Surveys in Games User Research

Chapter · March 2018

DOI: 10.1093/oso/9780198794844.003.0009

Florian Brühlmann
University of Basel

Elisa D. Mekler
ITU Copenhagen

Brühlmann, F., & Mekler, E. D. (2018). Surveys in Games User Research. In A. Drachen, P. Mirza-Babaei, & L. Nacke (Eds.), Games User Research (pp. 141–162). Oxford: Oxford University Press

```
@incollection{bruhlmann_surveys_2018,
        Author = {Br{\"u}hlmann, Florian and Mekler, Elisa D.},
        Booktitle = {Games {User} {Research}},
        Pages = {141--162},
        Publisher = {Oxford University Press},
        Address = {Oxford},
        Chapter = {9},
        Editor = {Anders Drachen and Pejman Mirza-Babaei and Lennart Nacke},
        Title = {Surveys in {Games} {User} {Research}},
        Year = {2018}}
```

Surveys in games user research

Florian Brühlmann, *University of Basel*

Elisa D. Mekler, *University of Basel*

Highlights

Surveys are an essential method of data collection that can deliver generalizable and actionable insights about the player's experience. In this chapter, we present practice-oriented guidance about when the method is appropriate, what constitutes a good questionnaire, and how to alleviate possible biases and issues with data quality.

1. Introduction

One of the main goals of Games User Research (GUR) is to evaluate the player-game interaction with the objective of using the results to improve the player experience. Surveys are a methodology that can help us achieve these goals. They are used to gather data from a subset of the population being studied. The results of the survey can then be generalized to the larger population. Surveys can capture players' opinions and self-reported gaming habits. They are a quick, easy, and cost-effective way to generate a sizeable amount of data to reveal more about the subjective experience of playing a game. This might make it seem relatively easy to put together a survey, but there are seemingly minor oversights that can severely limit the utility of your survey data. There are many different types of surveys, different ways to sample a population, and various ways to collect data from that population. Surveys can be administered via mail, telephone, or in person, as well as online. They have consistently been used in psychology, marketing, and HCI research to help answer a variety of questions related to people's attitudes, behaviours, and experiences (Müller, Sedley, and Ferrall-Nunge, 2014).

Drawing from our own work in player experience research, this chapter outlines the benefits and drawbacks of using surveys. In addition, this chapter provides a 'how-to' guide and best practices for practitioners and academics interested in applying surveys in GUR. We begin by outlining the types of research questions that are best addressed by surveys. We then review points that need to be considered when designing a survey, such as sampling and question types, followed by steps to be taken to ensure data quality.

2. What to research with surveys

The goal of a survey is to provide statistical estimates of the characteristics of the target population. Because surveying the whole population (e.g., all potential players of Dark Souls III) is usually impractical or even impossible, data from a sample of the population are collected instead. One central premise of survey research is that if you can provide a description of the sample, you can describe the target population (Fowler, 2013). The second premise states that the answers given by the participants in the sample accurately describe the characteristics of the respondents (Fowler, 2013). To retain these premises, researchers need to be aware of possible sources of error and bias and take measures to limit their influence. Only then can well-designed surveys provide insights into players' attitudes, experiences, motives, demographics, and psychographic characteristics.

Overall, there are two main categories of questions a researcher can ask respondents about: objective facts and subjective states. Objective facts are directly observable and can be verified by other people. For instance, objective facts can include demographic characteristics such as age and gender, as well as the number of hours spent playing games. In contrast, subjective states such as attitudes and emotions are not objectively verifiable. For instance, identification of a player with the protagonist of a game cannot be directly observed. Overall, surveys are very flexible when it comes to the research questions they can address: They can be used to assess characteristics of your player base, measure the differences between groups of players or different iterations of a design, and identify changes in players' attitudes and experiences over time. Surveys can be administered anytime during the game development cycle, depending on your research question.

Surveys in GUR can be useful to assess the following:

1. *Player attitudes and experiences.* Surveys can accurately measure and reliably represent attitudes and perceptions of a population. When assessing quantitative data, surveys provide statistically reliable metrics, which allow researchers to benchmark attitudes towards a game or an experience, to track changes in attitudes over time, and to tie self-reported attitudes to actual behaviour (e.g., via log data). Collecting qualitative data about player experience can also be used to understand players' interaction with a game or inform game design improvements.
2. *Motives.* Surveys can collect players' motives for playing a game at a specific time or in a specific situation. Unlike other methods, surveys can be deployed while a person is actually playing a game (e.g., an online intercept survey), minimizing the risk of imperfect recall on the players' part. It is important to note that the specific details and the context of one's intent may not be fully captured in a survey alone. For example, 'Why did you play Canabalt?' can be answered in a survey, but interviews may be more appropriate for determining a player's underlying motivations to engage with this game.
3. *Player characteristics.* Surveys can be used to understand a game's player base and to better serve their needs. Researchers can collect players' demographic information, their genre savyiness or overall gaming expertise, and psychographic variables such as personality traits. Such data enable researchers to discover segments of players who may have different needs, motivations, attitudes, preferences, and overall player experiences (e.g., Nacke, Bateman, and Mandryk, 2014).
4. *Comparisons.* Surveys can be used to compare players' attitudes, perceptions, and experiences. These comparisons can be made across player segments, time, competitor games, and between different iterations of game design aspects, such as interfaces. This allows researchers to explore whether or not players' needs and experiences vary across countries, assess a game's strengths and weaknesses among competitors, and evaluate potential game design improvements to make informed decisions.

Survey research is even more valuable when used in conjunction with other GUR methods. Players can be asked to self-report their behaviours, but gathering this information from log data (if available) will always be more accurate. This is particularly true when trying to understand precise behaviours, as players will struggle to recall their exact sequence of in-game actions. Combining surveys with more objective measures of player behaviour (see Chapters 16 and 19) helps to paint an even more detailed picture of the player experience. For instance, game analytics (see Chapter 19) might show that players are more likely to succeed in one version of the level, but the survey data may reveal that players were less challenged and experienced boredom (Hazan, 2013). Physiological measures (see Chapters 16 and 17) may be another complement to survey data. For example, player experience researchers have combined surveys of fun and immersion with measures of facial muscle and electrodermal activity (e.g., Nacke et al., 2010).

3. How to design a survey

As with all GUR methods, a survey is only as useful as its design. Before starting to write survey questions, researchers should first think about what they want to find out (Ambinder, 2014), what kind of data needs to be collected, and how the data will be used to meet their research goals. An overarching research goal may be to understand how challenging players find each portion of the game. Once research goals are defined, there are several other considerations to help determine whether a survey is the most appropriate method and how to proceed:

- Do the survey questions focus on the results, which directly address research goals and support informed decisions, or do they only provide informative data? An excess of questions increases the survey length and the likelihood that respondents will drop out before completing the questionnaire, diminishing the effectiveness of the survey results.
- Will the results be used for longitudinal comparisons or one-time decisions? For longitudinal comparisons, researchers must plan on multiple survey deployments without exhausting available respondents.
- What is the number of responses needed to provide the appropriate level of precision for the insights needed? Calculating the number of

responses needed (as described in the following section) will ensure

that key metrics and comparisons are statistically reliable. Once the target number is determined, researchers can then decide how many people to invite.

3.1. Determining the appropriate sample and sample size

The key to effective survey research is determining who and how many people should participate. The first thing that needs to be determined is the survey *population*—usually the target audience or the player base of a specific game or franchise. Depending on your research question, this may also encompass any set of individuals that meet certain predetermined criteria (e.g., novice players) and to whom you want your findings to apply to (e.g., what do novice players think of the new tutorial?). This is the population from which the *sample* for your survey should be drawn. Reaching everyone is typically impossible and unnecessary. However, if the sampling systematically excludes certain types of people (e.g., very disengaged players), the survey will suffer from *coverage error*, and its results will misrepresent the population.

While random sampling is the gold standard in scientific research, GUR is chiefly interested in capturing the attitudes and behaviours of relatively 'small' populations (i.e., the player base). Hence, it is acceptable to resort to *non-probability sampling* methods, such as volunteer opt-in panels, self-selected surveys (e.g., links on blogs, gaming forums, and social networks), and *convenience samples* (e.g., undergraduate psychology students). However, non-probability methods are prone to high selection bias and will reduce the representativeness of the results in comparison to random sampling methods.

It is important to carefully determine the target sample size for the survey. If the sample size is too small, findings from the survey cannot be accurately generalized to the entire player base or may fail to detect differences between groups. Therefore, calculating the optimal sample size becomes crucial for every survey.

There are many factors influencing the necessary sample size such as the frequency of a characteristic, the research question, effect size, the desired margin of error, and how accessible a certain demographic is. In general, a larger sample has less margin of error for an estimated characteristic of the population (e.g., enjoyment). However, it is important to remember that representativeness is related but not equal to sample size. This means it is not just small samples that are affected by sampling biases. Sampling bias can reduce the representativeness of even very large samples with small margins of error. There are various

formulas for calculating the target sample size (see Müller et al., 2014, for an extended discussion of sampling in HCI surveys), but there is no one-size-fits-all recommendation, as it depends on the specific situation.

As a rule of thumb, when a sample needs to be representative of a large population (e.g., the player base), we recommend a sample size of at least 500 survey respondents. Estimating a population parameter with this sample size would yield a margin of error of less than 5% at a 95% confidence level (Müller et al., 2014). Similarly, if you plan on conducting statistical significance testing (e.g., to compare two different game versions using an independent samples t-test), it is recommended to have at least 20–30 survey respondents per group to be able to properly assess the distribution of a statistic and have the necessary probability (statistical power) of detecting a true difference. However, estimates of sample size based on a priori power calculations (e.g., with the GPower program by Faul et al., 2007, which is available free of charge at http://www.gpower.hhu.de/) while taking the research context into account are preferable to rules of thumb.

3.2. Mode of survey invitation

There are four basic survey modes used to reach respondents: mail or written, phone, in-person, and online. These survey modes may be used independently or in combination with other modes. The survey mode needs to be chosen carefully, as each mode has its own advantages and disadvantages. For instance, surveys have different response rates, introduce distinct biases, require resources and costs, represent different scales of audience that can be reached, and offer respondents various levels of anonymity.

Today, many game research–related surveys are conducted online, as benefits often outweigh their disadvantages. Online surveys have the following major advantages:

- Easy access to large geographic regions (including international reach).
- Simplicity of creating a survey by leveraging easily accessible comer–cial tools.
- Relatively cost-effective to distribute (e.g., no paper and postage, simple implementation, insignificant cost increase for large sample sizes) and quickly analysed (returned data are already in electronic

format).

- Short fielding periods (i.e., time required to collect the answers), as the data are collected immediately.
- Lower bias due to respondent anonymity, as surveys are self-adminis- tered with no interviewer present.
- Ability to customize the questionnaire to specific respondent groups using skip logic (i.e., asking respondents a different set of questions based on the answer to a previous question).

3.3. Crowdsourcing

A promising mode for survey invitation are crowdsourcing platforms, such as CrowdFlower[1] or Amazon Mechanical Turk.[2] On these platforms, individuals, enterprises, and research institutions can post 'jobs', which may range from categorizing images, writing summaries of articles, to filling in surveys. The so-called crowd workers can then complete these jobs for a monetary payment set by the employer (Kittur et al., 2008). These platforms have the advantage of providing a large pool of readily available participants with a variety of backgrounds and interests. Crowdsourcing has been found to be reliable for behavioural research and user studies (Mason and Suri, 2012). Crowdsourced surveys are a low-cost avenue, even for small and indie game developers. People generally like to engage in crowdsourced micro-tasks, as fun and passing time are the main motivations for participation rather than monetary rewards (Antin and Shaw, 2012). This makes crowdsourcing platforms an ideal resource for participants when comparisons of interface variants or adjustments of difficulty settings are needed.

A few guidelines on what to keep in mind when distributing your survey over crowdsourcing platforms:

- Reduce the pool of participants to your target audience.
- Combine quantitative measures with qualitative data for a more detailed understanding of the collected data (e.g., have respondents explain in their own words why they rated the game as 'very easy').
- Include one or more control questions to reduce careless responses and identify fraudulent survey respondents. For example, include a control question such as 'Respond with "strongly agree" to this item' or ask for self-report of data quality (e.g., 'In your honest opinion, should we use your data?'; refer to Curran, 2016, and Meade and Craig, 2012, for an extensive review of careless response detection methods).

[1] www.crowdflower.com

[2] www.mturk.com

- Pay respondents only a small amount for completing the survey, but provide respondents who answered the questions carefully with a reward, and mention this in the job description.
- Examine data quality by combining objective measures such as completion time to identify outliers with reported technical difficulties.

Providing incentives is often effective for encouraging survey responses. As exemplified by crowdsourcing platforms, monetary incentives tend to increase response rates more than non-monetary incentives. If the crowdsourcing incentives are unavailable—for example, when social media is used as a means of recruitment—a lottery of monetary rewards or other prizes can be useful as an alternative.

4. The art of asking questions—questionnaire design

Once the research questions and the appropriate sampling methods are established, researchers can begin designing the survey questionnaire. Questionnaires allow researchers to gain information in an objective, reliable, and valid way:

Objectivity means that respondents' answers should not be influenced by who is conducting the survey. This is less problematic in online surveys but can have an influence in lab studies. For example, a respondent may answer survey questions overly positive so as to appeal to the researcher.

Reliability refers to the accuracy and consistency of the questionnaire. For example, a respondent's answer to a question about gaming habits should not drastically change from one day to the next. Reliability is more difficult to achieve when internal processes or attitudes are the focus of the study. While reliability is a concern for researchers that develop their own surveys, it is of special importance when measuring psychological phenomena (see deVellis, 2012, for an in-depth discussion).

Validity refers to the extent that the collected data represent the phenomenon of interest. A measure is valid if it measures what it is supposed to and not some other factors. Several aspects of validity are important for survey research:

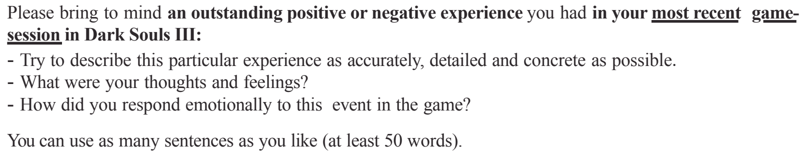
1. *Content validity* refers to the degree to which a survey or a question in a survey captures the phenomenon of interest in a representative manner. This means that all aspects of the phenomenon are somehow represented in the questionnaire. For example, asking participants how often they played a game in the last two weeks might not give the full picture of their behaviour, since the question does not capture how long these game sessions were.

2. *Criterion-related validity* refers to the association of a survey or a question in a survey with another characteristic or behaviour of the respondent. For instance, players who enjoy a particular game may be more likely to recommend the game to their friends. In this case, enjoyment can predict the likeliness of recommendation, and thus features criterion-related or, in another term, predictive validity.
3. *Construct validity* concerns the theoretical foundation of a facet measured by a questionnaire. A questionnaire shows construct validity if the answer behaviour of the respondent can be linked to the phenomenon of interest. For example, this means that a questionnaire for presence in games actually measures feelings of presence rather than feelings of enjoyment. Establishing construct validity is a complex process and beyond the scope of this chapter (but please refer to deVellis, 2012, for an in-depth discussion). Nevertheless, researchers should take content, criterion-related, and construct validity into account, and think about how participants' answers to the survey questions relate to their real-life behaviour. This can help with refining and selecting survey questions as well as identifying possible problems with question wording.

For most surveys there is only one opportunity for deployment, with no possibility for requesting further clarification or probing (cf. interviews, see Chapter 10). It is therefore important for researchers to carefully think through the design of each survey question, as it is fairly easy to introduce biases, which can have a substantial impact on the reliability and validity of the data collected. Survey questions should minimize the risks of measurement error that can arise from the respondent (e.g., lack of motivation, comprehension problems, deliberate distortion) or from the questionnaire (e.g., poor wording or design, technical flaws).

4.1. Types of survey questions

There are two main types of survey questions, open- and closed-ended. Open-ended questions (e.g., ‘Please describe a recent outstanding experience with a digital game’; see also Figure 9.1) require the respondents to produce their own answer in a text-based format, a video, or a mind map (e.g., Hillman et al., 2016). Closed-ended questions provide a set of predefined answers to choose from (e.g., multiple-choice or rating questions; see also Figure 9.2). The question type format that is the most appropriate for the research study is dependent on the research question. In general, open-ended questions make sense when not much is known about the phenomenon of interest and for discovering more about a topic in exploratory studies. In one of our own surveys, for instance, the use of open-ended questions allowed us to identify several game design aspects related to players’ emotional experience (Bopp et al., 2016). Open-ended questions are also valuable for studies where the research question is clear, and validated scales exist. These types of questions can help researchers understand why respondents answered questions in a certain way. For instance, it can make sense to ask participants to explain why they rated the aspect ‘control’ so low for a specific game.



Please bring to mind **an outstanding positive or negative experience** you had **in your <u>most recent</u> <u>game-session</u> in Dark Souls III:**
- Try to describe this particular experience as accurately, detailed and concrete as possible.
- What were your thoughts and feelings?
- How did you respond emotionally to this event in the game?

You can use as many sentences as you like (at least 50 words).

Figure 9.1 Example of an open-ended question (used for the study described in Petralito et al., 2017)

How long ago did the experience take place?

Figure 9.2 Example of a single-choice question (used for the study described in Bopp et al., 2016)

Open-ended questions are appropriate when:

- The range of possible answers is unknown (e.g., ‘What is your favourite game?').

- Measuring quantities with natural metrics (i.e., constructs with an inherent unit of measurement, such as age, length, or frequency). For instance, when researchers are unable to access information from log data, such as time, frequency, and length, they might ask open-ended questions to acquire this information (e.g., 'How many times do you access Steam in a typical week?'; using a text field that is restricted to numeric input).
- Measuring qualitative aspects of players' experience (e.g., 'What do you like best about this game?').

It may be tempting to use large text field entries throughout your survey, but it is important to note that these text fields may intimidate respondents and cause higher dropout rates. In one of our own surveys, we experienced a substantial dropout on the page that featured two large text fields. Nevertheless, the data quality was excellent for participants who persevered until the end of the survey (Bopp et al., 2016).

Closed-ended questions are appropriate when:

- The range of possible answers is known and small enough to be easily provided (e.g., 'Which devices do you use for gaming?'; the answers provided could include 'PC' and 'consoles').
- Rating a single object on a 7-point scale from 'Not fun at all' to 'Extremely fun' (e.g., 'Overall, how fun was the game?').
- Measuring quantities without natural metrics, such as importance. For example, 'How important is it to have your smartphone within reach 24 hour a day?' (on a 5-point scale from 'Not at all important' to 'Extremely important').

There are four basic types of closed-ended questions: single-choice, multiple-choice, rating, and ranking questions.

1. *Single-choice questions* work best when only one answer is possible for each respondent.
2. *Multiple-choice questions* (Figure 9.3) are appropriate when more than one answer may apply to the respondent. Frequently, multiple-choice questions are accompanied by 'select all that apply'.
3. *Ranking questions* (Figure 9.4) are best when respondents must prioritize their choices or express their preferences.
4. *Rating questions* (Figure 9.5) are appropriate when the respondent must

judge an object on a continuum on either a unipolar or a bipolar scale.

Rating questions should include midpoints to avoid having respondents, who actually feel neutral, end up making a random choice on either side of the scale. Also, include (optional) open-ended questions to encourage respondents

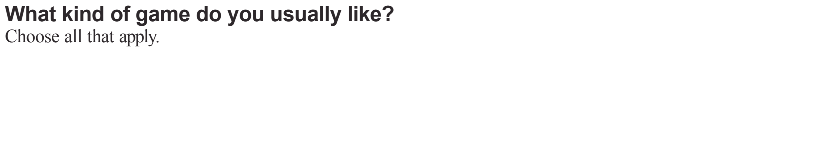

What kind of game do you usually like?
Choose all that apply.

Figure 9.3 Example of a multiple-choice question (used for the study in Bopp et al., 2016)

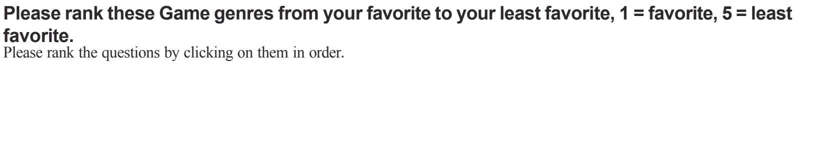

Please rank these Game genres from your favorite to your least favorite, 1 = favorite, 5 = least favorite.
Please rank the questions by clicking on them in order.

Figure 9.4 Example of a ranking question

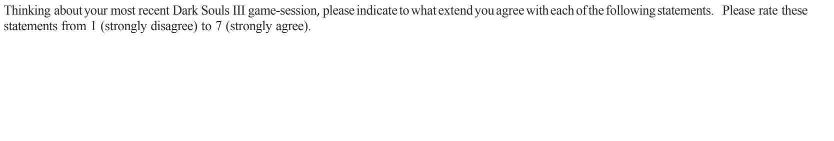

Thinking about your most recent Dark Souls III game-session, please indicate to what extend you agree with each of the following statements. Please rate these statements from 1 (strongly disagree) to 7 (strongly agree).

Figure 9.5 Example of a Likert-type scale rating question. Note the midpoint (4) on a scale ranging from 1 to 7

to comment on any points of confusion during the survey or to clarify their responses. These questions can be added at the end of each page or at the end of the entire survey.

4.2. Questionnaire biases and other pitfalls

After writing the first survey draft, it is crucial to check the phrasing of each question for potential biases. Consider the following:

- Avoid complex, difficult to understand questions.
- Avoid answer options such as 'no opinion', 'do not know', 'not applicable', or 'unsure', since respondents with actual opinions will be tempted to select this option to avoid spending time on thinking about their

opinion. A good survey question should be answerable by all respondents. Moreover, the analysis and interpretation of these 'no opinion' answers is often not straightforward.

- Instead of giving options that allow participants to opt out of responding, include an optional text entry field for respondents to justify and explain their answer at the bottom of each page.
- Avoid using the same rating scale in a series of back-to-back questions (e.g., 'How fun was X?', 'How fun was Y?').
- Be wary of cramming many questions into a single survey, as respondents will become bored or fatigued. This will increase your dropout rate, or worse, participants will rush through the questionnaire without paying attention to their answers. Both of these outcomes limit your data's usefulness and should be avoided.
- Social desirability occurs when respondents (e.g., fans of a franchise) answer questions in a manner they feel will be positively perceived by others (e.g., the game designers and developers). For example, respondents might give a more positive evaluation for popular publishers/studios or under-report unfavourable opinions (Ambinder, 2014). To mitigate these effects, respondents should be able to answer the survey anonymously. In addition, the survey should either be conducted by a third party or the survey should be self-administered.
- Halo and placebo effects by mentioning (allegedly) new game features can influence survey responses (Denisova and Cairns, 2015). Consider disclosing information about specific game features only sparingly.
- *Broad questions* lack focus and can be interpreted in different ways. For example, 'Describe the way you play on your smartphone' is too broad, as there are different games, motives, and situations for mobile gaming. These questions are too general and usually yield few actionable insights. A more focused set of questions for the example above include 'Which games did you play on your smartphone over the last week?' and 'Describe the locations in which you played on your smartphone last week.'
- *Leading questions* may introduce unwanted biases in the data by manipulating respondents into giving a certain answer. For example, 'This game has a Metacritic score of 93. How much fun did you experience with the game?'. The same holds true for questions that ask the respondent to agree or disagree with a given statement; for example, 'Do you agree

or disagree with the following statement: I use my smartphone

more often than my tablet for gaming.' Questions should be asked in a neutral way without examples or additional information that may bias respondents towards a particular response.

- *Double-barrelled questions* ask about multiple items while only allowing for a single response, decreasing the reliability and validity of the data. These questions can usually be detected by the existence of the word 'and'. For example, when asked 'How fun is it to play on your smartphone and your tablet?', a respondent with differing attitudes towards the two devices will be forced to pick an attitude that either reflects just one device or the average across both devices. Questions with multiple items should be broken down into one question per construct or item.
- *Prediction or hypothetical questions* ask survey respondents to anticipate or imagine future behaviours or attitudes in a given situation. Examples include 'Over the next month, how frequently will you access the PlayStation store?'; 'Which of the following features would make you have more fun with this game?'. Even if respondents have clear answers to these questions, their response may not predict actual future behaviours or experiences.

5. Established questionnaires in GUR

An alternative to constructing a new questionnaire is to employ a well-established questionnaire. Ideally, these questionnaires have been previously validated, which allows researchers to compare the results to other studies that have used the questionnaire. An existing questionnaire can be adapted to the specific study context as needed; however, this reduces the comparability between different studies. GUR is a relatively new field in comparison to other disciplines; therefore some questionnaires in GUR have not been extensively validated, and should be employed with caution (Brühlmann and Schmid, 2015). Some of the most commonly used GUR-related questionnaires are the following:

- Game Experience Questionnaire (GEQ). The GEQ by IJsselsteijn and colleagues (IJsselsteijn et al., 2008) incorporates seven different dimensions of player experience: sensory and imaginative immersion, tension, competence, flow, negative effect, positive effect, and challenge.

The GEQ is a self-report measure for a rather multifaceted investigation of game experience and is yet to be validated.

- Player Experience of Need Satisfaction Scale (PENS). The PENS is a proprietary questionnaire investigating the 'motivational pull' of video games (Ryan et al., 2006). It is based on self-determination theory and focuses on the three basic human needs: autonomy (volitional aspects of an activity), competence (experience of control and mastery), and relatedness (connection to others).
- Immersive Experience Questionnaire (IEQ). The IEQ was developed by Jennett et al. (2008) and measures the *player-related* factors cognitive involvement, real-world dissociation, and emotional involvement, as well as the *game-related* factors challenge and control.
- Positive and Negative Affect Schedule (PANAS). The PANAS is widely used as a measure for *strong* positive and negative affective reactions in a variety of contexts and was developed by Watson et al. (1988).
- Self-Assessment Manikin (SAM). The SAM measures the dimensions pleasure, arousal, and dominance of affective reactions nonverbally. It is composed of three rows of pictograms assessing the dimensions on a 5-point scale (Bradley and Lang, 1994).
- Intrinsic Motivation Inventory (IMI). The IMI is a multidimensional measurement device developed to measure different aspects of an experience (Ryan, 1982). It includes a subscale for interest/enjoyment that was originally intended as a measure for intrinsic motivation, but is also widely used as a measure of enjoyment in GUR (Mekler et al., 2014).
- Player Experience Inventory (PXI). Similar to the GEQ, the PXI aims to measure the player experience broadly and is conceptually linked to the MDA framework (Hunicke et al., 2004). The PXI measures functional consequences (dynamics) and psychosocial consequences (aesthetics) with various subscales as well as overall enjoyment from playing the game. The measure is still in development and soon to be validated (Vanden Abeele et al., 2016).

6. Conducting surveys

6.1. Survey implementation

There are several online survey tools available for implementing online surveys. Most of them can be used for free (with limited features), such as Google Forms, LimeSurvey, Questback, SurveyGizmo, and SurveyMonkey. When deciding on the appropriate survey platform, the functionality, cost, and ease of use should be taken into consideration. Depending on your study, the questionnaire may require a survey tool that supports functionality such as branching and conditionals, the ability to pass URL parameters, multiple languages, and a range of question types. Of course, there is always the option of preparing your online survey in-house and hosting it on your own servers.

In addition, the survey's visual design should be taken into account, since specific choices may unintentionally bias respondents. For example, progress bars can be misleading and intimidating in long surveys, resulting in increased dropout rates. In short surveys, progress bars are likely to increase completion rates, since substantial progress is shown between pages.

When launching a survey, check for common dropout points and long completion times, examine data quality checks, and review answers to open-ended questions. High dropout rates and completion times may point to flaws in the survey design, while unusual answers may suggest a disconnect between a question's intention and respondents' interpretation. Other survey data worth monitoring include the devices from which the survey was accessed and how many respondents dropped out on each page. It is important to monitor such metrics, so that improvements can be quickly applied before the entire sample has responded to the survey.

6.2. Data quality checks

Data quality checks have become a staple in empirical research, as they ensure a certain quality of respondents' answering behaviour and filter out responses that do not meet these standards. These qualities include attentiveness, honesty, and carefulness. Attentiveness refers to respondents reading all instructions and questions, without skipping over possibly important parts or missing a word. As there is no investigator present during online surveys, it is impossible to know whether or not respondents actually read the questions before answering without resorting to *attention-check questions*. These typically consist of bogus items, hidden within the main survey questions, which have only one correct answer. Example items include 'Yesterday while watching TV I had a fatal heart attack' and 'I read instructions carefully. To show that you are reading these instructions, please leave this question blank.' However, note that many survey respondents, particularly on crowdsourcing platforms, have learned to easily spot and circumvent these questions. Hence, we additionally suggest implementing a *seriousness check* at the end of the survey, which consists of a

(Aust et al., 2013). For example: ‘It would be very helpful if you could tell us at this point whether you have answered all questions seriously, so that we can use your answers for our analysis, or whether you were just clicking through to take a look at the survey.’ Respondents were able to choose one of two following answers: ‘I have answered all questions seriously’ or ‘I just clicked through the survey, please throw my data away.’

7. Data clean-up and next steps

7.1. Preparing and exploring the data

When exploring the collected survey data, you should always look for signs of low-quality responses. Low-quality survey data can either be left as is, removed, or presented separately from trusted data. If the researcher decides to remove poor data, there are three options: (1) Remove individual respondents’ data when of poor quality (i.e., listwise deletion), for instance, when they are identified as speeders or straight-liners (see following). (2) Remove individual questions or variables if the responses are of consistently poor quality (i.e., pairwise deletion), for instance, because respondents did not fully understand the question. (3) Exclude data beyond a certain point in the survey where respondents’ data quality has declined. The following are signals to look out for at the survey response level:

- *Duplicate responses.* Respondents might be able to fill out the survey more than once. Respondent information such as name, email address, or any other unique identifiers should be used to find and remove duplicate responses.
- *Speeders.* Respondents that complete the survey faster than what is expected under normal circumstances. Speeders may have carelessly read and answered the questions, resulting in arbitrary responses. Even if attention-check questions were implemented, examine the distribution of response times and remove any respondents that were suspiciously fast.
- *Straight-liners*. Respondents that always, or almost always, pick the same

answer option across survey questions are referred to as straight-liners. Grid-style questions (see Figure 9.5) are particularly prone to respondent straight-lining. Straight-liners tend to pick the first answer option when asked to rate a series of items or alternate between the first and

through the entire survey, consider removing the respondent's data entirely. If a respondent starts straight-lining at a certain point, consider keeping the data up until that point.

- *Missing data and dropouts*. Some respondents may finish a survey but skip several questions. Others may start the survey but break off at some point. Both result in missing data.
- *Low inter-item reliability*. When multiple questions are used to measure a single construct, respondents' answers should be consistent across this set of questions. Respondents that give inconsistent or unreliable responses (e.g., selecting 'very fast' and 'very slow' for separate questions assessing the construct of speed) may not have carefully read the questions and should be considered for removal.
- *Outliers*. Answers that significantly deviate from the majority of responses are considered outliers and should be examined. For questions with numeric values, we typically calculate outliers as anything outside of two or three standard deviations from the mean. Determine how much of a difference keeping or removing the outliers has on variables' averages. If the impact is significant, the researcher may either remove such responses entirely or replace them with a value that equals two or three standard deviations from the mean. Another way to describe the central tendency while minimizing the effect of outliers is to use the median rather than the mean.
- *Inadequate open-ended responses*. Open-ended questions may lead to low-quality responses due to the amount of effort required to answer these questions. Remove obvious nonsense answers, such as 'asdf'. After this, examine all of the other answers from the same respondent to determine whether all their survey responses warrant removal.

7.2. Data analysis and reporting

Data analysis and interpretation should be as objective as possible. Games user researchers usually have the benefit of not being as emotionally invested in the development of the game. This allows GUR to play the role of the player's advocate, who might have very different views than the developers. Nevertheless, it is beneficial to think about whether a different researcher would have come to the same conclusions.

To analyse closed-ended responses, *descriptive* and *inferential statistics* may be employed. Descriptive statistics describe the existing data set and help identify emerging patterns. They include measures such as the frequency distribution, central tendency (e.g., mean or median), and data dispersion (e.g., standard deviation). Inferential statistics can be used to draw inferences from the sample (your survey respondents) to the overall population (e.g., your player base). A comprehensive overview of statistical analysis methods is beyond the scope of this chapter, but we wholeheartedly recommend Andy Field's *Discovering Statistics* series on the topic. There are several packages available to assist with survey analysis: Microsoft Excel and certain survey platforms, such as SurveyMonkey and Google Forms, allow for descriptive statistics and charts. More advanced software such as SPSS, R, SAS, and Matlab can be used for complex modelling, calculations, and charting.

Analysing open-ended responses contributes to a more holistic understanding of the phenomenon being studied, as it reveals important insights that cannot otherwise be extracted from closed-ended responses. To do so, the qualitative data are analysed by applying a coding scheme established with regard to the objective of that survey question. We recommend Johnny Saldaña's (2009) book for a comprehensive overview and description of different coding approaches. After analysing all respondents' comments, researchers may begin to summarize the key themes of the data. These themes can be exemplified with representative quotes.

Once the question-by-question analysis is completed, findings need to be synthesized across all questions to address the goals of the survey. These findings will help identify larger themes and answer the initially defined research questions. Finally, these findings are translated into recommendations and design implications as appropriate.

Acknowledgements

Special thanks to Lena Aeschbach for assisting with the literature review and providing feedback on earlier drafts of this chapter.

References

Ambinder, M. (2014). Making the best of imperfect data: reflections on an ideal world. In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (pp. 469–469). New York: ACM.

Antin, J., Shaw, A. (2012). Social desirability bias and self-reports of motiva- tion: a study of Amazon Mechanical Turk in the US and India. In Proceed- ings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2925–2934). New York: ACM.

Aust, F., Diedenhofen, B., Ullrich, S., Musch, J., 2013. Seriousness checks are useful to improve data validity in online research. Behavior Research Meth- ods, 45(2), 527–535.

Bopp, J. A., Mekler, E. D., Opwis, K. (2016). Negative emotion, positive expe- rience? Emotionally moving moments in digital games. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2996–3006). New York: ACM.

Bradley, M. M., Lang, P. J. (1994). Measuring emotion: the self-assessment man- ikin and the semantic differential. Journal of Behavior Therapy and Experi- mental Psychiatry, 25(1), 49–59.

Brühlmann, F., Schmid, G. M. (2015). How to measure the game experience? Analysis of the factor structure of two questionnaires. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (pp. 1181–1186). New York: ACM.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology, 66, 4–19.

DeVellis, R. F. (2012). Scale development: theory and applications. Thousand Oaks, CA: Sage Publications.

Denisova, A., Cairns, P. (2015). The placebo effect in digital games: phantom perception of adaptive artificial intelligence. In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (pp. 23–33). New York: ACM.

Faul, F., Erdfelder, E., Lang, A. G., Buchner, A. (2007). G* Power 3: a flexible sta- tistical power analysis program for the social, behavioural, and biomedical sciences. Behavior Research Methods, 39(2), 175–191.

Fowler Jr, F. J. (2013). Survey research methods. London: Sage Publications. ISBN: 978-1-4522-5900-0.

Fowler Jr, F. J. (2014). Survey research methods. Thousand Oaks, CA: Sage Publications.

Hazan, E. (2013). Contextualizing data. In M. S. El-Nasr et al. (eds.), Game analytics (pp. 477–496). London: Springer.

Hillman, S., Stach, T., Procyk, J., Zammitto, V. (2016). Diary methods in AAA games user research. In Proceedings of the 2016 CHI Conference extended abstracts on Human Factors in Computing Systems (pp. 1879–1885). New York: ACM.

Hunicke, R., LeBlanc, M., Zubek, R. (2004, July). MDA: a formal approach to game design and game research. Proceedings of the AAAI Workshop on Challenges in Game AI, 4(1).

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A. (2008). Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies, 66(9), 641–661.

Kittur, A., Chi, E. H., Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 453–456). New York: ACM.

Mason, W., Suri, S. (2012). Conducting behavioural research on Amazon's Mechanical Turk. Behavior Research Methods, 44(1), 1–23.

Meade, A. W., Craig, S. B. (2012). Identifying careless responses in survey data. Psychological Methods, 17(3), 437–455.

Mekler E. D. , Bopp J. A., Tuch A. N., and Opwis K. (2014). A systematic review of quantitative studies on the enjoyment of digital entertainment games. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14) (pp. 927–936). New York: ACM.

Nacke, L. E., Bateman, C. Mandryk, R. L. (2014). BrainHex: a neurobiological gamer typology survey. Entertainment Computing, 5(1), 55–62.

Nacke, L. E., Grimshaw, M. N. Lindley, C. A. (2010). More than a feeling: measurement of sonic user experience and psychophysiology in a first-person shooter game. Interacting with Computers, 22(5), 336–343.

Petralito, S., Brühlmann, F., Iten, G., Mekler, E. D., Opwis, K. (2017). A good reason to die: how avatar death and high challenges enable positive experiences. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 5087–5097). New York: ACM.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. Journal of Personality and Social Psychology, 43, 450–461.

Ryan, R. M., Rigby, C., Przybylski, A. (2006). The motivational pull of video games: a self-determination theory approach. Motivation and Emotion, 30(4), 344–360.

IJsselsteijn, W., van den Hoogen, W., Klimmt, C., de Kort, Y., Lindley, C., Mathiak, K., …, Vorderer, P. (2008). Measuring the experience of digital game enjoyment. In Proceedings of Measuring Behavior (pp. 88–89). Maastricht, The Netherlands.

Vanden Abeele, V., Nacke, L. E., Mekler, E. D., Johnson, D. (2016). Design and preliminary validation of The Player Experience Inventory. In Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (pp. 335–341). New York: ACM.

Watson, D., Clark, L. A., Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology, 54(6), 1063.

10. Further reading

For an excellent, more in-depth overview of survey research in HCI, make sure to check out this paper, which also served as a template for this chapter:

Müller, H. Sedley, A., Ferrall-Nunge, E. (2014). Survey research in HCI. In J. S. Olson and W. A. Kellogg (eds.), Ways of knowing in HCI (pp. 229–266). Springer.

For all your statistics needs, we recommend any of the Andy Field 'Discovering Statistics' books, for instance:

Field, A., Miles, J. Field, Z. (2012). Discovering statistics using R. Thousand Oaks, CA: Sage Publications.

For a comprehensive overview of qualitative analysis and description of different coding approaches, we recommend:

Saldaña, J. (2009). The coding manual for qualitative researchers. Thousand Oaks, CA: Sage Publications.