Ehrenberg-Bass Institute Working Paper
Forthcoming in the *International Journal of Market Research*

"Understanding shopper transaction data: How to identify cross-category purchasing patterns using the duplication coefficient"

Authors:

Dr Arry Tanusondjaja - Ehrenberg-Bass Institute
Assoc Prof. Magda Nenycz-Thiel - Ehrenberg-Bass Institute
Assoc Prof. Rachel Kennedy - Ehrenberg-Bass Institute

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

Understanding shopper transaction data:

How to identify cross-category purchasing patterns using the duplication coefficient

Abstract

Shopper transaction data enables a deep analysis of what goes into shoppers' baskets; however, robust benchmarks are critical to see patterns in such rich data. This paper applies the *D* Duplication Coefficient from the Duplication of Purchase law as such a benchmark to help investigate patterns in simultaneous product category purchases. We demonstrate its application with data on 30,000+ UK and US supermarket transactions. The cross-category benchmarks allow meaningful deviations to be identified, isolating categories that are more or less intensely co-purchased than expected – which can then be used to guide decisions regarding store layout, prioritise in-store activations, and product category promotions.

Key words: Retail shopping, consumer behaviour, category buying

Introduction

Increasing availability of shopper transaction data provides new opportunities to investigate shopper behaviours, and advance shopper science. Due to its granularity, transaction data is considered as one of "*the* big data" to understand shoppers and assess marketing activities (Manyika *et al.* 2011; Breuer *et al.* 2013). To get a feel for its size, it is estimated that Walmart collect more than 2.5 petabytes (1 petabyte = 1 quadrillion bytes) of customer transaction data every hour in the US (McAfee & Brynjolfsson 2012). In the UK, Tesco uses the purchase data of their 16 million Clubcard holders to actively advertise additional services and guide shopper's purchases into healthier alternatives (Ferguson 2013). From a marketing perspective, it is vital to understand how we can extract meaningful insights from such data.

One of the ways in which transaction data can be used is basket analysis. Basket analysis is where we investigate the composition and size of shoppers' baskets in terms of categories and brands – i.e. the categories purchased together and those purchased on their own. This analysis can deliver useful insights for both manufacturers and retailers. For manufacturers, knowing how prominent their category is across consumers' baskets and how it is purchased with other categories – or competes – provides insights into the importance and growth potential of their categories, as well as guiding activations they may offer retailers (e.g. bundling promotions). For retailers, the information aids in determining what categories to stock, strategizing how to design the store as category location may impact competition (Sorensen 2009), and anticipating how promotion in one category affects other related categories (Kamakura & Kang 2007; Hruschka *et al.* 1999).

Although there are many research studies on the size and composition of shopping

baskets and how categories are purchased together in one transaction, little is known about what to expect from the category co-occurrence levels. To address this, we draw on the robust knowledge of brand competition encompassed in the *Duplication of Purchase Law* (see Ehrenberg *et al.* 2004) to investigate the expected levels of cross-category sharing from which meaningful exceptions can be systematically identified. This is obtained through the application of the *D* Duplication Coefficient from the Duplication of Purchase Analysis (analogous with Co-occurrence Analysis). The study investigates two sets of data with a total of more than 30,000 transactions from two supermarkets – one from the UK and the other from a US supermarket. The aim of the research is to pilot and demonstrate this approach to derive category co-occurrence benchmarks.

Background

The availability of store transaction data and the advance in the technology to efficiently process it has contributed to much research in the last twenty years – with varying foci, including those in cross-category relationships and their dependencies (e.g. Mild & Reutterer 2003; Russell & Kamakura 1998; Harlam & Lodish 1995; Ma *et al.* 2012; Niraj *et al.* 2008; Kamakura 2012; Chib *et al.* 2002; Seetharaman *et al.* 2005; Manchanda *et al.* 1999). For example, retailers can use market basket analysis to compare store performance (Berry & Linoff 2004); retailers and manufacturers can use it to understand more about the impact of product promotions in one product category on related categories (Walters 1991; Walters & Jamil 2002; Ma *et al.* 2012; Felgate *et al.* 2012).

Besides documenting how categories are purchased together in single transactions, store transaction data also enables researchers to track the number of baskets that

contain different categories. This measure, referred to as *category basket penetration* is considered to be the single most informative measure for category growth and decline (Fader & Lodish 1990) and is defined as the occurrence of a shopping basket containing the category over the total number of baskets (or transactions). For this purpose, a basket is defined as a set of items (or categories) purchased by a customer during one single shopping occasion (Mild & Reutterer 2003). While the items may be collected in a basket as the name implies, the data includes all transactions, so it can equally be collected in a shopping trolley, bag or held by the shopper. Knowing the "basket" compositions provides information on category substitutability (categories that can be purchased as alternatives) and complementarity (categories that are likely to be purchased together) (Russell & Kamakura 1998; Russell & Petersen 2000; Ma *et al.* 2012; Seetharaman *et al.* 2005).

Previous cross-category loyalty research

Cross-category purchase research has been a focus of numerous studies with researchers exploring how the purchase of one category impacts the propensity of purchasing another category either within the same transaction or across time (e.g. determined by studying previous purchases).

Cross-category loyalty research by Harlam & Lodish (1995) proposes a model to calculate the propensity of future purchases of items within a category based on past purchases. Their research builds on the premise that shoppers make a series of decisions when purchasing a group of items rather than one decision for the set of items. Given the model was only tested on one product category – powdered soft drink, its generalisability is unknown.

Other researchers (e.g. Chib *et al.* 2002; Seetharaman *et al.* 2005) also used panel data

to study cross-category purchase incidence and report high complementarity for product pairs such as *hot dogs* and *bacon*. The studies explored the usage of multivariate probit (MVP) and multivariate logit (MVL) models in cross-category correlations and documented positive cross-category correlations among all possible pairs from the product categories if they are different from zero. Despite these important contributions, we are still lacking empirically grounded benchmarks for cross-category sharing.

Another approach to tackle the relationship between one product category to the next is to document the sequence of the products that are purchased together in one basket, such as the study done by Kamakura (2012). However, even Kamakura (2012) warns that inferences of the purchase sequence from the basket composition can be misleading, unless actual observations of the shopping trip are also conducted which are not typically recorded with most transaction data.

Much of the previous research is limited in its scope with a narrow focus on specific pairings, such as cake mix + cake frosting (Ma *et al.* 2012) or frozen concentrate + refrigerated orange juice (Wedel & Zhang 2002) rather than looking at shopping baskets as an entity – clearly this has value to those involved in these individual categories but it only gives us part of the story of shopper behaviour. Thus, while basket analysis has been explored and modelled, previous models are rarely scalable beyond a few product categories and are therefore impractical for comprehensive market basket analysis (Kamakura 2012). As stated previously, past research also does not provide robust benchmarks for the level of co-occurrence of categories in a single transaction. Benchmarks allow us to know whether the co-occurrence is as expected or whether it deviates – as well as to predict future outcomes in similar

circumstances (Kennedy *et al.* 2014; Barwise 1995). This information is relevant for retailers and manufacturers, as it aids in monitoring how shoppers purchase product categories and allows insights into the impact of various in-store and out of store activations on cross-category purchases.

The application of Duplication of Purchase Law in cross-category purchase analysis

Outside of the shopper domain, generalised knowledge has been developed on brand buying behaviour that offers a different perspective in how we might tackle basket analysis and develop benchmarks to help build knowledge in this area. Knowing that buyers predictably shuffle between a number of brands within their repertoire, the *Duplication of Purchase Law* provides a benchmark of the likelihood of two brands being purchased, based on the size of the brands in the market (Ehrenberg *et al.* 2004). One benefit of this analysis is the computation of the expected level of sharing between two brands (or offers such as different pack sizes or flavours), using the following formula:

$$b_{X|Y} = D \times b_X$$

Where $b_{X|Y}$ is the percentage of buyers of brand (or offer) Y who also buy brand (or offer) X in the chosen period is proportional to X's penetration, and D (the Duplication Coefficient) is the average of the observed duplications for all pairs of brands, divided by the average penetration (Ehrenberg *et al.* 2004). Deviations to the expected levels of sharing can denote the presence of a partition, where brands share unusually higher or lower than expected (Ehrenberg *et al.* 2004). For example buyers of one diet soft drink may also purchase other brands of diet soft drinks, more than the level expected given their size in the market.

Taking the knowledge from the Duplication of Purchase (DoP) Law and the analysis that has been developed around it – this research aims to systematically explore if the method provides useful insights when applied to cross-category purchases and to evaluate the application of the *D* Duplication Coefficient in cross-category purchase analysis. The following section lists the specific Research Questions (RQ) of the study.

Research Questions

Building on the limitations of previous cross-category research and drawing on established knowledge on brand-buying behaviour, this research aims to address the following Research Questions:

RQ1: *Is the category co-occurrence pattern consistent with the Duplication of Purchase pattern?*

RQ2: *Based on the Duplication of Purchase analysis, do we see any meaningful deviations where categories co-occur more intensely than they should?*

Research Methodology

Data

We demonstrate and pilot the analysis on 15,862 UK supermarket transactions and 14,906 US supermarket transactions. The UK transaction data was collected from a supermarket in York over four weeks in 2010 (Week 22 – 25), whereas the US transactions were collected from a West Coast supermarket in Redwood City, California over a six-week period in 2006 (Week 34 – 39).

The data was collected as part of a project commissioned by the retailers to monitor shopper flow into the store as well as aisle navigation within the store. It captured all transactions from all product categories for both supermarkets during the relevant periods. The transaction data was downloaded from each supermarket data warehouse at the conclusion of each observation period. All transactions were included in the datasets reflecting normal purchasing which is less biased than data from loyalty cards, which skews heavy (i.e. to heavier users and bigger transactions), though the supermarket chains both have a loyalty card program. As such, the transaction information is not matched to any buyer information or profile. The data contains the items purchased in each unique transaction, which were then matched with a table with category aggregation details.

As background, both stores belong to a national chain and are of comparable size: 44,700 sq. feet (US) and 43,300 sq. feet (UK). The UK supermarket has a single right entry, whereas the US supermarket has dual entry to the store. The UK supermarket is situated in a middle-class area, whereas the US store is located in an upper-middle class neighbourhood. Both US and UK retail chains also have a number of other stores in the city.

There are three levels of category aggregation in the data (Level 3 is the most granular). As an example: *Bakery* (Level 1) contains *Prepacked Bakery* (Level 2) and then Level 3: *Cakes*, *loaves*, *morning goods*, *rolls, baps & baguettes* and *unclassified*. However, there are some differences in the categorisation between the supermarkets. For this research, Level 3 is used as the basis of the analysis as the granularity of the data allows us to pinpoint closer to the actual items, and lessons at this level are useful for store design purposes.

There are 233 Level 3 categories in the UK supermarket and 579 Level 3 categories in the US supermarket captured during the observation period. These categories are included in the descriptive analysis. For practical purposes, the Duplication of Purchase analysis as detailed below focuses on the top 46 categories with a basket penetration of 10% or more over the 4-week period in the UK, comprising 76% of all category purchases. For the US data, the analysis includes the top 35 categories with a basket penetration of 2% or more over the 6-week period in the US, accounting for 67% of all category purchases. The threshold was chosen arbitrarily for the pilot analysis for the ease of presentation.

Method of analysis

The research starts by describing the distribution of the number of product categories within each single shopping basket. This analysis helps to identify any patterns in the distribution, including any commonalities and differences across the data sets.

Our research also adopts the approach known as *affinity analysis* or *co-occurrence matrix*, which documents the coincidence of pairs of items in a market basket analysis (Russell & Petersen 2000; Berry & Linoff 2004). This is translated into the proportion of two categories sharing a basket together over the total number of baskets containing a particular category. For example in the UK baskets: 70% of baskets containing *cucumbers* also have *tomatoes*, and 34% of baskets containing *eggs* have at least one item from "*butter, margarine and spreads"*. Such analysis is not feasible using aggregate household data, as we cannot determine whether the categories are purchased together in one shopping trip.

The research utilises the approach used in Duplication of Purchase (DoP) analysis, which is consistent with the first step in the approach known as *affinity analysis*. DoP

analysis takes into account the size of the category, which in the context of this analysis is the category basket penetration. Adapting DoP analysis into cross-category analysis also allows us to calculate the *Duplication Coefficient (D)*, as described previously. Such as, when $D = 1$, buying one category makes the shopper no more or less likely than anyone else to buy another category. When $D > 1$, there is a pattern of higher co-occurrence (e.g. if D is 2, there is twice the likelihood that both categories will be purchased together compared to other categories of the same size).

DoP analysis also enables us to systematically identify deviations from expected levels – helping managers know where they need to focus. We highlight deviations that are 20% larger than the average duplication, to focus on deviations that may be managerially useful as an example in this paper.

Results

In this data, the UK baskets typically contain more product categories compared to US shopping baskets – with 7.3% of shopping baskets containing 41 or more product categories. Baskets containing one to three categories constitute 67% of the total transactions in the US supermarket, whereas they only represent 13% of the total transactions in the UK supermarket. The difference in the number of categories within the two data sets may be at least partly attributed to the differing category hierarchies in the two supermarkets, with the US data consisted of over 1,000 categories at Level 3 while the UK data used a more focused set of about 300 Level 3 categories. Furthermore, a much higher percentage of items in the UK supermarket data were assigned to a category, compared with the US where more items were *"Uncategorised"*.

The descriptive analysis of the two datasets reveals a familiar pattern in the distributions of categories that shoppers put in their baskets, which can be seen in **Figure 1**. Although the supermarkets display different distribution shapes, both patterns fit the Negative Binomial Distribution (NBD), specifically the *Truncated Negative Binomial Distribution* due to the lack of zero-class baskets (i.e. shoppers who enter the store without purchasing anything are not included). The lack of zero-class does not pose a concern for the fit, as the fit of the NBD is insensitive to this number (Chatfield *et al.* 1966). This distribution implies that for both supermarkets there is a heavy skew towards "light" baskets, containing only a few categories.

[Insert Figure 1 about here]

The different shape of the above distribution can be attributed to different dispersion parameters r: $r_{(US)} = 0.14$ (stdev: 0.03) vs. $r_{(UK)} = 0.42$ (stdev: 0.06). r measures the degree to which the variance differs in the sample: small dispersion parameters denote large variances within the sample (Jewell & Hubbard 2014). Next, the number of categories per basket and the composition of baskets over time are considered. The results show that UK shoppers put an average of 13.7 product categories (median: 11) into their baskets compared to 3.3 product categories (median: 2) in the US supermarket (**Table 1)**. The data reveals stability in the number of categories that shoppers put in the basket across weeks.

[Insert Table 1 about here]

The most prevalent categories are observed via the category penetration metric: the percentage of total baskets that contain the category at least once. **Tables 2a** and **2b** show that the top ten categories are very stable over time. The stability of the number of categories and their penetrations across the weekly shopping baskets is consistent

with brand stationarity, which has been comprehensively summarised in the NBD Dirichlet model (Goodhardt *et al.* 1984), and more recently demonstrated by Graham (2009). The composition of the top ten categories here is also similar across the two supermarkets, with both having high penetration categories such as milk, bread and bananas.

[Insert Table 2a about here]

[Insert Table 2b about here]

Product category composition of single category baskets

With the high prevalence of single-category baskets across both markets, we also analysed the common contents of such baskets. Both supermarkets exhibit similar patterns of the number of items in single-category baskets (**Figure 2**). The majority also only contain one single item (75% (US) and 67% (UK)).

[Insert Figure 2 about here]

The NBD Goodness of Fit tests also show very similar dispersion parameters for both markets: $r_{(US)} = 0.31$ (stdev: 0.10) vs. $r_{(UK)} = 0.33$ (stdev: 0.11).

With the larger number of product categories in each UK shopping basket, the incidence of single-category baskets is much lower than it is for the UK supermarket (4.4% vs. 31% in the US supermarket). However, knowing the content of the baskets is important, as the categories are strong enough to drive the shoppers to enter the supermarket for a single-category basket purchase.

Table 3 below shows similar patterns in terms of what categories are most commonly purchased in single-category baskets across the two supermarkets (across all product categories) – with *Alcoholic drinks*, *Milk*, *Drinking Water*, *Soft Drinks* and *Confectioneries* represented strongly.

[Insert Table 3 about here]

When one accounts for the slightly different classifications (e.g. *Citrus Soda* and *Cola* in the US vs. *Carbonates* in the UK; *French and Italian (Bread)* in the US vs. *Loaves* in the UK) the lists are remarkably similar and consolidate around beverages, treats and bread.

Product category co-occurrence and the application of the D Duplication Coefficient

The next stage of the analysis answers RQ1 and RQ2. After the basket penetration level for each product category is known, the next step is to determine whether certain product categories are more likely to co-occur in one transaction. This is obtained through the *Duplication of Purchase analysis* – by observing each possible product category pairing, and looking at the Average Duplication across all product categories.

An example of the resulting matrix is shown in **Table 4** (for the UK data). Only categories with basket penetration above 15% are shown for simplicity. From the table, *Penetration* figures show the prevalence of the product category across shopping baskets. For example 37% of all shopping baskets in the UK contain milk along with any other purchases, compared to only 16% for cereals. On the table, shaded lines are product categories that are more likely to be purchased in single-category baskets.

The results show that there is a higher degree of co-occurrence between large categories and a smaller degree with smaller categories. This finding is consistent with brand sharing patterns as predicted by the Duplication of Purchase Law: sharing is in line with size. Hence, this result answers RQ1 in that the co-occurrence of categories within baskets follows the expected patterns predicted by the Duplication of Purchase Law. Thus, buyers of less popular product categories are more likely to purchase large categories in one transaction. Such tables provide the useful benchmarks to compare the co-occurrence levels among various category combinations so we can ascertain whether certain pairings occur more or less intensely than expected (given the product category size).

[Insert Table 4 about here]

The co-occurrence analysis shows that buying one category does make a shopper more likely to purchase another product category, as evident by the *D* coefficient of 1.6 for the US supermarket and 1.5 for the UK. The greater the *D* coefficient is, the stronger the probability that purchasing one product category would lead to purchasing another product category.

By looking at the average duplication across the product categories, we can observe pairings where the levels are higher than expected. To answer RQ2, we select an arbitrary level of 20% deviation to highlight pairings where the co-occurrence levels are widely different to the expected. This figure has been derived with industry as it is likely to reflect managerial useful deviations. Co-occurrence levels that are less than expected are displayed in bold and italics, whereas levels that are more than expected are shown in bold with a rectangular border. A group of deviations may be indicative of a partition, where a group of product categories are purchased more

intensely together. For example, in the UK data, baskets containing *Crisps and savoury snacks* are also more likely to contain *Carbonated soft drinks*, whereas in the US, baskets containing *Regular Milk* are more likely to also contain *Refrigerated (products), Ice cream* or *Eggs.*

The ability to measure whether co-occurrence is stronger or weaker than expected is key in this research, which is missing from previous attempts – as well as establishing a benchmark for product category co-occurrence levels. Adapting the Duplication of Purchase method into co-occurrence matrices also explains why categories such as *Beer* or *Wine* are less likely to be purchased with any other product categories in the US data, and similarly product categories such as *Lager or Carbonates* in the UK data, as they were most probably purchased within single-category baskets as detailed in **Table 3** earlier.

Discussion and Conclusions

Through observing shopping baskets across the two supermarkets in the UK and the US, we have demonstrated that there are identifiable patterns in which categories are purchased and that benchmarks can be established for cross-category purchase analysis. The frequency of category occurrence within a basket follows the Negative Binomial Distribution with purchases skewing to few categories per shopping basket. The distribution provides the important benchmark from which to monitor in-store activations. It helps retailers and manufacturers to understand whether efforts should be directed towards increasing the number of product categories within each basket – by knowing which product categories to target – or by increasing the number of items *per* category). The average number of categories per basket and the top categories in terms of their

penetration show astonishing stability over time in both markets reflecting the habitual nature of shopping behaviour and loyalty to repertoires of categories.

Our findings on category competition extend the robust knowledge on brand competition (e.g. Ehrenberg *et al.* 2004; Goodhardt *et al.* 1984) to category level and provide benchmarks on what to expect in terms of category co-occurrence, by adopting the *D* Duplication Coefficient, thus filling a gap in the previous research. With this benchmark, we have demonstrated that the Duplication of Purchase Law holds across categories, as categories share shoppers with other categories in line with the category basket penetration. We have also demonstrated how to identify categories that compete more or less intensely than is expected based on their size alone.

Several implications emerge from the research. The most important for manufacturers and retailers is that we can predict the level of category co-occurrence within baskets and that there are predictable patterns for the frequency of category occurrence within baskets as well as their composition. Predictable patterns give benchmarks in terms of what to expect that are of great use in single and multiple category analysis and/or store layout management. The Duplication of Purchase is especially useful for retailers, to understand how – at the store level – categories are purchased together (i.e. complements) or compete (i.e. substitutes), due to the likes of layout. It provides a solid foundation to compare the effectiveness of interventions like store layout changes and cross-category promotions. The adoption of the *D* Duplication Coefficient also provides a metric to track whether shoppers are more or less likely to purchase category pairings over time. For manufacturers the ability to know how categories are purchased

together and compete helps define competition in a broader sense. For example, how much chocolate brands actually compete with other product categories such as crisps, nuts or fruit.

The ability to identify which categories compete more or less intensely at the store level (or across retailers or chains if broader data is used), gives manufacturers and retailers better information to make decisions. Such decisions may include promoting certain product categories together to increase the purchase propensity of other product categories with a closer affinity (as denoted by their higher sharing compared to the benchmark). For example, such analysis might lead retailers to promote activities that push for category combinations like *Salty snacks* and *Cola* together. Another application of this knowledge is guidance for manufacturers with brand presence in different categories to which categories to promote together to increase co-purchasing (for example, promoting cereal and children snacks, or pasta and pasta sauce from the same company).

Knowing the product categories that shoppers select in a single-category basket can also have implications for what is offered in key positions (and possibly at key times of the day) such as near the checkouts. For retailers, it provides a possible strategy to either increase the number of items purchased by encouraging multiple category purchases (i.e. multi-buys), or explore the possibility of increasing the number of categories purchased. The latter strategy does not necessarily require enticements or promotions, as it may simply mean placing complementary items in a closer proximity to categories often found in single category baskets so they are more salient and easier to notice and buy.

The causality between category placements and category co-occurrence requires

further study to determine whether shoppers purchase categories together because of the store placement or vice versa. The order of purchase is also outside the boundary of Duplication of Purchase analysis, so we cannot ascertain whether the shoppers put *Salty snacks* first and then *Cola*, or vice versa (or whether both were on a shopping list regardless of order, or both were purchased on impulse). This is consistent with the caution given by Kamakura (2012). However, the findings in this paper are still likely to be useful for category placements at store level. This approach also gives a benchmark for experiments to determine what activations best influence category co-occurrence. If both transaction and observation data are available, it will be possible to deep dive into order effects – with the extra knowledge to complement these benchmarks.

In summary, this study demonstrated the wealth of knowledge possible from basket analysis of transaction data compared with aggregate household purchasing data investigations. While household data analysis can provide great insights into consumer behaviour, it cannot reflect cross-category effects due to factors such as changes in store activities and traffic.

Further Research

This study is grounded in a total of over 30,000 transactions, to demonstrate the application of the Duplication of Purchase Law on cross-category purchases. Given many factors contributing to the result differences, such as seasonality, store format, and store location, it is desirable to further document the findings using more transaction data from additional stores and countries, looking for how far the patterns as well as the category co-occurrence *D* Duplication Coefficient generalise and in what conditions – ideally using data covering longer time periods and more varied

retailers, countries and the like. Future research could also look at how advertising or product promotions impact the benchmarks and the purchase patterns. Such information could explain any notable changes in basket penetration and refine the understanding at how product categories are purchased together and interact. The study has demonstrated an approach to establish benchmarks at how product categories interact with each other – so further research can focus on how these benchmarks move over time.

Future studies could also be conducted to further explore the application of the full NBD-Dirichlet model to product category buying behaviour. The NBD-Dirichlet model posits that each shopper has a certain propensity to buy a given brand at any time (Goodhardt *et al.* 1984) – the extension research can explore the premise that each shopper has a certain propensity to buy a particular category, *ceteris paribus*. While this paper has demonstrated that one of the laws associated with the NBD-Dirichlet model, i.e. Duplication of Purchase Law, holds at category level, future additional research can explore whether the other laws, such as the Natural Monopoly Law and Double Jeopardy, extend to product category purchases. This would sharpen our understanding on how product categories grow and compete.

References

Barwise, P. (1995) Good Empirical Generalizations. *Marketing Science*, **14**, No. 3, Part 2 of 2, pp.G29-G35.

Berry, M.J. and Linoff, G.S. (2004) *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.

Breuer, P., Forina, L., and Moulton, J. (2013) Beyond the Hype: Capturing Value from Big Data and Advanced Analytics. In *Perspectives on retail and consumer goods*, Vol. 1: McKinsey and Company, pp.76.

Chatfield, C., Ehrenberg, A.S.C., and Goodhardt, G.J. (1966) Progress on a Simplified Model of Stationary Purchasing Behaviour. *The Journal of the Royal Statistical Society Series A (General)*, **129**, Part 3, pp.317-67.

Chib, S., Seetharaman, P.B., and Strijnev, A. (2002) Analysis of Multi-Category Purchase Incidence Decisions Using Iri Market Basket Data. *Advances in Econometrics*, **16**, pp.57-92.

Ehrenberg, A., Uncles, M.D., and Goodhardt, G.G. (2004) Understanding Brand Performance Measures: Using Dirichlet Benchmarks. *Journal of Business Research*, **57**, 12, pp.1307-25.

Fader, P.S. and Lodish, L.M. (1990) A Cross-Category Analysis of Category Structure and Promotional Activity for Grocery Products. *Journal of Marketing*, **October**, pp.52-66.

Felgate, M., Fearne, A., and DiFalco, S. (2012) Using Supermarket Loyalty Card Data to Analyse the Impact of Promotions. *International Journal of Market Research*, **54**, 2, pp.214-40.

Ferguson, D. (2013) How Supermarkets Get Your Data – and What They Do with It. In *The Guardian*, online: Guardian News and Media Limited.

Goodhardt, G.J., Ehrenberg, A.S.C., and Chatfield, C. (1984) The Dirichlet: A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society*, **147**, 5, pp.621-55.

Graham, C.D.A. (2009) What's the Point of Marketing Anyway? The Prevalence, Temporal Extent and Implications of Long-Term Market Share Equilibrium. *Journal of Marketing Management*, **25**, 9-10, pp.867-74.

Harlam, B.A. and Lodish, L.M. (1995) Modeling Consumers' Choices of Multiple Items. *Journal of Marketing Research*, pp.404-18.

Hruschka, H., Lukanowicz, M., and Buchta, C. (1999) Cross-Category Sales Promotion Effects. *Journal of Retailing and Consumer Services*, **6**, 2, pp.99-105.

Jewell, N.P. and Hubbard, A. (2014) *Analysis of Longitudinal Studies in Epidemiology*. CRC Press LLC.

Kamakura, W.A. (2012) Sequential Market Basket Analysis. *Marketing Letters*, **23**, 3, pp.505-16.

Kamakura, W.A. and Kang, W. (2007) Chain-Wide and Store-Level Analysis for Cross-Category Management. *Journal of Retailing*, **83**, pp.159-70.

Kennedy, R., Scriven, J., and Nenycz-Thiel, M. (2014) When 'Significant' Is Not Significant. *International Journal of Market Research*.

Ma, Y., Seetharaman, P.B., and Narasimhan, C. (2012) Modeling Dependencies in Brand Choice Outcomes across Complementary Categories. *Journal of Retailing*, **88**, 1, pp.47-62.

Manchanda, P., Ansari, A., and Gupta, S. (1999) The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions. *Marketing Science*, **18**, 2, pp.95-114.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

McAfee, A. and Brynjolfsson, E. (2012) Big Data: The Management Revolution. In *Harvard Business Review*, online: Harvard Business School Publishing.

Mild, A. and Reutterer, T. (2003) An Improved Collaborative Filtering Approach for Predicting Cross-Category Purchases Based on Binary Market Basket Data. *Journal of Retailing and Consumer Services*, **10**, 3, pp.123-33.

Niraj, R., Padmanabhan, V., and Seetharaman, P.B. (2008) Research Note—a Cross-Category Model of Households' Incidence and Quantity Decisions. *Marketing Science*, **27**, 2, pp.225-35.

Russell, G.J. and Kamakura, W.A. (1998) Modeling Multiple Category Brand Preference with Household Basket Data. *Journal of Retailing*, **73**, 4, pp.439-61.

Russell, G.J. and Petersen, A. (2000) Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing*, **76**, 3, pp.367-92.

Seetharaman, P.B., Chib, S., Ainslie, A., Boatwright, P., Chan, T., Gupta, S., Mehta, N., Rao, V., and Strijnev, A. (2005) Models of Multi-Category Choice Behavior. *Marketing Letters*, **16**, 3/4.

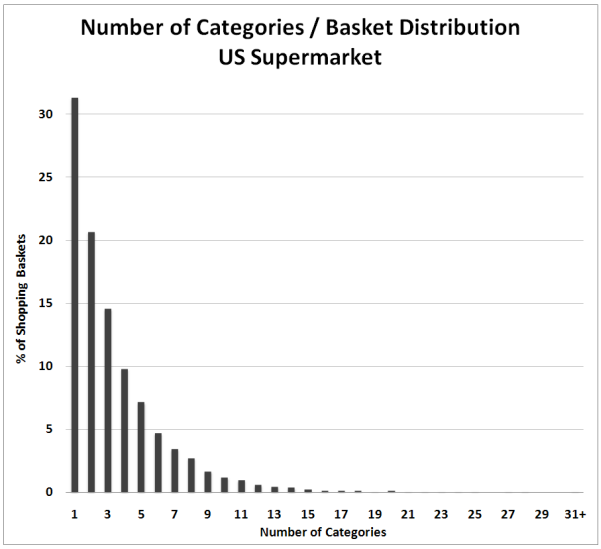
Sorensen, H. (2009) *Inside the Mind of the Shopper.* Vol. 1, Upper Saddle River, New Jersey: Pearson Education Inc.

Walters, R. and Jamil, M. (2002) Measuring Cross-Category Specials Purchasing: Theory, Empirical Results, and Implications. *Journal of Market-Focused Management*, **5**, 1, pp.25-42.

Walters, R.G. (1991) Assessing the Impact of Retail Price Promotions on Product Substitution, Complementary Purchase, and Interstore Sales Displacement. *Journal of Marketing*, **55**, April, pp.17-28.

Wedel, M. and Zhang, J. (2002) Assessing Cross-Category Impact from Store-Level Scanner Data. In *Working Paper Series*: University of Michigan Business School.

Figure 1 – Basket Composition: The Number of Categories (US vs. UK)



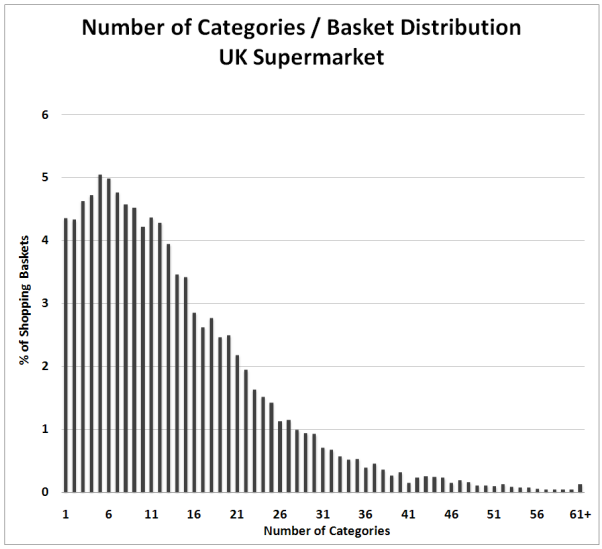


Table 1 – Number of Categories / Basket by Supermarket

Weeks	UK			US				
	22	**23**	**24-25**	**34**	**35**	**36**	**37**	**38-39**
Mean	14	14	14	3.1	3.1	3.5	3.5	3.3
Median	11	11	11	2	2	2	2	2

Table 2a – Top Ten Categories (% Number of Baskets) – US Supermarket

Top Ten Categories	Wk-34	Wk-35	Wk-36	Wk-37	Wk-38-39	Avg.	StDev
Regular Milk	12	13	13	13	13	13	0.4
Bananas	6.3	6.8	7.2	8.2	8.1	7.3	0.8
Cola	8.4	7.5	4.6	6.5	7.3	6.9	1.4
Wine	6.4	4.8	8.2	6.2	6.0	6.3	1.2
Loaf Breads	4.6	6.1	6.2	6.5	5.9	5.9	0.8
Beer	7.5	5.2	5.7	5.4	5.3	5.8	0.9
Drinking Water	4.8	4.6	4.6	5.2	6.0	5.0	0.6
Refrigerated	4.6	4.5	4.7	4.3	5.0	4.6	0.3
Chips	4.4	4.8	5.2	4.5	4.7	4.7	0.3
Poultry (Chicken & Turkey)	3.1	4.8	4.0	4.4	4.8	4.2	0.7

Data collected in August-September, 2006

Table 2b – Top Ten Categories (% Number of Baskets) – UK Supermarket

Top Ten Categories	Wk-22	Wk-23	Wk-24+25	Average	StDev
Loaves	41	41	38	40	1.6
Fresh Milk	37	36	37	37	0.6
Tomatoes	28	30	28	29	1.5
Cakes	28	27	29	28	0.9
Yoghurt & prepacked desserts	30	28	26	28	1.8
Bananas	27	25	27	26	1.1
Cheese - Prepacked	26	26	25	26	0.4
Potatoes	25	25	25	25	0.3
Apples & pears	26	23	24	24	1.6
Rolls, baps & baguettes	23	24	23	23	0.4

Data collected in May-June, 2010

Figure 2 – Number of Items in Single Category Baskets

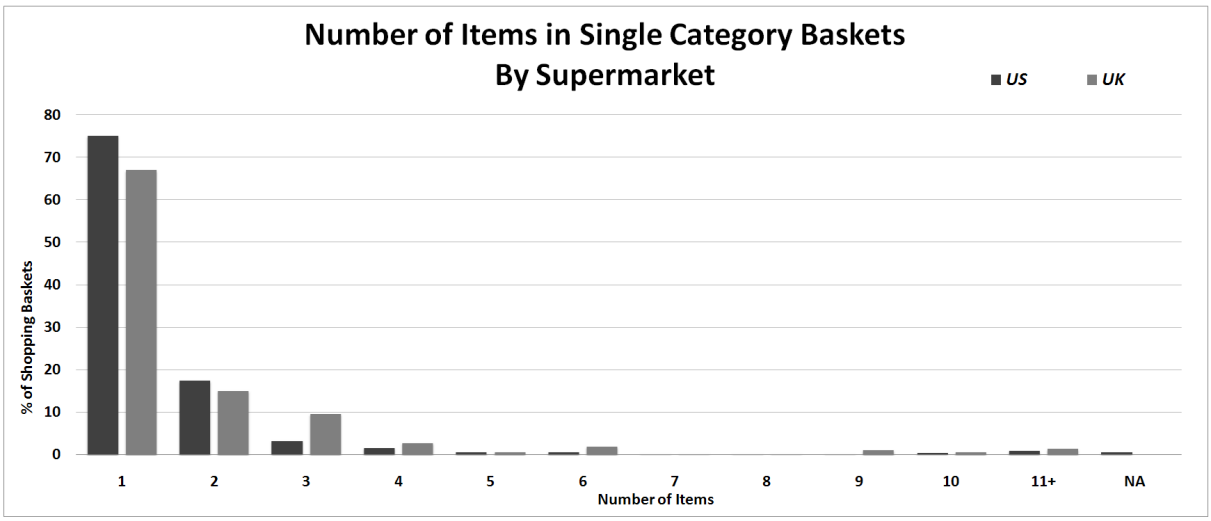


Table 3 – Top Ten Categories in Single-Category Baskets (% of Single-Category Baskets)

Top Ten Categories (US)	Average	Top Ten Categories (UK)	Average
Regular Milk	7.8	Lager	10
Beer	6.9	Carbonates	4.8
Wine	5.7	Fresh Milk	3.3
Drinking Water	4.0	Sweets	2.9
Cola	4.0	Chocolate	2.6
Energy Drinks	2.8	Water	2.3
Citrus Soda	1.9	Salad bar	2.3
Gum & Mints	1.8	Cider	2.2
Ice Cream & Premium Ice Cream	1.7	Loaves	1.7
French & Italian (Bread)	1.5	White Wine	1.7
Total	38		34

Table 4 – Co-occurrence Matrix (with 20% deviations identified) – UK Supermarket

	% Basket pentr.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AI
A	40		53	39	39	38	36	36	35	33	25	29	31	28	23	27	24	25	25	25	22	18	22	21	21	21	22	17	2(
B	37	58		40	38	41	39	36	36	34	29	30	29	27	24	27	25	25	25	27	23	18	23	20	21	20	23	16	21
C	29	55	51		36	42	42	41	42	41	32	38	33	27	23	28	**39**	23	26	25	28	17	**38**	27	**30**	21	25	16	28
D	28	54	49	36		39	35	33	34	31	31	29	29	29	25	26	***21***	29	24	22	***21***	21	***20***	21	***18***	***18***	***19***	20	***18***
E	28	54	53	43	39		44	43	37	41	28	37	34	30	23	28	27	26	24	29	25	19	25	23	23	23	23	18	23
F	26	55	55	46	37	47		38	39	**48**	30	39	31	29	22	27	28	26	24	28	27	19	25	21	23	21	23	17	23
G	26	56	51	45	36	47	39		37	38	32	33	36	33	25	30	27	26	28	28	24	19	26	25	25	26	25	17	25
H	25	56	53	48	39	42	41	38		39	30	33	31	28	24	31	31	26	28	26	**34**	18	27	23	24	23	25	17	29
I	24	54	51	48	36	47	**52**	40	40		28	**41**	32	28	21	27	30	26	24	29	29	17	27	21	24	22	24	17	25
J	23	44	46	40	38	***33***	34	36	32	***30***		27	32	29	24	25	24	22	26	***21***	***19***	19	22	23	22	***18***	19	17	***19***
K	22	51	49	48	38	46	46	39	37	45	29		31	27	22	24	30	24	22	27	26	18	27	22	26	20	22	17	23
L	21	58	50	44	39	44	38	44	37	37	36	33		35	25	29	26	27	29	27	23	20	26	26	26	24	25	19	22
M	21	54	48	37	40	39	36	40	34	33	32	29	35		**33**	28	23	**32**	28	26	**21**	**25**	23	25	21	24	21	**23**	**20**
N	19	47	45	***33***	37	***33***	***30***	33	***31***	***26***	29	***25***	28	36		24	***21***	25	25	21	***18***	23	***20***	21	***19***	20	19	21	***18***
O	18	60	55	44	40	42	38	42	42	36	31	29	33	32	25		26	28	31	27	27	21	24	22	20	22	27	18	24
P	18	54	51	**63**	34	42	41	39	43	41	32	38	31	27	23	26		22	25	24	32	16	**45**	28	21	21	25	15	29
Q	17	59	52	39	**48**	43	40	38	38	36	30	31	33	**39**	28	30	***22***		28	30	25	**24**	22	22	19	23	22	**24**	21
R	17	57	53	43	40	38	37	42	41	34	35	28	36	34	27	33	26	28		25	24	20	24	24	23	27	28	19	25
S	16	60	60	42	39	49	44	43	39	42	29	36	34	34	25	30	26	**31**	26		26	21	25	23	23	27	25	20	23
T	16	56	54	51	37	44	46	40	**54**	**45**	29	37	31	28	22	32	**36**	27	27	28		18	31	23	23	22	27	16	**33**
U	16	47	42	***31***	38	35	***31***	***31***	***28***	***27***	28	***26***	27	34	29	24	***18***	27	22	22	***18***		***18***	***17***	***18***	***18***	***19***	**28**	***1***
V	15	57	54	**70**	36	46	43	43	44	43	33	39	36	31	25	29	**51**	25	26	27	31	18		**33**	**34**	22	25	18	**32**
W	15	55	48	50	38	42	36	41	38	33	34	32	36	34	27	26	32	24	27	25	23	18	**33**		**34**	22	21	17	23
X	15	55	50	**57**	34	42	40	41	40	38	34	38	37	29	23	24	25	22	26	25	24	19	**35**	**34**		24	24	17	27
Y	15	56	50	40	34	42	37	45	38	37	28	30	34	34	26	27	25	27	31	30	24	19	23	22	24		24	17	29
Z	15	59	56	48	36	43	41	43	41	39	30	34	35	29	25	**34**	30	26	**32**	28	28	20	26	22	25	24		17	29
AA	15	46	***40***	***31***	37	***33***	***30***	***30***	***29***	***28***	27	***26***	26	32	27	22	***18***	27	***21***	22	***17***	**29**	***18***	***18***	***17***	***17***	***17***		***1***
AB	15	54	52	54	34	43	42	43	**48**	41	30	35	32	28	24	29	**35**	25	29	25	**36**	18	33	24	28	**29**	**29**	17	
Average Duplication		**54**	**51**	**45**	**38**	**41**	**39**	**39**	**38**	**37**	**30**	**33**	**32**	**31**	**25**	**28**	**28**	**26**	**26**	**26**	**25**	**20**	**26**	**23**	**23**	**22**	**23**	**18**	**24**

Note: *Product categories with at least 15% basket penetration are shown. Co-occurrence levels with greater than 20% deviation are shown in bold and bordered. Levels that are 20% below the expected are shown in bold and italics. Shaded product categories are those that are more likely to be purchased in a single-category basket.*

Legend: A: Loaves; B: Fresh milk; C: Tomatoes; D: Cakes: E: Yoghurt & prepacked desserts; F: Bananas; G: Cheese – repacked; H: Potatoes; I: Apples & pears; J: Rolls, baps & baguettes; K: Soft fruit; L: Packaged cooked meats; M: Crisps & savoury snacks; N: Carbonates; O: Butter spreads + margarine; P: Lettuce, celery & other salad vegs; Q: Children & everyday; R: Prepacked bacon & sausages; S: Cereals; T: Root vegetables; U: Chocolate; V: Cucumbers; W: Coleslaw & dips; X: Bagged salad; Y: Total cooking sauces; Z: Eggs; AA: Sweets; AB: Onions;