NatCen
Social Research that works for society

Scoping the use of industry data on category B gaming machines

Authors: Heather Wardle, Carla Seabury, Hashim Ahmed and Catherine Coshall
Date: 09/12/2013
Prepared for: The Responsible Gambling Trust

Contents

Executive Summary

Aims and methods

- This project aimed to understand and assess what types of data gambling industry operators hold and retain on category B gaming machines and how these data might be used for research purposes.
- Category B machines are those with the highest levels of stake and prizes currently available in Great Britain. In recent years, they have come under increasing scrutiny from policy makers and other stakeholders. Category B machines are available in range of venues including bookmakers, casinos, bingo halls and adult gaming centres (AGCs).
- Semi-structured interviews were held with 13 different gambling industry operators/machine suppliers to map the types of data they generate, the size of the resulting data, how long data are held for and what the potential benefits and limitations might be in using these data for research and policy purposes.
- In addition to semi-structured interviews, sample data, lists of data metrics and other supporting information were shared by operators with the research team.

Findings

Data types

- The interviews identified three different types of data generated and held by gambling industry operators. These were: transactional data, player tracking data and proxy session data.
- **Transactional data** are financial accounting data that monitor what money is put into the machine and what money is paid out. In some sectors, such as licensed betting offices (LBOs), every single financial transaction is recorded. In other sectors, such as bingo clubs and AGCs, these transactions are recorded at aggregate levels, such as once a day, once a week and so on.
- Transactional data are not linked to game system data. In other words, it is not possible to link game features (such as bonuses, feature games, near misses/wins) and respective player experiences with subsequent bet and stake size. For example, if someone 'nearly won' on a previous spin you cannot see this in the data and therefore cannot judge what the player's reaction was in terms of their next bet.
- **Player tracking data** are data that are generated from loyalty card programmes, similar to other reward card schemes. This records data in a unique session of play for that individual. However, the data recorded are still transactional; you can see how much someone staked and won/lost within a single session of play. Information is available on session length and date/time of play. Providing a player uses their card consistently, you can also link together multiple sessions of play.

- Availability and uptake of player loyalty card varies by sector. Players engage in these schemes on a voluntary basis. Three out of five LBOs included in this research had loyalty card data but some schemes were very new. Uptake was low, meaning that most players do not use these cards. Most casinos have loyalty card data with higher levels of uptake. This sector estimated that between one third to two thirds of all machine play was player tracked. Player tracking had only reached trial stage among the bingo operators included in this study and neither of the AGC operators had player tracking.
- **Proxy session data** are transactional level data that have been sliced up into discrete chunks based on rules of what might constitute the start and end of a session of play. Proxy session data were only available for the LBO sector.
- Rules governing identification of proxy sessions of play varied from operator to operator. Levels of accuracy, in terms of whether the proxy session really did identify a unique session of play, were largely unknown.
- Proxy session data do not allow different sessions of play for an individual to be linked they only identify single and unique session of play.

Gaps and limitations

- Across the category B estate in Great Britain, there is a great deal of inconsistency in the level and type of data collected. Some operators have transactional data which logs every financial transaction, player tracking data and proxy session data. Some only have aggregate levels of transactional information.
- There is also inconsistency within sectors. This relates to both the type of data collected and the naming conventions used.
- These inconsistencies create a logistical challenge for any researcher wanting to use these data for research purposes, as such inconsistencies have to be reconciled before any analysis can be undertaken.
- The level of industry-held data was described as ‘data rich; information light’. This means that there is a vast amount of data available but it provides insight only into a very narrow range of issues – namely those of financial transactions.
- Industry-held data on category B machines does not provide information about what is happening during the game, meaning reactions to game play can not be investigated.
- With the exception of casino player tracking data, there is no demographic information available about players. Information about whether the player generating the data is experiencing any form of gambling-related harm is not available.
- These data also lack contextual background which would be useful for explanatory purposes. This includes contextual information about the venue which might include information about machine layout, opening hours, availability of ATM machines etc, and also of the local area which might include metrics like location type (high street, out of town shopping centre etc), demographics of the local area etc.

Opportunities

- Some of these gaps can be supplemented with knowledge from other research. For example, broader contextual data could be generated from other administrative datasets (i.e., local area statistics, deprivation) and merged onto industry transactional data, so long as venue location was known. Surveys and qualitative research can also be used to broaden understanding.
- Industry-held data offers some unique opportunities for researchers and policy makers. Transactional data can be used to better examine volumes of play at different levels, for example, at a machine level, at venue level, region and so on. More detailed information about staking patterns could be generated also. This would be especially useful when new regulatory policies are introduced to examine impact pre-and post-implementation.
- Player tracking data, though in its infancy, offers an opportunity to examine within and between sessions play for some people. Here questions to be explored could be when do people stake up, stake down, change games and so on. This would need to be accompanied by methodological work to explore who uses these cards, why and under what circumstances.
- Finally, proxy session data could be used to better explore the sequence of staking events within a unique session of play. As this would be limited to machines in bookmakers, this could look at when within a session people stake at the maximum amount, trajectories of play within sessions or whether people play differently when playing with their own money or with money they have won and reinvested.
- These represent areas where there are currently significant gaps in knowledge. Using industry-held data could help to fill these gaps, though the challenges of doing so should be acknowledged.

Recommendations

- Given the complexities, inconsistency and gaps identified within industry-held data on category B machines, we recommend that the next step in this programme of research be a series of demonstrator projects. These projects should aim to further document the processes and challenges of using industry-held data for research purposes whilst seeking to demonstrate the analytic potential of these data.
- Industry-held data should be viewed as contributing to the researcher's methodological toolkit but does not and cannot answer all research and policy questions in this area alone.

1 Introduction

1.1 Policy context

In the last decade, the issue of gambling-related harm has gained increasing prominence with policy makers, media and the public alike. In Great Britain, the Gambling Act 2005 changed the way gambling was advertised, licensed and regulated. Since then, there has been increasing demand for evidence about gambling behaviour, including both individual and societal impacts.

Many research studies have been conducted all over the world exploring this issue. In Great Britain alone, there have been several recent studies looking at gambling; including routes into and out of gambling, gambling prevalence and problem gambling. One of the most prominent studies was the British Gambling Prevalence Survey (BGPS) series, a large-scale and nationally representative survey series looking at participation in gambling. The last study of this series was conducted in 2010. There have also been informative qualitative studies conducted and, in combination, these studies and others have contributed to the policy debate.

Whilst these studies, specifically the BGPS series, provide solid information about broad patterns of gambling behaviour and are useful in monitoring changes in behaviour and attitudes over time (i.e., the macro perspective), they are somewhat less useful in providing detailed insight into specific patterns of behaviour at a very fine-grained level (i.e., the micro perspective). A classic example is that of gambling expenditure data. Understanding who is spending what on gambling, and under what circumstances, is a classic area of policy interest. Knowing this could help to better estimate the financial impact of certain policies or identify those with (potentially) risky patterns of play. However, collecting information about gambling expenditure within a survey setting is notoriously difficult and resulting data tends to be inaccurate. This is because players often simply do not keep track of how much they are spending. When asked, they tend to overestimate winnings and underestimate losses. This is further complicated by issues of how to account for reinvested winnings (i.e., playing with the house's money) in expenditure calculations. However, this is the type of data gambling operators are routinely capturing. In short, if you want to know about patterns of expenditure on gambling, the best and most reliable data will be those generated by the operators themselves.

This is exactly the sort of administrative data that policy makers would wish to have access to, and there is growing demand for these types of data across many policy areas. Using administrative data has the further benefit of not incurring the level of costs usually associated with large scale ad hoc surveys, is less difficult to collect, and in some cases is arguably more reliable. This broader movement among social scientists and policy makers to utilise administrative data, combined with these potential benefits, provided impetus to assess how gambling industry-held data could

be used for research purposes. This was with the specific aim to assess if and how this could be integrated into the evidence base.

Interest into category B gaming machines (that is, those with the highest stakes and prizes currently available in Great Britain) has increased in recent years. In 2010 the Responsible Gambling Strategy Board made further examination of the relationship between gambling machines and behaviour a strategic priority.[1] In January 2013, recognising the thin evidence base upon which to build policy, the government also called for further exploration into the relationship between gambling behaviour and category B machines.

Whilst there have been some studies exploring the relationship between machines and behaviour in Great Britain, these have tended to be small scale in nature, conducted prior to the widespread introduction of the then-called 'Fixed Odd Betting Terminals' or provide only limited insight because of their approach. For example, secondary analysis of machine players interviewed in the BGPS was published in March 2013. This produced some insight into player types but did not contain a depth of detail about specific machine play behaviour.[2] Therefore, combining growing interest in using administrative data records for research purposes and the increasing policy focus on category B gaming machines became a strategic priority for the RGSB. This created the opportunity to work with Great Britain's leading operators and suppliers of machines to explore the extent to which data they capture and store could be used for research purposes.

It is within this policy context that the Responsible Gambling Trust (hereafter referred to as 'the Trust') has commissioned research into category B gaming machines with the broader aim to better describe and understand behaviour and to help develop policy which militates against gambling-related harm. This study is the first in this planned series of research to explore what contribution industry-held data might make to these broader aims and objectives.

1.2 Machine types

There are 12 categories of gaming machine in Great Britain which fall into distinct types, depending on their maximum stake and maximum prize. This project is concentrating on category B machines which are present in arcades, casinos, bingo halls and bookmakers. Table 1 shows the different machine types, as well as their maximum stake and maximum prize.

[1] See Responsible Gambling Strategy Board (2012) *Strategy 2012*. Available at: http://www.rgsb.org.uk/publications.html

[2] Wardle H, Sutton R, Philo D, Hussey D, Nass L. (2013) *Examining machine gambling in the British Gambling Prevalence Survey*. Gambling Commission.

Table 1 Gaming machine types		
Machine category	*Maximum stake*	*Maximum prize*
A	Unlimited	Unlimited
B1	£2	£4000
B2	£100	£500
B3	£2	£500
B3A	£1	£500
B4	£1	£250
C	£1	£70
D (non money prize- other than crane grab machine)	30p	£8
D (non money prize- crane grab machine)	£1	£50
D (money prize)	10p	£5
D (combined money and non money prize- other than coin pusher or penny falls machines)	10p	£8 (of which no more than £5 can be a money prize)
D (combined money and non money prize- coin pusher or penny falls machines)	10p	£15 (of which no more than £8 can be a money prize)

This project focuses on category B gaming machines: that is, on those machines which currently offer the highest stakes and prizes in Great Britain (category A machines are not currently available in the market.)

1.3 Research aims and objectives

The main objective of this scoping study was to understand and assess what types of data industry collect and retain on category B machines and how it might be used for research purposes. Our specific objectives were to:

- Investigate with different operators what data are captured, how they are currently used internally, what format, size and structure they are in (and how this might be manipulated), what data might be captured in the future and how this might be used for research purposes;

- Contextualise these data within broader knowledge and to demonstrate what these data add to the evidence base (i.e., to show where these data add value over and above what is already known, and the research gaps they fill);
- Outline how these data could be used in the future for research purposes and how they could be used to answer key policy questions;
- Provide recommendations about how data collection could be improved to facilitate easier access of this information by researchers in the future.

The machines research programme was designed to be conducted in a number of stages. The first stage involved a series of meetings with individual operators to obtain a better understanding of the types of data held, data structure and size. This was with the aim of documenting what could and could not be achieved using these data and hence to allow the Trust and the RGSB to strategically assess research priorities. In short, the aim of this first stage was to better understand what types of data exist, and in what form, so that recommendations about how to use this for research purposes could be made. This report details findings from this first stage of the research programme. The second stage of this research programme is likely to be a number of commissioned research projects, drawing on these findings and utilitising industry data to enhance our understanding of machine player behaviour.

1.4 Report structure

This report documents findings from the first stage of this broader research programme. Chapter 2 presents details about our data collection strategy and analytical approach. Chapter 3 provides an overview of the main findings, starting with mapping the data structures identified. This includes mapping data types by sector and also presents information on the availability and use of player tracking data and availability of proxy session data. Practicalities around data size and storage are also discussed. Finally, Chapter 4 discusses options for future stages of this research programme. These are broad suggestions so far but are likely to require refinement to ensure they meet strategic research priorities set out by the RGSB and the Trust.

This report has been peer-reviewed by the Trust's Machines Research Oversight Panel (MROP). MROP is a group of experts from academia, data analytics and industry appointed to oversee the aims, objectives and outputs of the Trust's machines research programme.

1.5 Disclaimer

Whilst every attempt has been made to ensure the information presented within this report is accurate, there are some instances where further clarification from the industry operator is needed. This should not alter the broad findings of this report but it does mean details are lacking in some areas. In some cases, examples of industry data were provided to us, meaning that more detailed understanding was obtained for

these operators. Others preferred to verbally describe the data and were unwilling to share further documentation until a later stage in the research programme (i.e., when non-disclosure agreements were in place). In short, the detail obtained in this phase of the study relied on how much detail the operators were willing to share. Finally, attempts have been made to independently check the accuracy of information shared as far as possible (for example, by researching individual loyalty schemes), but much of the data summarised in this report is based on what was shared in the interviews.

Finally, this is a fast-changing world and the results presented are based on interviews conducted between April and July 2013. Since then a number of new initiatives have been announced, such as the Association of British Bookmaker's new Code of Responsible Gambling and Player Protection. This contains the commitment to undertake more consistent central analysis of data to identify abnormal activity both in specific shops and, where possible, relating to individual customers[3] (ABB, 2013). Alongside this, the Gambling Commission has started to investigate whether centralised data for bookmakers could be provided to them. Therefore, there are many new initiatives which seek to make more use of industry data and have the support of the industry themselves. This was unknown at the time of the interviews. Therefore, some of the detail reported here may well become obsolete within a short time frame if these initiatives gain traction.

[3] See http://www.abb.uk.com/news/abb-code-for-responsible-gambling-and-player-protection/

2 Approach

Individual meetings were held with each operator who consented to take part in this study. The aim of these meetings was to carefully scope and fully understand the possibilities of what could, theoretically and practically, be done with the data that operators hold (be it player data or machine data). The table below shows the operators included and their sector. The sector referred to below is the sector in which these operators have land-based machines, it is not a definitive list of all sectors in which they operate.

Prior to meeting with the operators, ethical approval for this study was granted by NatCen's Independent Research Ethics Committee.

Table 2 Operators and suppliers

Name	*Sector*
Aspers	Casino
Betfred	Privately owned and independent Licensed Betting Offices
Coral	Licensed Betting Offices
Gala Coral	Bingo and Casino
Genting	Casino
Inspired	Machine supplier
Ladbrokes	Licensed Betting Offices
London Clubs International	Casino
Paddy Power	Licensed Betting Offices
Praesepe	Adult Gaming Centres, Bingo
Rank	Bingo and Casino
SG Gaming	Machine supplier
Talarius	Adult Gaming Centres
William Hill	Licensed Betting Offices

Broad assessment was made of the estimated market share these operators (in combination) have in their respective sectors. In the Licensed Betting Office sector (LBOs), participating operators represented over 90% of market share. Among casino operators it was c.90%, for bingo operators it was c.65% and among AGCs it was between 25-30%.[4] This study therefore included the majority of category B machine operators in most sectors.

[4] These assessments were typically made based on number of venues rather than revenue, turnover or footfall.

A semi-structured interview schedule was developed for these interviews. This was developed based on helpful feedback from MROP and reviewed and agreed with the Trust prior to using. A copy is provided in Appendix A. The schedule started with the operator or supplier providing some background and contextual information about themselves; looking at their size, number of venues, types of venues as well as numbers and types of machine. Information was also obtained about demographics of their customer base and market share, as knowing the profile of their customers and their share of the market would help to contextualise any data they had.

Information was gathered about the different types of data that they held, both for individual players and for machines generally. The player data section concentrated on whether any player tracking or other type of loyalty data was available.[5] This included asking about the types of data that may be recorded on players from the time of sign up to loyalty schemes and also during play, as well as how the operator used those data internally. Obtaining estimates about what proportion of each operator's customer base had signed up to the loyalty scheme was attempted, though not every operator was willing to share this information. In terms of machine data, the interview sought to examine the type and metrics of data that were recorded directly by the machines, including limitations and whether it included time stamps. The following set of questions were about the structure of the data, where and how it is held, how easy it would be to interrogate as well as its size and granularity. Finally, practicalities of the process, the resources the operator had to help us and whether they would be willing and able to share these resources for the purposes of the study were discussed. This was also an opportunity for operators to raise any confidentiality and data security concerns.

The meetings were audio recorded to ensure no information or action point was missed. Full verbal consent was obtained from each operator before proceeding with the recording.

Following the meetings, the research team at NatCen set up a charting framework in Microsoft Excel, covering the main themes that were discussed. This is called 'framework analysis'. Framework analysis is a qualitative data management technique using a matrix approach where data are summarised into cells with a row representing an individual case (or operator) and a column representing a common theme across the data. The advantage of this approach is that it facilitates the analysis of different aspects of an operator's experiences as well as enabling analysis of particular themes across different operators. In this report, the full diversity of operator practice and opinion is presented and this has been summarised into key themes.

[5] Loyalty card data (whereby a customer uses a card to track play and earn rewards) was identified early in the project as being of particular interest because of the potential to examine individual patterns of play at a fine-grained level and to track play over time.

The framework structure followed the key headings of the interview schedule and included charts on company background, player tracking or loyalty schemes, machine data, data structure and systems, and finally the practicalities of accessing their data and any follow-up points. Data from each meeting were summarised into this framework. Audio recordings of each meeting were listened to twice by different members of the team to populate the charts with the corresponding information and to clarify any points of confusion. This provided the research team with a written record for each operator and supplier.

Using the charts as well as the meeting recordings and notes, formal summaries were then drawn up and provided to each operator. These summaries included any concerns raised around confidentiality and the operator's commercial sensitivity. They were sent to each operator for their approval and/or comments. The purpose of the summary was to provide a written record of the meeting but also to clarify our main take away points or to highlight points for further clarification with each operator. Some operators returned with amendments or points of clarification while others were satisfied that the summaries covered all relevant information. A number of follow-up emails were sent and phone calls conducted to clarify issues, and in some cases, a second meeting was held. This process helped to ensure accuracy of the information collated.

Finally, on 18th June, an all-day workshop was held at the NatCen offices in London. In attendance were the project team as well as a representative from the Trust. The workshop sought to bring all the information together to enable us to develop a model of machine data, looking at data size, structure, technical requirements, player tracking as well as sessional and transactional data. The overall picture of what data looked like and how these varied or did not vary between sectors was discussed, and data formats reviewed to assess what would be most practical to work with. This process was repeated for the three different types of data identified (transactional, player tracking and proxy).

The discussions from this workshop form the basis of this report. To further ensure clarity of findings, a number of helpful operators have reviewed key (draft) sections of this report and provided feedback and expert opinion to ensure accuracy.

3 Findings

This chapter presents findings about the types of data, the specific metrics and the different levels of data available for category B machines. The potential limitations of using various forms of data are discussed, as are issues pertaining to data size, structure and consistency.

3.1 Types of data available

3.1.1 Transactional data and game data

The main type of data which all sectors across the machine industry hold is transactional accounting data. These are financial data which monitor what money is put into the machine and what money comes out. Their intended purpose is to provide financial accounts and records of how much money a machine is making, which is required for accounting, audit and tax purposes. These data are obtained from the accounting and occurrence meters within the machines. These meters are mandated by the Gambling Commission's technical requirements standards to be a 'primary metering system which is independent of the main control system'. This is required across all category B machines.

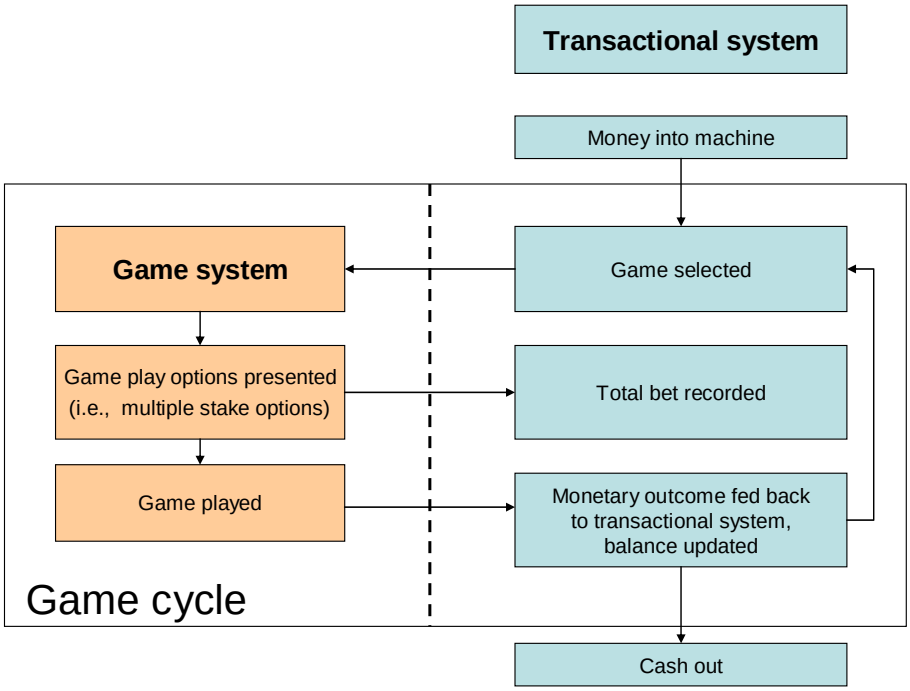
These transaction data are collected in a variety of ways and the level of detail recorded depends on a) the sector, b) the age of the machines and related systems and c) the accounting protocol to which the sectors adheres. All of these issues will be discussed in greater detail in this report.

Machines contain two systems: the transactional accounting system and the game system. The transactional accounting system (also known as the 'platform') records the credit balance which is increased by credit being added (notes in, coins in etc) and decreased by credit being withdrawn (cash out / voucher out). Different games also interact with the platform by decreasing the platform credit balance for money staked and increasing the platform credit balance for any player winnings.

The platform and the individual game inter-relate but, crucially, the game level system does not output information about activity which takes place in the game in a consistent way. The only data that are recorded consistently by the platform is detail of the stake and any amount returned to the player. The level at which this information is outputted also varies by machine. For example, some machines output this at a daily, weekly or every game cycle level.

Figure 1 shows a simplified diagram of how the two systems relate to each other.

Figure 1: Diagram of the transactional and game systems within a machine



This is a simplified diagram but shows the broad structure of how the two systems relate to each other. The transactional system records the money going into the machine. At the point at which a game is selected, the game system takes over operation of the machine. Using an example like roulette, the game system presents all functionality associated with that game. Therefore, the opportunity to place one large bet on red or many smaller bets on specific numbers is embedded within the game system. Once the player has chosen their bets and stakes (either multiple lines, multiple numbers etc), the total value of the bet is fed back to the transactional system which records the total bet and alters the credit display. The game is then played and the monetary outcome from the game system is also fed back to the transactional system to update the balance.

This process of placing a bet on one game and one outcome is called the **'game cycle'.**

A critical finding is that the only communication between the game system and the transactional system is to update the latter with information about total money wagered and money lost and/or won per game cycle. However, whilst this information is passed to the transactional system, it does not necessarily mean that every machine is recording game cycle data in a systematic way, and the way in which machines record these data can differ considerably. Some machines systematically record information within the transactional platform at a game-cycle level. This game

cycle information will typically show the date / time of the game cycle event, the game name, the amount staked and any amount returned to the player's credit balance as winnings for the game cycle. If the player chooses to play again, this is then captured as a new game cycle. Other machines do not record data at this minute level of detail and whilst the process of communication between the game and the transactional system is still happening in the same way, the information is not logged in this level of detail. These issues are discussed in Section 3.1.3.

A second critical finding is that data from the game system are not available, stored or recorded in any systematic way. Game system data are held on the machine in log files specific to each game / game version on the terminal. Multi-game terminals may contain up to 60 different game specific log files. These log files can be accessed manually through the machine to review precisely what type of bet a player made and what the exact outcome was. Industry staff reported that they only use this functionality and information to settle a dispute. This level of game data is written into a text file format and is logged in this format (for some machines, these log files are centrally stored, for others they are not). This means this game system data are not currently captured in a way where it can be easily used for research purposes. An example of the text format, stored on machines, is shown in Figure 2.

Figure 2: Example of text file game data on B2 terminals

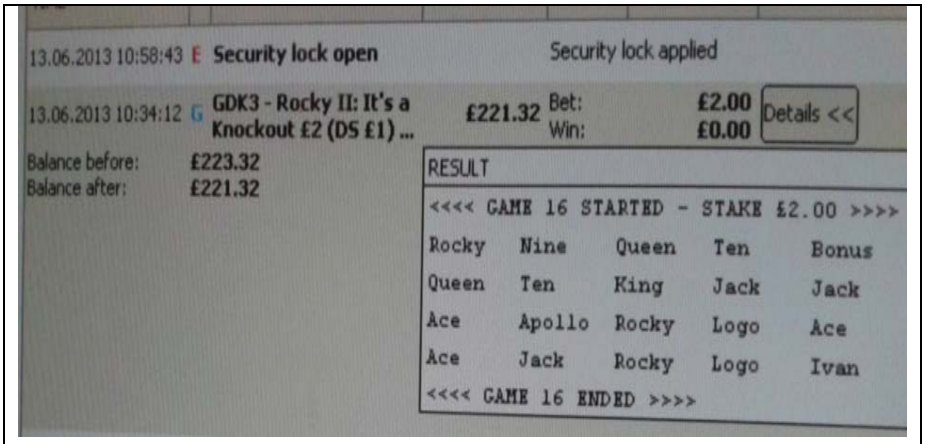


Figure 2 (cont.)

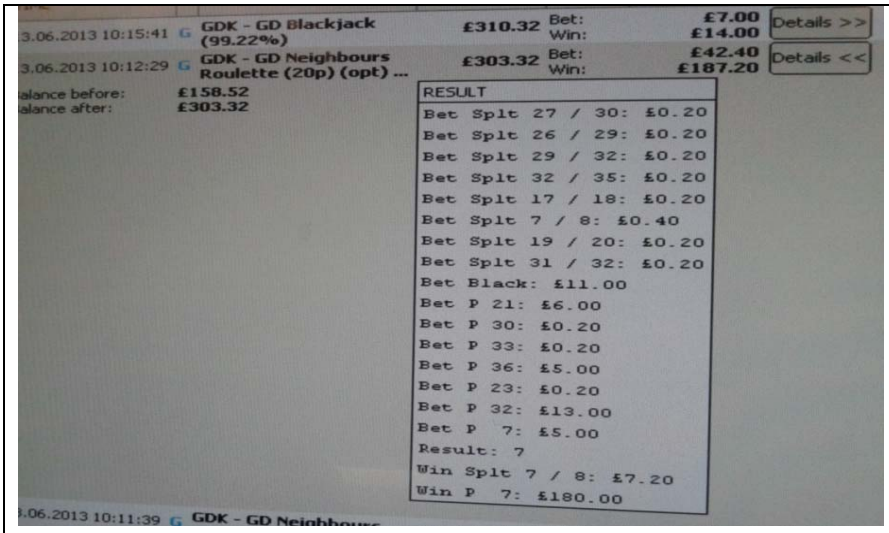


Figure 2 shows some of the different types of text file output displayed on machines when this information is manually sought. Here a member of staff has specifically requested for the log details of each game to be displayed. The first display is from a slot machine style game in which various combinations of symbols in a 4x5 grid equals a winning line. The display simply shows what the result for that game was. The second example is that of roulette. This is potentially more informative as it displays what the specific bets were at different staking levels and what the result is, and, in this case, what the winning combination is. This shows a winning outcome. However, in a losing case, this would show what the outcome was, and *theoretically* it would be possible to determine whether this was a 'near miss' or not.

However, these data are not routinely captured and outputted into log text files. To use them would require converting the text files into a database format where it could be analysed and matched against the transactional detail for that game cycle. This conversion is not done by any of the operators or suppliers we have met with so far. We have also been advised that there is a significant level of variation between log file outputs for different games (and game variants) and so attempting to use these data for research purposes would be difficult and laborious. That said, it is the relationship between game data and transactional data that has the most potential for research purposes.

These data have clear research potential, as it would be possible to see exactly how people are making their bets and their staking patterns on some games, notably roulette. With the roulette example, it may be possible to determine near misses. However, the current way the data are captured and stored precludes use of them without significant researcher effort to restructure, clean and link the data into some useable format. Different approaches would be required for different game types; e.g., roulette and blackjack would need a different method to link their game data to transactional data, and for some games, the detail recorded is not very informative without understanding the specifics of what the outcome means (as can be seen by the very different outputs displayed in Figure 2). The game system log data, specific to

each game variant, is stored in rich text format and is either held by operators and/or suppliers. However, operators from AGCs and other venues have advised that this information is not routinely available without manually pulling it from the machines.

3.1.2 What is not included

As the previous section set out, data from the specifics of the game platform are not consistently recorded in a way that researchers could easily access. This means that a variety of information about certain machine features are not captured and it would not be possible, with current systems, to build this into research. A few operators did say they were currently investigating whether more could be done to link game data and account data together. The timescales on this were, however, uncertain. Therefore, with information that is currently available, there are a number of aspects of machines features and machines play which industry held data **cannot** help with. These include:

- Use of the autoplay button – this is not recorded in the transactional data, which is simply recording money going through the machine, and it is our understanding that use of this button is not stamped onto the data.
- Use of credit transfer and banking features – how players use these features and how they transfer money between credit and banking systems is not recorded in the data.
- Staking patterns – on games with multiple staking pattern options (i.e., where players can play multiple lines or multiple numbers such as roulette) these data are not captured. Only the total value of the stake is recorded by the transactional system, not the component parts of the stake.
- Browsing activity – on terminals with multiple game offers, menu browsing activity such as looking at different games, using the help button, etc, is not captured.
- Any game feature activity, such as progress towards bonuses, use of holds/nudge buttons or engagement in bonus games is not captured.

In addition to these aspects of machine play which are not routinely captured, there are a multitude of different levels (and formats) at which the more 'routine' transactional data are held. Whilst not mentioned by all operators, this is likely to be a direct consequence of the Gambling Commission's technical standards. These set out which metrics are required to be recorded but do not give specific instruction about how often these data are required to be reported (i.e., are the data required for every game cycle or just aggregated at a terminal level by day or week?) .[6] This is discussed in more detail in the next section.

[6] Since the original interviews were conducted, the Association of British Bookmakers has announced plans to address some of these issues. This relates to specifying and formalising data structures between LBO operators. This may help improve the level of data routinely captured in this sector.

3.1.3 Levels of data – terms and definitions

Having established that the main type of data available are transactional and are not related to game information, we can summarise the different levels of data available. These are defined below. These terms are important to clarify as they will be used throughout this report. Where possible, we replicate the most common language used by the operators interviewed. However, where different definitions or words were used by operators, we have used terms that we believe are simplest and easiest to follow by a non-technical audience.

Transactional data

In this report, 'transactional data' refers to financial account data (and related contextual variables) that are recorded by machines and exported to operators so they can track money in and money out of the machines. Notably, it is recorded at the machine or in some cases venue level, rather than at the individual level. The metrics included in these data are largely governed by the Gambling Commission's technical standards for each machine category type. For B3 and B2 machines, the minimum reporting requirements are:

- Cash in
- Cash out
- Total value of plays
- Winnings (on multi-game machines).

In addition, there is a requirement that these are logged for each game type on multi-game machines and that a return to player percentage is calculated. This makes transactional data the most consistent form of information available across the category B estate, as all operators are required to have some basic levels of information. There are still significant variations between operators: these are discussed in Section 3.2

Whilst this list represents the basic requirements, in reality some machines record more information than this. This can be split into two types:

a) Accounting data – these include cash in, cash out, total value of plays, return to player, stake, note in, coin in, voucher in, etc.

b) Behaviour data – these include data about number of bets made, name of game played, length of session (where available).

The behaviour data are used to help interpret and better understand the patterns of play illustrated by the accounting data. However, for the purposes of clarity for this report, the term 'transactional data' is used to include both types.

The level of transactional data recorded varied between operators and in this report we distinguish between two levels of data captured: atomic data and aggregate data.

'Atomic' data refers to information that is collected for every single game cycle. This means that for each game cycle it is possible to track how much money was staked and what the monetary outcome was. Typically we can see what game type was played and how the money was inserted into the machine (i.e., as notes or coins). Because this is 'tracking' every transaction for a game cycle, of which there are millions per day, we have called this atomic level data. This term was also used by some of the operators interviewed.

'Aggregate' data refers to one of two things:

1) This is either atomic level data aggregated to different levels. This could be by time period (hour, day, week) or it could be by context (game, terminal, venue, estate). There are also interactions between the two (venue by hour, game by week, etc).
   Or
2) Aggregated information retained and recorded directly by the machine. Some operators *only* hold aggregate information as their machines only transmit information at certain intervals. For this type machines can range from aggregating information at 15-minute intervals to aggregating information on a weekly basis. Here atomic level data are not available and the only data available are that aggregated by the respective machine/system.

Therefore, there are two types of aggregate data – those drawn and summarised from atomic level data (where atomic level data is also available) and those for which the level of aggregation is governed by how frequently the information is transmitted and collected from the machine.

Player tracking data

'Player tracking data' refers to data which are linked to a loyalty or membership card system. Of those operators who had player tracking functions, the broad systems were the same. Insertion of a card into a machine opens a new player record or instance. Transactional information about the game cycles within that instance is recorded. Removing the card from the machine signals the end of the session and the player record from that instance is closed.

However, not all data recorded within the start and end of the session are stored at the atomic level. Some operators only record aggregate transactional information for that player session, whereas others have atomic level data and associated aggregates derived from these.

Interestingly, some operators who did not have atomic level information for their whole transactional data did have atomic data for their player tracked information.

In addition, standard practice across the industry is to make card usage optional: a player with a card could use a machine but choose not to insert his/her card. Therefore this play cannot be linked to that player's information. This has implications for missing data, which is discussed in Section 3.3.

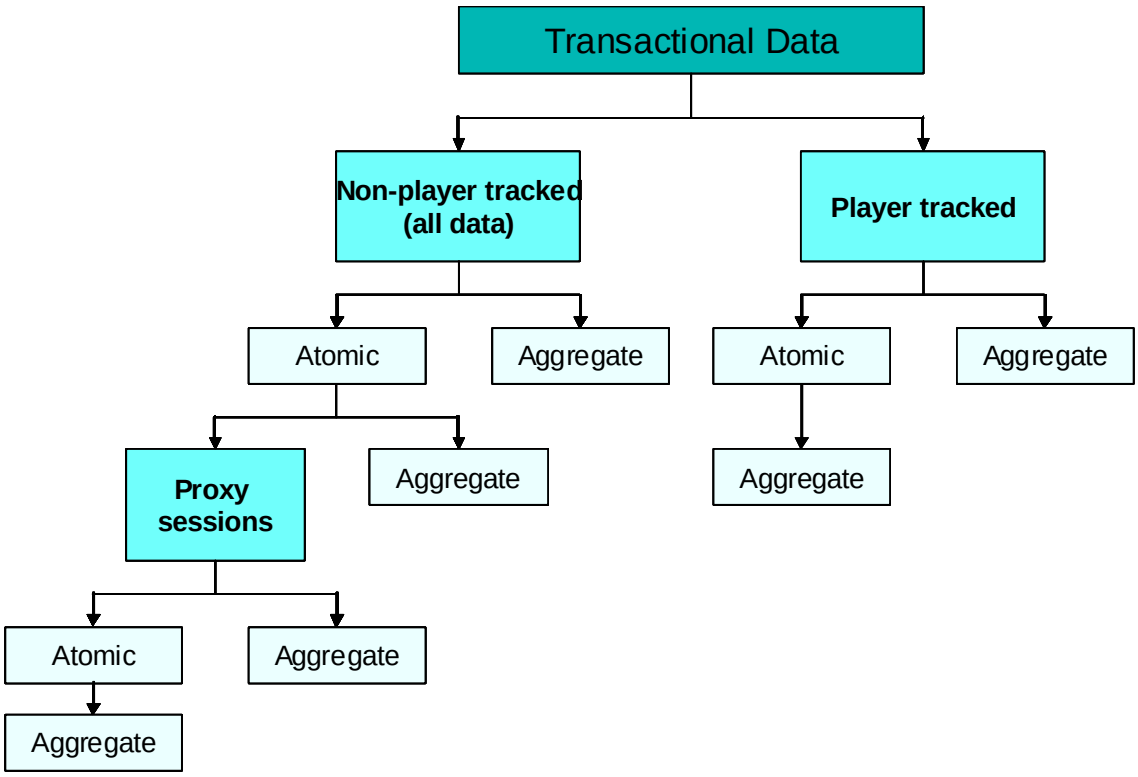
Proxy session data

'Proxy session data' refers to (non-player tracked) atomic level transactional data which have been split into sessions according to a set of pre-defined rules. Identification of proxy sessions varied between operators but was largely predicated on a) the starting balance (either being zero or less than the minimum stake), combined with b) cash being put into the machine and c) the length of time the machine was dormant. These definitions and algorithms are discussed in Section 3.4. The main consideration here is that proxy sessions can only be identified using atomic level data, and rely on the application of algorithms to identify what was a *likely* start and end point of a session. The potential limitations associated with this method are also discussed in Section 3.4.

Relationship between data levels

Figure 3 visualises how these different data levels relate to one another.

Figure 3: Data structure diagram



This simplified diagram shows the various levels of data available for category B machines in Great Britain. Some operators have all of these levels available, some only have part of the data available. However, it demonstrates the relationship between different levels of data. For example, it clearly illustrates that you can only have proxy session information *if* you have atomic level transactional data. However, once proxy session data are generated they can either be stored at an aggregate or atomic level; this depends on the operator and their systems. It also demonstrates that just because you have playing tracked records does not necessarily mean you

have full detail of exactly what happens within each game cycle, as some data are only kept at an aggregated session level. Some operators have a complex array of data available at different levels whilst others record a basic level of detail, having only aggregated, non-player tracked, transactional information.

Key points

- The system which records routine transactional data from machines operates independently from the game system.
- Only the transactional system records data in a consistent way about machine play.
- The transactional record system and game system only speak to each other to pass information about stake and money won/lost. This conversation happens at an aggregate level – so the total amount staked for a game cycle is recorded by the transactional system.
- The game system data are unstructured, stored as free text and not captured consistently or linked post-hoc to transaction data.
- This limitation means there is a lot of potentially useful information that is not present in the data – the data available are loosely based around what is mandated by the Gambling Commission's technical standards for accounting and occurrence events.
- The level of detail recorded by the transactional system also varies. It can be atomic level data, aggregate level data or both. Some operators have player tracking, others do not. Those with atomic level transactional data can also have proxy session information.
- In order to obtain atomic level transactional information, machines have to be server-based and connected to a network. Other machines stand alone and transactional data are collected manually from internal meters.

3.2 Transactional data

As noted above, all operators have transactional data which record the money going into and out of the machine. What data the machine records and at what level of detail depends on a variety of factors. These include:

- What type of machine it is
- What sector the machine is based in (and therefore what systems the operators have), and
- What accounting protocol the sector adheres to.

In the sections that follow, information about machine data availability within each sector is summarised. Within sectors, there are some common themes about the level of detail held largely because each sector, typically, has a common category of machine.

However, we first summarise information about the different methods and systems of extracting data from machines and how this results in different levels and types of data being available.

3.2.1 Machine data extraction systems – overview

Server-based and automated

The level and detail of data exported from machines depends on two factors; first, whether the data extraction system is automated, and second, whether the machine is a server-based machine or not. Here, a 'server-based' system means that the machine is fitted with the technology to transmit transactional level data at regular intervals to a server housed either onsite or offsite (and crucially, is connected to a network). The most common form of server-based machines are those found in bookmakers. Supplied by either Inspired Gaming or SG Gaming, game cycle data from these machines are transmitted back to the supplier. The bingo industry also has some server-based machines. For example, one operator has approximately 1700 out of 1880 B3 machines provided by one of the main suppliers with data transmitted and stored in a similar way to machines found in bookmakers. In AGCs, some machines are also server-based and transmit information back to their supplier, though this is not the norm.

The key metrics that these server-based machines collect include:

- Amount staked
- Game name/machine type
- Financial outcome
- Date and time of the start of the game cycle

- Type of transaction (note in, coin in, remote in, manager loaded cash, receipt in)
- Total value of transaction
- Cash out
- Date and time of the transaction

The first four metrics on this list refer to items of data captured for each specific game cycle. Because they are based on each game cycle, it is possible to calculate the total number of bets placed within specific periods. The next four metrics are transactional that are recorded as and when they occur – most obviously, cash in (typically) at the start of play session and cash out at the end of play session (further discussion of the identification of sessions is provided in Section 3.4).

There are also other items of data that the machines record: these include whether the terminal door is open, whether there has been an error or blockage on the terminal, and so on.

From the metrics listed above, data about the value of total plays, return to player etc, can be calculated for each game type available. For B2 machines, this is mandated by the Gambling Commission's technical standards.

Whilst these are the common metrics recorded by these machines, different suppliers have different naming conventions and collect a different range of data relating to machines. These issues are discussed from Section 3.2.2 onwards.

Analogue and manual

Other operators have what we are describing as 'analogue-based' systems to capture data from machines. This means that the data are recorded by the meters within the machine, to the standards required, but the process of extracting these data from the machine is not an automated one. Newer machines have the capability of electronic monitoring and one of the AGC operators has the capability to switch across, but that would mean running conflicting processes within a single venue. In the instance of manual collections, the process of recording metered transactional data is a manual process conducted once a week where members of staff either open the machine and manually record the data into a spreadsheet or, in the case of digitised machines, pull the information from the account tab on the machine service menu (an example of how this looks is provided at Figure 4). This process occurs for at least one of the casinos included and for at least one of the AGC operators. This, of course, introduces the possibility of individual error and also means the data are only recorded at a weekly aggregate level, when staff do the weekly counts.

Figure 4: Example of data pulled from machines with analogue transfer systems

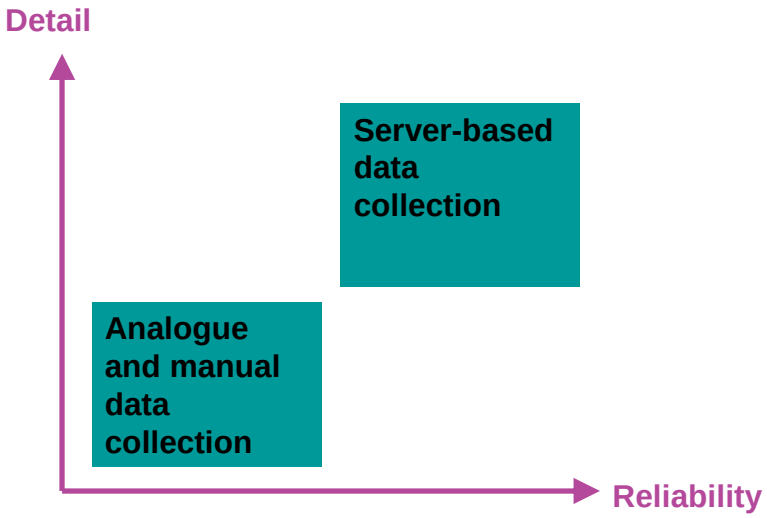
In Figure 4, it can be seen that the only metrics recorded are total cash in, total cash out, number of notes in, games played and value of total plays. These are presented for distinct periods of time. In this example, there are two accounting periods – short term and long term. We presume that short term presents information from the date last accessed but need industry operators to confirm this.

Mixture

There are also variants in between this scale of fully automated and manual machine data capture. For some machine types (and for certain operators), whilst the machine data capture system is not routinely automated, hand-held devices are used to capture data from the machine meters, rather than the manual recording system. Other operators have bespoke systems in which data are transmitted via wifi to a system in the venue and then collated across all their venues.

Summary

In terms of how useful the resultant information is, capture of data on machines can be visualised as follows:



Here, the spectrum of data capture ranges from server-based to analogue and manual data collection systems, the former of which offer greater detail and depth of data as well as greater reliability. In terms of the detail and depth of data available, this can range from including a) a greater number of metrics being recorded and/or b) data being collected at a more granular level. The opposite of this is where manual data collection systems are used. These systems offer less depth and breadth of data and are, by nature, likely to be less reliable because of the manual or semi-manual processes involved in data capture. Some operators have a mix of data system types, with some machines within their estate being server-based in terms of data capture and others relying on manual processes.

This demonstrates the significant level of variation evident between machine types and operators, which influences the depth and breadth of transactional data recorded.

What follows is a summary of these issues by sector.

3.2.2 Licensed Betting Offices

Five different LBO operators were interviewed during this phase of research. These operators each have the same type of machine (B2 machines) in their venues. These machines are provided by two main suppliers who were included in this study.

Data capture system

All category B2 machines in major bookmakers have a server-based data capture system. The machines transmit transactional data back to their respective suppliers and record atomic level detail for game cycles, representing the most complete information available. The data are held by the respective supplier, who undertakes a range of analytics on the data, but they are owned by their clients.

Each supplier also offers their clients different levels of aggregation from these atomic data, whilst some operators also undertake their own analysis of atomic level data. Clients can either do this by running their own reports through writing queries to a database made available to them for this purpose, or have a regular feed of pre-specified reports sent to them by the supplier. It seems that the larger bookmakers tended to do more of their own analytics and report running, whereas smaller ones tended to rely on data and reports sent from suppliers.

Metrics

Consistent metrics

The listing below gives an overview of metrics available which are broadly consistent between the two main suppliers of LBO machine data.

For each game cycle, the following metrics are collected:

- Transaction date
- Transaction time
- Note in
- Coin in
- Total money in
- Money out
- Ticket in/Ticket out[7]

For various levels of aggregation, there are further metrics available. These include:

- Number of games played
- Average stake
- Gross win
- Return to player (inverse of gross win)
- Value of total plays
- Total coin in
- Total note in
- Total cash out
- Total returns (money returned to players balance by way of winnings)
- Total voucher in (where applicable)
- Total voucher out (where applicable)

These metrics are the ones that, to our knowledge, are common across both suppliers. Both suppliers have provided us with simplified listings of their key measures and we used these documents to identify common metrics. We also identified that these metrics can be summarised to different levels of aggregation. For example 'gross win' (defined as total stakes minus total prizes) can be calculated for a

[7] Money won is typically paid out on a printed ticket that is then cashed with the cashier. In some cases, tickets with money credit from previous sessions can be inserted into machines to start play.

venue, for a machine, for an hour period and so on. Review of these listings has also allowed us to identify the inconsistencies between suppliers and hence the different levels of data available.

Inconsistent metrics

From review of the data listings shared so far, we can see that there are differences in the level of data collected on the vouchers and tickets used during play. For example, metrics here include total value of tickets returned to the player, total value of recycled tickets (where a ticket is printed and inserted back into the machine), number of tickets printed, number recycled, number not recycled and so on. Further investigation is needed to clarify whether this information is taken into account in the main transactional variables (i.e., whether the value of these tickets are included in figures like gross win, etc).

A range of other inconsistent metrics were identified but cannot be listed here because the information is commercially confidential.

Compatibility

In addition to recording different types of metrics, a review of each supplier's data listings allowed us to assess how compatible data are between them. For key metrics such as notes in, coins in, average stake and gross win, the data and definitions used by each supplier appear to be broadly consistent.

However, for some of the more detailed metrics, the terminology used by each supplier varies. For example, one supplier records the value of total plays under a metric called 'turnover' whilst the second supplier records these data under a metric called 'total stakes'. A further example is that of the total amount of cash loaded into a machine by a manager. For one supplier, this is called 'manager loaded cash in' but for the other it is called 'total remote in'.

Therefore, in order to work with data from these two suppliers (covering the five LBOs taking part in this study) we would need to create a combined data dictionary which reconciles these differences.

Finally, both suppliers confirmed that they could output data to our specification in a .csv file format. Therefore, the main compatibility challenge is ensuring that we fully understand the naming conventions and definitions used for each metric.

3.2.3 Casinos

Five different casino operators were included in phase 1. There is less consistency between casino operators than LBO operators in terms of the data they collect, how often they collect it and how it is stored. This is because each operates independently and there is no common supplier. Furthermore, one operator noted that because the number of machines allowed in some casinos is small, it does not always make

financial sense to set up a complex data capture system to systematically track information at a granular level.

The majority of machines in casinos were B1 machines, though most operators had a few B2 and/or B3 machines. In casinos with mixed categories of machines (B1, B2, B3), there was also a mix of what levels of data were captured and stored. Given the very small numbers of B2 and B3 machines identified within casinos, we concentrate on providing information about B1 machines *only* in the sections that follow.

Two operators reported that they collected data at an atomic level; however, one of these operators stores this for only three days before it is erased from their system. It was unclear for the other operator whether the data were collected for each game cycle or not. The operator stated that data was collected and transmitted in 20-30 second bursts but noted that if customers had more than 30 key strokes per minute it was difficult to identify individual stakes.

Most of the casino operators interviewed confirmed that they use IGT's Slots Accounting System (SAS) protocol. This is a casino industry standard and this protocol dictates what information the machine stores and transmits.

As with the Gambling Commission's technical standards, the SAS protocol dictates what metrics should be recorded but does not make recommendations about the frequency or regularity of data capture, though the protocol (v4 onwards) does allow for real time event reporting. It seems that how often data are recorded depends on the internal systems of each operator. Finally, as one operator noted, adherence to this protocol governs what level of data are recorded. Essentially, what is captured is what is in the protocol - no more and no less.

Data capture systems

The data capture systems used by the casino sector varied. Some casinos relied on manual readings of meter data, whereas others had developed their own bespoke systems. Interestingly, there was at least one operator who used a mix of these two approaches, collecting some data in an automated way and some based on manual weekly counts of meter data by staff. This means that little consistency between casino operators was identified.

Metrics

Because we are currently uncertain about whether one operator collects atomic level data or not, we present the list of metrics commonly held by most casino operators. Most casino operators only collect aggregate levels of transactional data.

Aggregate level metrics

For various levels of aggregation, these metrics were commonly available across the casino sector:

- Number of games played
- Value of total plays
- Total coin in*
- Total note in
- Total cash out
- Total bet
- Turnover (value of total plays)
- Tickets in
- Tickets out
- Win
- Loss
- Handpay[8]

*One operator identified how coin in may be interpreted differently as either total value of plays or literally as total coin in.

One operator also collects Outlet ID and Terminal ID, but no similar contextual metrics were collected by others, though it may be possible to request some contextual details be recorded by operators on a venue by venue basis.

Consistency

First, not all of the metrics mentioned above are recorded by a single system for each operator. As noted above, one of the casino operators collects these metrics using two different systems. One system is used to collect basic information, such as total bet and turnover at an hourly basis which is analysed at the daily level by the operator. This is an automated system with the regular output of data. However, for other metrics, such as total coin in, note in, tickets in, tickets out, these data are captured at a weekly level when casino staff do the weekly counts from the machine meters. This is then manually entered into an accounting database.

[8] Handpay refers to the situation where a member of staff pays out the prize money rather than the machine.

At least one other casino operator reported only collecting data on a weekly and manual basis.

Some operators have outsourced their data capture and have a third party who host and analyse it for them. In this case, the operator has a contractor whose systems collect basic machine data from each venue. These data are collected at 15-minute intervals, meaning this is the lowest level of aggregation available for this operator.

This demonstrates significant inconsistency between casino operators. Not only is each casino operator using their own bespoke systems, some are based on manual recording of metered data, whereas others are based on automated transfer of information. The level of data recorded ranges from 15-minute intervals to weekly. This means that attempting to reconcile data between casino operators is likely to be a complex task.

3.2.4 Bingo

Three bingo operators were included in phase 1 of this project: the main two bingo operators in Great Britain and one operator with a small stake in the bingo sector (with only nine clubs nationwide). As with casinos, there was a great deal of variation between what data are collected, at what frequency and how they are stored and managed.

Data capture systems

Two of the three bingo operators use the Playsafe data management system to capture and store machine data. This data management platform has four levels at which it records information:

- Level 1: data are transmitted by wifi connection and data stored in hourly aggregates
- Level 2: data are transmitted within venue only and one needs to be in the venue to read these data
- Level 3: data are read from machines using a hand-held device and captured at weekly level
- Level 4: data are manually read from meters and inputted into the playsafe system.

Within one venue all four levels might be used, depending on the type and age of the machine.

We are awaiting confirmation from each operator about the Playsafe data capture levels used in their bingo clubs and the extent to which B3 machines use one level

over another. One operator has confirmed that whilst their data are stored locally, they are also collated centrally and that most clubs are on the full (i.e., level 1) data management system.

The other main bingo operator has the majority of their B3 machines provided by one of the main LBO suppliers. This means that these machines are server-based and capture more detailed transactional information than the other bingo operators. For this operator atomic level detail is available.

Metrics

For machines on the playsafe system, the following metrics are collected:

- Cash in
- Cash out
- Net cash*
- Value of total plays
- Percentage usage per week*

*for both of these metrics, further definition is needed.

The number of games played is not recorded. There are also some 'static' metrics which the Playsafe system records. These essentially describe the machine and include:

- Machine ID
- Price of play
- Jackpot level
- Manufacturer

For other machines which are served-based, the assumption by bingo operators is that the machines are capturing more game-cycle information (note here that the operators themselves were not certain of what was captured). Suppliers of the server-based B3 machines have confirmed that, broadly, some information is being captured at a game-cycle level. The suppliers noted that there may be some nuances around this. For example, these machines need to be connected to a network to transmit data, whereas many machines are configured to work in 'stand alone' mode and have no connection to send data. In short, whilst some machines may have this functionality, they may not be configured to capture these data. Where machines are networked, this means that a similar range of metrics as those collected for LBO machines are available. These include:

For each game cycle:

- Transaction date
- Transaction time
- Note in
- Coin in
- Total money in
- Money out

For various levels of aggregation:

- Number of games played
- Average stake
- Gross win
- Return to player (inverse of gross win)
- Value of total plays
- Total coin in
- Total note in
- Total cash out
- Total returns (money returned to players balance by way of winnings)

As with casino data, no further contextual data are available, though these could be created.

Consistency

As observed in the casino sector, consistency both within and between bingo operators is a major issue. Not all bingo operators have the same data capture systems. This means that data exist at different levels of aggregation and that a different range of metrics are captured. Notably, there is considerable variation within operators about how the levels of data are recorded. This is most clearly illustrated with the Playsafe database.

With the bingo sector, it is clear that a lesser amount of information, in terms of metrics collected, is available and that these data are inconsistent between the different bingo operators.

3.2.5 Adult Gaming Centres

Two AGC operators were included in this research. There was a great deal of variation between them and the range of data collected, in terms of metrics, was much narrower in detail.

Data capture systems

One operator also used the Playsafe data management system in their AGCs. As noted above, this has four different levels of data capture which vary across the operator's estate. For example, one venue may use level 1 data capture whilst another may use level 3. This makes the level of data captured for this specific operator variable between venues.

The other AGC operator stated that they largely used a manual recording system whereby data were manually recorded from the machine meters on a weekly basis and entered into a database. They did, however, note that some their machines captured information electronically and that a few B3 machines had server-based functionality but were not connected to a network. We are waiting for confirmation from this operator about how many of their B3 machines have electronic data capture and how many rely on the manual weekly process of capturing information.

Metrics

Only a very limited range of metrics are available for AGCs. These are all aggregate data metrics as game cycle information is not available.

Metrics commonly available are:

- Coin in
- Note in
- Cash out
- Value of total plays
- Return to player (calculated from data above)
- Venue code/location
- Cabinet type/machine name

The operators did not mention that the number of games played is routinely captured; though we know this is available electronically on some B3 machines (see Figure 4).

In addition, one operator noted that they were able to record what level of promotions (i.e., free plays) were also being put into the machine. We assume this is recorded as a separate metric on the metering system.

No contextual variables were available but it could be explored with operators how easy or difficult these would be to create.

Consistency

Whilst there is only a fairly limited range of metrics captured by AGC operators, they are fairly consistent in terms of definition and what they are collecting. The main inconsistencies appear about the level to which these data are recorded (daily, weekly etc) and the method used to capture this information. Even within an operator's estate, different venues may be using different methods. This has resultant issues for accuracy as some data are recorded using a manual process whereas other data are collected electronically. One operator did say that, where possible, they attempted to reconcile electronic and manual data records.

Key points :

- Transaction data are most detailed and, potentially, most consistent for LBOs. Here atomic level data are available for all and the main consistency issue is the varying definitions and terms used by each supplier. This is potentially the most useful set of data available for research purposes.
- All other sectors have a variety of systems to record information. This means atomic level data are available for some but not for others. This means there are significant consistency issues for these sectors.
- Some sectors, like AGCs, have very limited levels of data available, only in weekly aggregates, a very narrow range of metrics and data are collected using manual methods. This calls into question the usefulness of these data for research purposes.

3.3 Player Tracking Data

This section details the player tracking data collected across the gambling industry. The amount, type and level of detail of player tracking data available from the gambling industry differs by sector and can also differ by individual operator within each sector.

Taken at face value, it appears that the industry holds rich, insightful data on machine players. However, deeper scrutiny highlights a multitude of limitations, confirming that these data can be inconsistent and unrepresentative of the player base. Use of such data for research purposes will require standardisation across the sector of interest or industry as a whole.

Player tracking in all sectors is achieved through use of a loyalty card system. These schemes vary between operators but the broad principle is that players attain various rewards for play. To do this, players must insert their card into the machine at the start of the play session and remove it at the end. This triggers a player-tracked session record to be created in the data. Therefore, in this report reference to 'player tracking' means data generated by the presence of a loyalty card scheme. One of the main challenges identified by this study is that of attempting to use data generated for different purposes for a research purpose.

3.3.1 Licensed Betting Offices

Opportunities and limitations

Player tracking in the LBO sector of the gambling industry is in its infancy, with three out of the five operators interviewed running any kind of player tracking scheme linked to machine play. These schemes are not yet well established and are of variable quality. Two of the three operators launched their player tracking schemes in 2013; the third has been in operation since 2011.

Despite this, the schemes in this sector do show potential to be of use. All three operators collect data at both an atomic and aggregate level and record which games customers are playing within their play session. This means that transactional data for players, showing their total stakes and returns for each game cycle, are available for the session of play. In short, it is possible to track the journey of player using transactional data within a session. Assuming that players use their cards consistently, play between different sessions can also be linked.

However, there are limitations. Card usage is optional for all schemes and the LBO with the longest running scheme has low levels of take-up (here measured as the proportion of machine play that is tracked). Whilst low levels of take-up themselves are not overly problematic as it still equates to a vast amount of data, the possibility of self-selection bias is a concern. Certain types of players with particular patterns of play may display a preference for using these cards, with resultant research and

analysis generated from these accounts potentially being unrepresentative of wider patterns of play. This is something which needs further investigation.

Furthermore, anecdotal evidence from the LBO operators with newer schemes predicts low levels of take-up also. This is based on an understanding of which types of players are more likely to choose to use the card, the level of rewards attached to machine play being low, and also customer suspicion of the cards, with players believing them to affect how much and how often they win on the machines. However, as the schemes roll out and are more heavily promoted, these levels of take-up may improve. Indeed, one operator is heavily promoting their new scheme. Therefore, more time is needed before we can ascertain if this low take-up is the case in reality.

It was also clear that the way in which individual operators implement loyalty schemes and promote them may lead to differential take-up among the machine player population. Some schemes offered simple point rewards for play, whereas others offered this and access to certain preferential games (i.e., roulette with no zero) for loyalty scheme members. Most loyalty cards were not just for machines but also for over the counter bets, where points rewards are greater, leading to the possibility that people may use their card for one type of gambling within a bookmakers and not for the other. Each operator sets their own level of rewards, meaning that some schemes may appeal to certain groups more than others. All of this highlights the need to understand the player-tracking data in context of the scheme in which it is offered to assess whether differential take-up might be observed within and between these operators.

Finally, anecdotal evidence from operators suggested that the people who used their cards to track machine play were not the 'big-spenders' on machines. This assessment has been made by comparing average staking levels of tracked and non-tracked play. If this is the case, any data drawn from player tracked information may give a much skewed view of behaviour. This needs further investigation. The operators also noted further limitations of using these data, in that people can put their cards into machines at any point in play and so the records are unlikely to represent a full account of that person's machine play (i.e., they may use their card in one session but not in another, they may put cash into the machine and then insert their card meaning the cash input is not attached to their session, etc.). Customers may also have more than one card and more than one account. For some operators, it is difficult to reconcile these multiple accounts.

Scheme types

In LBOs, there were two types of player tracking scheme evident:

- One which requires registration of some demographic information to obtain a card, and
- One which is an anonymised registration scheme, requiring only an email address or unique customer pin. Crucially this is not linked to any named registration system.

Only one operator (with the more established scheme) runs a scheme which can link player demographic information to sessional information; the other two running anonymous schemes which link only to an email address or a customer pin number. All schemes record key session metrics, such as:

- Machine type
- Start time / end time
- Length of session
- Stake
- Money in / money out
- Amount won
- Number of games
- Number of spins
- Game details
- Outcome of game
- Frequency of visit

In addition, the operator running the registered scheme has the ability to link their data to personal player details:

- Name
- Address
- Sex
- Age

Knowing basic demographic details is important as it would at least allow us to look at different patterns of play among different age/sex groups of machine players.

From these data, it is possible to calculate average stake (on a sessional basis or for an aggregated time period) and, as noted above, game cycle journeys through a session can be tracked.

Consistency

The consistency issues noted for LBOs and transaction data also apply here as it is the same range of metrics, just with a start point, end point and player record stamped on the data. In addition to these issues, for two of the operators, anonymised player data for their whole estate are stored on and accessed through their supplier's server. The third operator has anonymous atomic level data from their supplier's server and matches these themselves to their demographic player data stored on their own server for their whole estate. This creates some inconsistency between the systems in which the data are stored and the level of data stored (i.e., for one operator, some demographic data are held).

3.3.2 Casinos

Under the Gaming Act 1968, casinos were required to maintain a membership list of all their customers, and to record any non-member guests accompanying a member. To this end they developed a culture of holding and using a membership database, which at one time was ingrained in the way they operate as a business and was the norm to their customers. Although the Gambling Act 2005 removed this obligation, a culture of collecting basic player details is still prevalent within the sector.

All five casino operators interviewed hold player tracking data, though the ability to track individual players is a relatively new phenomenon and in varying stages of development within the sector. Four of the operators have held such a system for two to five years.

Similarities are evident across the sector with all casino operators' systems having the ability to link player information to sessional data. Homogeneous player metrics captured across the sector include:

- Player name
- Age
- Sex

and these are linked to machine play and sessional data metrics such as:

- Machine type

- Start time / end time
- Length of session
- Stake
- Money in / money out
- Amount won
- Number of games
- Number of spins
- Machine occupied percentages
- Usage of points (in some cases).

Consistency

There are a number of inconsistencies between operators. For two of the four operators, atomic level game cycle data are available for player-tracked sessions. For the other three operators, data are recorded at a more aggregate level, either in session summaries (mainly to be able to allocate points earned) or in 20-30 second bursts. Therefore, the level of data available varies between casino operators.

There are also varying levels of take-up observed. As with LBOs, it is the sector tendency to make card usage optional. These different systems mean that take-up and usage of cards when playing machines differs across the sector. Reports of take-up varied from around one third to over two thirds of play being player tracked, though measurement of this between operators may vary. Whilst this is a vast improvement on the LBO sector, it still raises questions about representation for some operators. The same limitations as observed for LBOs apply; a voluntary system means that people can choose to use the card or not, meaning incomplete records of machine play. Likewise, players may also have more than one card, though one operator has recently completed a process to reconcile accounts.

Some operators highlighted that there are issues in how they have implemented and developed their loyalty scheme which will affect the data recorded. For example, one operator has recently introduced a new loyalty scheme for 'high-end' players (i.e., their big spenders) and efforts have been made to encourage them to use the new rather than existing scheme. As a result, data from the pre-existing loyalty scheme for this particular operator will be skewed towards the lower-end players because high-end players have been encouraged to stop using this scheme.

3.3.3 Bingo

Player tracking for category B machines in bingo halls has, at best, reached trial stage to date and is not yet embedded within the sector. Of the three participating bingo operators, two run player tracking schemes, both on a trial basis at present. One operator has been trialing player tracking for a little over a year, since May 2012, and the other since February 2013; both on less than 25% of venues across their bingo estate.

These operators collect both atomic and aggregate level information. Each operator is able to link machine sessional data to player specific profiling information, capturing:

- Name
- Sex
- Age

and linking these to machine metrics such as:

- Machine type
- Start time / end time
- Length of session
- Stake
- Money in / money out
- Number of games
- Number of spins
- Frequency of use
- Length of total play
- Usage of points.

Information per game is available if there is only one game on the machine, but for multi-game machines it is not possible to identify which game(s) the player was playing during their session.

Card usage for machine play is optional and the proportion of machine players who are using the card is unclear. Anecdotal evidence from operators suggests that take-up is low at present, due to a lack of real incentive to use the card, and that the ability to own more than one card will create unwanted noise in the data.

Consistency

For one operator, data are stored by each of their individual venues. However, it is collated in one place by their supplier and stored for the whole estate on the supplier's server. The other operator stores data on their own system.

3.3.4 Adult Gaming Centres

Two AGCs were interviewed, neither of which operate any kind of player tracking scheme linked to machine play. However, one operator has the player tracking functionality available in other group companies, should it become commercially viable.

Key points

- Increasing implementation of industry loyalty schemes means that more playing tracking data for machine play are becoming available. This allows us to analyse some information about play sessions, frequency of play and, in some cases, player journeys within sessions of play. However, there are numerous limitations to consider.
- Most gambling industry schemes to link player data to machine sessional data are in their infancy. Therefore, this project is timely, as industry is increasingly exploring this, but also a little premature in terms of what is currently available.
- The main limitation to the use of industry player tracking data is that it is simply not representative of the player base of any sector of the industry. It is a voluntary system, raising questions about completeness of data recorded and who chooses to use these cards and under what circumstances. Take-up of operator schemes is low. Any analysis based on this would need

to be heavily caveated with regard to issues of representation.

- The player data collected at present is limited, with some operators running anonymous schemes, and even those conducting registered schemes mainly collecting only age and sex information. This limits what can be done with the data in terms of linking player demographic detail to sessional information and obtaining an understanding of how different groups of people behave.
- The level of player data recorded differs by sector and also by individual operator within each sector, creating an industry which lacks a standardised way of running player tracking schemes and collating the data produced from them.
- The most promising sector for future use of such data is the casino sector, which have the best take-up rates and more sophisticated schemes. However, further developments to LBO schemes and improvement in take-up rates could increase their worth.

3.4 Proxy session data

As discussed in Section 3.1.3, proxy session data are non-player tracked information generated from atomic level data. What we mean by 'proxy session' is atomic level transactional data that have been split into sessions based on various criteria, including the starting balance, cash being put into the machine and the length of time the machine was dormant. It is not identifiable to a person, it simply a way of carving up the transactional data into chunks that *look like* unique sessions of play. Because the player is unknown, repeated sessions for a person cannot be linked together and a person's journey through different sessions of play is unknown.

As the next sections will show, the level of proxy session data held by LBOs is considerably more advanced than that held by operators in other sectors of the industry.

3.4.1 Licensed Betting Offices

Defining a session

To create proxy session data, a number of parameters must be identified and applied to the data. These parameters and definitions were similar for all LBO operators. Operators tend to agree that a session starts when money is first put into the machine and the starting balance was either zero or less than the minimum stake (and has been for a period of 30 seconds or more). However, identifying the end of a session is rather more difficult. It was commented that the easiest way to establish a session end event is when a player prints a ticket. However, players sometimes continue playing after printing a ticket. Another option is to look at time delays between games. A commonly used definition of session end was when the credit balance was zero or less than the minimum stake and the machine was dormant for a period of 30 seconds or more. These definitions are likely to be refined and developed over time.[9]

However, any session rules can be susceptible to error. For example, if a new player starts on a machine immediately following the departure of a previous player, it is not possible to identify whether this is a new person or the same person playing with additional money. In this example, speed with which players churn may mean that many small sessions look like one long session in the data because the machine has not been left dormant.

[9] Since these interviews, the Association of British Bookmakers has announced that they intend to work with the machines suppliers and major bookmakers to generate a common set of rules for identifying a proxy session. Once this initiative is complete it should improve consistency of data between LBO operators.

Data

Proxy session data are held by all but one of the LBO operators we spoke to. This operator stressed that, while it is theoretically possible to get these data, it would require a large number of assumptions to be made, which may or may not be correct. Furthermore, the process of drilling down into the logged data would be highly resource-intensive and impractical. The point about accuracy is an important one. This was reiterated by another LBO operator who said that identifying sessions involves making broad assumptions about what constitutes a session, such as the machine balance going to zero and the machine being dormant for 30 seconds or more. These assumptions were not considered to be standardised. Also, while a session may be identified, one operator commented on the difficulty of capturing the intensity of the play through the session (i.e., staking level on certain games like roulette, for example whether a player made a large number of small bets or fewer large bets or a combination). Essentially, what this operator noted was the lack of information on what happens within games, and that identifying proxy sessions are just another way of slicing the transactional data.

The same operator felt that proxy transactional data have value, as the proxy session data tracks every transaction; including the length of the session, the amount staked, the gross win, game(s) played, margin and stake level. It is therefore possible for this operator to see a profile of a customer moving between games (on the same machine), or to determine that within their session they only played one game. They claimed that these data can also tell us when the following happens mid session:

- A player transfers all of the money / winnings to the credit meter
- A player cashes out all of the money / winnings
- A player withdraws some of their winnings from the machine but banks the rest and continues to play.*

*This was not our understanding when consulting with other operators but warrants closer inspection.

This has the potential to be particularly useful as, in theory, it opens the possibility of identifying when a player takes back the money they originally staked and continues to play with the house's money. Looking at whether people play differently with their own money and the house's money was identified as an area of interest by many LBO operators.

Compatibility

Of the five LBOs we spoke to, three currently have compatible proxy session data that could be combined by their machine supplier. The other two operators also potentially had compatible proxy session information, as their machines are provided by the same supplier. However, this may not be compatible with the other three and careful understanding of how sessions were identified and defined would need to be made.

Interestingly, two operators also have their own systems for identifying proxy sessions, which are not compatible with others. Decisions would need to be taken as to which data (those from the supplier or from the operator) to use and consideration given to differences between them. This process would need to be conducted with input from these operators to attempt to find common ground, so long as operators were willing to share this information.[10]

3.4.2 Casinos

Only one of the five casino operators we spoke to potentially had proxy session data available. This is because this casino also allows the use of pre-loaded (anonymous) cash cards. When used, data for that unique session of play are recorded. The session will start when the player enters the cash card into the machine and will end when they remove it. Like the loyalty card system, atomic level information is logged against the cash card, which therefore records session information that is anonymised and not linked to personal registration cards. However, this is an optional system and whilst it is possible to use the cards like this in theory, in practice most play starts with cash being put into the machine rather than the cash-loaded card.

As with LBO proxy session data, individual and unique sessions of play can be identified. However, also as with the LBO proxy session data, they cannot be linked together for a person, so a person's journey in terms of repeated sessions of play cannot be tracked.

3.4.3 Bingo Halls

One of the three bingo operators we spoke to has proxy session data. This is because the vast majority of B3 machines for this operator are provided by the same firm who supply machines to LBOs. This supplier has confirmed that largely the same level of proxy sessional information that they provide for LBOs can be provided for these bingo machines.

3.4.4 Adult Gaming Centres

AGCs do not have proxy session data.

[10] As noted previously, the Association of British Bookmakers is leading an initiative to do this.

Key points :

- Proxy session data are only consistently available across the LBO sector. This is because atomic level transactional data are needed to generate this.
- One casino and one bingo operator have the potential to have proxy session data, though it is unclear if they do this.
- Proxy session data are simply another way of slicing transactional data into chunks based on assumptions of what looks like discrete sessions. How sessions are defined varies and how accurate they are is unknown, which raises questions about use of these data for research purposes.
- However, it represents comprehensive data for the LBO sector. Therefore, examining how accurate this information is, facilitating improvements across the sector to make data more consistent, and exploring what these data can tell us about within-session play is still useful.

3.5 Summary of data availability across operators

In Sections 3.2-3.4 we have summarised the various types of data (transactional, player, proxy) available by sector. In the tables that follow, this information is presented for each operator included in the research process. This summary data matrix therefore shows how many operators have certain types of information available by sector and serves to further highlight the disparity between them.

Operator & Sector	Transactional data			Player data					Proxy session data			Other
Licensed Betting Offices												
	Atomic	**Aggregate**	**Compatibility**	**Available**	**When**	**Type 1**	**Type 2**	**Compatibility**	**Available**	**Type**	**Compatibility**	
1	Y – game cycle data	Can aggregate to specification. Options include: • By hour, day, week, month etc • Venue, machine, game, estate	Data for operators 1-3 compatible and supplier can combine for us.	Y	2013 onwards	Anony-mised	Atomic & aggregate	System and data compatible with LBO operator 2	Y	Atomic & aggregate	Data for operators 1-3 compatible and supplier can combine for us.	
2	Y – game cycle data	Can aggregate to specification. Options include: • By hour, day, week, month etc • Venue, machine, game, estate	Data for operators 1-3 compatible and supplier can combine for us.	Y	2013 onwards	Anony-mised	Atomic & aggregate	System and data compatible with LBO operator 1	Y	Atomic & aggregate	Data for operators 1-3 compatible and supplier can combine for us.	
3	Y – game cycle data	Can aggregate to specification. Options include: • By hour, day, week, month etc • Venue, machine, game, estate	Data for operators 1-3 compatible and supplier can combine for us.	N	Trial only in July 2012				Y	Atomic & aggregate	Data for operators 1-3 compatible and supplier can combine for us.	

Operator & Sector	Transactional data			Player data					Proxy session data			Other
Licensed Betting Offices cont.)												
	Atomic	**Aggregate**	**Compatibility**	**Available**	**When**	**Type 1**	**Type 2**	**Compatibility**	**Available**	**Type**	**Compatibility**	
4	Y – game cycle data	Can aggregate to specification. Options include: • By hour, day, week, month etc • Venue, machine, game, estate	Data for operators 4-5 compatible, supplier can combine. Not currently compatible with operators 1-3	Y	Since 2011	Register-ed	Atomic & aggregate	Data system unique to this operator	Y	Atomic & aggregate	Operator has own systems to identify – not compatible with others	Age, sex from membership records. Contextual information about venue potentially available
5	Y – game cycle data	Can aggregate to specification. Options include: • By hour, day, week, month etc • Venue, machine, game, estate	Data for operators 4-5 compatible, supplier can combine. Not currently compatible with operators 1-3	N					Y	Atomic & aggregate	Operator has own systems to identify – not compatible with others	Contextual information about venue potentially available

Operator & Sector	Transactional data			Player data					Proxy session data			Other
Casinos[11]												
	Atomic	**Aggregate**	**Compatibility**	**Available**	**When**	**Type 1**	**Type 2**	**Compatibility**	**Available**	**Type**	**Compatibility**	
1	Y – 20-30 second recording – not necessarily game cycle	Y – Data recorded at hourly basis.	Operator's own system	Y	2013	Register-ed	Atomic (game cycle) & aggregate	Operator's own system	Y	From cash card. Better than proxy	Only operator to have cash card	
2	N - tbc	Y – basic data recorded by hour	TBC	Y	Since 2009	Register-ed	Atomic (game cycle) & aggregate	Operator's own system	N			
3	N	Y – Data recorded at hourly basis	Operator's own system – hosted by external partner	Y – B1 machin-es only	2011/2012 approx	Register-ed	Aggregate (15 minute cycles – tbc)	Operator's own system	N			
4	Y - but transient - only stored for 3 days	Y – by recorded by machine per day	Operator's own system – hosted centrally	Y	TBC	Register-ed	Atomic (game cycle) & aggregate (although atomic only stored for 3 days)	Operator's own system	N			

[11] For reasons relating to commercial confidence, one casino operator did not wish to be listed in this table.

Operator & Sector	Transactional data			Player data					Proxy session data			Other
Bingo												
	Atomic	**Aggregate**	**Compatibility**	**Available**	**When**	**Type 1**	**Type 2**	**Compatibility**	**Available**	**Type**	**Compatibility**	
1	Y (game cycle data)	Y – data held at venue level; can be combined by supplier; weekly aggregation	Operator's own system + supplier's system	Y – trial in 20 clubs	From Feb 2013	Registered	Atomic (game cycle) & aggregate	Operator's own system + supplier's system	Y	Atomic (game cycle) & aggregate	Suppliers system (waiting for confirmation if compatible with LBOs)	
2	Unknown – operator to confirm	Y – waiting confirmation on levels of aggregation	Uses Playsafe system with some metrics from supplier where machines are server based	Y – trial in aprox 20 clubs	From May 2012	Registered	Atomic (game cycle) & aggregate	Operator's own system	Un-known – oper-ator to confirm	Unknown – operator to confirm	Unknown – operator to confirm	
3	N (though potentially if Barcrest machines in bingo halls – need to confirm)	Y – data recorded at hourly levels for some machines, otherwise at weekly level.	Uses Playsafe system – unique to them	N					N			

Operator & Sector	Transactional data			Player data					Proxy session data			Other
Adult Gaming Centre												
	Atomic	**Aggregate**	**Compatibility**	**Available**	**When**	**Type 1**	**Type 2**	**Compatibility**	**Available**	**Type**	**Compatibility**	
1	Y – only for Barcrest machines	Y – different levels for different machines. Some daily, some weekly	Operator's own system	N					Un-known potentially for Bar-Crest machines		Operator's own system	
2	Waiting for confirmation	Y – different levels for different machine. Some electronic, some manual	Operator's own system	N					N			

3.6 Data size

3.6.1 Crude data size estimates

In addition to looking at data availability and structure, a key aim of this study was to gain a better understanding of data size. This is driven, largely, by practical considerations for future research, so that we can gain a better understanding of how much server capacity may be required to house the data. Furthermore, software requirements should also be considered in terms of how easily or otherwise the data may be analysed.

Not every operator knew how large their different datasets would be. Whilst we have asked for clarification, this information has not been easy to generate. In some cases, it would require the data to be pulled from the system in order to answer this. However, the table below summarises our broad understanding of data size by operator and sector. Where data size is unknown, we estimate this using information from similar operators.

Operator & Sector	Data types	Data size	Notes
Licensed Betting Offices			
Operator 1	Atomic data Aggregate	On average, 0.5 gigabyte of data per day for atomic level game cycle data Unknown	Size may vary based on day of week/time of year. Estimated that 1 daily extract equals 1 million transactions Unknown
Operator 2	Atomic data Aggregate	On average, 0.5 GB of data per day for atomic level game cycle data Unknown	Size may vary based on day of week/time of year. Estimated that 1 month extract equals 175 million rows of data. Estimated that 1 month aggregated extract of loyalty card data equals 867,000 rows of data
Operator 3	Atomic data Aggregate	Atomic data estimate unknown, but likely to be similar to operators 1 & 2 Estimate that aggregated transaction data are around 24 MB per year/ 2 MB per month	
Operator 4 & 5	Atomic data Aggregate	Atomic data estimate unknown, but as these are smaller operators, likely to be less than 1 GB of data between them Unknown	

Casinos			
Operator 1	Atomic – player tracking Aggregate	Aprox 8 MB per day, 240 MB per month Estimate on average 8 MB per club per day	This is for all machine data output on a daily/monthly basis. File is output in text file format which is why it is smaller than expected and processed in a database. Stated other operators don't have this system so their data will be larger
Operator 2	Atomic – player tracking Aggregate	Estimate that a few months of data are around 500 MB	Operator unsure, information obtained from third party supplier
Operator 3	Atomic – player tracking Aggregate	Unknown	Waiting for IT department confirmation
Operator 4	Atomic – player tracking Aggregate	Unknown	Waiting for IT department confirmation
Operator 5	Atomic Aggregate	N/A Approx 100 GB of aggregated data	Atomic data are transient (only stored for 3 days) therefore file size not relevant
Bingo			
Operator 1	Atomic – player tracking Aggregate	Unknown	Waiting for IT department confirmation
Operator 2	Atomic – player tracking Aggregate	Unknown	Waiting for IT department confirmation
Operator 3	Aggregate	Unknown Likely to be small as this operator only has a small number of clubs.	Waiting for IT department confirmation
Adult Gaming Centres			
Operator 1	Aggregate	Total size of playsafe database is 8 GB. Therefore estimate this to be 1 GB per year/85 MB per month	
Operator 2	Aggregate	Unknown but estimate will be similar to the other AGC operator. Therefore estimate 1 GB per year/ 85 MB per month	

Looking at LBOs first, we can see that around 2.5 GB of atomic data are being generated on a daily basis. Therefore, one month's worth of data at this level would be approximately 75 GB in size.

Other sectors were much less specific about the size of their data, and at the time of writing this report we were still waiting for clarification from some operators. However, we can assume around one month's data from AGCs would total around 200 MB and equivalent estimates from casinos may be around 2.2 GB per month.[12] We would estimate the size of data held by bingo clubs to be in excess of casinos given their greater number of venues. But we would also estimate this to be lower than LBOs, as they do not record data in as much detail. Therefore, a reasonable estimate for bingo clubs might be around 10 GB per month.

Taking this together, single monthly extracts of data across all sectors would be around 88 GB per month. However, this is a very crude approximation and does not take into account the variety of data types at different levels of aggregation that may be required.

The operators and suppliers interviewed also noted that even with their large and powerful data warehouses, they are not keeping data indefinitely. Atomic level data are generally kept for a six-month period (though some are keeping this for longer, up to 18 months), whereas aggregate information is kept for a number of years.

3.6.2 Data size considerations

The data sizes estimated above are for one data file extract only. As we have illustrated throughout this report, there are many different types and levels of data available. It is likely that researchers working on future projects would wish to obtain further data at varying levels of aggregation. This could be having both atomic data for player tracking records and having those data aggregated by session per player, for example. Or they may wish to look at data aggregated by venue, by region, by game type played and so on. All of this will increase the size of the data used.

A major consideration for researchers is their process when cleaning, reconciling and preparing data. For quality control purposes, when cleaning data or deriving new variables it is standard to save new copies of the resultant data that are generated. This is so that there is always a master copy available and steps can be retraced if errors have been made. In addition, it is likely to be a researcher's intention to merge data together, thereby creating new datasets from the originals. Furthermore, standard disaster recovery policy means that (typically) all networked data needs to be backed up to a separate server. This automatically doubles the size of server space

[12] We have calculated estimates for two operators based on known data from their most similar counterpart who reported this information. At the time of interview, this operator has three casinos, albeit one with a larger number of machines, and estimates their monthly extract of data to be 100 MB. This equates to 33 MB per venue. The other two operators have 59 venues between them. Therefore, we estimate the volume of data to be 33*59 = 1947 MB for these two operators. This is in addition to the c.340 MB reported for the other giving a total of 2287MB or 2.2GB. This is only a crude approximation and we will update these estimates once further information is obtained.

needed.[13] Therefore, erring on the side of caution, we anticipate that researchers would need at least ten times more capacity on their systems than the actual size of the data initially requested, to ensure that they could work with the data in a sensible way.

Finally, we need to be cognisant of the size of data that standard analytical packages can work with. For example, the size of data that packages like SPSS can cope with will depend on the depth and breadth of the datasets: what this means is how many variables/metrics there are (i.e., columns) and how many cases (i.e. rows). We envisage that the requested data will have only a few metrics (possibly less than 20) but will have many, many cases (using the example from one operator, around 5.8 million rows/cases per day). It is our understanding, that depending on the overall file size, SPSS can handle many millions of cases but functionality of processing may be slow. Whilst researchers are not constrained to using these packages, we are cognisant of the need to recommend action for future research that can be delivered swiftly and within budget. Packages such as SAS, SPSS and STATA are those used routinely by analytic researchers and therefore the skills (and licenses) for using these packages already exist. We would be keen to hear suggestions from the Trust and MROP about alternative software and methods to a) manage and b) analyse these potentially large data.

[13] There are ways to reduce this capacity requirement, such as backing up to tape rather than servers. This means that in the event of a disaster it would take approximately one week to recover the data as opposed to 24 hours, but less server space is required.

4 Discussion

Having described the range of data available, and their size and structure, this section summarises the key themes emerging so far and discusses some options for future stages of research. This also includes discussion of the various challenges and limitations identified. These are our initial thoughts and are intended to generate further discussion and thought from key stakeholders about what appropriate next steps might be.

In summarising our findings and presenting options we are mindful of the original total budget set aside for this programme of work. We are also mindful of the need to proceed quickly with any further stages of research (should they be agreed).

In the sections that follow, we first summarise our findings with specific consideration of questions that these data can contribute towards. We then summarise some of the main challenges and limitations identified, followed by potential opportunities presented by these data. Finally, we outline some suggestions for what might realistically and sensibly be done next.

4.1 'Data rich, information light'

'Data rich, information light' was a phrase used by one of the operators interviewed to summarise the level and state of industry data available on category B machines. This phrase effectively summarises our main findings from this study. It describes how there are a vast amount of data available but the information and insight they currently provide is narrow in focus. What is meant by this is summarised below:

- The main source of information available is transactional, which limits the types of research and policy questions that can be asked of the data.
- This transactional data are held at various different levels, meaning that some operators have much more detail on specifics of game cycles whereas others hold data at an aggregate level.
- There is a great deal of inconsistency within and between operators and sectors in the level and type of data that is captured. However, because of data size, atomic level data are not stored indefinitely and likely to be only available for a rolling six-month period.
- All data does include a temporal dimension, so changes over time can be tracked. But, as noted above, only aggregate level data are being kept for a long time period.
- There is a very limited range of contextual or demographic information available. Demographic information is only available for some operators who

track player data using a person-registered system and even then the demographic information collected is sparse. Little contextual information is available (though this could be created by merging other administrative records together at a geographical level, i.e. crime data, deprivation, high street types, etc.)

- The casino sector offers the most promising data on player tracking, as levels of loyalty scheme uptake are higher among this sector and this information is linked to registration records, providing some demographic information.
- Proxy session data are most complete for the LBO sector, though the accuracy and consistency of these data are unknown and its classification is likely subject to error. It may be possible to validate some proxy session assumptions against player-tracked data.
- This project is timely as operators are undertaking more investigation about how best to use this information for themselves and how to supplement this with other meaningful metrics. For example, one operator described how they plan to look at how to integrate transactional data and game data.
- Some operators are also in the process of expanding player tracking and loyalty card systems following trials. In the forthcoming year, we would anticipate coverage of loyalty schemes to increase. Operator activity in this area as well as levels of uptake should continue to be monitored.

All of the above means there is a potential wealth of data that could be used (making this source data rich). However, there are many limitations to consider and the range of information that these data include as a standalone source is somewhat narrower in focus than originally anticipated.

4.2 Managing expectations

Prior to undertaking this phase of work, there was a broad perception that the industry has significant insight into machine behaviour from analysing their data internally. This, in turn, led to a great deal of optimism among regulators and policy makers that industry-held data could answer many key questions. Information from this study suggests that the extent to which this is true varies between operators and sectors. Suppliers of machines in LBO's have significantly more insight and more finely developed tools than some of the individual operators. This is in their interest so that best performing machines/games can be identified with analytics used to help them better understand the market. Some individual LBO operators rely on suppliers for data and/or intelligence, whereas others had fairly sophisticated independent analytic teams. That said, all operators noted the difficulty in using this data and the narrow focus of the information it provides.

In other sectors, the extent to which machine data were used by operators was much more variable. This may be (potentially) related to lower levels of detail available and because in sectors like casinos, machines are not as prominent in

terms of revenue. Where further internal analytics were available, these were not generally shared with us as they were seen as proprietary and commercially sensitive.

However, the broad opinion of most operators was that now was a good time to start looking at these issues as they themselves were starting to look at how they could use these data more systematically. Therefore, whilst industry-held data offer an opportunity to generate new types of research knowledge, they are not a panacea. Utilising this potential (and dealing with the challenges raised in this report) is likely to be an ongoing process rather than a one-off specific event.

4.3 Challenges and limitations

We have identified a number of challenges in attempting to use industry data for research purposes. Whilst these are not insurmountable, they are important as they may limit either the range of what can be achieved or affect the accuracy of resulting data. Some challenges are practical and pertain to what can be reasonably achieved within this scoping project.

4.3.1 Challenges raised by operators

During each interview we discussed any concerns that operators may have about the process and progressing to the second phase of this research. Operators were also asked to review written summaries of these meetings which provided a further opportunity to discuss concerns. Two main issues were raised. These were:

1. Protection of commercially confidential information

This concern was raised by nearly all of the operators consulted. By and large, all operators stated that they were willing to share their data but had concerns about how they would be reported. The main concern was that their proprietary and commercially confidential information could become known by their competitors. This comprises a number of factors:

- That information about their systems, metrics or intentions to expand in certain areas may become common knowledge beyond what they would typically share. For example, some operators and/or suppliers did not want to advertise dates for launching new initiatives (which could have impact in terms of improving the level of data they collect) to competitors.
- That competitors might be able to 'reverse engineer' any statistics published and extract their individual figures from this. The logic here is that each operator will know what their own figures and data are and could subtract these from any estimates published to extract results for competitors.
- That data should be published at a sufficiently aggregated level to prevent reverse engineering but also that care be taken not to use labels (such as a

'large casino' which is currently identifiable to Aspers Stratford or Aspers Milton Keynes) that will identify data from individual companies.

Only one operator stated they were happy to share data and have data published that named them as the source.

These issues are critical as operators were clear that they would not share data unless these points could be addressed. In the meetings with operators, a range of solutions were discussed to allay these concerns. First, we offered to include a clause in the confidentiality agreements that covered not just identifiable but also potentially identifiable data. This issue was discussed with the Trust and it was suggested that so long as any data published comprised aggregated information from three or more operators, this should help prevent reverse engineering (though this will limit what analysis can be done). Finally, consultation with NatCen's legal team has highlighted that, in order to keep to the terms of any confidentiality agreement, operators have the right to review any document or detail that will be made public to ensure that commercial sensitive material can be redacted. It should be noted that this right to review is only in relation to the identification of commercially sensitive material and not for substantive review. This represents a logistical challenge as separate reports only covering information for each operator will need to be produced for review in the first instance. On the whole, operators indicated they would be willing to proceed on this basis though it is anticipated that these issues will have to be revisited in any future stages.

1) Resources and practicalities

Some operators noted that committing various time and staff resources to produce the data needed to the specification required might be difficult. Suppliers to the LBOs suggested they would need between a four-week to two-month lead-in period to be able to generate any data required. It was also suggested that some financial costs may be associated with this. We envisage that the lead-in times required by operators will vary between these time frames. In our experience, these lead-in periods can quickly become extended as further clarification and checking of resultant data are undertaken.

One operator noted that because of internal restructuring, the personnel with expertise in this area will no longer be based at the company. Therefore, they may find it difficult to produce the data we need. This operator does, however, have some data they have already provided to an external academic that could be made available to us.

Finally, for some operators, the data are not held or stored with them but with an external third-party organisation (not currently included in the scoping exercise). We have made contact with one such organisation which has been helpful in clarifying information but has taken a long period of time to do so. Therefore, there may be additional complications due to the more complex nature of these arrangements.

We need to be aware of these issues when thinking about what can reasonably be achieved in future research aiming to use industry data. Clear specification of what we

require in what format will be crucial to ensuring this process runs smoothly, as will identifying what data are required to implement future research successfully. It may be that asking for smaller samples of only certain types of data helps address some of the resourcing concerns raised.

4.3.2 Challenges raised by the research team

A number of challenges were identified by the research team in using these data for research purposes. These are summarised below.

Gaps in data and explanatory potential

This phase of research highlighted a number of significant gaps in using industry data for research purposes. These relate to the following areas:

- Game data
- Demographics
- Contextual information.

As noted throughout this report, the data provided are transactional data, tracing the money in and out of machines. There is no systematic link to what happens within the game. This closes down a whole range of potential research about how people play and interact with certain game characteristics or features of machines. One operator stated that they were currently thinking about how better to capture this level of information. At the time of interview, this was not something that was high on operators' agenda but we have since been informed that this is increasing in importance.

However, efforts could be taken to explore the extent to which this development could take place. Rather than rely on verbal reports of its complexity, this could be examined empirically. The challenge this represents should not be underestimated. Given the number of game providers and game types involved, each of whom has different log file data outputs, this is likely to be extremely complex. Exploring the feasibility of this and working with industry to assess this could be part of the development of a broader machines data strategy.

Demographic data were only collected among those operators who ran a player registered loyalty card scheme. Even then, the level of detail recorded was typically limited to age and sex. It is only on the registered schemes that metrics like self-exclusions could be merged with account data, though the operators were unsure if this occurred. This means there is a very limited range of additional information to use to explore patterns of behaviour among different groups of people, which may be of use in terms of developing risk profiles. Notably some loyalty card schemes were chosen by operators precisely because they were anonymised and did not require any demographic information to be recorded. These operators stated that this was what their customers wanted. Yet, within the same sector, other operators have

demonstrated that it is possible to run a registered scheme which is broadly acceptable to players. This is likely related to various hardware and system requirements of operators and machine suppliers (the anonymised loyalty scheme is a product developed by one of the two LBO suppliers).

To collect demographic information from those who currently have an anonymised scheme will require a wholesale shift in approach for those operators. This is likely to require new systems and possibly new machines (or software, at least) to be developed. Strong justification of the need for and utility of resultant data would be likely be required for this to happen. However, this is also something that could be explored if a more strategic machines data strategy was to be developed by policy makers/regulators.

Finally, very little contextual information about the venues or the machines themselves was available. Only one supplier of LBOs noted that they felt this was important and had a strategy for obtaining this information in the future, though it was not currently captured. Among other suppliers, there were no intentions to collect these data, though one operator has since expressed interest in using operational reports from staff in the development of a risk assessment strategy. This would require development of a coherent and consistent strategy to collect this contextual information which could be rolled out to other operators.

Analytical challenges

We have touched on some of the analytical challenges already. In many cases this pertains to how accurate the data are (in the case of proxy session data), to how representative they are (in the case of player tracking data) or whether they represent a useful focus (in the case of transactional data). However, one further aspect was raised by operators. This is the need to fully understand the background of what operators were doing at certain points in time to be able to interpret data accurately.

For example, operators noted how periodic in-house promotions could show up in the data as large upturns in turnover, where actually this turnover is largely 'comped' play from the operator. Here, contextual understanding to what was going on across an operator's estate is vital in understanding the data correctly. This raises two issues, first, the difficultly of disentangling complimentary play or offers from the transactional data, and second, the need to know about schemes such as this to interpret the data correctly. Context therefore needs to be expanded to include operator practice which may be needed on a venue-by-venue basis. To our knowledge, this is not systematically recorded.

Consistency and standardisation

Phase 1 of this scoping study has demonstrated the lack of consistency within and between operators in terms of the level of data they collect and how they collect it. The LBO sector arguably has the most consistent information, as data for five operators are available from just two suppliers. Therefore, data from just two sources,

rather than five, have to be reconciled. However, some LBO operators have more detailed information in-house, especially around player tracking, meaning further variation for some data types. That said, some operators have noted that they would be willing to work on refining outputs to increase consistency further.

In other sectors, each operator tends to have a wide variety of systems some of which are bespoke to the individual operator. There is also inconsistency about how data are reported with aggregated data being reported in a range of 15-minute intervals to one week intervals.

This inconsistency is problematic as it means combining data from many sources will be difficult. Working on a sector by sector basis seems the most appropriate approach but then further challenges are raised by industry concerns regarding the reverse engineering of results. If this concern can not be resolved, it means future research will be limited to producing work only where there is consistent information from at least three individual operators. This means that any potential analysis will be limited to the least detailed level of information available. For example, if two operators record information on a daily basis but another only records data on a weekly basis, then any analysis has to be run at the weekly level to ensure there is consistent information for at least three operators. Here, operator concerns are (potentially) limiting the scope of what type, detail and level of information can be produced. Furthermore, some potentially useful data, such as the LBO player tracking data with demographic information, could not be used under the 'rule of three' as any resultant data would be identifiable to the single operator who has this. The same is true of the cash card data used by a single casino operator. Likewise, use of player tracking data from LBOs will be limited to data generated from June 2013 onwards, as it is only from this date that at least three operators had player tracking schemes.

These considerations aside, to progress with using these data, a detailed data dictionary will need to be produced. A critical issue is that whilst there are common metrics which are recorded, there is no standardised protocol about naming conventions or definitions. This gives rise to a great deal of variety between operators. Therefore, the data dictionary will need to list the naming conventions of various metrics from each source, record their definition from the source and create a common identifier for a combined dataset. This is likely to be time consuming but is critical to understanding this information and creating a uniform and standardised set of data to work with.

Data size

Final challenges relate to data size. As noted in Section 3.5, the size of data available is large. Just for atomic level transactional data across the estate, we estimate a monthly supply to be in excess of 88 GB. Across the course of a year, this would be over 1 terabyte of data. Whilst this is not, perhaps, as large as might be expected, we cautiously estimate that around ten times this capacity would be needed to allow all data management processes to be completed. This may limit what can be achieved in future research stages.

4.4 Potential opportunities

Despite the challenges noted in the previous section, there are a number of potential opportunities regarding how these data might be used and/or supplemented to provide further knowledge about machine play in Great Britain.

4.4.1 What value can these data add?

Whilst there is a wealth of data available for analysis, a key question is identifying what would be most useful for the Responsible Gambling Strategy Board, the Trust and policy makers across government organisations. Any further research should have utility for policy and practice. This, therefore, raises questions about where to focus resources and what types of information the Trust and stakeholders value most.

The overarching aim of the whole category B programme of research is to describe, understand and militate against potential for gambling-related harm. The industry data scoping project is the first part of this programme of work looking at what contribution industry-held data could make to these objectives.

What we have identified is that use of these data would allow us to better describe a variety of machine-related behaviours. The most standardised data are collected and stored at a machine level. This provides an opportunity to look at volume of play at different levels of aggregation – for example by game type, by machine, by venue, by region, but it does not tell us much about individuals. This has to come from player tracked data. Therefore, whilst this offers a unique opportunity to map volume of play at different levels, there are questions to consider about how much value data viewed through the lens of the transaction, and not the person, can have for policy makers.

We believe these data will be very helpful in better describing the market, and changes in the market, in which individuals conduct their play. It also offers the opportunity to assess volumes of play, as defined by spend, at different levels of aggregation and time periods. All of this provides useful metrics which could help us to better understand the context of individual play.

There is an opportunity also to examine how certain regulatory policies might affect volume of play (for example, extended licensing hours or changes in stakes and prizes). This would be extremely useful in terms of helping researchers to better understand these issues and assess impact. For example, the triennial review of stakes and prizes on gambling machines highlighted how little is known about patterns of staking behaviour. On B2 machines, there was a suggestion that the maximum stake could be reduced from £100. Whilst anecdotally the industry have stated the average (notably not median) stake is likely to be around £12-£15, the actual shape of the distribution of staking behaviour on B2 games is not known.[14]

[14] Though in the intervening time between when this project was completed and this report published, the Gambling Commission, supported by the Association of British Bookmakers, has gathered some further information on this: see: http://www.gamblingcommission.gov.uk/pdf/letter to alison pritchard re b2-b3 gaming machine analysis.pdf

Discussing this with operators suggests that the distribution curve has a long tail to £100 with an upturn in transactions at the £100 mark. If the shape of distribution were known, it would be easier to theorise about the potential impact of changing stake levels and the number of transactions this would be likely to affect. This could be easily generated from transactional data. Furthermore, any changes to stake and prize levels could be monitored pre-and post-implementation through transactional data to help assess impact in a robust way.

A further policy area of increasing interest (especially among Local Authorities) is the impact of extended licensing hours on gambling behaviour. Transactional data could be used to assess how levels of transactions vary at different points in the day and night and what patterns of play are evident. For example, are the sessions conducted late at night longer or shorter than day time sessions, are stakes higher, are sessions more volatile, etc.[15] Furthermore, as there are individual venues which are receiving permission to extend licensing hours, the data could be used to make some comparisons pre-and post-implementation.

To extend this further, there has been much debate about the impact of so-called 'clustering' of gambling premises. There are important questions about the cumulative impact of such clustering. For example, if a new premises opens in an area does this mean there is an overall upturn in turnover for all businesses (here the argument would be that more total demand is generated) or does it mean that some players swap their place of play, meaning that at an area level, overall levels of turnover are fairly stable but that consumers have more choice about where they play. The reality is likely to be some combination of the two. However, to date, looking at transaction level information within an area (and hence modelling the effects of increases or decreases in provision) has not been possible in any robust way.

Likewise, the Responsible Gambling Strategy Board has identified advertising and marketing as a key area of interest. There is an opportunity that transactional data could contribute to an understanding of the impact of these initiatives.

Therefore, just because the information is largely transactional and the central focus is the machine/venue/estate, this does not mean that the data cannot provide useful insight for policy. In particular, industry responses to certain regulatory reforms often centre around issues of impact on business. Transactional data offer the policy and research community an opportunity to better assess these issues independently. Furthermore, in Great Britain currently, regulatory approaches to machines tend to focus on regulating the machine rather than the individual (i.e., in terms of stake, prizes, speed of play, etc). A body of research which focuses on the machine as the central object of interest aligns with this current regulatory perspective and may provide some insight into these issues.

[15] This would likely require us to draw on all three types of data: transactional, proxy and player, with suitable caveats.

However, there are some limitations, and transactional data are one step removed from providing information about individuals and individual patterns of play. Transactional data are less helpful in understanding this and its relationship to individual risk. No information about the individual is available and therefore they are not of central focus. Player tracking data are of primary importance in focusing on the experience of individuals. In some sectors this information does not exist, in others it is in its infancy and there are many issues to consider about how reliable this information is.

Despite the many challenges associated with using player tracking data, they still provide the best opportunity to learn more about patterns of play of specific individuals. Any resultant data would need to be heavily caveated, especially around issues of representativeness and these issues would need to be investigated fully. However, knowing more detail about a certain segment of machine players is preferable to knowing nothing. There has been increasing interest shown in individual-led regulatory models whereby regulation and policy is more targeted and bespoke to the individual. The opportunity to better understand patterns of play among a certain segment of machines players should not be underestimated. This could be focused on research questions which examine both within-session play and how the player responds to certain events, and also on between session play in terms of frequency of play and sequences of sessions. If combined with other sources of information (such as assessment of whether players are experiencing problems), this could provide a building block to help understand and assess risky patterns of play. At the very least, some parameters for prevention initiatives might be established. Of course, given the attendant issue of self-selection and low levels of take-up, a key question is whether this is robust enough for policy purposes. In the short term, this information is likely to be very unrepresentative and not suitable for policy development, but it could serve as a useful building block for knowledge and highlight future avenues of investigation. This offers new opportunities to generate knowledge not previously available. As player tracking rolls out across the industry and uptake increases, the value of these data for research purposes will increase. Knowing how to use the data and, potentially, having an opportunity to shape how and what is collected and stored, is likely to be both useful and increasingly important.

4.4.2 Filling gaps and supplementing knowledge

Although significant gaps in the data were identified, there are opportunities to supplement this with information from alternative sources. For example, to fill the demographic information void, an option would be to conduct some work with machine players recruited from venues.

Here, a greater range of demographic and contextual detail could be collected (for example in a survey) and permission to link responses to industry-held data obtained. With appropriate consent, permission could be asked to use their log in ID for their loyalty card(s) to extract this information from industry records.[16] Development work would be needed to test the feasibility and acceptability of this to both players and the industry. Alternatively, if observational work was being done at a specific venue, information from observations could be matched with machine data so long as the date, time and machine ID were recorded. These are just some options to explore how to generate a greater level of demographic information on players and how to use it alongside industry held data.

In terms of generating contextual data, the postcode of each individual venue is available, and creating a database of contextual variables is something that could fairly easily be produced by researchers using GIS techniques. Here a range of metrics could be merged together with venue location to look at things like a) number of other venues in the area, b) type of location (high street, other etc), c) various metrics of deprivation of the surrounding area, d) any local area crime statistics, d) information about the local populace (demographic profile etc). Similar techniques were used on the machines 1 study, which mapped the geographic location of venues with slot machines. However, this study used the Gambling Commission's register of licensed premises to do this. This relies on Local Authorities returning this information and some authorities had failed to do so. Therefore, obtaining this information direct from operators would be more accurate.

In additional to contextual details about the environment in which the venue is situated, a machines data strategy could recommend that specific information about each venue is recorded. This could include metrics such as opening hours, whether there is an ATM in the venue, number and layout of machines, etc. This would need to be systematically recorded and updated for each venue but could help to add explanatory power to the transactional and player tracked data.

Finally, where data are player tracked, there is the potential to situate an individual's machine play in context of their broader gambling engagement. Assuming that the loyalty card was being used consistently across all products (not necessarily a given) one could identify different player types (machines only; machines and table games etc) for any given day. Then volumes of play could be examined for those who do integrate products to assess what the contribution of machine play is to their broader gambling engagement. Whilst it is not possible to link play between different operators, this would at least give greater insight into play with a single operator.

[16] It is standard procedure on many social surveys to ask for permission to link data to administrative records. For example, the Health Survey for England asks for permission to link to the Hospital Episodes Statistics Register and the Cancer register. This procedure has been undertaken for well over a decade. Here, the principle is the same but instead of NHS records permission would be sought to link to player tracked records. How acceptable this is to players (if at all) would need to be investigated.

There are a range of potential solutions to the issues raised above, some of which may be more feasible than others. However, this does highlight the broad gaps in industry data and the need to think creatively to resolve these. In all cases, it involves a more systematic development of a data strategy and supplementing the transactional data with further information, drawn from other sources, to increase the potential of what these data add.

4.5 Recommendations

Our recommendations for next steps take into account the challenges noted in Section 4.2. They are also based on the recognition that in the majority of cases the size, type and scale of data were described to us rather than examples being shared. Some operators did share examples of data outputs and showed us their systems. However, the majority did not. Whilst we are confident that we have collated enough information to summarise the broad themes presented within this report, there is likely to be more work to do in terms of understanding structures, once actual data are provided.

We recognise the broad potential of industry data for research purposes and the variety of areas to which these data could contribute. Using industry-held data across a range of sectors represents a unique opportunity to explore a range of issues where hitherto there has been no or little information. This potential should not be under-estimated. However, we also recognise the need to better understand the potential contribution of these data to the evidence base more generally and how it might be combined with other research approaches to provide a more rounded understanding of machine gambling behaviour. All research methods have their limitations and it seems that industry held machine data are no exception. The data are potentially rich but somewhat narrow in focus, and would benefit from additional metrics being collected to enhance their explanatory potential.

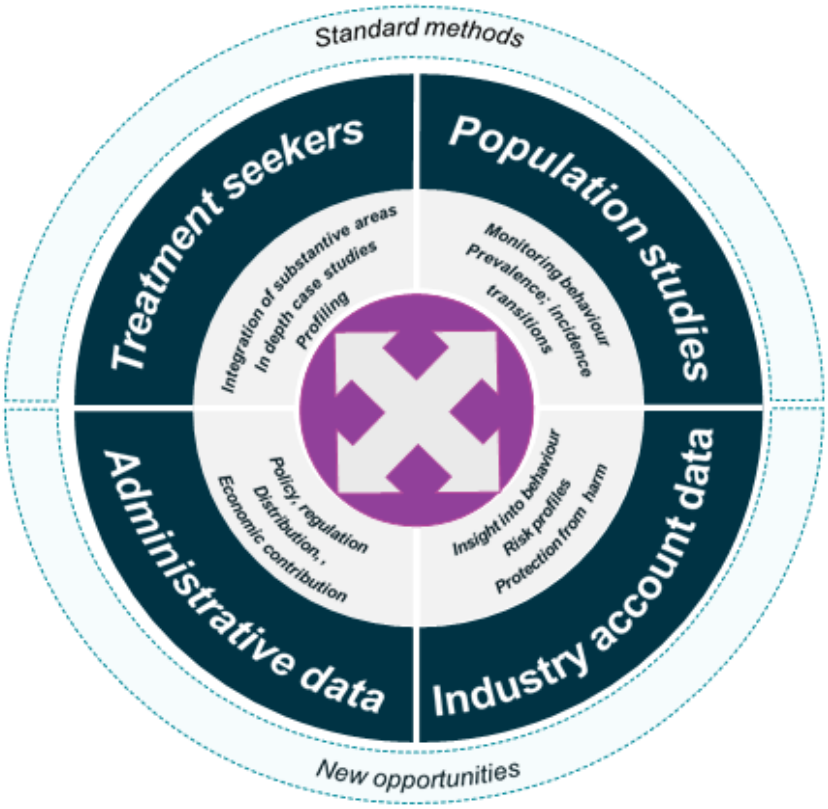
Our view, based on what we have learned so far, is that there is the opportunity to use industry data to examine certain policy questions but its real potential comes by adding this to researchers' methodological toolkit. Hence, we do not believe, having spent time with operators across the gambling industry, that operator-held data can provide a holistic view of machine play and players. Rather, it is as a supplement to other forms of information that these data will provide most use.

Careful thought is needed to explore how best to integrate this source with evidence from other methods and how to link these approaches together in a way that allows us to promote a more integrated approach to evidence. For example, we have already briefly outlined opportunities for matching administrative data to venue records to boost the level of contextual information available. Other options include surveys of loyalty card members to obtain demographic information we are missing and so on.

Thinking through these new opportunities and making links between different types of information to generate a more integrated understanding of machine behaviour is, in

our opinion, important. How an integrated approach to understanding gambling might be visualised is shown in Figure 5.

Figure 5: Diagram of integrated approach to gambling research evidence



In this diagram, industry-held data make a notable contribution to knowledge, and provide significant new opportunities to learn more about machine player behaviour but cannot provide all the answers alone. It has to be viewed in conjunction with other sources of knowledge and, in all likelihood, these sources have to be linked together to enhance understanding.

Given the various (likely) challenges and limitations documented thus far in using industry-held machine data, we believe this diagram is useful to help orientate our next steps. We would argue that any further phases of this research should be focused on better exploring industry data potential within this integrated model, whereby the power of industry-held data is explored in conjunction with other sources of evidence and information.

This is a slight reorientation in focus whereby the main objectives of future phases would be to explore, demonstrate and document the analytic potential of these data. We would recommend this be approached using a series of small-scale case studies. Each of these would form individual demonstrator projects.

These could include (but are not limited to):

1) Exploring use of player account data in casinos and LBOs.

2) Further interrogating transactional level data and providing basic overviews of volume of play per sector.

3) Exploring within-session play and patterns of behaviour.

4) Examining how game data may, or may not, be linked to transactional records.

For each demonstrator project, the focus would be around further exploring small sub-sets of industry data and documenting the process and challenges of using these. We would anticipate that this would mean taking very small slices of data – potentially no longer than a week at atomic level – to better understand what these data tell us, what processes are needed to prepare the data across operators for analysis and, resources permitting, produce some small scale analytic output to demonstrate this. Because of the small scale and exploratory nature of this work, we do not anticipate that the analytical findings from these studies would be applicable for policy use. This programme of research should be viewed as a process, rather than event. These demonstrator projects would contribute valuable understanding to this process and would document the analytic potential of these data. This, in turn, would help to inform development of a coherent data strategy for machines in which the full potential of industry-held data could then be realised.

These 'demonstrator projects' could focus on the following areas:

Project option 1: Exploring use of player account data in casinos and LBOs
Given that player account data are only collected systematically in casinos and LBOs, this project would be limited to these venues. A first and important objective, in our opinion, would be to better quantify who uses these schemes and to assess what type of player they are. This would require comparison of key metrics from player account data with those from non-player accounts (either proxy sessional data or aggregate transactional data). This would allow us to assess how representative, or otherwise, this information is before undertaking further analysis.

Once established, samples of these data could be examined to explore within-session patterns of play. This could look at behavioural aspects relating to how and when people stake up, stake down, change product, walk away and so on. These data could also be examined to look at how play patterns vary across the day and across a week. Describing if and how machine play late at night varies from machine play during the day may yield important insight for policy questions surrounding issues of extended opening hours, for example. Building on this, it may be possible to start identifying some foundations for metrics which could help to identify sub-groups of machine players (i.e., long players, low stakers; sporadic/intermittent players etc). However, all of this demonstrator analysis would need to be contextualised to be clear that these are not representative data, and therefore

results are not intended for policy use at this stage. The focus is demonstrating how this type of analysis could be done to contribute to the evidence base.

Depending on the scale of the demonstrator project, it may be possible to merge some contextual variables onto the data. For example, venue density using Gambling Commission licensing records could be merged onto the data or other geo-demographic data which could be purchased for this purpose.

Before any of this is attempted, there is further scoping work required to map data from different sources against each other, to reconcile differences and to produce usable datasets. This is likely to be a significant undertaking, with a combined data dictionary needed that maps and tracks definitions from differing data sources.

Project option 2: Further interrogating transactional level data and providing basic overviews of volume of play per sector

The Machines Expert Group, the Responsible Gambling Strategy Board and the Gambling Commission have stated that very little is known about broad transactional level information on machines. For example, what is the shape of the stake distribution on B2 machines, how much play on B2 terminals is on B3 content, etc. A project work package could focus on identifying these key questions and working with industry to examine how easy they are to answer in a systematic and ongoing way. This would most likely focus on aggregate level data to give a better picture of the market and the value added from this process would be thinking through how to develop this more systematically so that changes over time could be tracked.

Project option 3: Exploring within-session play and patterns of behaviour

This has similarities to some aspects of project option 1, use of player data, but is expanded to include examination of proxy sessional data. Here it may be possible to examine to what extent proxy session data are accurate (using a small sample of data) by comparing with player tracked sessions. The inclusion of proxy session data means that behaviour from a greater number and type of players can be reviewed, although limitations of accuracy and linking sessions together needs to be considered. These data are really best used to explore micro sessional behaviour in more depth. Here the focus would be on attempting to describe what happens within a session and different patterns of play.

As with potential project 1, the same issues of further scoping and managing the data would need to be undertaken first and may represent a fairly significant time investment. We would advise that consideration be given to whether this investment is worthwhile compared with the potential learning gained.

Potential option 4: Examining how game data may, or may not, be linked to transactional records

This could include further scoping of the potential use of the game log file data to examine if and how they may be linked to transactional records. We would also recommend that this include further consultation with industry members, especially those interested in developing this area, to look at how this might be achieved. Further consideration of

technological advances in machine software and accounting systems should be made to examine whether this is likely to be more feasible in the future.

Of course, each potential case study would need to be fleshed out in more detail and careful consideration given to which project/projects are likely to provide the best value for money. Feedback from the Trust, the RGSB and MROP will be vital to ensure that any future research has practical utility for policy purposes.

Appendix A: Detailed interview schedule: initial meetings

Introduction

- Introduce team (both sides), purpose of research, what we hope to gain out of session
- Assure confidentiality (especially around recording)
- Background from operator:
    - Size
    - Number of venues
    - Types of venues (brand names)
    - Location of venue types – geographic distribution
    - Numbers of machines, cat B specific
    - Types of machines
    - Insight into demographics of customer base, (overview of play times/busy times potentially by venue types)
    - Market share

Data Types

- **Player data**
    - Whether any card based player tracking (if no, any plans to introduce?)
        - If so, mechanisms of how this works
        - How many people have them
        - Any estimate of the proportion who have them from player base
        - How often used
        - Circumstances in which they are not used (multi-machine play)
        - Can a person have more than one account? How monitored?
        - Is this brand specific – can it be linked across companies
        - What types of data recorded:
            - At sign up (demographics etc)
            - During play
            - How is this used internally?
            - What impediments for sharing? (i.e., data protection, small print at sign up)
    - What other types of loyalty card data available?
        - If so, mechanisms of how this works
        - How many people have them
        - Any estimate of the proportion who have them from player base
        - How often used?
        - Circumstances in which they are not used (multi-machine play)
        - Can a person have more than one account? How monitored?
        - Is this brand specific – can it be linked across companies
        - What types of data recorded:
            - At sign up (demographics etc)
            - During play
            - How is this used internally?

        - What impediments for sharing? (i.e., data protection, small print at sign up).
    - What else?
        - Any other types of data on players collected? What method, what does it tell us? How is it used?

- **Machine data**
    - What level of data does machine hold internally?
        - Does it record stake, prizes, bonus games, game type, use of auto transfer, credit transfer, money in/money out etc.
        - Are data linked to game choices and what happens within games?
            - If not, what is the possibility of linking this
            - Is this detail about game choice and game play even recorded?
        - At what level is this data recorded and stored? Polling data, i.e., every button press or aggregate data.
            - If aggregate, aggregated over what period or geography (i.e venue, region etc)?
            - If polling, can this be linked to what was happening in the game?
        - Does it time and date stamp data?
            - If so, could this be used to try to identify breaks in session?
        - What are likely limitations of using this type of data?

Data structure

- What sort of structures are the data held in?
    - Server based machines vs. non server based
    - What type of databases is used by operator?
    - How easy to interrogate and reformat
    - Does this vary for different companies, if so, in what ways?
- How far back does data go? (i.e., historical records for different types)
- How granular is the data – is it typically used/reconfigured at an aggregate level. What are these levels?
- Likely size of extracted data – data for 1 day, 1 week, 1 month?
- Where data are held, who is responsible for control of Data Warehouse?

Data Metrics

- Overview of data metrics wish list – confirmation of what we can and can't get.
- What are minimum requirements for dataset – what are the limitations/challenges in the view of this operator?
- What other metrics might be available?
    - Anything else that is readily available that we can use?
    - Contextual information, i.e., intelligence about timing of play/business models. Other contextual information from staff, how to link data together? What self-exclusion/responsible gambling strategies exist. How are they recorded?

Practicalities

- What resources do the operators have to be able to help us?
- Likely levels of support from Data Warehouse and analysts?
- Likely length of time between first data request and delivery of usable data? How to work together to improve this, make it quicker?
- What's the internal data language for this operator?
- What we need to do to set up confidentiality agreements
    - Timings for these?
    - Other issues/concerns

Internal use of data

- Uses of data internally
- Analytic techniques used (i.e., whether session times can be extrapolated from non-account data etc)
- General advice about use of data
- Get a formal position about publication and how we can use it?

Next steps

- Resources
- Working with data warehouses
- Further contact/meetings/queries

Data metrics wish list

Top - level

Loyalty card data - individual	**Machine level data - individual**
Main metrics to understand include: • Prevalence of use • Profile of users • Frequency of use • Incidence of uptake • How cards used • Multiple card use? • Multiple machine play? • Level of data collected • Whether card data can be linked to machine/game • Are card used cross operators/companies?	Main metrics to understand include: • How data collected? (aggregate for machines/venues; polling data of button presses etc) • Type of data collected • How long data held for • Whether date and time stamped • Whether sessions can be approximated

Loyalty card data - context	Machine level data - context
Main metrics to consider include: • Marketing initiatives • Comps • Patterns of play in venues at different times • Other events within the venue • New machines/games introduced • Non-category B play – how integrated • Interaction with staff • Responsible gaming, self-exclusion measurement • Potential follow-up	Main metrics to consider include: • Marketing initiatives • Comps • Patterns of play in venues at different times • Other events within the venue • New machines/games introduced • Interaction with staff • Responsible gaming, self-exclusion measurement • Potential follow-up

Specific metrics

Loyalty card	Machines
Specific metrics include per session: • Cash in • Cash out • Time and date stamps • Game choice • Game change • Within game features (i.e., bonuses, style of play) • Use of credit transfer/bank options • How dealt with money in the machine • Use of autoplay • Use of hold buttons • Stake – variability: monetary size, number of lines etc • Machine ID • Machine settings – i.e., speed, bonus features	Similar metrics but will become clearer once we know how data collected and stored. We may be able to get the following at an aggregate or machine level? • Cash in • Cash out • Time and date stamps • Game choice • Game change • Within game features (i.e., bonuses, style of play) • Use of credit transfer/bank options • How dealt with money in the machine • Use of autoplay • Use of hold buttons • Stake – variability: monetary size, number of lines etc • Machine ID • Machine settings – i.e., speed, bonus features

Loyalty card – context:	**Machines – context:**
Metrics include: • ID • Demographics • Length of membership • Frequency of use • Other play (other gambling activities, other machines – integration) • Venue type • Opening hours • Marketing (need to be date stamped) • Comps (need to be date stamped) • Self-exclusion • Customer service notes • Universal/cross operator ID?	Metrics include: (first few could come from other sources….) • Demographics • Length of membership • Frequency of use • Other play (other gambling activities, other machines – integration) • Venue type • Opening hours • Marketing (need to be date stamped) • Comps (need to be date stamped) • Self-exclusion • Customer service notes