich set of metrics with convergent and divergent validity in our empirical context of fast moving consumer goods. First, the multiple versions of the hierarchy-of-effects model gave us several measures of awareness, preference, purchase intention, trial, affect/liking and satisfaction(Lavidge and Steiner 1961; Ray et al. 1973; Smith and Swinyard 1982). Moreover, consumer perception should matter on several attributes such as product quality, value-for-money and trust(Kotler and Keller 2007).

<Table 1> Overview of Dashboard Metrics Widely Tracked for Fast-moving Consumer Goods

Concept	Operationalization (absolute or relative to competition)
Consumer price	What is the net price paid for brand?
Advertising Awareness	Did respondent remember seeing an ad about brand?
Top-of-mind brand awareness	Is brand the first evoked when mentioning the category only?
Unaided brand awareness	Is brand in the set evoked when mentioning category only?
Aided brand awareness	Is brand recognized when mentioned to respondent?
Taste rating	Given aided awareness, does brand have “Taste I love”?
Quality rating	Given aided awareness, does brand have “high quality”?
Value rating	Given aided awareness, does brand offer “good value”?
Cost rating	Given aided awareness, does brand “cost more”?
Healthy rating	Given aided awareness, is brand “healthier for me”?
Satisfying rating	Given aided awareness, is brand “satisfying”?
Liking | Aware	Given aided awareness, how much does respondent like brand?
Liking | Tried	Given ever tried, how likely will respondent buy brand?
Good feeling rating	Given aided awareness, is it a “brand I feel good about eating”?
Fun rating	Given aided awareness, is brand “fun to Eat”?
Trust rating	Given aided awareness, is it a “brand I trust”?
Favorite	Is brand the respondent's favorite?
Purchase intention | Aware	Given aided awareness, how likely is respondent to buy brand?
Purchase intention | Tried	Given ever tried, how likely is respondent to buy brand?
Trial	Did respondent ever try brand?
Last Week User	Did respondent use brand within the last week?
Last Quarter Purchaser	Did respondent purchase brand within the last three months?
Last Month Purchaser	Did respondent purchase brand within the last four weeks?
TV Watching occasion	How likely is respondent to use brand “while watching TV”?
Home occasion	Likelihood to use brand “while hanging out at home”?
Entertaining occasion	Likelihood to use brand “while entertaining friends or family”?
Relaxing occasion	Likelihood to use brand “while relaxing by yourself”?
Afternoon Lift occasion	Likelihood to use brand “when needing a lift in the afternoon”?
On-the-Go occasion	Likelihood to use brand “while on the go”?
Sport Watching occasion	Likelihood to use brand “while watching a sports event on TV”?
Satisfaction | Aware	Given aided awareness, how satisfied is respondent with brand?
Satisfaction | Tried	Given ever tried, how satisfied is respondent with brand?
Future-needs rating	How likely is brand to fulfill respondent needs in the future?

Last but not least, consumers may associate the brand with specific usage occasions, such as entertaining friends and family versus relaxing by oneself. An increase in usage occasions for the brand should increase sales performance. 〈Table 1〉 provides an overview of the dozens of metrics that may result from such assessment.

Obviously, many more metrics could be put forward. Indeed, Ambler(2003) recommends that dashboard metrics should give information on every likely failure cause, and thus be comprehensive enough to enable decision makers to recognize deviant patterns and dis-

cover new problems and opportunities. At the same time though, attention at the executive level is "brief, fragmented and varying"(Mintzberg 1973) and "data is prolific but usually poorly digested and often irrelevant" (Little 1970, p. B466). Therefore, the need for reducing complexity – and thus the number of metrics – is key to the concept of a marketing dashboard (NPV 2004; LaPointe 2006; Pauwels et al. 2009). The exact number of metrics is still subject of debate(Ambler 2003), with US companies preferring 6-10 metrics (similar to the simple dashboard used by a car driver), and UK companies often comfortable with 10-20 metrics (for which a comprehensive airplane pilot dashboard may be a more appropriate analogy). Across all applications, 10 metrics is most typical; that is also the number of "general" metrics that are claimed to matter across companies and industries(Ambler 2003).

In any case, even the high end of this range requires a substantial reduction compared to the (non-exhaustive) list of potential metrics in 〈Table 1〉. While company employees and industry experts may be ideally suited to come up with a comprehensive set of metrics, they are not well placed to reduce them to a manageable number. Different departments and senior managers often hang on to 'their metrics' and obstruct the necessary simplification(Marketing NPV 2005). Thus, academic research could help by providing and comparing tools for reducing metrics (Pauwels et al. 2007).

The first step in reducing metrics considers the statistical properties of each metric separately. Specifically, Ambler (2003) recommends deleting metrics (1) which show little variation over time or (2) that are too volatile to be reliable. These two criteria suggest a univariate time series analysis for each metric. Next, he recommends deleting metrics which (3) are not a leading indicator of market outcomes and (4) add little in explanatory power to existing metrics. Econometric literature provides several ways of establishing these criteria, including temporal precedence tests (e.g., Granger Causality tests) and multiple variations of regression analysis. Marketing practitioners often favor stepwise regression for its fast and automatic selection of variables, while academic researchers recommend more sophisticated methods such as reduced rank regression and forecast variance error decomposition (based on a Vector Autogressive Model). To the best of our knowledge,

the performance and results of these different methods have yet to be compared.

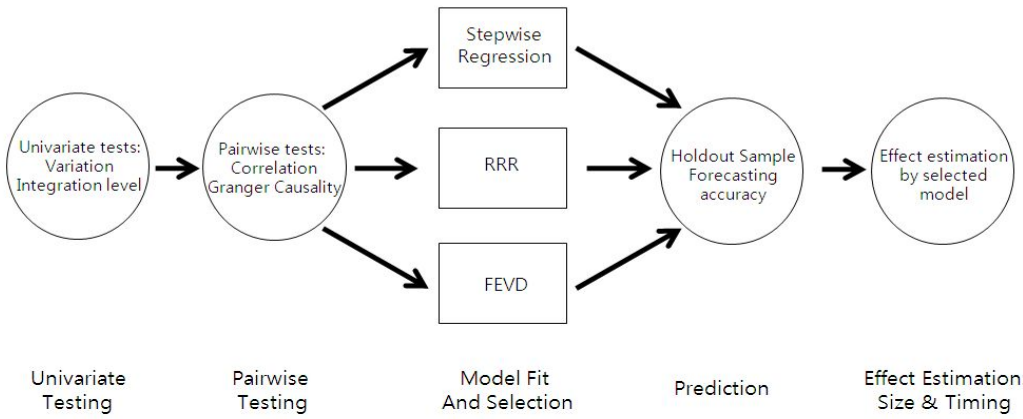
The other important question that arises when creating dashboards is which performance variables (for instance, short term or long term; market share or stock price, etc) should be considered when populating the dashboard with metrics, since different dashboards can be created depending on the target performance metric. However, in keeping with the context of fast-moving consumer goods, and in line with the extant research in this area that has mainly looked at sales performance metrics and revenue premium, this research concentrates on dashboards for the above performance variables.

Given the preference for 10 original metrics, we can now apply established econometric methods for the selection / elimination of metrics to create a dashboard that performs 'best' according to some established criteria, such as adjusted R2, Akaike information criterion, or Bayesian information criterion (Miller 1989; Blattberg et al. 2008). Ambler(2003) suggests deleting metrics that (1) are not leading indicators of market outcomes and that (2) add little in explanatory power to existing metrics. Econometric literature provides several ways of establishing these criteria, including temporal precedence tests (e.g., Granger causality tests) and multiple variations of regression analysis. To the best of our knowledge, though, the literature has yet to combine and apply these methods to the metric selection problem for marketing dashboards and to compare the forecasting accuracy of the metric sets selected by these methods.

Ⅲ. Metric Selection Framework

〈Figure 1〉 summarizes a proposal for a general empirical framework for metric selection. While this study restricts empirical demonstration to the fast-moving consumer goods category, the proposed methodology is generalizable to other categories and industries.

The first task involves selecting a reduced set of metrics from among the full set of variables. This research suggests proceeding in logical order from simple to more elaborate approaches. First, apply univariate tests to check for the variability of each metric and its time series properties (stationary versus evolving).Second, examine in pair wise tests the correlation among candidate



<Figure 1> Methodology

metrics and the Granger causality of each metric with the performance variable. Third, estimate a set of proposed econometric methods, preferably selected from different research traditions, which will yield competing sets of metrics. Having selected these competing sets of metrics, proceed to examining the predictive validity of these sets in the holdout sample. Finally, use the most successful set of metrics and estimation model to assess the relative size and wear-in period of the effect of each metric on the performance metric. These two issues are key to managers: effect size indicates how much of an impact a change to the metric has on performance, while the wear-in period indicates how much time managers have for remedial action before performance itself is affected.

Univariate and Pairwise Tests for Metric Selection

This research proposes two univariate tests to begin the metric selection process. The variability of a measure is assessed by its coefficient of variation, which is more appropriate than the standard deviation in the general case when variables are measured in different units or have very different means (Wilkinson 1961). Next, unit root tests verify that variables have the same level of integration, which is needed to avoid spurious relation in econometric models(Granger and Newbold 1986; Dekimpe and Hanssens 1999). Pair wise tests include the (contemporaneous) correlation between the variables, and their dynamic relation by means of Granger cau-

sality (GC) tests(Granger 1969; Hanssens et al. 2001).

In essence, Granger causality implies that knowing the history of a variable X helps explain a variable Y, over and above Y's own history. This 'temporal causality' is the closest proxy for causality that can be gained from studying the time series of the variables(i.e., in the absence of manipulating causality in controlled experiments).

These univariate and pairwise tests should substantially reduce the candidate variables for the dashboard. Indeed, in certain situations managers may stop here if they feel that enough variables have been eliminated for a dashboard suitable for their particular needs. The steps below help in further reducing metrics by observing their ability to forecast performance out of sample.

Econometric Models for Metric Selection

Numerous models in econometrics and statistics can be used to winnow down a large number of variables to a smaller number. This study organizes them along two dimensions: (1) whether or not they create new constructs(Stone and Brooks 1990), and (2) whether or not they consider endogeneity among performance and potential mindset metrics in explaining the dynamic variation in the performance variable(Nijs et al. 2007). The first dimension separates methods that only use the original variables from methods that also create new constructs(factors). Examples of the former include all-subset and stepwise regression. Examples of the latter include partial least squares, principal components, and reduced-rank regression. The second dimension separates all these mentioned techniques from dynamic system models that account for potential endogeneity among the candidate metrics in calculating how much they explain the dynamic variation in performance. This property could matter because candidate dashboard metrics and, in general, customer mindset metrics may affect each other over time in complex feedback loops (Lehmann and Reibstein 2006; Pauwels et al. 2009; Srinivasan et al. 2009). Combining both dimensions thus yields the following groups of models: (1) regression models that use the original variables, (2) models that create new constructs and (3) dynamic system models (dynamic system models that create new variables have yet to be developed in literature).

Rather than using just one econometric method, it is important to examine

several for two reasons. First, given the limited knowledge on metric selection for marketing dashboards, this study aims to be inclusive rather than exclusive. The likelihood of eliminating an important variable by one econometric method but selecting another is high. Managers, therefore, should have an overview of all feasible candidate metrics at this stage. Second, it is easy to foresee that different econometric models may be best suited for different situations. In the interest of generalizability, it appears best to consider a menu of econometric methods from which to select the best suited for the situation. In the interest of conciseness though, this study recommends the selection of one method from each of the identified groups. For the empirical analysis in this study, we choose stepwise regression from the first group because it is widely used in marketing practice(Meri and Zahavi 2005).

The study includes using reduced-rank regression(hereafter RRR) from the second group because it is the data reduction technique that maps the identified factors back to the original variables, thus allowing cost savings from tracking fewer metrics. Finally, from the third group forecast error variance decomposition(hereafter FEVD), a recently popular technique (Nijs et al. 2007; Srinivasan et al. 2008) that is derived from vector autoregressive(VAR) models, is chosen. Specific information on the econometric specification and estimation is available in the Technical Appendix. While the specific econometric methods may change based on industry and context, the broad selection criterion for methodologies ought to follow the path outlined above. The decision on which method is best suited for a particular situation can be made empirically on the basis of predictive accuracy, as detailed below.

Comparison of Predictive Accuracy of the Selected Metric Sets

For each method used, managers should evaluate the selected set of metrics based on how accurately they forecast sales performance in a holdout sample. This is the proper comparison basis because a dashboard's value to management is determined by how well it predicts future performance and by its usefulness in what-if analyses that help improve future decisions(Krauss 2005; LaPointe 2006). Sophisticated time series methods may perform well in-sample due to curve fitting but may

fare worse out-of-sample. As a result, econometric model comparisons should be based on their accuracy in a holdout sample (e.g., Stone and Brooks 1990; Levin and Zahavi 1998; Neslin et al. 2006). The specific procedure is described in the Technical Appendix.

IV. Data

The U.S. data are provided by a market research firm that gathers key performance indicators for a national brand manufacturer, both of which prefer to remain anonymous. The data period runs for 156 consecutive weeks in the early 2000s. The product category is a frequently purchased snack (an impulse purchase good with relatively low involvement) with household penetration in the upper 90% range and a purchase cycle of about seven weeks. Consistent with past research(Petersen et al. 2009), this study uses weekly measures on all variables to allow for the inclusion of short, medium and long-term impact metrics in our dashboard for both the national as well as store brand. The national brand in our analysis is the clear market leader; in fact its only competitive threat comes from the store brand, operationalized as the composite of all store brands in our US-wide dataset. Calculation of the annual revenue premium of the national brand(Ailawadi et al. 2003) reveals that this important market outcome measure went down from $162 million in the first year to $152 million in the second and $111 million in the third year of our data. The main decline derives from a reduction in volume premium (from 21% to 5%) rather than price premium (from 24% to 19%).

We perform analyses for two very different brands: the mature national brand (the market leader) and the growing store brand (the market challenger). While prices are expected to matter for both brands, customer-based brand equity theory(Keller 2003) implies substantial differences in the predictive power of, say, awareness measures (important for the market challenger) versus say, special usage occasions (important for the market leader). Thus, analysis for each brand may uncover both similarities and differences into (1) which of the methods performs best (of key interest to researchers) and (2) which specific metrics are selected (of key interest to managers).

The market research firm gathers weekly data on the measures in 〈Table 1〉 for both the national brand and a

national composite of the store brands. With the exception of average price, all measures are obtained by telephone surveys of 100 category buyers (the first 100 contacted people who indicated they purchase in the category, out of a random sample of U.S. households from the phone directory). 〈Table 2〉 shows some representative examples of the questions asked.

<Table 2> Representative Telephone Survey Questions and Measures

Variable	Survey Question	Measure
aided advertising awareness	For which of these brands have you seen, heard, or read any advertising in the past few months? (Respondent is read a list of brands, and indicates YES or NO to each)	percentage of respondents indicating "yes"
purchase intention given awareness	How likely would you be to purchase or have someone in your household purchase [brand]? In the next three months, would you definitely buy it, probably buy it, might or might not buy it, probably not buy it, or definitely not buy?	percentage of respondents indicating "definitely buy it" or "probably buy it"
value rating	I'll ask you to rate the brands on each characteristic using a scale from 1 to 5, where 5 means you strongly agree, 3 means you neither agree nor disagree, and 1 means you strongly disagree. You may use any number between 1 and 5. Let's start with the value a brand provides. On this scale from 1 to 5, how would you rate[brand]?	percentage of respondents indicating 4 or 5 on the 5-point scale

Because dashboards often include relative as well as absolute metrics (e.g., Ambler 2003), the empirical application arrives at a total of 99 candidate metrics: the absolute measures for each brand and the difference between the scores for the brand and the store brand composite on each measure (hereafter "relative" or "diff").

Marketing managers are often overwhelmed by this large amount of data, that is sold to them as "key performance indicators" and have expressed strong interest in retaining a smaller set of met-

rics, which truly are leading indicators of performance. This would both render the metrics more actionable and allow the company to reduce the cost of regularly (in our case weekly) collecting these metrics.

As noted above, the choice of performance variable is an important one, as all the variables in the dashboard will depend on the choice of this variable. Depending on the application, managers may select either top-line or bottom-line performance variables. Within the data limitations (as typical for marketing datasets, no cost or stock price information is available), the empirical application analyzes both national-and store-brand sales with revenue premium as performance variables. While brand sales is an absolute performance metric, revenue premium is a relative performance metric, calculated as the difference in revenue between the national brand and the store brand.

V. Empirical Results

Results from the univariate and pairwise tests are reported first, as they demonstrate the ability of these tests to reduce the number of candidate metrics, and the results from the econometric model selection are discussed subsequently.

Metric selection from Univariate and Pairwise Tests

All variables show variability, with the coefficient of variation ranging from 3.91(aided awareness for the national brand) to 295(national-brand consumption while watching TV). Second, the unit root tests reveal that not a single variable is classified as evolving, with consistent results for different test versions (detailed results available upon request). As the performance variables (national-brand and store-brand sales) are also stationary, this study concludes that all variables are of the same integration level. This situation is typical for top brands in fast moving consumer goods (Nijs et al. 2001; Slotegraaf and Pauwels 2008).

The next step in the empirical analysis is to perform pair wise Granger causality tests with each variable and national-brand sales, store-brand sales and revenue premium, respectively. 〈Table 3〉 displays the candidate metrics that Granger-cause each performance variable. The Granger causality tests enable a fast reduction of the number of candidate metrics from 99 to 17 (for sales) or 18 (for revenue pre-

mium). A list of 17-18 variables may be right for several firms; in fact, as mentioned earlier, British companies typically prefer 10-20 metrics(Ambler 2003). However, this analysis aims to further reduce the list to 10 metrics, as noted above.

<Table 3> Granger Causality Test Results: Leading Performance Indicators

Metric Type	Granger Cause Revenue Premium	Granger-Cause National-Brand Sales	Granger-Cause Store-Brand Sales
market	price of national brand(pricenb) price of store brand(pricest)	price of national brand (pricenb) price of store brand(pricest)	price of national brand (pricenb) price of store brand(pricest)
awareness	relative top-of-mind awareness(awaretomdiff) top-of-mind awareness, national brand(awaretomnb)	unaided awareness, national brand(awareunnb)	relative top-of-mind awareness(awaretomdiff) top-of-mind awareness, national brand(awaretomnb) relative unaided awareness (awareundiff) unaided awareness, national brand(awareunnb) unaided awareness, store brand(awareunst)
knowledge	National brand costs more(costnb) Quality nat brand(qualnb) valuediff	national brand is satisfying(satisfynb) relative taste(tastediff) relative quality(qualdiff)	relative cost(costdiff)
Liking	like given aware national brand(likeawarenb) Feelings for store brand(feelst) Relative brand trust(trustdiff)	like given tried national brand(liketriednb) Feelings for store brand (feelst) Relative fun perception (fundiff) Relative brand trust (trustdiff)	like given aware national brand(likeawarenb) relative feelings(feeldiff)
preference	favoritest	purchase intention given awareness, national brand (piawarenb)	
purchase	tried in the last four weeks, national brand (tried4wnb), afternoon lift occasion, national brand(afternoonnb) afternoon lift occasion, store brand (afternoonst) relative relax occasion (relaxdiff) relative sports-watching occasion (sporttvdiff)	tried in the last four weeks, national brand(tried4wnb), relative trial in the last three months, (tried3mdiff) afternoon lift occasion, national brand(afternoonnb) entertain friends occasion, national brand (entertainnb) on the go occasion, national brand(onthegonb)	relative trial in the last three months(tried3mdiff) afternoon lift occasion, store brand (afternoonst) relative entertain friends occasion(entertaindiff) relative relax occasion (relaxdiff) relative sports-watching occasion(sporttvdiff) relative TV-watching occasion(tvdiff)
reinforcement	satisfied given tried national brand(satistriednb) satisfied given used national brand(satisusednb)	satisfied given tried national brand(satistriednb)	satisfied given tried national brand(satistriednb)

Metric selection from Econometric Methods

For the sake of brevity, this research focuses on comparing the metric selection for national brand versus store brand sales, displayed in tables 4-5 with the in-sample fit statistics for each model.

<Table 4> Selected Metrics and Their Explanatory Power for National-Brand Sales

	Stepwise Regression	**Reduced-rank Regression**	**FEVD**
Market	pricenb, pricest	pricenb, pricest	pricenb, pricest
Awareness	awareunnb	awareunnb awarenunst	awareunnb
Knowledge	satisfynb	tastediff	satisfynb qualdiff
Liking	feelst	fundiff trustdiff	liketriednb trustdiff
Preference	piawarenb		
Purchase	tried3mdiff afternoonnb entertainnb	tried3mdiff entertainnb	tried4wnb afternoonnb, entertainnb
Reinforcement	satistriednb	satistriednb	
R^2(Adjusted R^2)	94.25 (78.93)	92.71(73.27)	92.86(77.30)
Akaike Info Criterion (Bayesian Info Criter)	27.98(29.61)	28.21(29.82)	28.08(29.65)

<Table 5> Selected Metrics and Their Explanatory Power for Store Brand Sales

	Stepwise Regression	Reduced-rank Regression	FEVD
Market	pricenb pricest	pricenb pricest	pricenb pricest
Awareness	awareunnb awareunst	awareunnb awareunst	awareunnb awareunst awaretomdiff
Knowledge	costdiff	costdiff	costdiff
Liking	feeldiff	likeawarenb	likeawarenb
Preference			
Purchase	tvdiff sportdiff relaxdiff	tried3mdiff sportdiff afternoonst	tvdiff sportdiff relaxdiff
Reinforcement	satistriednb	satistriednb	
R^2(Adjusted R^2)	82.37(65.09)	92.71(73.27)	92.86(77.30)
Akaike Info Criterion (Bayesian Info Criter)	27.38(28.68)	28.21(29.82)	28.08(29.65)

In each case, the selected metrics explain a large part of the of the performance variance in-sample. Using stepwise regression results in a higher R^2 for national brand sales (0.9425) and using FEVD yields higher R^2 for store brand sales (0.9286) for our case study. As to the specific metrics selected, several observations stand out from tables 4-5. None of the preference items are included in the sets selected for store-brand sales. For national-brand sales, only "purchase intention given aware" is included by stepwise regression. Further analysis shows that these preference variables load high with recent trial (in an exploratory factor analysis) and/or do not add sufficient explanatory power (in a model with recent trial) to warrant selection. This dominance of purchase behavior/occasion metrics over stated-preference/purchase intention metrics is consistent with reports in recent marketing literature (Kumar, et al. 2006; Rust et al. 2004).

Second, all three methods tend to se-

lect a similar number of metrics from each conceptual category, but often differ in the specific metric they select. For instance, in the purchase category, stepwise regression and RRR select the difference in trial over the last three months (tried3mdiff), while FEVD selects the national-brand trial over the last four weeks (tried4wnb). Overall, four metrics are selected by all three methods for national-brand sales: prices for both brands, unaided awareness of the national brand, and entertainment usage of the national brand. For store-brand sales, three metrics are selected by all econometric methods: price for both brands and unaided awareness for the national brand.

These selected metrics are appealing in the context of this low-involvement, impulse purchase product category. First, price is important as a managerial control variable, as is trial–a key objective for both national-as well as store-brand managers(Kumar and Steenkamp 2007). Second, unaided brand awareness is key for the national brand given its price disadvantage compared to the store brand. Finally, the selection of both absolute and relative metrics indicates the extent to which brand managers and retailers have their (sales) fate in their own hands. On the one hand, metrics such as "entertainnb" (likelihood of using national brand when entertaining) and "afternoonnb" (likelihood of using national brand to get an afternoon 'lift') suggest perceived optimal usage occasions that national-brand managers can use in their advertising messages to improve brand sales. On the other hand, many metrics are relative and thus depend on the perceptions of both the national brand and the store brand. For national-brand sales, both the trust gap and the perceived quality gap are selected. For store-brand sales, the perceived cost gap and the relative brand appropriateness for relaxing (in front of the TV or sports) matter.

Predictive validity of selected Dashboard Metrics

〈Table 6〉 provides the out-of-sample predictive validity of the selected metric sets. All four forecasting criteria yield consistent results. First, stepwise regression, while showing the best in-sample explained variance for national-brand sales (〈Table 4〉), has the worst out-of-sample forecasting accuracy for both national-brand and store-brand sales. This finding is consistent with the decades-long criticism of stepwise regression, which has been called "unwise regression"(Leamer

1985), but in sharp contrast with a recent research paper declaring stepwise regression "the winner" in a forecasting accuracy contest with other methods in three data sets(Meiri and Zahavi 2005, hereafter MZ). Besides differences in candidate metrics and forecasting criteria (MZ used R^2 in the holdout sample and the Gini and ML coefficients), this difference is likely due to two main factors: first, MZ did not compare stepwise regression with reduced-rank regression and FEVD, and second, MZ allowed each method to select a different number of metrics, with stepwise regression always selecting the smallest number, about half as many as most competing models. It is well known that a smaller number of variables produce better out-of-sample forecasting accuracy (Armstrong 2001), which is the reason this study keeps the number constant across methods. Future research should examine the robustness of these conclusions by investigating model performance for different numbers of metrics.

<Table 6> Out-of-sample Forecasting Accuracy of Selected Metrics

	Stepwise Regression	Reduced-rank Regression	FEVD
National-brand Sales			
Root Mean Squared Error	626,487	530,501	485,277
Mean Absolute Error	516,751	424,022	396,132
Mean Absolute Percentage Error	19.76	15.84	14.98
Theil's Inequality Coefficient	0.1079	0.0906	0.0845
Store-brand Sales			
Root Mean Squared Error	358,979	343,163	303,972
Mean Absolute Error	283,048	275,630	230,637
Mean Absolute Percentage Error	10.63	10.54	8.71
Theil's Inequality Coefficient	0.0708	0.0676	0.0594

Turning to the forecasting accuracy of the remaining selection procedures, the set of metrics based on FEVD outperforms the set selected by reduced-rank regression for each criterion. For national-brand (store-brand) sales, the improvement in Theil's inequality coefficient is 16% (4.5%) moving from stepwise to reduced-rank regression and 7% (12%) moving from reduced-rank re-

gression to FEVD. The rank order and relative magnitude of the forecasting accuracy results are robust to alternative specifications of Equation 5, such as adding contemporaneous effects of dashboard metrics and adding an autoregressive term for brand sales. Moreover, allowing stepwise regression and reduced-rank regression to select lagged terms of candidate metrics results in metric sets that perform even worse out-of-sample.

For the dataset, the best forecasting performance results from the set of metrics selected by FEVD, as derived from a vector autoregressive model that includes all variables that Granger-cause performance. Analysis rules out several possible explanations, including the pre-model specification Granger Causality tests (which are used as the starting point for all three econometric models), in-sample curve fitting (since the analysis compares out-of-sample forecasting accuracy) and model specification issues such as the number of lags included (all models are allowed to use lagged variables in their metric selection and only 4 lags are used in the same OLS regression model to compare the performance of selected metrics). Therefore, after eliminating other possible alternative explanations, the most plausible explanation is that FEVD accounts for the endogeneity among the candidate metrics in calculating how much they explain the dynamic variation in performance. This property should be important for metric selection if candidate dashboard metrics affect each other over time in complex feedback loops (Lehmann and Reibstein 2006; Pauwels et al. 2008; Srinivasan et al. 2009).

While these observations are important to researchers and dashboard builders, managers and dashboard users may be more interested in the size and timing of the effects of the selected metrics on performance and on the extent to which they can control them (Reibstein et al. 2005). This is done next.

Managerial Implications: Timing and Magnitude of Dashboard Metrics' Impact

Based on the vector autoregressive coefficients of the selected metric model, this study estimates impulse response functions(Dekimpe and Hanssens 1999; Pauwels et al. 2004) to track the over-time impact that unexpected changes (shocks) in the dashboard metrics have on the performance variable (national-brand sales). For the calculation of the short-term (same-week) effects, the generalized, simultaneous-shocking ap-

proach(Pesaran and Shin 1998; Dekimpe and Hanssens 1999) is useful. To enable calculation of long-term effects significant effects are accumulated over time(Pauwels et al. 2002). Finally, this research calculates 'wear-in' as the number of weeks before the peak effect (largest effect in absolute value) is reached, and the 'wear-out' as the number of weeks the effects remains significant after the peak(Pauwels 2004). 〈Table 7〉 displays the size and timing of the impact of each selected dashboard metric on national-brand sales.

<Table 7> Size and Timing of Each Dashboard Metric's Effect on National-brand Sales*

	Short-term	Long-term	Wear-in	Wear-out
PriceNB	−161,794	−84,417	0	8
PriceST	71,561	121,997	0	1
AfternoonNB	32,129	32,129	0	0
TrustDiff	23,707	23,707	0	0
QualDiff	0	66,963	1	0
LiketriedNB	0	40,420	1	0
Tried4wNB	0	72,481	2	4
EntertainNB	0	60,071	2	3
SatistriedNB	0	46,788	2	0
AwareUnaidedNB	34,164	68,537	3	4

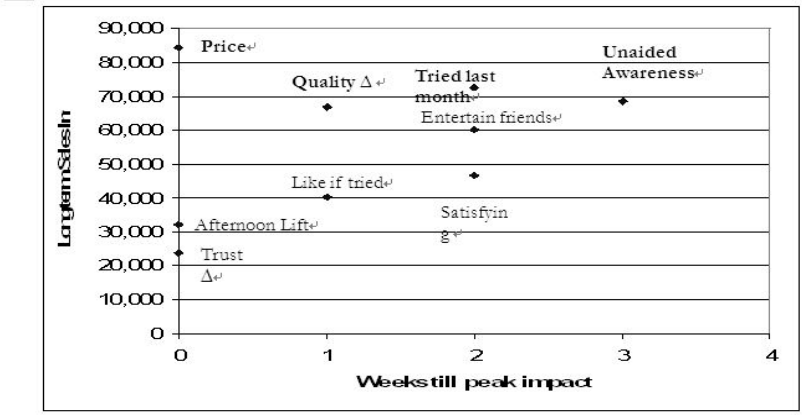
* Changes to sales units resulting from a one-standard-deviation change in the dashboard metric. "Wear-in" is the number of weeks it takes until the peak sales impact is reached (with 0 meaning the peak impact is immediate), and "wear-out" is the number of weeks with significant effects after peak impact. The variables are ordered first by impact wear-in time, then by impact size (absolute value).

For short-term effects, both the sign and the relative magnitude of the effects are in line with expectation from previous research: own price has the largest immediate impact on sales, followed by competitive price. In the long run however, store-brand price has a larger effect on national-brand sales than the national brand's own price! This reversal is driven by both the post-promotion dips after

national-brand price discounts and by the slow decay of the harm that store-brand price cuts inflict on national-brand sales. The likely reason is increased consumer price sensitivity, which hurts long-term sales of the more expensive brand, consistent with Wathieu et al.(2004), Pauwels and Srinivasan(2004) and Van Heerde et al.(2008). As a result, national-brand managers are correct to worry about the growth of store brands, and they have a very limited ability to turn the situation around with price changes.

In contrast, the large effects of the other metrics point to several levers national-brand managers can pull. First, the national brand derives a strong immediate benefit from its trust gap with the store brand and from national-brand usage as an afternoon lift. The immediate effect of these survey metrics is only two to three times smaller than the competitive price effects and may be influenced directly by national-brand marketing communication. After one week's delay, the quality difference between the national brand and store brand is important, followed by the metric "liking if tried the national brand." Thus, the national brand's excellent quality and consumers' emotions associated with the brand are main weapons in the battle with store brands, consistent with recent discussions in literature(Kumar and Steenkamp 2007).

After a two week delay, the metrics "tried the national brand in the last four weeks", "national-brand usage to entertain friends", and "national brand is satisfying" become important. Free samples and other trial inducements may complement marketing communication focusing on how perfect the national brand is for entertaining friends and how satisfying it is to eat. Priming consumers to consume the product in public benefits the national brand, consistent with its dominance in usage occasions with a higher social expressive or sign value(McCracken 1986). Finally, the importance of unaided awareness, at three weeks' delay, suggests that even popular brands still must ensure that the brand name comes to mind first when the consumer is thinking about the product category. Overall, national-brand managers can influence 9 out of the 10 metrics selected in the dashboard to some extent (with store-brand price being the only exception). 〈Figure 2〉 visualizes these findings by juxtaposing the size and timing of the effects of each of the nine dashboard metrics that are (at least partially) under managerial control.



<Figure 2> The size and timing of the effects of each of the nine dashboard metrics

For store brand sales, 〈Table 8〉 displays the size and timing of the impact of each metric. Consistent with asymmetric price effects (e.g., Blattberg and Wisniewski 1989), the store brand has a smaller own-price elasticity (−0.56) and is relatively more affected by competitive price. However, the long-term own-price elasticity increases to −1.77 due to positive adjustment effects. In other words, price changes by the store brand appear to increase consumer price sensitivity, which helps the less expensive brand in the case of price promotions (Wathieu et al. 2004).

<Table 8> Size and Timing of Each Dashboard Metric's Effect on Store-brand Sales*

	Short-term	Long-term	Wear-in	Wear-out
PriceST	−44,641	−140,880	0	3
PriceNB	37,637	1,735	0	1
Costdiff	25, 328	25, 328	0	0
TVdiff	20,544	20,544	0	0
Relaxdiff	−14,075	−14,075	0	0
Sportdiff	−13,627	−8,368	0	7
Likeawarenb	−25,678	−86,383	1	2
Awareunst	31,662	87,793	2	1
Awareunnb	0	−60,579	2	0
Awaretomdiff	−21,606	−43,322	4	0

* Changes to sales units resulting from a one standard deviation change in the dashboard metric. "Wear-in" is the number of weeks it takes until the peak sales impact is reached (with 0 meaning the peak impact is immediate), and wear-out is the number of weeks with significant effects after peak impact. The variables are ordered first by impact wear-in time, then by impact size (absolute value).

New results emerge for the effect of the remaining metrics. Store-brand sales immediately benefit from a larger perceived cost difference with the national brand and from increased usage while watching TV (also relative to the national brand). These results reflect consumers' primary motivations for buying the store brand: lower cost and private (i.e., non-social) consumption (Kumar and Steenkamp 2007; McCracken 1986). In contrast, consumers' relative usage of the product while relaxing and watching sports on TV is negatively associated with store-brand sales. These usage occasions appear closely related to the "afternoon lift" and "entertaining friends" metrics that benefit national-brand sales. With one week's delay, consumer affect for the national brand (liking given awareness) exerts a strong negative effect on store-brand sales. This metric is not under the control of store-brand managers, though they may attempt to find ways around the national brand's hold on consumer hearts. In the longer run, there are three awareness metrics that are important for store-brand sales and that reflects the store brand's status at the bottom of the customer-based brand equity pyramid (Keller 2003) and the continuing need to make the target market aware of its existence. With two weeks' delay, unaided awareness of the store brand helps (and unaided awareness for the national brand hurts) store-brand sales. Finally, with four weeks' delay, relative top-of-mind awareness affects store-brand sales. Again, most metrics are at least partially under control of store-brand managers (arguably including the price of the national brand, which is set by the retailer). The two exceptions are unaided awareness and affect for the national brand.

Finally, 〈Table 9〉 presents the effect size and timing results for revenue premium. When revenue premium is the key performance metric national brand managers should take special care in managing both the actual price premium and the cost perception vis-à-vis the store brand. Quality perceptions and liking given awareness provide a boost to the revenue premium. Interestingly, the magnitudes of these effects are dwarfed by that of metrics that operate with at least a week delay. Top-of-mind awareness for the national brand provides a huge benefit, while customer positive feelings for and especially preference (favoritism) for the store brand are harmful. As for usage occasions, the national brand should promote the product

as an afternoon snack. Increases of customer associations of the national brand with this usage benefit the revenue premium. Interestingly, with a longer delay, customer associations of the store brand with afternoon snack usage also boost the price premium. This indicates that this usage occasion triggers increased attention to the full category in the store, at which point many consumers buy the national brand.

<Table 9> Size and Timing of Each Dashboard Metric's Effect on Revenue Premium*

	Short-term	Long-term	Wear-in	Wear-out
PriceNB	-2,301,017	-1,430,495	0	2
PriceST	2,343,032	4,448,196	0	1
Costnb	-19,634	-19,634	0	0
Qualnb	13,397	13,397	0	0
Likeawarenb	9,626	9,626	0	0
Awaretomnb	39,972	81,193	1	0
Feelst	-3,134	-8,210	1	0
Afternoonnb	4,912	14,891	1	0
Favoritest	0	-77,402	2	0
Afternoonst	0	10,890	2	1

* Changes to revenue premium resulting from a one standard deviation change in the dashboard metric. "Wear-in" is the number of weeks it takes until the peak revenue premium impact is reached (with 0 meaning the peak impact is immediate), and wear-out is the number of weeks with significant effects after peak impact. The variables are ordered first by impact wear-in time, then by impact size (in absolute value).

VI. Conclusions

As more marketing data are available to managers, and as managers are held more accountable for marketing efforts' outcomes, tracking the right metrics becomes increasingly important. This study addresses this important issue by proposing a general empirical framework for metric selection in marketing dashboards. Based on past academic research and managerial insights, this research lays out a step-by-step approach that can enable managers to populate a dashboard across a range of industries and situations. The research demonstrates how a combination of various statistical techniques can identify a manageable list of metrics that are leading indicators

of brand performance. For both the analyzed national brand and store brand, the most useful procedure combines a preliminary reduction of metrics (achieved in this case through Granger causality tests), which creates a shortlist of dashboard variables, with econometric methods that establish relationships between those metrics and identify the ones with the greatest impact on performance.

This study proposes and tests a method, with an application in a fast moving consumer goods category, for reducing a large number of metrics (99 in this case) to a smaller number (10) most suitable for a dashboard predicting sales performance for two very different brands, along with the revenue premium for the leading brand. Though the focus is on the usefulness of different selection tools in the process of creating the dashboard (i.e. "how to fish"), this research also generate valuable managerial insights through the selection of variables in this specific category.

Limitations in this work point to avenues for future research. While the selection procedures discussed here themselves generalize, their forecasting performance will vary across situations, so replication across time, countries and industries is required. The method introduced here clearly performs well for the product category and performance metrics in the case study at hand, but the application to a dataset with a wider range of product categories is a promising area for future research. From marketing theory and practice, one would expect rather different leading indicators across various product categories, for example, for drugs (physician attitudes, patient trial, number of new prescriptions), office furniture providers (information requests, bid requests and bids won), or financial investment services (financial product performance, lead generation, retention). Second, this study lacks information on marketing's power to affect these leading indicators, nor of the costs of their data collection. Nowadays though, online data collection allows for more cost effective and faster data gathering on many candidate metrics, such as coverage in consumer reviews, blogs and social media.

Likewise, the specific application only contained sales and revenue premium information, not firm earnings or stock market capitalization. Using the approach described above, connecting to such firm-level financial performance is straightforward when these data are available, as demonstrated by Pauwels

et al.(2004). Third, other variable selection techniques should be developed for and/or applied to dashboard creation, to improve forecasting accuracy over and above that obtained by the methods discussed in this paper. Such methods could also be suitable to analyze time series of short duration for companies that have not (yet) gathered candidate metric data for a long time. Finally, regular updates of marketing dashboards will be required, based on the changing predictive power and managerial usefulness of the metrics(La Pointe 2005). Separating a calibration and a holdout sample, as done here, is just a start. With longer datasets, rolling windows estimation enables managers to drop less relevant older data while adding new information(Pauwels and Hanssens 2007). The frequency of major updates is likely higher for turbulent times and industries versus more mature ones, which offers additional opportunities for future research.

In conclusion, this research represents a first step in developing methodologically robust and practically relevant techniques for reducing a large number of candidate metrics to a few leading indicators that may be included in a marketing dashboard. Our hope is that selecting such predictive metrics in an objective way(Stewart 2008) will help marketing managers to "realize full accountability and strategic status in the Boardroom as reliable forecasters and achievers of consistent growth(Blair 2007).

<Recived: 17 August 2011>
<Revision Recived: 15 October 2011>
<Final Version Received: 23 October 2011>

References

Ailawadi, Kusum, Donald R. Lehmann, and Scott A. Neslin(2003), "Revenue Premium as an Outcome Measure of Brand Equity," Journal of Marketing, 67(October), 1-17.

Aldrin, Magne(2002), "Reduced Rank Regression," Encyclopedia of Environmetrics, Volume 3, 1724-1728.

Ambler, Tim(2003), "Marketing and the Bottom Line," second edition, Financial Times Prentice Hall.

Anderson, T. W.(1951), "Estimating linear restrictions on regression coefficients for multivariate normal distributions," Annals of Mathematical Statistics, 22, 327-351.

Armstrong, Scott J.(2001), Principles of Forecasting: A Handbook for Researchers and Practitioners, International Series in Operations Research & Management Science, 30, Springer.

Blair, Meg(2007), "What will be different? The MASB Vision," August 16-17, presentation available at: http://www.themasb.org/

Blattberg, Robert C. and Kenneth J. Wisniewski(1989), "Price-Induced Patterns of Competition," Marketing Science, 8(4), 291-309.

Blattberg, Robert C., Byung-Do Kim and Scott Neslin(2008), Database Marketing: Analyzing and Managing Customers, International Series in Quantitative Marketing, Springer.

Churchill, Gilbert A. Jr(1979), "A Paradigm for Developing Better Measures of Marketing Constructs," Journal of Marketing Research, 16(1), 64-73.

Clark, Bruce H., Andrew V. Abela, and Tim Ambler(2006). "Behind the Wheel," Marketing Management, 15(3), 18-23.

CMO Council(2004), Chief Marketing Officers Council's Marketing Measures Performance Audit, White Paper.

Dekimpe, M. and D. Hanssens(1999), "Sustained Spending and Persistent Response: A New Look at Long-Term Marketing Profitability," Journal of Marketing Research, 36(November), 397-412.

Draper, N. and Smith, H.(1981), Applied Regression Analysis, 2nd Edition, New York: John Wiley & Sons, Inc.

Granger, Clive W.(1969), "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," Econometrica, 37(3), 424-438.

Granger, Clive W. and Paul Newbold(1986), Forecasting Economic Time Series, 2nd edition, Harcourt Brace Jovanovich, NY.

Hanssens, Dominique M.(1998), "Order Forecasts, Retail Sales and the Marketing Mix for Consumer Durables," Journal of Forecasting, 17, 327-346.

Hanssens, Dominique M., Leonard J. Parsons and Randall L. Schultz(2001), Market

Response models: Econometric and Time Series Analysis. 2nd ed, Kluwer Academic Publishers.

Hocking, R. R.(1976) "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32(1), 1-49.

Hyde, Paul, Edward Landry and Andrew Tipping(2004), "Making the Perfect Marketer," Strategy+Business, Winter, Booz Allen Hamilton, http://www. strategy-business.com/press /16635507/04405.

Izenman(1975), "Reduced-Rank regression for the multivariate linear model," Journal of Multivariate Analysis, 5, 248-264.

Judd, Charles M. and Gary H. McClelland(1989) Data Analysis: A Model Comparison Approach" Sand Diego: Harcourt Brace Jovanovich.

Keller, Kevin Lane(2003), "Conceptualizing, Measuring, and Managing Customer-Based Brand Equity," Journal of Marketing, 57(January), 1-22.

Kohavi R.(1995), "A Study of Cross -Validation and Bootstrap for Accuracy Estimation and Model Selection," In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada. San Francisco: Morgan Kaufmann, 1137-1143.

Krauss, Michael(2005), "Marketing Dashboards Drive Better Decisions," Marketing News, 39(16), 1.

Kumar, Nirmalya. and Jan-Benedict Steenkamp (2007), Private Label Strategy: How to meet the Store Brand Challenge, Harvard Business School Press.

Kumar, V., Rajkumar Venkatesan and Werner Reinartz(2006), "Knowing what to sell, when, and to whom," Harvard Business Review, 84(3), 131-7, 150.

Lapointe, P.(2005). "Marketing by the Dashboard Light," Marketing NPV/Association of National Advertisers.

LaPointe, Pat(2006), "For Better ROI, Think Sailing, Not Driving," Brandweek, 47(5), 17-18.

Lavidge, Robert J. and Gary A. Steiner(1961), "A Model for Predictive Measurement of Advertising Effectiveness," Journal of Marketing, 25, 59-62.

Leamer, Edward(1985), "Sensitivity Analyses Would Help," The American Economic Review, 75(3), 308-313.

Leeflang, Peter S. H. and Jan C. Reuyl(1984), "On the Predictive Power of Market Share Attraction Models," Journal of Marketing Research, 21(2), 211-215.

Lehmann, Donald R. and David J. Reibstein(2006), "Marketing Metrics and Financial Performance," Marketing Science Institute Monograph.

Levin, Nissan and Jacob Zahavi(1998), "Continuous Predictive Modeling – a

Comparative analysis," Journal of Interactive Marketing, 12(2), 5-22.

Little, John D. C.(1979), "Decision Support Systems for Marketing Managers," Journal of Marketing, 43(3), 9-26.

MarketingNPV(2005), "Timken Rolls Out a Marketing Dashboard for Industrial Bearing Group", 3(1), 3-6.

Marketing Science Institute(2006), 2006-2008 Research Priorities: A Guide to MSI Research Programs and Procedures, Cambridge, Mass.

McCracken, Grant(1986), "Culture and consumption: A theoretical account of the structure and movement of the cultural meaning of consumer goods," Journal of Consumer Research, 13, 71-84.

Miller, A.(1989), Subset Selection in Regression, London: Chapman and Hall.

Meiri, Ronen and Jacob Zahavi(2005), "And The Winner Is…Stepwise Regression," Extended Summary of the paper presented at the Direct Marketing Education Foundation Conference, http://www.the-dma.org/dmef/proceedings05 /AndtheWinner-Meiri.pdf

Mintzberg, H.(1973), "The Nature of Managerial Work," New York: Harper & Row, Myers and Mullet.

Naert, Philippe A. and Marcel Weverbergh (1981), "On the Prediction Power of Market Share Attraction Models," Journal of Marketing Research, 18(2), 146-153.

Neslin, Scott, Sunil Gupta, Wagner Kamakura, Junxiang Lu and Charlotte H. Mason(2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," Journal of Marketing Research, 43(May), 204-211.

Nijs, Vincent R., Marnik G. Dekimpe, Jan-Benedict E.M. Steenkamp, and Dominique M. Hanssens(2001), "The Category Demand Effects of Price Promotions," Marketing Science, 20(1), 1-22.

Nijs, Vincent, Shuba Srinivasan, Koen Pauwels(2007), "Retail-Price Drivers and Retailer Profits," Marketing Science, 26(4), 473-487.

Palda, Kristian S.(1966), "The Hypothesis of a Hierarchy of Effects: A Partial Evaluation," Journal of Marketing Research, 3(1), 13-24.

Pauwels Koen(2004), "How Dynamic Consumer Response, Competitor Response, Company Support and Company Inertia Shape Long-term Marketing Effectiveness," Marketing Science, 23(4), 596-610.

Pauwels, Koen, Tim Ambler, Bruce Clark, Pat LaPointe, David Reibstein, Bernd Skiera, Berend Wierenga, and Thorsten Wiesel(2009), "Dashboards as a Service: Why, What, How and What Research is Needed?," Journal of Service Research, 12(2), 175-189.

Pauwels, Koen, Dominique Hanssens, and S. Siddarth(2002), "The Long-term Effects of Price Promotions on Category Incidence, Brand Choice, and Purchase Quantity," Journal of Marketing Research, 34(November), 421-439.

Pauwels, Koen, Jorge Silva-Risso, Shuba Srinivasan and Dominique M. Hanssens(2004), "New Products, Sales Promotions and Firm Value: The Case of the Automobile Industry," Journal of Marketing, 68(October), 142-156.

Pauwels, Koen and Shuba Srinivasan(2004), "Who benefits from store brand entry?," Marketing Science, 23(3), 364-390.

Petersen, Andrew J., Leigh McAlister, David J. Reibstein, Russell S. Winer, V. Kumar and Geoff Atkinson(2009), "Choosing the Right Metrics to Maximize Profitability and Shareholder Value," Journal of Retailing, 85, 1, 95-111.

Pesaran, Hashem H. and Yongcheol Shin(1998), "Generalized Impulse Response Analysis in Linear Multivariate Models," Economic Letters, 58(1), 17-29.

Ray, Michael L., Sawyer, Alan G., Rothschild, Michael L., Heeler, Roger M., Strong, Edward C., and Reed, Jerome B.(1973), "Marketing Communications and the Hierarchy of Effects," in New Models for Mass Communication Research, ed. Peter Clarke. Beverly Hills, CA: Sage Publishing, 147-76.

Reibstein, David, David Norton, Yogesh Joshi, and Paul Farris(2005), "Marketing Dashboards: A Decision Support System for Assessing Marketing Productivity," presentation at the Marketing Science Conference, Atlanta, GA.

Reinsel, Gregory C. and Raja P. Velu(1998), "Multivariate Reduced-Rank Regression," Springer-Verlag. Roecker, Ellen B.(1991), "Prediction Error and Its Estimation for Subset-Selected Models," Technometrics, 33(4), 459-468.

Rust, R. T., Ambler, T., Carpenter, G. S., Kumar, V., and Srivastava, R. K.(2004), "Measuring Marketing Productivity: Current Knowledge and Future Directions," Journal of Marketing, 68(4), 76-89.

Rust, Roland T. and Henderson, Pamela W.(1985), "Should Emotional Advertising Response Models Reflect Brain Physiology?," Proceedings of the American Academy of Advertising, NR204-7.

Simon(1973), "Applying information technology to organization design," Public Administration Review, 268-278.

Slotegraaf, Rebecca and Koen Pauwels (2008), "The impact of brand equity and innovation on the long-term effectiveness of promotions," Journal of Marketing Research, 45(June), 293-306.

Smith, Robert E. and William R. Swinyard(1983), "Attitude-Behavior Consistency: The Impact of Product Trial versus

Advertising," Journal of Marketing Research, 20(3), 257-267.

Srinivasan, Shuba, Koen Pauwels and Vincent Nijs(2008), "Demand-Based Pricing Versus Past-Price Dependence: A Cost-Benefit Analysis," Journal of Marketing, 72(March), 15-27.

Srinivasan, Shuba, Koen Pauwels and Marc Vanheule(2009), "Do Mindset Metrics Explain Brand Sales?," Marketing Science Institute, Report 08-119.

Steckel, Joel and Wilfried Vanhonacker(1993), "Cross-Validation Regression Models in Marketing Research," Marketing Science, 12(4), 415-427.

Stewart, David W.(2008), "How Marketing Contributes to the Bottom Line," Journal of Advertising Research, March, 94-205.

Stone, M. and R.J. Brooks(1990), "Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression," Journal of the Royal Statistical Society. Series B (Methodological), 52(2), 237-269.

Van Bruggen, Gerrit H. and Berend Wierenga(2000), "Broadening the Perspective on Marketing Decision Models," International Journal of Research in Marketing, 17(2-3), 159-168.

Van Heerde, Harald, Els Gijsbrechts and Koen Pauwels(2008), "Winners and Losers in a Major Price War," Journal of Marketing Research.

Wathieu, Luc, A.V. Muthukrishnan, and Bart J. Bronnenberg(2004), "The Asymmetric Effect of Brand Positioning on Post-Promotion Preference," Journal of Consumer Research, 31(3), December, 652-657.

Webster, Jr., Frederick E., Alan J. Malter and Shankar Ganesan(2005), "The Decline and Dispersion of Marketing Competence," Sloan Management Review, 46(4), 35-43.

Wierenga, B., Van Bruggen G. H., and Staelin, R.(1999). "The Success of Marketing Management Support Systems," Marketing Science, 18(3), 196-207.

Wilkinson, G. N.(1961) "Statistical estimation in enzyme kinetics," Biochemical Journal, 90, 324-332.

Wind, Yoram(2005), "Marketing as an engine of business growth: a cross-functional perspective," Journal of Business Research, 58(7), 863-873.

Technical Appendix

Econometric model specification, estimation and forecasting accuracy evaluation

Stepwise Regression

Stepwise regression allows automatic selection of a limited set of regressors, based on various statistical criteria(Hocking 1976; Draper and Smith 1981). The researcher selects a number of regressors that are always present (in our case, a trend and seasonal patterns) and the procedure whereby added regressors will be considered. We apply several versions: unidirectional, stepwise, and combinatorial, each in its forward and backward version. All versions yield the same set of metrics in our empirical analysis. Stepwise regression offers the benefits of computational efficiency and the automatic selection of variables. However, its pitfalls are well documented(Roecker 1991): inflated R^2, F-statistics, and p-values (due to pretesting issues) and severe problems in the presence of collinearity, which are not much alleviated with increased sample size. Due to the risk of capitalizing on chance features of the data, statisticians strongly suggest demonstrating the explanatory power of the selected variables in a holdout sample(Judd and McClelland 1989) and deplore that they are rarely tested that way. As a result of its drawbacks, stepwise regression all but disappeared from top marketing journals in the 1990s.

However, recent applications in direct marketing demonstrated that stepwise regression performs well out-of-sample, both in predicting direct mail response(Meiri and Zahavi 2005) and customer churn(Neslin et al. 2006). We note though that both papers focus on predicting tactical performance in the form of 0/1 dummy variables, for which logistic regression is the appropriate modeling choice(Neslin et al. 2006). In contrast, marketing dashboards focus on strategic performance variables such as sales and profits, whose continuous nature yields a different set of competing models. These relate to two drawbacks of (stepwise) regression: it does not directly utilize the co-relatedness of explanatory variables, and it does not address the endogeneity among candidate metrics and performance measures.

Reduced-rank Regression

Reduced-rank regression (hereafter RRR) directly addresses the co-related-

ness of candidate metrics with the assumption of a lower rank in the matrix of regression coefficients (Anderson 1951; Reinsel and Velu 1998). RRR can be considered a generalized version of the more common canonical analysis (Izenman 1975). Indeed, if the covariance structure is known, results from RRR and canonical analysis or partial least squares (PLS) will be identical. However, while PLS analysis results in factors, RRR can be used to map the factors to original variables (with the analysis providing both the coefficient and standard error of the mapping) and thereby to select a small number of metrics that can best explain the dependent variable . Thus, if the classical linear regression is represented as:

$$Y_i = X_i' C + \epsilon_i \quad (1)$$

where Y_i is m x 1 and X_i is n x 1, then, for RRR, the matrix C (m x n) is assumed to have a rank r such that:

$$\text{rank (C)} = r \leq \min(m, n). \quad (2)$$

This implies that there exist (m-r) linear restrictions on the coefficient matrix C.[ii] The solution to the RRR problem is obtained by minimizing the well-known linear regression criterion under Restriction 2. Thus, if the solution to the classical linear regression model can be represented as:

$$B_{OLS} = \sum_{k-1}^{n} B_k \quad (3)$$

where the first term (k=1) is the rank-one matrix that explains the maximum possible variance in Y_i, the second term (k=2) is the rank-two matrix, and so on; then the solution to the RRR problem can be represented as:

$$B_{RRR} = \sum_{k-1}^{n} B_k \quad (4)$$

RRR has traditionally been applied as a shrinkage regression(Aldrin 2002), with the intent of reducing the number of estimated coefficients when insufficient observations are available. However, this method applies equally well to our task of selecting the best combination of explanatory variables from a set of correlated variables. To the best of our knowledge, this is its first application to solve a marketing problem.

Forecast Variance Error Decomposition

While the RRR methodology is uniquely suited to address our problem of correlated explanatory variables, it

does not account for the endogeneity among candidate metrics and performance measures, nor can it easily handle complicated dynamic interactions. These weaknesses are addressed by vector autoregressive models and the derived forecast error variance decomposition (FEVD) of the performance variable. In essence, FEVD provides a measure of the relative impact over time of shocks initiated by each of the candidate metrics on the performance variable (Hanssens 1998; Nijs et al. 2007; Srinivasan et al. 2008). Analogous to a "dynamic R^2," it calculates the percentage of variation in the performance variable that can be attributed to both contemporaneous (i.e. same-week) and past changes in each of the candidate metrics. As standard in FEVD applications (Ibid), we allow the FEVD up to 26 lags to settle on the dynamic percentage of performance variation that is explained by a particular variable.

Previous literature is mute on a cut-off rule for candidate metrics that explain too little in the FEVD of performance. As a rule of thumb, we propose to drop candidate metrics that explain less than 2% of the forecast error variance of performance in any of the 26 lags considered. Moreover, from each pair of highly correlated metrics (more than 0.5 in absolute value), we propose to drop the variable that explains the lesser amount of the forecast error variance.

Estimation procedure

For all three econometric models, we start with the set of variables that Granger Cause the performance variable, and then investigate on a rotating basis whether other variables should instead be included in the final set of 10 metrics. Our choice to start with this set of variables reflects the sequential nature of our research design and our research purpose to select leading sales indicators. However, it may reduce the potential benefits of Forecast Error Variance Decomposition compared to the other methods, because only FEVD usually includes Granger causality tests into its selection of model variables(Hanssens et al. 2001). Indeed, the computational gain from this procedure is especially important for Reduced Rank Regression, which is infeasible to implement with several tens of metrics. The very high correlation of relative unaided awareness with national brand unaided awareness (0.98) requires us to exclude one of the two variables from analysis (also on a rotating basis), leaving us with 16

Granger Causing metrics for store brand sales.

Out-of-sample evaluation of predictive accuracy

Consistent with recommendations and standard practice(Kohavi 1995; Steckel and Vanhonacker 1993), we use two-thirds of our sample for model estimations and the remaining one-third for the holdout to assess predictive accuracy. We estimate the same ordinary least squares regression model for each set of metrics in order to maintain comparability. After estimating this model on the estimation sample, we also create a dynamic forecast for the third year. This procedure yields a comparable assessment of the predictive validity of each set of metrics, irrespective of the model used to select them. An important choice is how many lags to include of a selected metric. Several metrics (e.g., awareness) can be expected to have their largest sales impact several weeks after they themselves change. We choose to start with four weeks of lags and use specification searches to investigate whether more lags are necessary. Moreover, in our base model, we choose to exclude the contemporaneous (same-week) effects of each metric. Knowing the impact of a metric at the same time as performance is only informative for diagnostic purposes, not for predictive purposes. For example, managers may be interested to learn that sales went down this week because consumer awareness went down this week, but can do more with the knowledge that sales will go down next week because awareness went down this week. Thus, predictive accuracy without contemporaneous effects reveals the leading-indicator nature of the selected metrics (see Neslin et al. 2006 for similar arguments). In sum, we estimate the model in Equation 5 for each of the three sets of metrics:

$$BrandSales_t = \alpha + \delta^* t + \sum_{i=2}^{13} r_i^* SD_i + \sum_{j=1}^{10}\sum_{k=1}^{4} \beta_{jk}^* Metric_{j,t-k} \quad (5)$$

with t representing a time trend and SD representing 4-weekly dummies to account for seasonal factors. We examine the robustness of our results to researcher choices on the predictive versus diagnostic nature of methods and forecasting criteria. First, we amend the stepwise and reduced rank regression method by entering up to 13 lags of each candidate metric in the metric selection

procedure. This alleviates the original non-dynamic nature of the methods and allows them to identify metrics that take some time before affecting performance. Second, we expand equation (5) by including the contemporaneous performance effects of the selected metrics. This corresponds to the more typical specification of market response models. As for forecasting accuracy criteria, we compute the root mean squared error, mean absolute error, mean absolute percentage error, and Theil's inequality coefficient. The latter measure is scale-invariant and thus generally preferable when comparing the predictive ability of competing models (e.g., Naert and Weverbergh 1981; Ghosh, Neslin, and Shoemaker 1984; Leeflang and Reuyl 1984).