Behavioural analytics: 
Exploring judgments and choices in large data sets

This item was submitted to Loughborough University’s Institutional Repository by the/an author.


Additional Information:

- This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the Operational Research Society on 5 March 2018, available online: http://www.tandfonline.com/10.1080/01605682.2018.1434400.

Metadata Record: https://dspace.lboro.ac.uk/2134/36665

Version: Accepted version

Publisher: © Operational Research Society 2018. Published by Taylor and Francis

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the published version.
Behavioural Analytics: Exploring Judgments and Choices in Large Data Sets

Ian N. Durbach\textsuperscript{1,2} and Gilberto Montibeller\textsuperscript{*3}

\textsuperscript{1}Centre for Statistics in Ecology, the Environment and Conservation, Department of Statistical Sciences, University of Cape Town
\textsuperscript{2}African Institute for Mathematical Sciences
\textsuperscript{3}Management Science and Operations Group, School of Business and Economics, Loughborough University

Abstract

The ever-increasing availability of large data sets that store users’ judgments (such as forecasts and preferences) and choices (such as acquisitions of goods and services) provides a fertile ground for Behavioural Operational Research (BOR). In this paper we review the streams of Behavioural Decision Research that might be useful for BOR researchers and practitioners to analyse such behavioural data sets. We then suggest ways that concepts from these streams can be employed in exploring behavioural data sets for (i) detecting behavioural patterns, (ii) exploiting behavioural findings and (iii) improving judgments and decisions of consumers and citizens. We also illustrate how this taxonomy might be utilised in practice, in three real-world studies with behavioural datasets generated by websites and on-line user activity.

Keywords: behavioural operational research, behavioural data mining, behavioural analytics.

1 Introduction

The explosive growth of sophisticated websites and online services that store users’ information and preferences has generated an increasing availability of large datasets to Operational Research (OR) practitioners and researchers (Mortenson et al., 2014). One important subset of such databases is what we term \textit{behavioural databases}. A behavioural database records judgments of users (for example estimates about a given variable, such as predictions of results in sport analytics (Wright, 2013) or behaviour of bank users with their bank accounts (Adams et al., 2001)) as well as their choices (for instance the services that citizens utilize in a public service program (De Cnudde and Martens, 2015)).

Behavioural databases have been explored for both behavioural research (e.g. Gilovich et al., 1985; Jullien and Salanié, 2000; Ateca-Amestoy and Prieto-Rodriguez, 2013) and

\textsuperscript{*}Corresponding author: g.montibeller@lboro.ac.uk
Information Systems research (e.g. Burtch et al., 2013; Goh et al., 2013; Provost et al., 2015), as well as for commercial purposes (e.g. Davenport and Harris, 2007; Davenport et al., 2010). However, guidelines for mining these datasets from an OR perspective, i.e., with a focus on improving decision making, are lacking.

On the other hand, the emergence of Behavioural Operational Research (BOR) (Hamalainen et al., 2013; Franco and Hamalainen, 2016; Kunc et al., 2016) has emphasised the relevance of behavioural issues for OR models and modellers. Some of these issues might require the exploration of large behavioural data sets. For instance, understanding the behaviour of bank customers with their money may help an OR analyst to optimise a bank’s balance sheet. The efficiency of delivering a public service might be maximised if a better understanding of users’ choice patterns, recorded in a behavioural data base, enables the operational researcher to fine tune the estimates of future demands.

The extensive developments of Behavioural Decision Research (BDR), since the breakthrough paper by Tversky and Kahneman (1974), provide a rich source of models to analyze behavioural data sets. BDR encompasses the well-known literature on behavioural biases (Kahneman et al., 1982; Gilovich et al., 2002), but also the growing interest on intertemporal choices (Read, 2007; Scholten and Read, 2010), on heuristics (Gigerenzer and Gaissmaier, 2011), on the aggregation of individual judgments (Kerr and Tindale, 2011), and on debiasing (Fischhoff, 1982; Larrick, 2007; Montibeller and von Winterfeldt, 2015). However, the field is somehow fragmented, making it difficult for BOR researchers and practitioners to fully deploy its insights and findings.

The aim of this paper is thus to conceptualize different modes for mining behavioural datasets from a BOR perspective, rooted in well-established developments in BDR. We review the streams of Behavioural Decision Research that might be useful in this context, and suggest ways that such behavioural streams might be valuable for BOR researchers and practitioners as a means of detecting behavioural patterns, exploiting behavioural findings, and improving judgments and decisions. We illustrate how the taxonomy may be employed in practice, presenting three real-world studies with behavioural datasets generated by websites and on-line user activity.

The remainder of the paper is divided into four sections. The next section reviews the different streams of BDR. The subsequent section conceptualizes the taxonomy for exploring behavioural datasets that we propose. This is followed by an illustration of its use in conceptualising three real-world cases employing behavioural datasets generated by websites and on-line user activity. The paper concludes with some final remarks, discussion of limitations and suggestions for further developments on this front.
2 Streams in Behavioural Decision Research

The study of how unaided subjects make their judgments and choices, and how such decisions can be improved, is the realm of Behavioural Decision Research (BDR). In this section we identify six streams of research within BDR for BOR research to explore: individual cognitive biases, prospect theory, intertemporal choice, debiasing, heuristics, and collective judgments. Each of these streams is a vast sub-field; here we just describe them briefly and provide pointers to further literature.

2.1 The Cognitive Biases Stream

After the inception of Behavioural Decision Theory in the early 1960’s (Edwards, 1961), two major approaches have emerged: one based on the cognitive illusions paradigm and another based on the engineering psychology paradigm (von Winterfeldt, 1999).

Researchers adopting the cognitive illusions paradigm compare a normative decision principle (such as Bayes’ rule) with actual behaviour in judgment and choice (such as the actual updating of prior probabilities). Discrepancies between the normative principle and the observed behaviour characterize a cognitive bias (such as conservatism in Bayesian updates, one of the first identified biases (Phillips and Edwards, 1966)). From the seminal work of Tversky and Kahneman (1974), a large body of knowledge has emerged (e.g. Kahneman et al., 1982; Gilovich et al., 2002) that investigates the mental heuristics that lead to systematic cognitive biases. We thus refer to this as the individual cognitive biases stream.

A large number of these biases have been identified (see Bazerman and Moore (2008) for an introductory overview). Some of the most well-known biases, which are not only observable in experimental settings but also in real-world decision making (von Winterfeldt, 1999; Montibeller and von Winterfeldt, 2015), are: anchoring (the a priori presentation of an unrelated numerical value influences the estimate of a given parameter (Tversky and Kahneman, 1974)), framing (the way a prospect is presented, either as a loss or as a gain, has an impact on the choice made (Tversky and Kahneman, 1981)), and overconfidence (estimates of ranges of a variable that are too narrow or of probabilities of future events occurring that are not realistic (Lichtenstein and Fischhoff, 1977)).

2.2 The Prospect Theory Stream

Another major contribution of the cognitive illusions paradigm was the development of prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), which suggests a descriptive model for decisions under uncertainty. The three main components of such a model are: a reference point (zero utility) demarcating gains and losses; a two-part S-shaped utility function that is concave for gains and convex for losses; and an an
inverse S-shape weighting function for probability estimates.

A prospect theory model thus allows one to represent different decision attitudes to gains and losses, as well as over-weighting of low-probability events and under-weighting of high-probability events. It also highlights the importance of framing effects (whether a prospect is presented as a loss or gain) and of loss aversion (the asymmetry between losses and gains, with the former looming larger than the latter). These aspects are frequently observable in unsupported decision making (for a comprehensive coverage of this theory see Wakker (2010) and Kahneman and Tversky (2000)). We term this the prospect theory stream.

2.3 The Intertemporal Choice Stream

The third behavioural stream, intertemporal choice, which is also based on the cognitive illusion paradigm, analyses the relative value that subjects attribute to benefits and costs at different points in time. The normative model for intertemporal choice, proposed by Fisher and by Samuelson in the 1930s (for a historical review see Loewenstein (1992)) uses exponential discounting with a constant discount rate, and is widely employed in net present value economic analysis. It reflects both opportunity cost and time preferences, where a payment of a given value is preferred earlier than later (Read, 2007).

In a similar way as with the previous two streams, empirical research into time preferences has shown several deviations from this normative standard, such as: time inconsistencies (with the use of a hyperbolic discount functions instead of exponential ones), sign effects (lower discounting for losses than for gains), magnitude effects (lower discount rates for large amounts than for small amounts), and framing effects affecting such choices (Read, 2007; Scholten and Read, 2010; Attema, 2012).

2.4 The Debiasing Stream

While the cognitive illusions paradigm has mainly been focused on developing descriptive models of decision making, the main aim of those embracing the engineering psychology paradigm in Behavioural Decision Theory is to improve decision makers’ judgments and choices. Researchers in this paradigm identify problems decision makers might have in implementing a normative rule and provide suitable tools to overcome such problems (von Winterfeldt, 1999), debiasing judgments and value/probability elicitation (Montibeller and von Winterfeldt, 2015; Morton and Fasolo, 2009). For instance, the elicitation of estimates for a given variable, which could be too narrow due to the overconfidence bias, might be stretched by using counterfactual questions as debiasing tools. The elicitation of preferences over a set of alternatives might be debiased against the framing bias by using adequate elicitation protocols that emphasise the equivalence of gains and loss prospects.

The debiasing literature has focused mostly on probability judgments, and although
early efforts proved ineffective (Fischhoff, 1982) a rich theory and practice has since developed, most notably including the use of adequate incentives, suitable cognitive strategies, and adjusted elicitation procedures, the latter both for probability (e.g. Spetzler and Staël Von Holstein, 1975; Johnson and Bruce, 2001; Garthwaite et al., 2005) and preference elicitation (e.g. Wakker and Deneffe, 1996; Delquie, 1997; Carlson and Bond, 2006). In more general terms, a better understanding of the mechanisms that generate biases can in certain cases provide suggestions for debiasing (Arkes, 1991; Larrick, 2007; Montibeller and von Winterfeldt, 2015). We call this field the debiasing stream.

2.5 The Heuristics Stream

In the cognitive illusions paradigm, the role of mental heuristics is seen as negative: they constitute sub-optimal judgmental strategies in comparison to a normative rule. Furthermore, its focus is on externally observable cognitive biases generated by such sub-optimal heuristics, instead of the heuristics themselves (Gigerenzer, 1991, 1996).

An opposite view about the role of heuristics is the one Gigerenzer and his associates have championed. In his perspective heuristics play the central role, being defined as “strategies that ignore information to make decisions faster, more frugally, and/or more accurately than more complex methods” (Gigerenzer and Gaissmaier, 2011, p. 453).

This field of research, which we refer to as the heuristics stream, places its main emphasis on understanding in depth which heuristic is being used in a given judgment (for an overview see Gigerenzer and Gaissmaier (2011)). Heuristics they identify include: the recognition heuristic (if only one of two alternatives is recognized then it has higher value), the fluency heuristic (if both alternatives are recognized, the one recognized faster has higher value), the hiatus heuristic (the customer is classified as inactive after not purchasing for a given number of months), simple non-compensatory decision rules (such as the lexicographic rule, in which alternatives are evaluated on the most important attribute, in case some of them have the same performance on this attribute, the performance of these remaining alternatives on the second most important attribute is considered and so on until a choice is made), and simple compensatory decision rules (such as counting the number of cues favoring an alternative). Sharing a positive agenda about decision making capabilities with the engineering psychology paradigm, it sees many of these heuristics as efficient ways of reaching accurate judgments.

2.6 The Collective Judgments Stream

The sixth stream in BDR is the collective judgments steam. This has a much longer history of formal research than the previous five, dating back to Galton (1907) demonstrating that the median of estimates produces a highly accurate estimate itself (for a discussion see Levy and Peart (2002); for a recent review on collective judgments see Kerr and Tindale
On the aggregation of individual preferences, early scientific works started with de Borda and Condorcet in the 18th century, followed by Arrow’s impossibility theorem and the field of social choice theory (see Gaertner (2009)). More recent developments have suggested the aggregation of individual cardinal utility rather than ordinal preferences (e.g. Keeney and Nau, 2011). The main aims of the collective judgment stream are therefore to: i) understand how individual estimates or preferences from a group should be aggregated; ii) assess how precise the aggregate estimates are, or in case of preferences how fair the aggregation rule is; iii) identify what cognitive biases may impact on such judgments.

Behavioural OR practitioners and researchers usually have a clearly prescriptive perspective for their analysis (for a distinction between prescriptive, normative and descriptive approaches see Bell et al. (1988)). This is in contrast to BDR findings and frameworks in the cognitive biases, intertemporal choice, and prospect theory streams (as well as most of the heuristics stream) which follow a descriptive approach. Most of the work on the collective judgment stream is either normative or descriptive, again fitting with BDR research interests. Only the debiasing stream has a clear prescriptive emphasis. Furthermore, with the exception of the collective judgment stream, the others rely heavily on evidence from relatively small-scale laboratory experiments (Bardsley, 2005; Robinson et al., 2017).

The question thus is how this vast body of knowledge in BDR could be employed by BOR researchers and practitioners who want to explore behavioural datasets. We suggest ways of exploring them in the next section, and review studies where analyses of behavioural issues have been conducted using real-world behavioural data sets, although not necessarily with a prescriptive approach.

3 Exploring Behavioural Datasets

We suggest here that the process of analyzing behavioural data sets can be conceptualized as having three main purposes for the analysis: (i) detecting users’ behavioural trends and patterns, (ii) exploiting behavioural findings, and (iii) improving users’ judgments and choices. Such a process must lead to behavioural insights which help BOR researchers and practitioners to improve systematic behavioural data collection, increase its learning about users’ judgments and behavioural patterns, and support evidence-based decision making.

This definition highlights three main modes for exploring behavioural datasets: detecting behavioural patterns, exploiting behavioural findings, and improving users’ judgments. We now suggest how each one of these modes can be employed within the BDR stream identified in the previous section and describe examples where such analyses have been employed using behavioural databases.

The detection of behavioural trends and patterns in users’ judgments and choices is
by some way the most developed of the three modes in BDR research. The cognitive bias stream provides a plethora of biases that can be investigated in behavioural databases. Several applications to sports data sets have been reported (e.g. Kocher et al., 2012), most notably the search of evidence for or against the “hot hand” hypothesis in basketball and other sports (Gilovich et al., 1985; Bar-Eli et al., 2006). Prospect theory provides a useful framework for identifying risk attitudes and reference points of users, for instance Jullien and Salanié (2000) provide an application using British racetrack patrons. In a similar way, descriptive frameworks for intertemporal choices can be used to identify what kind of discounting function users are employing, as well as biases in comparison with the normative model, for example the effect of framing in borrowing and lending decisions of private households conducted by Breuer and Soypak (2012). The detection of heuristics in behavioural data sets is to date rather more limited, but an interesting example is the comparison of the hiatus heuristic against normative decision rules for three industries, conducted by Wübben and Wangenheim (2008). Following the collective judgment stream, behavioural analysis can help in describing which social decision scheme a group is using (Kerr and Tindale, 2011), or how accurate a forecast is (Clemen and Winkler, 2007), as exemplified by the extensive analysis of accuracy of weather forecasts conducted by Clemen and Winkler (1990). As these examples illustrate, despite a relatively large number of analysis of behavioural issues already available in the literature, they are typically descriptive.

The second mode we suggest for analysing behavioural datasets is to exploit behavioural findings. This of course has already been done in marketing and sales in practice (Bazerman, 2001), but again we believe that behavioural data sets are rich sources for BOR studies and applications. There are several ways of exploiting cognitive biases: for example, after detecting inadequate time discounting for debt repayment Capital One targeted customers that borrowed large amounts on credit cards and repaid them slowly (Davenport and Harris, 2007). Prospect theory can help in framing products in a positive way by changing the reference point, such as in the marketing of financial products (Shefrin and Statman, 1993). The UK Government Cabinet Office’s Behavioural Insights Team have used large experimental trials to show how framing effects can be successfully exploited to increase the prevalence of salary-based charitable donations (Behavioural Insights Team, 2013a) and registration for organ donation from members of the public (Behavioural Insights Team, 2013b). The heuristics stream may be employed to identify the heuristic guiding choices of users and then exploit this finding, such as the use of recognition heuristic for assembling investment portfolios (Ortmann et al., 2008). The intertemporal choice stream can support the understanding of time preferences of customers, such as their trade-off between delivery time versus delivery cost in online shopping (Zaubereman et al., 2009). Analysis about collective judgments can be used to increase the appeal of
products and services offered to a group of users. Of course, ethical issues arise when BOR practitioners are exploiting such behavioural trends, and those must be carefully considered in every application.

The third mode for exploring behavioural datasets is to improve the judgments and choices of users involved in transactions stored by these databases. This can be an aim for customers of private organizations, as advocated by Bazerman (2001), as well as for citizens in public contexts (Thaler and Sunstein, 2008; Keeney, 2008). Patterns identified by the cognitive biases and prospect theory streams can often be reduced or eliminated by the use of adequate tools, such as proper incentives and engineered elicitation tools (Fischhoff, 1982; Arkes, 1991; von Winterfeldt, 1999; Larrick, 2007; Montibeller and von Winterfeldt, 2015). In a similar way, the understanding of intertemporal choices and preferences can help either to guide individual decisions or to design better public policies, such as changing health-related behaviours or long-term health planning, respectively (Attema, 2012). An understanding of efficient heuristics can help to improve the speed and accuracy of judgments (Gigerenzer and Gaissmaier, 2011). Collective judgments can be improved by using adequate aggregation rules of individual estimates (Clemen and Winkler, 2007; Kerr and Tindale, 2011) and debiasing individual estimates. With the exception of “nudges” (Thaler and Sunstein, 2008; Johnson et al., 2012), for instance making opting-in the default for pension planning, we are not aware of such debiasing efforts in behavioural databases. There are thus opportunities for BOR practitioners on this front.

Table 1 summarizes the BDR streams and the possible modes that they can be used for exploring behavioural datasets. We see two ways of employing this taxonomy in practice. The first one is descriptively, to conceptualise studies in which behavioural data sets were explored and root it within BDR stream(s) with a clear mode of practice. The second one is prescriptively: the BOR analyst might use it to help identifying which issues she may want to explore and locate them within BDR stream(s), then decide what is her purpose in the data exploration. Next we show how the taxonomy might be employed in practice.

4 Exploring Behavioural Datasets in Action

In this section we present three illustrative examples, using real-world behavioural datasets, which showcase the three modes of exploration we suggested (detecting in Study 1, exploiting in Study 2, improving in Study 3). We have attempted to give broad coverage in terms of BDR streams (collective judgments in Study 1, prospect theory and heuristics in Study 2, and cognitive biases and debiasing in Study 3). The cases are essentially a convenience sample based upon which datasets and findings were available to us and not limited by confidentiality agreements. Our aim in these three examples is to show the potential that behavioural data mining can provide, and to give a sense of the kinds of
questions that might be tackled in the various modes of behavioural analysis, rather than a comprehensive analysis of a particular decision problem or an attempt of trying every possible combination in our taxonomy.

4.1 Case I: Assessing the “wisdom of the crowd” for non-convex loss functions (Detect mode)

Our first case is drawn from the collective judgments stream, and adopts a detect mode to evaluate the quality of collective judgments in a sports prediction game.

4.1.1 Overview

The “wisdom of the crowd” is a popular aphorism used when a collective judgment performs well relative to the individual judgments that comprise it (Surowiecki, 2005; Vul and Pashler, 2008; Önkal et al., 2012). Studies have shown that when averaging is used both to aggregate individual judgments and errors, collective judgments are more accurate for any convex loss function (Larrick and Soll, 2006), typically by between 3% and 24% (Armstrong, 2001).

Our study is motivated by an application where collective judgments were often found to perform relatively poorly in a sports prediction game. Unusually, in this application poor collective judgments are not necessarily viewed negatively. Users are given post-game feedback about how their predictions compared to the collective prediction. Poor collec-
tive predictions thus potentially paint individual players in a favourable light, giving them the impression that their own predictions were relatively good, and thus that they possess above-average forecasting ‘skill’. Our clients wanted to explore this lack of “crowd wisdom” and better understand how it was affected by their choice of function for rewarding predictions.

Our analysis shows that the apparent poor performance is due to the non-convexity of a loss function that is routinely used to evaluate sports predictions, and that the collective judgment in fact performs well under more conventional loss functions like absolute error (AE).

4.1.2 The Context: Sports Prediction Games

The case is based on research conducted for Superbru, an online sports prediction game in which players arrange themselves into groups for the duration of a sports tournament (http://www.superbru.com). Players are awarded points for close predictions and compete within their groups to maximize the number of points earned. Players may belong to multiple groups, typically one or two small groups of friends or work colleagues, and two large groups consisting of supporters of the same team and the entire user-base. Playing the game is not incentivised by the company with financial rewards (except for the most accurate player in a tournament wins a prize with a cash value between $500 and $2,000, with over 80,000 users this is not considered to be a significant cash incentive), limiting the degree to which findings can be extended to more common sports betting contexts.

For the tournament we analyzed, predictions consist of two components: a winning team and a margin of victory. Users earn one point for correctly predicting the winning team, and half a point if their margin prediction is within five points of the actual score margin.

4.1.3 Data and Methodology

Our data consists of 8.7 million predictions made by 87,899 users over the course of the Super Rugby tournament held in 2011, as well as the country of origin and team supported by each user. The Super Rugby tournament is the largest rugby union tournament in the Southern Hemisphere; in 2011 it consisted of 15 regional teams from New Zealand, Australia and South Africa. The league phase of the tournament consisted of 119 games, followed by 5 knock-out games.

Individual judgments constitute our basic data, while collective judgments (the ”crowd prediction”) is given by the mean of the predictions made for each game. For each game then, we can assess the accuracy of both individual and collective judgments using the following three loss functions:

1. *Superbru error* (SBE). From a baseline of zero, a prediction incurs penalties (i.e.
errors) of $p$ points if the winning team is incorrectly predicted and $0.5p$ points if the predicted points margin differs from the actual points margin by more than five points. These errors are added together to form the SBE. The choice of $p$ is arbitrary and does not affect results when examining the metric in isolation (i.e. the extreme left boxplots in Figure 2), but because we later form weighted combinations of the SBE and AE we select $p$ so that the mean SBE over games is equal to the mean AE. This is a heuristic means of providing rough comparability between the scales of the error metrics. The mean SBE is $0.74p$ and the mean AE is 12.55, so $p$ is set to 17.

2. **Absolute error (AE).**

3. **Bounded absolute error (BAE).** This loss function returns the AE within the error margin $[-5; +5]$ and a constant error of 5 outside of these bounds. As for the SBE, this function can also be scaled arbitrarily by a factor $r$ that has no effect on the performance of the BAE metric but ensure comparability (on average) when forming convex combinations with the SBE. The mean BAE is 4.35 and we therefore set $r = 2.9$.

Figure 1 shows the loss functions with $p = 17$ and $r = 2.9$.

![Figure 1: Loss functions used to evaluate collective judgments, assuming an actual points margin of +9. SBE = error used by the Superbru game; AE and BAE = unbounded and bounded absolute error respectively.](image)

4.1.4 **Results**

The top set of boxplots in Figure 2 shows distributions of the relative improvement returned by collective judgments for convex combinations of the SBE (left-most boxplot) and AE (right-most boxplot). The distributions show that, under the kind of non-convex loss functions often used in sports prediction games, collective judgments (a) return on
average smaller errors than the average errors made with individual judgments, but (b) return substantially worse forecast errors relatively often – note that the lower quartile for a SBE-only loss function shows that in one quarter of games the error from the crowd prediction is at least 8% larger than the average user error.

If all weight is placed on the AE, collective judgments improve forecast errors by on average 25%. Although there is substantial between-game variability (improvements vary from less than 0.01% to 96%) and strong positive skewness (the median improvement is 11%), the results return the expected “crowd wisdom” effect and agree in broad terms with the effect sizes reported in Armstrong (2001). Under the non-convex loss function currently used by the game (i.e. with zero weight placed on the AE), the mean and median improvement are actually larger than under AE loss (29% and 16% respectively), but more notably in 51 of the 125 games played the collective judgment is shown as inferior, at times substantially so.

![Figure 2](image)

**Figure 2**: Boxplots showing distributions of reductions in errors obtained using collective judgments (relative to the average errors of individual judgments) for convex combinations of the SBE and AE (top panel) and SBE and BAE (bottom panel). Boxes show medians and quartiles; whiskers show non-outlying minima and maxima; and means are indicated with crosses.

The bottom set of boxplots in Figure 2 shows that the “wisdom of the crowd” is substan-
tially reduced as the BAE (right-most boxplot) is incorporated into the SBE (left-most boxplot), both in mean and median terms. In small amounts, the BAE decreases the variability of the crowd effect by drawing in the right-tail (good crowd performance) far more than the left-tail (poor crowd performance). The net effect is that although mean and median crowd wisdom decline substantially as the weight placed on AE decreases, the frequency of inferior collective judgments remains relatively constant until a substantial weight is placed on the BAE (in our example, exceeding 30%). As more weight is placed on the BAE the median and mean crowd wisdom eventually become negative. While basing the game purely on BAE is not considered to be a desirable option because of the enjoyment derived from predicting winning teams, a reward function that includes a small BAE component is practically feasible. One such function (effectively placing a one-third weight on BAE) would simply augment the current margin system by offering between 0.5 and 1 points for a predicted-actual margin difference of between five and no points.

4.1.5 Behavioural Insights

Here the analysis of the behavioural dataset showed that “the wisdom of crowds” is heavily dependent on how judgmental estimates are evaluated, and hence something that is at least partly under the control of the Superbru’s developers. The frequent lack of “crowd wisdom” observed in the game was shown to be a desirable but unintended side effect of the choice of loss function, and something that could even be reduced further by small modifications to the currently-used loss function. A related concern was that some users might be cheating by creating multiple profiles and using these to gather information on collective judgments (shown to users once they have submitted their own predictions). Our behavioural analysis suggested that there is relatively little to be gained from such a strategy, particularly under the proposed loss function, and that the issue of if and when to show information about collective judgments is predominantly one of user experience (including perceived fairness) rather than actual fairness.

4.2 Case II: Determinants of charitable giving between users of an online slot machine game (Exploiting mode)

Our second case is drawn from the heuristics stream, but employing also concepts from the prospect theory stream. It illustrates how the exploiting mode can be used to determine a strategy for increasing the likelihood of charitable giving between users of an online slot machine game.

4.2.1 Overview

Previous research suggests that people’s propensity for giving can be influenced by, among other things, asset class, how the assets have been acquired, recipients and the potential
for reciprocity, and past behaviour (e.g. Andreoni and Miller, 1993; Hoffman et al., 1994, 1996; Soyer and Hogarth, 2011). Fairness and reciprocity have been studied by operations researchers working in supply chain management, where it has been shown to play an important part in empirical contracting mechanisms (Loch and Wu, 2008; Katok and Wu, 2009). Several researchers have viewed fairness and reciprocity from a “heuristic” perspective (Axelrod, 1984; Todd et al., 2008), demonstrating how simple heuristic strategies (“tit-for-tat”, for example) can perform well in interactive, iterative decision contexts (Kraines and Kraines, 1993; Messick and Liebrand, 1995). Furthermore, several of the empirical findings around charitable giving can be explained as reference-point effects i.e. they influence what are viewed as gains and losses, with subsequent impacts on decision making (Tversky and Kahneman, 1992; Arkes, 1991). While these effects are essentially heuristic in nature, they are also one important element of prospect theory. Hence, there are some insights from that stream in this case.

Our client wanted to encourage the “gifting” (giving without incurring personal loss) of tokens by its users. From the perspective of the online casino, gifting is desirable because it targets potential new users and introduces a social aspect into the game. We find that, despite the fact that users are giving away the casino’s tokens (rather than their own), users are more likely to gift if they have themselves been the object of gifting, either from other users or by redeeming daily offers made by the casino.

4.2.2 The Context: Social Slot Machines

The case is based on research conducted for Online Skill Ltd.’s Clickfun Casino application, a suite of online slot machine games that are hosted on the web and Facebook (http://www.clickfun.com and http://apps.facebook.com/clickfuncasino). The games are not played for any monetary reward. Users receive a certain base quantity of free tokens each day when they log onto the game, and in addition have the option to claim further “bonus” tokens (a small quantity for each hour that they are logged on for, and a larger quantity by following an email link sent out once per day). Further tokens may be purchased for cash. Tokens can only be used to play the games i.e. cannot be redeemed or used in any other way.

Users can also “gift” parcels of tokens to any contact in their social network. The “gifted” tokens do not reduce the user’s balance, but each user is limited to giving one parcel of 50 tokens (or 10 “spins”, worth approximately 150 tokens) per contact per day.

4.2.3 Data and Hypotheses

Data is continuously collected in the form of time-stamped game events (spins, wins, coupon activity, sending and receiving tokens, etc.) for each user. Our current analysis is based on event data obtained for some 6.3 million events obtained for a subsample of
12,000 users over a five-day period. From this data, we extract information on the number of token and spin parcels gifted by each user. We wish to relate these outcomes to the following inputs:

- \( x_1 \): an indicator of whether spin gifts were received from other users;
- \( x_2 \): an indicator of whether token gifts were received from other users;
- \( x_3 \): the number of bonus tokens claimed (centered and scaled to unit variance to facilitate interpretation);
- \( x_4 \): an indicator of whether bonus tokens were actually played;
- \( x_5 \): the “win percentage” (total number of tokens won divided by total number of tokens played);
- \( x_6 \): the total number of spins made (log-transformed).

On the basis of previous research, we expect that being the recipient of charitable gifts will exert a stronger positive influence on the propensity to gift than personally “earnings”, whether measured by success or time expended (Hoffman et al., 1994; Rutström and Williams, 2000). Because exchanges are between social network contacts and parties are personally identified, we expect user-to-user giving to exert a stronger influence than relatively anonymous company-to-user giving (Majolo et al., 2006). We do, however, still expect a significant company-to-user effect (Khouja et al., 2013, e.g.).

4.2.4 Results

Table 2 shows odds ratios (OR) and related significance results for logistic regression models fitted to our data. All fit indices indicated that our models fit reasonably though not exceptionally well. Likelihood-ratio tests against null models are highly significant (spins: \( \chi^2_6 = 3351.61, p < 0.001 \); tokens: \( \chi^2_6 = 1581.92, p < 0.001 \)), and pseudo-R2’s are fairly high (spins: 32.3%; tokens: 24.3%). Hosmer-Lemeshow goodness-of-fit tests returned non-significant i.e. favorable results (spins: \( \chi^2_8 = 12.26, p = 0.14 \); tokens: \( \chi^2_8 = 9.95, p = 0.27 \)). At optimal cut-off values model sensitivity and specificity are 76% (spins, cut-off = 0.16) and 75% (tokens, cut-off = 0.07). Areas under the ROC curves are 0.85 (spins) and 0.83 (tokens) models respectively.

Both spin and token models show substantial agreement. Receiving and redeeming gifts from other users shows the largest positive association with gifting propensity. Relationships are strongest where the received and sent gifts take the same form. Receiving spins increases the likelihood of sending spins more than receiving coins does, and vice versa.

A weaker but still strongly significant positive association exists between gifting propensity and receiving and redeeming gifts from the company itself. Here, where we can discern
the effect of receiving and actually using the gifts, it appears that the act of using a gift has a stronger influence on gifting than merely claiming and storing it. The two measures of “earned” winnings (effort and success) play the smallest (in the case of winnings, insignificant) role. The non-significance of winnings may be due to these being non-monetary and related to chance in a transparent way, although in mitigation of this a number of studies have found that winnings from gambling are usually interpreted as “earned” (see Delfabbro (2004), or Biner et al. (1998) for a non-monetary example).

4.2.5 Behavioural Insights

Online Skill’s case involves a number of peculiarities. Users are encouraged to donate gifts that do not belong to them, have no monetary value, and do not form part of a zero-sum transaction (in the sense that nobody is disadvantaged by the transaction). Nevertheless, some familiar behavioural patterns emerge. Users are more likely to initiate the gifting transaction if they have been recipient of a gift than if they have been successful through their own doing. Gifts received from a (usually known) individual exercise a larger effect than gifts received from the company itself, but the latter still elicits a positive response. Spending time playing also positively influences gifting, although less strongly; winning, apparently, plays no significant role. This suggests that gifting arises predominantly out of a sense of fairness and obligation, and that this sense is unaffected by success.

From Online Skill’s perspective a positive feature of the gifting process is that it appears to be self-reinforcing, so that the company’s primary aim should be on initializing rather than maintaining gifting. The two basic options available for this are to increase the win percentage (since no money is at stake) or to offer more (or more attractive) bonuses. Our analysis suggests that the latter would be far more productive. As the number of bonuses on offer in the game is already large, users should ideally simply be encouraged to take up and use the existing offers. This applies particularly to users who are not currently gifting. Although these results do not directly suggest how this goal could best be achieved, Behavioural analysis clearly provides a means for customizing and continuously assessing the effect of any potential strategies.
4.3 Case III: Debiasing insufficient adjustment in online sports predictions (Improving mode)

Our third case is a hypothetical application of the improving mode and uses concepts from the cognitive biases and debiasing streams of BDR. Although we have not implemented any debiasing procedures in reality, this final case hopefully provides some illustration how BOR practitioners might use this mode, in this case to improve the predictions made in the context of an online sports prediction game.

4.3.1 Overview

Forecasts, like many kinds of judgments, are prone to biases arising from use of heuristics. A number of previous studies (Lawrence and O’Connor, 1992; O’Connor et al., 1993; Lawrence and O’Connor, 1995) have shown that people over-react to the feedback provided by the last actual value when setting their new forecasts. Other studies have shown that affective or emotional responses can also play a significant role in the setting of forecasts (e.g. Wilson and Gilbert, 2003).

The particular type of forecast we consider here is a prediction of the outcome of a competitive match between two teams in a league setting. That is, each contest involves two teams, teams compete regularly over the course of a relative short season (once per week for 14 weeks in our setting), and each team plays each other team only once (or a small number of times) over the course of a season, so that “learning opportunities” are limited.

The aim of the current case is to demonstrate the existence of a systematic tendency to under-react to historical information for teams of which one is a supporter, using the same Superbru predictions examined in Case 1, and to suggest ways in which these forecasts may be debiased, in the sense of reducing the magnitude of the observed biases.

4.3.2 Data and Hypotheses

Our data consists of 8.4 million predictions made by 87,603 users in 119 Super 15 rugby matches over the course of a single season in 2011 – a more detailed description was provided in Section 4.1.

One plausible mechanism for creating forecasts in the given context is to start with an initial assessment of the general quality of the two teams, and then adjust this assessment by taking into account recent information about the teams – most notably, how they performed in the previous week’s matches. Standard heuristics-and-biases research suggests that the strength of the anchor will be partly determined by affective or emotional responses to the teams involved (e.g. Wilson and Gilbert, 2003). This leads directly to the hypothesis that the sensitivity of predicted outcomes to past performance will be lower for fans of a team than for non-fans. We test this hypothesis by fitting a linear model to
the predicted winning margin $y_{i,t}$ given by subject $i$ in game $t$:

$$y_{i,t} = \beta_0 + \beta_1 h_{t-1} + \beta_2 a_{t-1} + \beta_3 F_i^{(h)} + \beta_4 F_i^{(a)} + \beta_5 h_{t-1} F_i^{(h)} + \beta_6 a_{t-1} F_i^{(a)} + \epsilon_{i,t}$$

where $h_{t-1}$ and $a_{t-1}$ denote the actual margin of the home (away) team in its previous game, $F_i^{(h)}$ and $F_i^{(a)}$ are binary variables indicating whether subject $i$ is a self-reported supporter of the current home (away) team, and $\epsilon_{i,t}$ is a random error term.

Because two teams are involved in each game (one of which plays “at home” and the other “away”) the sign of $y_{i,t}$ is arbitrary and we denote positive (negative) values to mean a win for the home (away) team. Note though that positive value for $h_{t-1}$ and $a_{t-1}$ indicates a win for that team i.e. $h_{t-1} > 0$ and $a_{t-1} > 0$ mean wins for the current home team and current away team in their previous matches (irrespective of whether these were played home or away).

Our hypothesis about the role of affect implies that, in the above linear model:

- $\beta_1 > 0$ (a positive association between past and forecasted performance of the home team); $\beta_3 > 0$ (a positive fan bias for the home team); and $\beta_5 < 0$ (smaller adjustment i.e. a weaker relationship between past and forecasted performance for fans of the home team),

- $\beta_2 < 0$ (a positive association between past and forecasted performance of the away team, noting the sign change), $\beta_4 < 0$ (a positive fan bias for the away team) $\beta_6 > 0$ (smaller adjustment i.e. a weaker relationship between past and forecasted performance for fans of the away team).

In order to also make some assessment of the sensitivity of actual (rather than predicted) performance to past performance, we fit a second linear model to the actual winning margin $h_t$ (again, orientated so that a positive value denotes a win for the home side). The observed winning margin $h_t$ in game $t$ is modelled by:

$$h_t = \beta_7 + \beta_8 h_{t-1} + \beta_9 a_{t-1} + \epsilon_t$$

This model is fitted to data gathered on all observed game results from the inception of the Super Rugby tournament in 2006 up to and including the 2011 season. This gives a total of $n = 575$ games (note that in this case games are the unit of observation rather than forecasts made by a user for a game – hence the much smaller sample size). The results obtained from this second model provides the historical sensitivity of actual performance to past performance in Super Rugby games and thus provides a basis for an approximate comparison of whether forecasts are more or less sensitive to past performance than actual results.
4.3.3 Results

Table 3 shows coefficients (with associated standard errors in parentheses) for the two linear models described above. All effects are highly significant at $p < 0.001$ although the models explain only a moderate proportion of the variation in their respective outcomes: $R^2 = 18\%$ for forecasted margins and $R^2 = 10\%$ for actual margins. We find that, on average, home ground advantage conveyed an actual benefit of 4.6 points, but that predicted outcomes only favour the home side by 2.8 points – that is, the effect of playing at home is systematically underestimated by Superbru users. In contrast to previous research (e.g. Lawrence and O’Connor, 1992; O’Connor et al., 1993; Lawrence and O’Connor, 1995), we find that forecasts i.e. predicted outcomes, show the same sensitivity to previous performance as do actual outcomes – no significant differences are observed. Turning to our hypothesised relationships, values for $\beta_1$ through $\beta_6$ are all in the hypothesised direction and are highly significant, demonstrating that affective responses result in systematic biases in favour of the supported team, and a tendency to underweight empirical data in the form of past performance (more so for the home team than the away team).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Forecasts</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{t-1}$</td>
<td>Last performance (home)</td>
<td>$\beta_0$ 2.779 (0.004)</td>
<td>$\beta_7$ 4.604 (0.708)</td>
</tr>
<tr>
<td>$a_{t-1}$</td>
<td>Last performance (away)</td>
<td>$\beta_1$ 0.217 (0.001)</td>
<td>$\beta_8$ 0.252 (0.044)</td>
</tr>
<tr>
<td>$F(h)$</td>
<td>Fan (home)</td>
<td>$\beta_2$ -0.187 (0.001)</td>
<td>$\beta_9$ -0.184 (0.041)</td>
</tr>
<tr>
<td>$F(a)$</td>
<td>Fan (away)</td>
<td>$\beta_3$ 5.228 (0.015)</td>
<td>n/a</td>
</tr>
<tr>
<td>$h_{t-1}F(h)$</td>
<td>Last perf. $\times$ Fan (home)</td>
<td>$\beta_4$ -6.147 (0.015)</td>
<td>n/a</td>
</tr>
<tr>
<td>$a_{t-1}F(a)$</td>
<td>Last perf. $\times$ Fan (home)</td>
<td>$\beta_5$ -0.068 (0.001)</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\beta_6$ 0.016 (0.001)</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 3: Results from two linear models show the sensitivity of forecasted and actual performance to past performance and fan status, demonstrating the importance of affective responses to forecast behaviour.

4.3.4 Behavioural Insights

The analysis above clearly outlines three systematic biases in users predictions: a tendency to underestimate the advantage gained by playing at home, a tendency to overestimate the performance of the team one supports, and a tendency to under-react to the past performance of the team one supports. While it is not in the interest of Superbru management to attempt to convince their users of these biases, or to correct for them, it is not difficult to envisage the kinds of debiasing interventions that might lead to better forecasts. Indeed the literature around debiasing (e.g. von Winterfeldt, 1999; Larrick, 2007; Montibeller and von Winterfeldt, 2015) suggests that a simple message informing users about the established biases when they are asked to make a prediction, or the use of counter-factual statements warning them against distorted predictions, might persuade
at least some users to reconsider their intuitive judgments and in doing so might achieve some reduction in the magnitude of the observed biases. Accordingly, while we have not attempted debiasing procedures in this particular case, we hope that readers will see how behavioural analysis might fruitfully be applied in an improving mode to further these kinds of aims.

5 Conclusions

In this paper we suggested different modes that can support Behavioural Operational Research (BOR) researchers and practitioners in using Behavioural Decision Research (BDR) for exploring behavioural datasets. We identified six main streams of BDR (cognitive biases, prospect theory, intertemporal choices, debiasing, heuristics and collective judgments) and suggested how they can be used in three modes: describing behavioural trends, exploiting behavioural findings, and improving user’s judgments and choices.

We illustrated the use of this taxonomy in three real-world studies, each one covering a mode of exploration. These six BDR streams are not mutually exclusive, and BOR researchers and practitioners may well explore more than one in a single study, for instance by detecting a cognitive bias and then improving judgments with a suitable de-biasing tool, as illustrated in Case III, or by drawing concepts from one stream (prospect theory) to analyze a problem under another stream (heuristics), as done in Case II.

While the behavioural literature is already widely available to BOR researchers and practitioners, our contribution is to summarize the large set of developments in BDR in a coherent and hopefully useful way for BOR research and practice. Although behavioural issues have been regularly analyzed in the past, particularly in marketing research, we hope that our review and taxonomy provides a way of organizing general behavioural issues in behavioural datasets, as well as helping to provide structuring tools to unstructured behavioural databases.

We recognize that our classification of BDR streams and modes of practice, in common with any taxonomy, simplifies a large body of literature and several types of application areas. In addition we do not make any claim that the taxonomy, despite our best efforts, is completely exhaustive. There are therefore several areas open to further research. From a descriptive point of view, systematic surveys could investigate how practitioners are using behavioural datasets in practice and if such deployment matches our taxonomy, thus validating or refining/redesigning it. A research agenda with multiple-case studies, in which the taxonomy is prescriptively employed to guide the choices of streams and modes, also would be welcome and in line with calls for practitioners to develop Behavioural OR case studies (Ormerod, 2017). From a conceptual point of view, research could be conducted on further formalization of the modes of use we suggest and their links with each stream. In addition other BDR streams might be incorporated into the taxonomy,
given how large the field is and how fast it develops. Finally, our literature review showed that the improving mode is currently underdeveloped; there is clearly much scope for BOR practice to improve user’s judgment and decision making, particularly in public decisions (Tsoukias et al., 2013), and to BOR researchers to assess the impact on users and organizations when this mode is employed.

Acknowledgements
The authors wish to thank those at Superbru and Online Skill Ltd. for granting permission to report results obtained using their data.

References


