

Spring 6-10-2017

GENERATING CONSUMER INSIGHTS FROM BIG DATA CLICKSTREAM INFORMATION AND THE LINK WITH TRANSACTION-RELATED SHOPPING BEHAVIOR

Daniel Schellong

RWTH Aachen University, Aachen, Germany, schellong@time.rwth-aachen.de

Jan Kemper

RWTH Aachen University, Aachen, Germany, kemper@time.rwth-aachen.de

Malte Brettel

RWTH Aachen University, Aachen, Germany, brettel@time.rwth-aachen.de

Follow this and additional works at: http://aisel.aisnet.org/ecis2017_rp

Recommended Citation

Schellong, Daniel; Kemper, Jan; and Brettel, Malte, (2017). "GENERATING CONSUMER INSIGHTS FROM BIG DATA CLICKSTREAM INFORMATION AND THE LINK WITH TRANSACTION-RELATED SHOPPING BEHAVIOR". In Proceedings of the 25th European Conference on Information Systems (ECIS), Guimarães, Portugal, June 5-10, 2017 (pp. -). ISBN 978-989-20-7655-3 Research Papers.
http://aisel.aisnet.org/ecis2017_rp/24

This material is brought to you by the ECIS 2017 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

GENERATING CONSUMER INSIGHTS FROM BIG DATA CLICKSTREAM INFORMATION AND THE LINK WITH TRANSACTION-RELATED SHOPPING BEHAVIOR

Research Paper

Schellong, Daniel, RWTH Aachen University, Aachen, Germany,
schellong@time.rwth-aachen.de

Kemper, Jan, RWTH Aachen University, Aachen, Germany, kemper@time.rwth-aachen.de

Brettel, Malte, RWTH Aachen University, Aachen, Germany, brettel@time.rwth-aachen.de

Abstract

E-Commerce firms collect enormous amounts of information in their databases. Yet, only a fraction is used to improve business processes and decision-making, while many useful sources often remain underexplored. Therefore, we propose a new and interdisciplinary method to identify goals of consumers and develop an online shopping typology. We use k-means clustering and non-parametric analysis of variance tests to categorize search patterns as Buying, Searching, Browsing or Bouncing. Adding to purchase decision-making theory we propose that the use of off-site clickstream data—the sequence of consumers’ advertising channel clicks to a firm’s website—can significantly enhance the understanding of shopping motivation and transaction-related behavior, even before entering the website. To run our consumer data analytics we use a unique and extensive dataset from a large European apparel company with over 80 million clicks covering 11 online advertising channels. Our results show that consumers with higher goal-direction have significantly higher purchase propensities, and against our expectations - consumers with higher levels of shopping involvement show higher return rates. Our conceptual approach and insights contribute to theory and practice alike such that it may help to improve real-time decision-making in marketing analytics to substantially enhance the customer experience online.

Keywords: clickstream analysis, online consumer journey, big data, e-commerce

1 Introduction

Synonymous with the rise of B2C purchases shifting online (\$1.6 trillion or 7% of total retail spending in 2015), marketers are moving more and more advertising budgets into digital channels and formats – today almost 30% of \$600 billion worldwide total media spend is already digital (eMarketer, 2015). Developments in information technology and internet analytics have provided practitioners and researchers with an unprecedented ability to track and analyze consumer choices according to their individuals’ shopping journey touch points in great detail (e.g., Winer 2009; Kauffman et al. 2012). But naturally, in addition to its many promises, big data also raises new challenges – most prominently online companies struggle to use the vast amounts of available consumer data to systematically create actionable marketing insights on an ongoing basis (Bharadwaj et al., 2013). Therefore, it is ever more success-critical to generate relevant insights out of consumer and market data to understand the drivers and characteristics of online shopping in order to improve marketing decision-making leading to more conversions and ultimately more loyal customers (Bucklin and Sismeiro, 2009; Lambrecht and Tucker, 2013; Yadav and Pavlou, 2014).

One of the main aims of our study is to find out if online shopping types can already be identified before the consumer enters the website. To address this, we run a data-driven cluster analysis approach to develop a typology of online shoppers by operationalizing off-site clickstream metrics – more specifically via the level of consumer involvement and consumer search behavior. We use large-scale off-site clickstream data—the sequence of advertising channel clicks towards a website over time—as it provides information on the exposure and effects of online advertising on consumers and their subsequent transaction-related behavior (Li and Kannan, 2014; Nottorf and Funk, 2013).

Existing research has shown that purchasing behavior may differ depending on the visiting pattern of the individual in question (Moe and Fader, 2004; Montgomery et al., 2004; Sismeiro and Bucklin, 2004). Bonfrer and Dreze (2009) in their study on e-mail performance posit that in future applications and with better data availability, optimizing purchase behavior should be possible through insights on the link of clicks with purchase behavior per user type. Therefore, we conjecture that at the underlying link between shopping type affiliation and purchase propensity is manifested in the navigation path the user takes in the way that the path unveils the user's underlying search goal. The empirical results and insights from our study can help marketers and advertisers to better manage their consumer traffic online by understanding the role advertising channel choice plays in e-commerce purchase and post-purchase behavior in order to substantially enhance the customer experience online, e.g., by personalizing the website design and more targeted and relevant advertising campaigns.

From a marketer and advertiser standpoint, off-site advertising click path data offers multiple opportunities to generate insights but has not been looked at as part of a consumer behavior study (Bucklin and Sismeiro, 2009). Therefore, we follow several calls to use clickstream data to infer shopping goals of users—finding out if the user is planning to make a purchase, or just in for exploratory reasons in order to understand consumer decision-making (Bucklin et al., 2002; Rohm and Swaminathan, 2004). In order to leverage the wealth of available information, big data needs to be managed intelligently requiring new methods and diagnostics to filter out the relevant information to subsequently turn event data into valuable insights for marketing decision-making (Leeflang et al., 2014; Lilien, 2011), allowing e-commerce firms to make more accurate and predictive forecasts in order to more profitably target and market to customers.

Our research contributes to theory and practice in at least three ways. First, we extend the literature on the interlink between consumer behavior and advertising effectiveness research (Ganesh et al., 2010; Moe, 2003; Rohm and Swaminathan, 2004) by presenting a novel shopping type framework developed by and interdisciplinary metrics from clickstream data. Second, we generate new empirical consumer insights through development of significant shopping type clusters, and provide strong evidence for clear link with transaction-related metrics, extending related prior studies on effects on purchase (Van den Poel and Buckinx, 2005; Sismeiro and Bucklin, 2004), and post-purchase behavior (Bechwati and Siegal, 2005; Kang and Johnson, 2009). Third, our large-scale data-driven approach shows how to develop new methods to analyze existing online advertising channel data to generate consumer insights for individual-level and dynamic marketing decision-making.

2 Conceptual Framework and Shopping Typology Development

We seek to advance purchase decision-making theory as part of the overall consumer journey online in operationalizing consumer shopping involvement and user search behavior to develop specific shopping types by empirically analyzing patterns of individual-level off-site clickstream data. More concretely, we analyze how consumers react to advertising appeals in different advertising channels and collect all revisits to an online retailer's website within a 30-day timeframe preceding a potential purchase or non-purchase decision (see Data section for more detailed information on the definition off-site clickstream). For illustrative purposes, Figure 1 presents an example of an off-site clickstream.

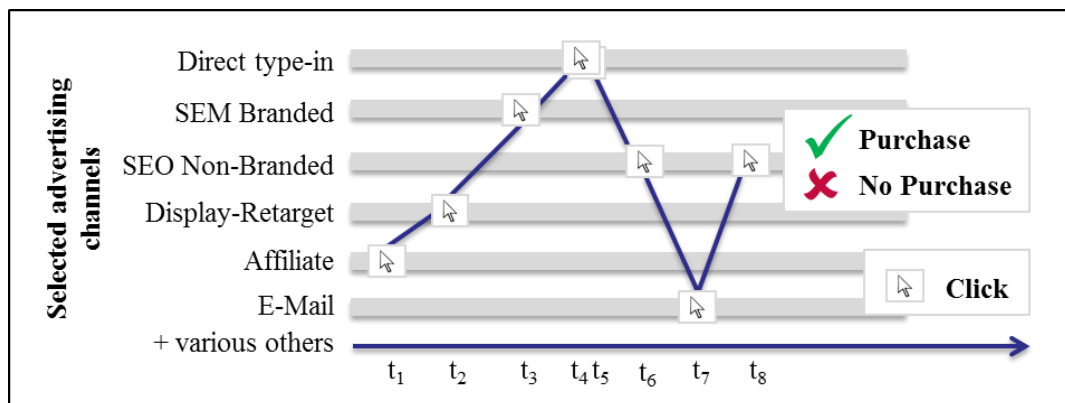


Figure 1. Example of an off-site clickstream

In this example, the first touch point was a click on an affiliate link at t_1 , which subsequently forwarded the user to the advertiser's landing page or product detail page. At t_2 , the user clicked on a display retargeting banner advertisement (content of the ad based on previous browsing behavior on the website), before searching for a branded keyword string on a search engine and clicking the respective sponsored search advertisement on which the e-commerce firm was bidding on. The user continues his shopping process on various other channels until ultimately choosing whether to purchase or not at t_8 .

The remainder of this chapter is structured as follows, we describe our online consumer shopping type framework based on two dimensions and present the respective theoretical grounding thereof. After that, we specify the characteristics of the emerging shopping types before discussing the expected off-site clickstream patterns and the anticipated effect on purchase and post-purchase behavior.

2.1 Online Shopping Motivation

Consumer behavior literature highlights the importance to better understand online shopping motivation for effective marketing campaigns (Bettman et al., 1991; Janiszewski, 1998). Each consumer runs through multiple decision stages and browses through various online channels, e-commerce websites and multiple brands and product detail pages on his path to purchase (Hauser and Wernerfelt, 1990). However, to the best of our knowledge, little attention has been paid to the inherent intention of consumers to select specific advertising channels to access online shops.

In this context, the question concerning how shopping goals can be conceptualized using real data forms a major challenge to scholars and practitioners and has not been researched much apart from a few studies (Klapdor et al., 2015; Moe, 2003; Puccinelli et al., 2009). Existing research on consumer shopping goals looks at consumers' general orientation and specific goals, mostly based on experimental or survey data (Ganesh et al., 2010; Rohm and Swaminathan, 2004). A key element of this paper is exploring the possibilities on how to uncover search types of browsing and shopping behavior by defining and analyzing off-site clickstream metrics to offer new insights into consumer behavior online.

To address this issue, we follow the general research framework from Moe (2003), which was just recently replicated for off-site clickstream data by Schellong et al. (2016). Furthermore, we use a much larger clickstream data set and combine it with all relevant purchase and post-purchase information.

2.2 Typology Dimensions

The conceptual framework of our typology is based on two dimensions, consumer shopping involvement, and consumer search behavior.

2.2.1 Shopping Involvement

Consumer shopping involvement reflects the degree to which the user is involved with the selected website helping to form his consideration set, pursuing information acquisition activities about a brand or product while considering alternatives before making a purchase or non-purchase (Puccinelli et al., 2009; Zaichkowsky, 1985). “Depending on their level of involvement, individual consumers differ in the extent of their decision process and their search for information” (Laurant and Kapferer, 1985). Therefore we recap, the amount of information necessary to satisfy a purchasing or information goal is determined by the shopping involvement level.

The broad and popular term referred to as engagement is a phenomenon that subsumes varied forms of user interaction and involvement with media and is constituted of psychological as well as behavioral elements (Brodie et al., 2011; Mollen and Wilson, 2010). Calder et al. (2009) argue that engagement is antecedent to shopping behavior outcomes such as usage, affect, and responses to advertising. Our research on involvement most prominently falls into the classification of website engagement, or more concretely engagement employing internet websites, and shopping involvement is a dimension of website engagement (Hyder, 2015). Involvement and more specifically purchase-decision involvement can be defined as a state of mental readiness that influences the “the extent of interest and concern that a customer brings to bear on a purchase decision task” (Mittal, 1989, p. 150). Therefore, we use visits and the duration of each clickstream as a proxy to elicit the level of involvement per shopping type.

2.2.2 Search Behavior

Following Janiszewski (1998), we differentiate between goal-directed and exploratory search as forms of consumer search behavior.

That’s why we introduce a unique and interdisciplinary approach to operationalize between goal-directed and exploratory search in integrating information retrieval insights to categorize advertising channels for the clustering analyses. Based on Broder's (2002) work on classifying web search queries regarding the users’ degree of goal-direction, we follow a similar approach as Klapdor et al. (2015) and categorize advertising channels in informational or navigational (see metrics section for more details on the classification per channel). This categorization matches to what we want to extract out of advertising channel choice patterns – the specific user's goal when entering an e-commerce website in order to categorize shopping types.

To summarize, the level of shopping involvement (frequency and horizon) and search behavior in the form of type of channels used (navigational or informational) characterize the framework of shopping types. We distinguish between four specific shopping types: Buying, Searching, Browsing, and Bouncing.

Table 1 summarizes the framework of online consumer shopping types.

Consumer Shopping Involvement	Consumer Search Behavior	
	Directed	Exploratory
High Involvement	Buying	Browsing
Low Involvement	Searching	Bouncing

Table 1. Online consumer shopping types

2.3 Shopping Types

In clustering shopping types of users it is important to note that we segment modes rather than people; given that a consumer can have multiple shopping modes or types over time - e.g., a consumer can be a Browsing Type in one month and a Buying Type in the next month. Based on the theoretical deriva-

tion of both dimensions from our shopping type framework, we expect the selected off-site clickstream metrics (consumer involvement and search behavior) to show the following sign characteristics as presented in Table 2.

	Shopping Involvement		Search Behavior	Purchase Behavior	Post-Purchase Behavior	
	Channel Involvement Frequency	Channel Involvement Horizon	Channel Focus	Conversion Rate	Basket Size	Return
	High/Mid/Low	Short/Mid/Long	Navigational/Informational	High/Mid/Low	High/Mid/Low	High/Mid/Low
Buying Type	High	Short	Navigational	High	High	Low
Searching Type	Mid	Mid	Navigational	Mid	Mid	Mid
Browsing Type	Mid	Long	Informational	Mid	High	Low
Bouncing Type	Low	Short	Informational	Low	Mid	Mid

Table 2. Expected Online Shopping and Purchase Pattern

When visiting the e-commerce store, the “Buying Type” (Directed Search/High Involvement) has a specific shopping goal in mind, namely obtaining information regarding a product or the need to purchase it in a relatively short timeframe. Being a “Buying” shopping type is also typified by frequent involvement within a rather short time horizon, collecting the most relevant shopping information to satisfy the need to make a purchase or non-purchase decision.

Similar to the aforementioned Buying Type, the “Searching Type” (Directed Search/Low Involvement) is also goal-directed. Not as focused as the Buying Type, the Searching Type is in the process of forming a consideration set in order to satisfy his information or shopping need, with the difference in the level of involvement. However, the Searching Type is not as focused, being in the process of forming a consideration set to satisfy his information or shopping need, albeit with a different involvement level: the Searching Type is involved with a medium frequency during a mid-horizon.

The “Browsing Type” (Exploratory Search/High Involvement) differs from both previous types in the way that Browsing is characterized by an exploratory search behavior. Accordingly, the search behavior is rather unplanned and without a specific utilitarian goal in mind, which can also be called experiential shopping: the online shopping experience by themselves is enjoyment or entertainment. Furthermore, a Browsing Type tends to focus more on informational channels, given that they offer more inspiration from a broad set of e-commerce websites.

The “Bouncing Type” (Exploratory Search/Low Involvement) is exploratory in the way that if engaged in a shopping process it is very much likely that the channels used are unfocused and that his involvement level is low. Bouncing translates into a single-visit with no further interaction or returns to the e-commerce website. Another explanation for the occurrence of a Bouncing Type is, as the name states, consumers that come to the site and leave the site (bounce) right after - this can be due to the fact that the consumer didn’t find what he was browsing for.

2.4 Expected Effects on Transaction-Related Behavior

Our data set allows us to connect each individual off-site clickstream with the resulting transaction events in case of a purchase. Most existing research on advertising effectiveness uses click-through-rates on advertising campaigns as success variable, our path data allows us to further monitor if, when and to which degree advertising appeals lead consumers to convert at the website. Furthermore, we analyze post-purchase events in the form of return behavior.

2.4.1 Purchase Behavior Effects

Based on purchase decision-making theory several studies argue that there is profound evidence that frequent visitors to a store will also eventually make a purchase at some point in time (Laurant and Kapferer, 1985; Murray and Häubl, 2007). Janiszewski (1998) renowned works on search behavior

theory analyzing store routines also follows this argumentation. Therefore, we argue that the more focused a consumer's search process the higher the likelihood to purchase. To map this to our shopping types, we would expect the Buying Type to have the highest conversion rate due to having a purchase goal in mind when engaging with the website, followed by the Searching Type that is also goal-directed. The Browsing Type undergoes more exploratory shopping should have a lower conversion rate because they do not have a purchasing goal in mind but non-zero due to the occurrence of infrequent impulse purchases. In sum, we argue the more directed a shopping type, the higher the conversion propensity. Regarding the size of the basket, in terms of items and order value, the argumentation follows consumer shopping involvement theory (see theoretical grounding in typology dimension section). We speculate that the higher the level of shopping involvement, the larger the basket size. Buying and Browsing shopping types should have higher order sizes than the Searching Type. For the Browsing Type this is due to the fact that if a purchase occurs as part of the consumer journey, it is likely to be an unplanned or impulse purchase (Koufaris et al., 2002). Hedonistic shopping leads to increased irrational purchasing, especially for non-utilitarian goods like fashion (Childers et al., 2001).

2.4.2 Post-Purchase Effects

Bechwati and Siegal (2005) examine the link between pre-choice process and post-choice product returns and find clear evidence that the generation of different thoughts (nature of cognitive responses and the type of disconfirming information) at the pre-choice stage result in a different likelihood for product returns. Adding to this line of argumentation Kang and Johnson (2009) posit in their study that purchases who are not accompanied by a rational in-depth evaluation before making a purchase, called impulsive or hedonistic consumption, have a higher likelihood to experience post-purchase regret and hence higher return propensities compared to consumers who undergo an extensive evaluation of alternatives before making a purchase or non-purchase. Therefore we argue that consumers who undergo a thorough brand and product consideration process – depicted in our study by shopping involvement frequency and horizon – will try to make sure that their purchase decision has been conducted in a sound and systematic way in order to decrease the possible reversal of their buying decision (Laurant and Kapferer, 1985). In other words, high shopping involvement leads to less frequent product returns compared to consumers with low or mediocre shopping involvement. In regards to shopping types we would expect the Buying Type to have a lower, and the Searching Type a higher average return rate and amount of goods returned.

3 Data

This research paper makes use of a large-scale data set covering almost 30 million clickstream-based on over 81 million advertising channel clicks provided by our partner firm, a leading European online-only fashion retailer (context setting). Our research partner utilizes a broad range of available online channels – most notably display reach, display retargeting, paid search (SEM), organic search (SEO), comparison, affiliate, social networks and e-mail campaign links.

The unit of analysis of our study are off-site clickstream journeys, which are based on granular cookie information and represent the sequence of advertising channel clicks through which consumers visit and convert at the firm's website (Li and Kannan, 2014; Nottorf and Funk, 2013). In other words, off-site clickstreams represent the chronological sequence of advertising channel choice clicks that lead to subsequent visits on the respective website(s) under investigation (for all users, not just existing customers). In this context "Off-site" does not mean across-site data - information of users between and on other (potentially competitive) sites (Bucklin and Sismeiro, 2009). We define a converting off-site clickstream as follows: If the consumer has not yet placed an order within a 30-day timeframe (average cookie lifetime and industry-standard), every subsequent visit within this timeframe is an extension of the initiated shopping process. If the user journey ultimately leads to a purchase, it forms a converting clickstream, as reflected by 5.6% of cases in our sample (not to be mistaken with the "standard" conversion rate metric, which is usually based on visits, and should, therefore, be lower).

By contrast, if no conversion occurs within a 30-day period (from the first advertising channel click), all visits made within this period represent a non-converting clickstream.

The large clickstream data set was obtained from the retailer's web server log file database that directly captures each and every click for all web properties of the partner firm. Clickstream data are stored in semi-structured website log files and include details such as timestamp of each request, cookie information that uniquely identifies the website visitor (e.g., IP information and internet provider), the source link with information on the respective advertising type clicked (e.g., keyword information from search engine marketing), and the destination URL. The web server information then needs to be matched with the respective channel information characteristics.

The extensive data set consists of a total of 81,412,696 rows of advertising channel clicks resulting in 29,939,213 off-site clickstream journeys. The observation period covers the full months of March and April of 2014, therewith covering the most non-seasonal months of the year (the month March and April do typically not hold any large sale or new season stock periods). Since our research partner is active in most European countries, our data set is based on advertising consumer clicks from Germany, Italy, Poland, Sweden and the United Kingdom.

4 Consumer Journey Metrics and Methodology

4.1 Categorizing Shopping Behavior

Because consumers have different reasons for visiting a retail site, it is important to understand and account for various patterns in the relationship between visiting and purchasing (Moe and Fader, 2004). Studies have shown that in many cases, consumers build up to a purchase (Moe, 2001; Putsis and Srinivasan, 2014). In other words, consumers will make a series of non-purchase visits before making a purchase visit. In order to find patterns in clickstream data, we follow Schellong et al. (2016) and define several clickstream metrics to operationalize consumer shopping behavior in order to link shopping types of consumers with purchase and post-purchase behavior.

Table 3 provides a description of the metrics used.

Metrics	Unit	Description of Metrics
Involvement Frequency metric		
CLICKS	in # clicks	Total number of channel clicks to the e-commerce website
Involvement Horizon metrics		
TOTDURATION	in days	Time between first and last channel click of customer clickstream journey
CLICKGAP	in days	Average time of gaps in between each visit of clickstream journey
Channel Focus metrics		
NAVISHARE	in %	Share of navigational channel clicks as part of clickstream journey
INFOSHARE	in %	Share of informational channel clicks as part of clickstream journey
Channel Variety metric		
UNIQUECH	in #	Number of different channels used during clickstream journey
Conversion metric		
CONVERSION	in %	Indicator variable: "1" if CS ends with a purchase, "0" otherwise
Purchase Behavior metrics		
SOLDITEMS	in days	Number of items purchased per order
GROSSSALES	in days	Total value purchased per order
Post-Purchase Behavior metrics		
RETURN	in %	Indicator variable: "1" if order contains at least one returned item, "0" otherwise
RETURNEDITEMS	in # items	Number of items returned per order
GROSSRETURN	in € value	Total value returned per order
Onsite metric		
PURCHASESESSIONDUR	in min	Session time: Time between purchase channel click and actual purchase time

Table 3. Summary of clickstream measures

We operationalize shopping involvement via frequency and horizon metrics, therewith providing insights in the involvement level of consumers based on the number of clicks and length of each clickstream. This offers a strong understanding of the individual's involvement level with the e-commerce website for the specific time period. **CLICKS** reflect the frequency of clicks (or visits) per clickstream, independent of the type of channels used. **TOTDURATION** is the overall length of the clickstream, up to 30 days in duration. **CLICKGAP** another involvement metric represents the average time in between each click, also called intra-visit times. Taken together the involvement metrics reveal the consumer's shopping involvement level and information acquisition motivation. The High/Mid/Low scale in Table 2 reflects the level of involvement, ranging from high to low.

Next to the level of shopping involvement, we depict search behavior based upon the user's shopping goal via the respective choice of advertising channel, thus enabling us to identify directed vs. exploratory shopping types. While our research partner uses 11 different advertising channels (including direct type-in), using each channel as a separate variable would make the calculations complex and the number of attribute variables (columns) would grow exponentially (course of dimensionality). Based on Anderl et al. (2016) recent work we consider and control for most existing channel-categorizations that have been used in multi-channel advertising research include contact origin, separating between customer- and firm-initiated channels (Haan et al., 2013; Li and Kannan, 2014), and branded vs. generic channels (Jansen et al., 2011; Rutz and Bucklin, 2012). However, we rely on the classification logic developed in information retrieval research on user intention and purchase decision-making theory by Broder (2002). Seeking to understand the underlying goal of search, he proposes to categorize web search according to navigational and informational queries. Applying this approach to our data set to deduce users' goals, we define a channel interlink with search goal affiliation – classifying every advertising channel as navigational or informational.

Table 4 provides an overview of the online advertising channels and respective goal categorization.

	Shopping Goal Categorization
Channel	Navigational vs. Informational
CRM (E-Mail, etc.), Display - Retarget, DTI (Direct Type-In), SEM Branded, SEO Branded, Social Media Owned	Navigational
Affiliate, Comparison, Display - Reach, SEM Non-Branded, SEO Non-Branded	Informational
Note. No Offline Channels (such as TV or Print) included	

Table 4. Goal categorization of online marketing channels

The share of advertising channel clicks that are navigational or informational per clickstream is reflected by the metrics **NAVISHARE** for navigational clicks and **INFOSHARE** for informational clicks, describing the degree of goal-direction in terms of website choice per clickstream. If the consumer is certain about his website choice as part of his shopping process, **NAVISHARE** should be higher, whereas **INFOSHARE** should be higher if the consumer has a broad consideration set rather than a clear website goal in mind.

In order to validate the link of off-site clickstreams with purchase behavior, we were able to collect information if an actual **CONVERSION** occurred within the respective time period. Further, if a purchase has been made we have all the information regarding the amount of items sold (**SOLDITEMS**) and value (**GROSSSALES**) per order. To see how our shopping type clusters relate to post-purchase, specifically return behavior, we know if any item of an order has been returned (**RETURN**), how many items (**RETURNEDITEMS**) and for how much value (**GROSSRETURN**).

Table 5 shows the descriptive statistics for each selected clickstream metric.

Metrics	Unit	Mean	SD	Min	Max
Involvement Frequency metric					
CLICKS	in # clicks	2.770	4.870	1.000	8954.000
Involvement Horizn metrics					
TOTDURATION	in days	4.727	8.746	0.000	30.000
CLICKGAP	in days	1.136	2.408	0.000	15.000
Channel Focus metrics					
NAVISHARE	in %	0.511	0.469	0.000	1.000
INFOSHARE	in %	0.364	0.449	0.000	1.000
Channel Variety metric					
UNIQUECH	in #	1.488	0.928	1.000	12.000
Conversion metric					
CONVERSION	in %	0.057	0.232	0.000	1.000
Purchase Behavior metrics					
SOLDITEMS	in # items	1.000	0.932	0.000	17.929
GROSSSALES	in € value	1.000	0.950	0.000	34.079
Post-Purchase Behavior metrics					
RETURN	in %	1.000	0.866	0.000	1.749
RETURNEDITEMS	in # items	1.000	1.403	0.000	28.345
GROSSRETURN	in € value	1.000	1.427	0.000	57.582
Onsite measures					
PURCHASESESSIONDUR	in min	26.739	332.139	0.000	42927.700
Note: Full dataset comprises 29,392,629 clickstreams based on a total of 81,412,696 clicks from March and April of 2014. For the clustering procedure all non-binary involvement and channel metrics have been winsorized (5% of outliers on both ends have been replaced with the value of the 95%-quantiles respectively). All purchase and return behavior metrics have been indexed with the mean value (=1) for each variable respectively.					

Table 5. Descriptive statistics of clickstream, purchase and return behavior

4.2 Methodology

We conduct a data-driven analysis of the customer online journey by understanding the role advertising channel choice plays in e-commerce purchase behavior. The goal of this paper is to identify and analyze online consumer shopping type patterns and their effects on transaction behavior based on search behavior and involvement metrics. To address this, we decided not to apply a predictive modeling or data mining approach to our clickstream data set, since framework and metric development are an important step by itself. Hence, we decided to use cluster analysis for our research, in particular, we use k-means clustering algorithm for our large clickstream data set. We selected clustering for several reasons: a) clustering methodology is frequently used in marketing research, in particular in the field of segmentation and behavioral studies (Wedel and Kamakura, 2000) and remains a useful tool in online consumer behavior research – especially in the field of clickstream research (e.g., Moe 2003; Dias and Vermunt 2007), b) k-means partitioning based clustering works especially well for large data sets (Wedel and Kamakura, 2000), and c) we address the call from practitioners as well as marketing research to develop new and rather simple methods to analyze online data for direct implementation (Lilien, 2011). Most importantly, we build on and extend the proven concept initially developed by Moe (2003), who has developed an approach using clustering to segment onsite behavior patterns.

Based on our large data set (almost 30 million clickstreams), we apply k-means partitioning-based cluster analysis to segment shopping types based on advertising channel choice metrics. Cluster methodology segments search types based on a set of specified clickstream variables – in our case CLICKS, TOTDURATION, CLICKGAP, NAVISHARE, INFOSHARE, and UNIQUECH. All values are winsorized to set all outliers to the 95%-percentile in order to correct for both ends of the data but keep the values in the data set. Furthermore, all non-binary variables are standardized for the clustering procedure using z-scores, in order to have similar significance in the clustering procedure. We observed no multicollinearity issues between the variables (no correlation larger than 0.6). The full correlation and covariance tables are available upon request.

K-means, being an a priori clustering method needs a pre-defined number of clusters before running the segmentation (Ter Hofstede et al., 2002). For the segmentation process, we use the clustering algorithm developed by Hartigan and Wong (1979). To come up with the most optimal number of clusters we run the algorithm 20 times for a range of realistic number of cluster solutions and compare the local minima values of the sum of squared distance errors (SSE), and keep the solution providing the best fit to the data (Wedel and Kamakura, 2000). We plot the values and look for the “elbow” criterion—number of clusters at which the SSE decreases abruptly—indicating the optimal number of clusters (Wedel and Kamakura, 2000). Figure 2 visualizes the individual results per number of clusters, which nicely shows the trend of SSE over a different number of clusters, providing the graphical basis for the selection of the 4-cluster solution.

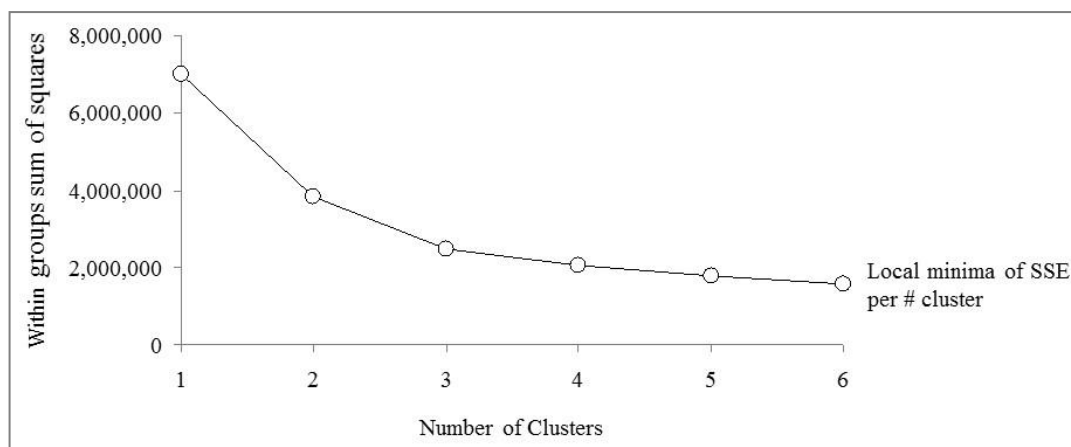


Figure 2. Plot of within cluster variance of cluster solutions

In order to research the link with transaction-related variables we calculate the mean values per cluster of the final cluster solution and compare the differences (incl. robustness checks) – this process is called profiling (Wedel and Kamakura, 2000). For the clustering procedure, we use the statistics software R Version 3.0.2 (via RStudio Version 0.99.486) and run the calculations in a AWS EC2 cloud-based instance set-up, as available computational resources locally were not enough in running the statistical calculations on our big data click-stream data set in an efficient manner.

5 Results

Table 6 shows the optimal 4 cluster solution.

Cluster	Unit	1	2	3	4	Kruskal-Wallis-Test*
		BUYING	SEARCHING	BROWSING	BOUNCING	
N	in CS	2,942,244	3,936,826	3,318,507	19,195,052	
	in %	10.0%	13.4%	11.3%	65.3%	
Involvement Frequency metric						
CLICKS	in # clicks	7.941	3.248	3.176	1.196	***
Involvement Horizn metrics						
TOTDURATION	in days	19.851	2.476	19.656	0.160	***
CLICKGAP	in days	2.187	0.746	5.622	0.066	***
Channel Focus metrics						
NAVISHARE	in %	0.710	0.596	0.597	0.449	***
INFOSHARE	in %	0.287	0.393	0.362	0.370	***
Channel Variety metric						
UNIQUECH	in #	2.690	2.142	1.813	1.000	***

Note. N = 29,392,629. All shopping involvement and channel metrics were used for the k-means clustering procedures. All non-binary measures have been winsorized (5% of outliers on both ends have been replaced with the value of the 95%-quantiles respectively). For the clustering procedures all metrics have also been scaled (centering variables and creating respective z-scores) in order to guarantee standardized variable distances for the clustering algorithm.
* = p < .05, ** = p < .01, *** = p < .001

Table 6. Result of Cluster Solution

With the cluster solution we can answer the key research question if there are distinct search characteristics that can be found in advertising channel choice data indicating the predominant shopping type. These findings build initial evidence that off-site clickstream data can greatly contribute to the understanding of individual consumer behavior online. Findings on distinct search characteristics are also significant enough to assume that off-site clickstream data are an excellent source for real-time user goal identification. The shopping types and respective patterns do overall validate our framework presented in Table 1.

The Cluster 1 “Buying Type” shows the highest number of clicks (7.9) and the highest share of navigational channels (71%), with average clickstream duration of almost 20 days. This is in line with our expected pattern except for the shopping involvement horizon which was estimated to be comparably lower since a directed shopping type has a clear goal in mind and wants to satisfy his need in a rather short timeframe. A possible explanation for this could be that most Buying Types are frequent customers who are loyal to the website and continuously fill their wish list until they at some point convert on the website on a regular basis (19.4% conversion rate – see Table 7 for all purchase effects).

The Cluster 2 “Searching Type” utilizes on average 3.2 advertising channels as part of their clickstream journey during a timeframe of 2.4 days. With a utilitarian task in mind gathering information to narrow the consideration set process with more navigational (0.6) than informational channels (0.4).

The Cluster 3 “Browsing Type” being an exploratory shopping mode with a rather long average clickstream time period (19.6), converting or not converting on an average number of 3.1 visits. The Browsing and Searching Types are quite similar aside from the length of the clickstream (19.6 Browsing compared to 2.4 days for the Searching Type). This points to a possible explanation that the classification between the two is more of a continuum rather than a binary decision.

The Cluster 4 “Bouncing Type” which covers almost two-thirds of all clickstreams of our study (65.3%) has the lowest level of involvement (frequency and duration) compared to the other Types as well as the lowest navigational share (0.4) and by far the lowest share of converting clickstreams (2.6%). This cluster shows similar characteristics as an “Others” cluster and most likely entails users that regularly delete cookies which makes it impossible to track journeys or consumers.

To check for the robustness of the cluster solution we run a non-parametric Kruskal-Wallis-Test (Hollander and Wolfe, 1973) for each variable with the result of no indication of non-significance. After evaluating the results of the cluster solution we analyze the effects of the four shopping types in regards to transaction-related variables by profiling the cluster solution with the respective purchase and post-purchase metrics. Table 7 shows the cluster results for the shopping types by transaction-related variables.

Cluster	Unit	1 BUYING	2 SEARCHING	3 BROWSING	4 BOUNCING	
N (CONVERSION = 1)	in CS	571,370	490,694	387,419	222,301	Kruskal-Wallis- Test*
	in %	34.2%	29.4%	23.2%	13.3%	
Conversion metric						
CONVERSION	in %	0.194	0.098	0.067	0.026	***
Purchase Behavior metrics						
SOLDITEMS	in # items	1.044	0.932	1.107	0.954	***
GROSSSALES	in € value	1.048	0.950	1.076	0.949	***
Post-Purchase Behavior metrics						
RETURN	in %	1.026	0.957	1.070	0.971	***
RETURNEDITEMS	in # items	1.069	0.901	1.158	0.927	***
GROSSRETURN	in € value	1.083	0.920	1.117	0.914	***
Onsite measures						
PURCHASESESSIONDUR	in min	0.851	0.921	1.079	1.200	***
Note. N = 1,671,784. All purchase and return behavior metrics have been indexed with the mean value (=1) for each variable respectively. * = p < .05, ** = p < .01, *** = p < .001						

Table 7. Result of cluster solution profiling

We can confirm our assumption regarding the positive correlation between goal-directed search behavior and conversion propensity, which is in line with Moe (2003). The directed shopping types Buying and Searching show higher conversion rates (19.4% and 9.8% respectively) compared to the exploratory shopping type Browsing (6.7%). In regards to the basket metrics, number of items and value of order, we argue that the higher the level of involvement the larger the average shopping cart of customers. We find the same pattern in the results: Buying and Browsing show higher order values compared to the Searching and Bouncing Types.

Regarding post-purchase behavior, the data rejects our argumentation based on findings from Bechwati and Siegal (2005) and Kang and Johnson (2009) – the higher the level of involvement per shopping type the lower the return probability of consumers. Based on our results the Browsing Type has the highest percentage of returns (1.07 – values indexed), followed by the Buying (1.03) and Searching Type (0.96), therewith opposing our initial assumption based on shopping involvement theory. We argue that the findings can potentially be attributed to the effect that impulse purchases satisfying hedonic needs quickly lose their value after receiving the order, and eventually be returned to retailers (Kang and Johnson, 2009; Rook, 1987).

6 Implications and Discussion

This paper delivers valuable input for academia and practice and adds to the fields with at least three contributions: First, our results offer e-commerce firms consumer behavior insights to better manage customer journeys as well as increase online marketing effectiveness. We are the first study using off-site clickstream data outside of advertising effectiveness and attribution studies. Therewith serving as an initial application example and building a starting point for further research on optimizing consumer journey touch points.

Second, we extend the existing literature on diverging effects on purchasing behavior depending on the visiting pattern (Moe and Fader, 2004; Montgomery et al., 2004; Sismeiro and Bucklin, 2004) by presenting our empirical results: We find that shoppers with directed search behavior show higher purchasing propensities, the same is true for shopping types with higher level of shopping involvement. Our results further reveal, surprisingly, that shopping types with a high level of involvement show higher levels of returns compared to the other shopping types.

Third, despite the potential clickstream data analysis holds for research and practice in marketing, these data are almost surely being underutilized (Bucklin and Sismeiro, 2009; Kauffman et al., 2012). We show that advertising channel choice behavior is a valid and unexplored source to generate consumer insights – this should stimulate more research with consumer journey data. Furthermore, we contribute to the request in marketing research to develop new and rather simple methods to analyze online data for direct practicability – therewith convincing practitioners without in-depth statistical knowledge to implement the methods and actually apply the insights (Lilien, 2011).

As our work is based on a single-company research approach, studies using data from different companies, industries or customer groups would be of high interest. Enhancing the understanding of the consumer journey would further be the opportunity to investigate clickstream activities by product characteristics (e.g., search or experience goods) or product price. Furthermore, understanding if there are cross-cultural effects in online advertising click behavior would help multi-national companies in forming their international online marketing strategies. After having identified shopping types of consumers and understanding the link with transaction-related behavior, a logical next step for e-commerce companies would be to run A/B tests to find out which marketing activities work best for each type and adjust their automated marketing systems (e.g., online marketing real-time bidding systems) and dynamic website systems accordingly.

References

- Anderl, E., Schumann, J.H. and Kunz, W. (2016), "Helping Firms Reduce Complexity in Multichannel Online Data: A New Taxonomy-Based Approach for Customer Journeys", *Journal of Retailing*, Vol. 92 No. 2, pp. 185–203.
- Bechwati, N.N. and Siegal, W.S. (2005), "Process on of the Prechoice The Impact Returns Product", *Journal of Marketing Research*, Vol. 42 No. 3, pp. 358–367.
- Bettman, J.R., Johnson, E.J. and Payne, J.W. (1991), "Consumer decision making", Handbook of consumer behavior.
- Bharadwaj, A., El Sawy, O. a., Pavlou, P. a. and Venkatraman, N. (2013), "Digital Business Strategy: Toward a Next Generation of Insights", *MIS Quarterly*, Vol. 37 No. 2, pp. 471–482.
- Bonfrer, A. and Dreze, X. (2009), "Real-time evaluation of e-mail campaign performance", *Marketing Science*, Vol. 28 No. 2, pp. 251–263.
- Broder, A. (2002), "A taxonomy of web search", *ACM SIGIR Forum*, Vol. 36 No. 2, p. 3.
- Brodie, R.J., Hollebeek, L.D., Juric, B. and Ilic, A. (2011), "Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research", *Journal of Service Research*, Vol. 14 No. 3, pp. 252–271.
- Bucklin, R.E., Lattin, J.M., Ansari, A., Gupta, S., Bell, D., Little, J.D.C., Mela, C., et al. (2002), "From Clickstream Choice and the Internet: to Stream", *Marketing Letters*, Vol. 13, pp. 245–258.
- Bucklin, R.E. and Sismeiro, C. (2009), "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing", *Journal of Interactive Marketing*, Vol. 23 No. 1, pp. 35–48.
- Calder, B.J., Malthouse, E.C. and Schaedel, U. (2009), "An Experimental Study of the Relationship between Online Engagement and Advertising Effectiveness", *Journal of Interactive Marketing*, Direct Marketing Educational Foundation, Inc., Vol. 23 No. 4, pp. 321–331.
- Childers, T.L., Carr, C.L., Joann, P. and Carson, S. (2001), "Hedonic and utilitarian motivation for online retail shopping behavior", *Journal of Retailing*, Vol. 77, pp. 511–535.
- Dias, J.G. and Vermunt, J.K. (2007), "Latent class modeling of website users' search patterns: Implications for online market segmentation", *Journal of Retailing and Consumer Services*, Vol. 14, pp. 359–368.
- eMarketer. (2015), Worldwide Retail Ecommerce Sales.
- Ganesh, J., Reynolds, K.E., Lockett, M. and Pomirleanu, N. (2010), "Online Shopper Motivations, and e-Store Attributes: An Examination of Online Patronage Behavior and Shopper Typologies", *Journal of Retailing*, New York University, Vol. 86 No. 1, pp. 106–115.
- Haan, E. De, Wiesel, T. and Pauwels, K. (2013), "Which Advertising Forms Make a Difference in Online Path to Purchase?", Marketing Science Institute Working Paper Series, Vol. 13 No. 13.
- Hartigan, J. a. and Wong, M. a. (1979), "A K-Means Clustering Algorithm", *Journal of the Royal Statistical Society*, Vol. 28 No. 1, pp. 100–108.
- Hauser, J.R. and Wernerfelt, B. (1990), "An Evaluation Cost Model of Consideration Sets", *Journal of Consumer Research*, Vol. 16 No. 4, p. 393.
- Ter Hofstede, F., Wedel, M. and Steenkamp, J.-B.E.M. (2002), "Identifying Spatial Segments in International Markets", *Marketing Science*, Vol. 21 No. 2, pp. 160–177.
- Hollander, M. and Wolfe, D. (1973), *Nonparametric Statistical Methods*, John Wiley & Sons, New York.
- Howard, J.A. and Sheth, J.N. (1969), *The theory of buyer behavior*, Wiley, New York.
- Hyder, A. (2015), *Customer Engagement: Contemporary Issues and Challenges*, Taylor & Francis.
- Janiszewski, C. (1998), "The Influence of Display Characteristics on Visual Exploratory Search Behavior", *Journal of Consumer Research*, Vol. 25 No. December, pp. 290–301.
- Jansen, B.J., Sobel, K. and Zhang, M. (2011), The Brand Effect of Key Phrases and Advertisements in Sponsored Search, *International Journal of Electronic Commerce*, Vol. 16.

- Kang, M. and Johnson, K. (2009), "Identifying characteristics of consumers who frequently return apparel", *Journal of Fashion Marketing and Management*, Vol. 13 No. 1, pp. 37–48.
- Kauffman, R.J., Srivastava, J. and Vayghan, J. (2012), "Business and data analytics: New innovations for the management of e-commerce", *Electronic Commerce Research and Applications*, Vol. 11 No. 2, pp. 85–88.
- Klapdor, S., Anderl, E., Schumann, J.H. and von Wangenheim, F. (2015), "How to Use Multichannel Behavior To Predict Online Conversions", *Journal of Advertising Research*, Vol. 55 No. 4, pp. 433–442.
- Koufaris, M., Kambil, A. and Labarbera, P.A. (2002), "Consumer Behavior in Web-Based Commerce : An Empirical Study", *International Journal of Electronic Commerce*, Vol. 6 No. 2, pp. 115–138.
- Lambrecht, A. and Tucker, C. (2013), "When Does Retargeting Work? Information Specificity in Online Advertising", *Journal of Marketing Research*, American Marketing Association, Vol. 50 No. 5, pp. 561–576.
- Laurant, G. and Kapferer, J.-N. (1985), "Consumer Involvement Profiles", *Journal of Marketing Research*, Vol. 22 No. 1, pp. 41–53.
- Leeflang, P.S.H., Verhoef, P.C., Dahlström, P. and Freundt, T. (2014), "Challenges and solutions for marketing in a digital era", *European Management Journal*, Vol. 32 No. 1, pp. 1–12.
- Li, H. and Kannan, P.K. (2014), "Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment", *Journal of Marketing Research*, Vol. 51 (February No. 1, pp. 40–56.
- Lilien, G.L. (2011), "Bridging the Academic–Practitioner Divide in Marketing Decision Models", *Journal of Marketing*, Vol. 75 No. 4, pp. 196–210.
- Mittal, B. (1989), "Measuring Purchase-decision involvement", *Psychology and Marketing*, Vol. 6 No. 2, pp. 147–162.
- Moe, W.W. (2003), "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream", *Journal of Consumer Psychology*, Vol. 13 No. 1–2, pp. 29–39.
- Moe, W.W. and Fader, P.S. (2004), "Dynamic Conversion Behavior at E-Commerce Sites", *Management Science*, Vol. 50 No. 3, pp. 326–335.
- Moe, W.W. and Fader, P.S. (2001), "Uncovering Patterns in Cybershopping", *California Management Review*, Vol. 43 No. 4, pp. 106–177.
- Mollen, A. and Wilson, H. (2010), "Engagement, telepresence and interactivity in online consumer experience: Reconciling scholastic and managerial perspectives", *Journal of Business Research*, Vol. 63 No. 9–10, pp. 919–925.
- Montgomery, A.L., Li, S., Srinivasan, K. and Liechty, J.C. (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data", *Marketing Science*, Vol. 23 No. 4, pp. 579–595.
- Murray, K.B. and Häubl, G. (2007), "Explaining Cognitive Lock-In: The Role of Skill-Based Habits of Use in Consumer Choice", *Journal of Consumer Research*, Vol. 34 No. June, pp. 77–88.
- Nottorf, F. and Funk, B. (2013), "The economic value of clickstream data from an advertiser's perspective", In: *Proceedings of the 21st European Conference on Information Systems*.
- Van den Poel, D. and Buckinx, W. (2005), "Predicting online-purchasing behaviour", *European Journal of Operational Research*, Vol. 166 No. 2, pp. 557–575.
- Puccinelli, N.M., Goodstein, R.C., Grewal, D., Price, R., Raghubir, P. and Stewart, D. (2009), "Customer Experience Management in Retailing: Understanding the Buying Process", *Journal of Retailing*, Vol. 85 No. 1, pp. 15–30.
- Putsis, W. and Srinivasan, V. (2014), "Buying or Just Browsing? The Duration of Purchase Deliberation", *Journal of Marketing Research*, Vol. 31 No. 3, pp. 393–402.
- Rohm, A.J. and Swaminathan, V. (2004), "A typology of online shoppers based on shopping motivations", *Journal of Business Research*, Vol. 57 No. 7, pp. 748–757.
- Rook, D.W. (1987), "The Buying Impulse", *Journal of Consumer Research*, The Oxford University

- Press, Vol. 14 No. 2, p. 189.
- Rutz, O.J. and Bucklin, R.E. (2011), "From Generic to Branded: A Model of Spillover in Paid Search Advertising", *Journal of Marketing Research*, Vol. 48 No. 1, pp. 87–102.
- Rutz, O.J. and Bucklin, R.E. (2012), "Does banner advertising affect browsing for brands? Clickstream choice model says yes, for some", *Quantitative Marketing and Economics*, Vol. 10 No. 2, pp. 231–257.
- Schellong, D., Kemper, J. and Brettel, M. (2016), "Clickstream data as a source to uncover consumer shopping types in a large-scale online setting", In: *Proceedings of the 24th European Conference on Information Systems, ECIS 2016*.
- Sismeiro, C. and Bucklin, R.E. (2004), "Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach", *Journal of Marketing Research*, Vol. 41 No. 3, pp. 306–323.
- Wedel, M. and Kamakura, W. (2000), *Market Segmentation: Conceptual and Methodological Foundations*, Kluwer Academic Publishers, Boston.
- Winer, R.S. (2009), "New Communications Approaches in Marketing: Issues and Research Directions", *Journal of Interactive Marketing*, Direct Marketing Educational Foundation, Inc., Vol. 23 No. 2, pp. 108–117.
- Yadav, M.S. and Pavlou, P.A. (2014), "Marketing in Computer-Mediated Environments: Research Synthesis and New Directions", *Journal of Marketing*, American Marketing Association, Vol. 78 No. 1, pp. 20–40.
- Zaichkowsky, J. L. (1985), "Measuring the Involvement Construct," *Journal of Consumer Research*, 12 (3), 341–52.