Chapter 5
Measuring Experiences

Paul Cairns and Christopher Power

Abstract The science of HCI in the third wave is intended to understand user experiences through the filter of the values and contexts of individuals using systems and moreover as filtered through the values and contexts of individual researchers. This is not to neglect the importance of measurement to science and the challenges of measuring user experience (UX). This chapter will discuss how HCI can draw on the methods of modern psychometrics to provide tools for measuring user experiences. In particular, we will introduce bifactor analysis as a way to examine both the conceptual coherence of a questionnaire for measuring UX and also the distinct influences of different facets of the core concept. Further, through looking at modern methods of analysis, in particular treatment of outliers, we also consider how modern statistics are not to be treated as black boxes but require researchers to think more deeply about the people behind the data. Drawing on our work in player experiences, we make the case that psychometrics used well as a tool in UX has an important role to play in HCI as a successor science.

5.1 Introduction

In its early days and its first wave (Harrison et al. 2007), HCI was concerned with engineering systems to make people working with machines more effective (Long and Dowell 1989). Typically, more effective meant people (not machines) were faster and made fewer mistakes. This engineering conceptualisation of HCI relied on measurement as key: to engineer a good system it was necessary to measure the performance outcome of interest and then refine the system to improve the measure (Dowell and Long 1998). However, as HCI, and indeed the world, progressed from computers as limited workplace tools to widespread, everyday devices, so the emphasis in HCI moved from engineering systems to developing richer

P. Cairns (✉) · C. Power
University of York, York, UK
e-mail: paul.cairns@york.ac.uk

M. Filimowicz, V. Tzankova (eds.), *New Directions in Third Wave Human-Computer Interaction: Volume 2 - Methodologies*, Human–Computer Interaction Series,
https://doi.org/10.1007/978-3-319-73374-6_5

understandings of people's relationship with digital technology and the interventions that might lead to new possibilities. This has been characterised as the third paradigm (Harrison et al. 2011) or the third wave (Bødker 2006).

In the third wave, the emphasis has moved away from a focus on the individual interacting with a system to a more holistic view of interactions. People interact with technology within a context of the physical space, their social situation, their goals and more importantly the values and meaning of the interactions. Technology is no longer just about getting things done but it is a tool to enable people to have meaningful experiences to the point where the focus on technology may be secondary (Baumer and Silberman 2011). The difference in emphasis is the difference between the best way for people to input a text message (Cox et al. 2008) to the way in which texts bring people together and enables social and political outcomes (Vieweg et al. 2010).

For HCI as a discipline to make progress in this third wave, it can be envisioned as a successor science (Harrison et al. 2011). Successor science is a term growing from feminist philosophy which has identified that scientific practice and hence the resulting science embodies gender, class and racial biases. A significant example of this is well discussed and analysed in Stephen Jay Gould's classic book The Mismeasure of Man (Gould 1996) on the way in which measures of intelligence, in particular IQ, have promoted the intellectual superiority of white men on the back of weak, misleading or wrong scientific evidence. Successor science instead sees science as epistemologically situated in society, culture and history. This does not invalidate the scientific knowledge produced but it requires that researchers are not blind to the inherent bias in their methods and that a fruitful line of research is to seek out how the knowledge found might differ from taking a different epistemological stance.

Even while it is acknowledged that it is not possible to engineer experiences (Wright et al. 2003: 52), the experiences that people have still form a valuable focus for science. In digital games, for example, game developers look to bring about both short and long-term engagement with games (Cairns 2016) and do so through designing for a range of intended experiences such as challenge (Denisova et al. 2017), fun (Lazzaro 2009), flow (Chen 2007), social presence (Hudson and Cairns 2014) and so on. But what exactly are these experiences? How does the design of games influence them? What else in the context of players and their playing influences these experiences? And how do these different experiences influence each other? Being able to define and measure experiences allows us to begin to answer such questions. Or rather, measuring player experiences at least allows research to isolate the potential phenomena of player experiences (Hacking 1983) in ideal conditions ("in the lab") before looking for the richer experiences found "in the wild."

It should also be noted that the third wave is not intended to supplant the focus on individual interactions. There is still a specific focus on Interaction within HCI, such as gestural, wearable and tangible interactions (Reeves 2015; Kuutti and Bannon 2014) and therefore a need to quantify aspects of these interactions, including the experiences they offer. However, at the heart of measuring experience there

seems to be contradiction in terms. Wright and McCarthy (2004) see experiences with technology, amongst other things, to be:

- Holistic: experiences can only be understood as happening to the whole person
- Situated: experiences arise pre-linguistically out of engagement with a specific situation
- Singular: highly specific to the person having the experience
- Becoming: experiences make new the world from which they arise and so are a process of redefining themselves.

By contrast, measurement is inherently intended to be:

- Reductionist: dividing a whole into parts which can be separately understood
- Abstract: a measurement means the same thing separate from both the context of measurement and the instrument used
- Averaged: individual measurements of experience are not as important as the aggregation of several measurements
- Definite: a measurement is fixed in both value and meaning at the time of measurement

How then is it possible to claim to measure experiences? Here we do not claim to remove or ignore this contradiction. Instead, we aim to show that by being explicit about the limitations of any measurement it is still possible to do science but it very much has to be a successor science where the epistemological stance of any findings are always open to negotiation. This does not prevent HCI from making progress in at least some aspects of knowledge but moreover forces us to acknowledge and even seek out the limitations of what we learn.

The most common approach to measuring experiences used in HCI is questionnaires. As a discipline, HCI has recognised the implicit and subjective nature of user experiences and drawn on the work in psychometrics, particularly the methods of questionnaire development, to produce instruments specific to measuring experiences of interactions. Such questionnaires cover a wide range of general facets of user experiences including engagement (O'Brien and Toms 2010), aesthetics (Hassenzahl 2004) and spatial presence (Witmer and Singer 1998) as well as ones specific to particular contexts such as digital games (Jennett et al. 2008) or mobile interfaces (Ryu and Smith-Jackson 2006).

Any sort of measurement, including with questionnaires, necessarily operationalises the concepts to be measured with the risk that they become identified as the concepts themselves. That is, there is a risk of false positivism that the only meaningful experience of, for example, spatial presence is that defined by Witmer and Singer (1998) in their questionnaire. This flies in the face of the situated, personal and emergent nature of experiences proposed by third wave HCI. If HCI is to function as a successor science then it must acknowledge the epistemological biases inherent in any form of data gathering and therefore the inherent limitations to any questionnaire. A questionnaire used to measure experience is epistemologically situated in both the context of use of the questionnaire and the processes which generated the questionnaire in the first place.

Another risk of quantifying user experiences, particularly in experiments, is that there is then the move to consider only average behaviour as captured in statistical averages of the measurements. This neglects the variations that constitute the experiences of individuals.

In this chapter, we describe how psychometric methods can be employed in HCI yet still maintain a view on the contingent and situated knowledge these methods generate. We also discuss new methods of statistical analysis that bring a richer interpretation of questionnaires. Specifically, bifactor analysis (Reise 2012) considers both the unifying concept of a questionnaire and where there are nuances and deviations from the unified concept. These methods and the challenges of third wave HCI to these methods are discussed with reference to our own work in the development of questionnaires for measuring player experiences and in particular with reference to our recent development of a questionnaire to measure the feelings of uncertainty people have when playing digital games.

Furthermore, when it comes to analysing data from questionnaires, modern statistical methods force explicit consideration of the assumptions underpinning tests and how concern for the underlying distributions leads to examination of possible features in the data such as bimodality and outliers. Whereas traditional statistical methods might consider these features as problems to be avoided (or worse, ignored), modern methods view them as requiring further investigation and understanding. As such, we make the case that modern statistical methods for psychometrics are appropriate to a vision of a successor science suitable for the third wave of HCI.

5.2 Questionnaires for User Experience

The goal of using psychometric methods in user experience is to develop a questionnaire that participants complete and can be used to assign a value, a number, to the level of the experience had by participants. Each item of the questionnaire is typically a Likert item (Likert 1932), that is, a statement to which respondents are required to rate their level of agreement from Strongly Disagree to Strongly Agree. Such Likert items typically have 5 response options, (though sometimes 7) and these are simply scored from 1 to 5. Where necessary these scores are sometimes reversed to take account of the direction of the statement, for example, “I did not understand the game mechanics” is scored in reverse from “I understood the game mechanics.” These item scores are then summed or more often averaged either across the whole questionnaire or across subscales from the questionnaire depending on the questionnaire structure. For example, the Game Engagement Questionnaire (GEQ) (Brockmyer et al. 2009) is a single scale and a measure of engagement is obtained by averaging across all of the items in the questionnaire. By contrast, the social presence in gaming questionnaire, the CCPIG (sea-pig) (Hudson and Cairns 2014), has two separate subscales, one for measuring social presence between players on opposing teams and another for measuring social presence between players

on the same team. It would not make sense to have a single notion of social presence across these two contexts and so the subscales are scored separately.

In order to develop a questionnaire, the first stage is to generate items for the potential questionnaire and to iteratively refine the items. The second stage is to statistically validate the proposed set of items to see if they have coherence and also to identify their structure in terms of subscales.

The basic steps are therefore:

1. Define the concept to be measured
2. Generate and refine an item pool
3. Trial the items with target participants
4. Administer the questionnaire to a large number of participants
5. Conduct factor analysis to identify weak items and the factor structure (subscales) of the questionnaire.

The following sections will consider the basic activities of these steps and the challenges of producing a meaningful measure of user experience using them.

5.2.1 Uncertainty in Games

To make the discussion in this chapter concrete, we will use as a running example our development of a questionnaire to measure uncertainty in games. This is, in part, because this the most recent work on questionnaire development that we have been involved in. It is also because in our development we set out to use the factor analysis methods described here rather than re-analyse a questionnaire that had been produced using different methods.

Our interest in uncertainty in games arose from two sources. First, it was clear that uncertainty is a common experience for people involved in information seeking, for example finding historical documents in an archive (Pugh and Power 2015). However, the feeling of uncertainty comes both from the challenge of finding documents that may or may not exist and from interactions with the search tools where the failure to find documents may be more about the idiosyncrasies of the search tools. Secondly, uncertainty is already recognised as an important constituent in the experience of playing digital games (Salen and Zimmerman 2004; Costikyan 2013). Games are a good context in which to study user experiences because the purpose of games is to generate experiences for players and those experiences are an end in themselves (Cairns 2016: 90) unlike information seeking where a user must have a task for which the interaction with a search tools is not the primary goal.

Perhaps the most interesting aspect of these two really different domains is that when working with users, we often encountered the same descriptions regarding the experience. Users in information seeking would describe not knowing where to look next, using phrases like “being overwhelmed” and “going in circles” when they were awash with information spread across multiple archives. In digital games, players would use similar phrases when trying to solve problems, or deciding which

actions would lead them to the best outcomes. This is particularly important, as it means that this experience is one that users can not only identify in different contexts, but also one they can describe with clarity that they are feeling. This means it is a good candidate for measuring with a psychometric scale.

For these reasons, we set about investigating players' experiences of uncertainty in digital games. Our first analysis was lightweight and represented an initial report on this area using traditional statistical techniques (Power et al. 2017). Our subsequent analysis however aimed to apply the most modern techniques in order to get a more situated account of our data, as will be discussed (Power et al. to appear).

5.3 Grounding the Concept

In order to generate items, there must first be some notion of what the experience to be measured actually is. Where this notion comes from can be quite vaguely defined but then it probably needs to be investigated further to provide a more concrete concept.

In looking at uncertainty in games (Power et al. 2017), the motivation came from a confluence of the concept in the two different domains of information seeking and player experience. In the domain of information seeking, uncertainty had not just arisen from our own work but was also well reflected in the literature and models of information seeking (Kuhlthau et al. 2008). In digital games, uncertainty was recognised and discussed in the literature but had more recently been more deeply analysed by Costikyan (2013), where uncertainty was mapped to different sources both in and around a game. Beyond these very domain specific views of uncertainty, we also found discussions of uncertainty being a contributing factor in cognition (Kahneman and Tversky 1982) and specifically related to decision making processes (Fox and Ülkümen 2011; Ülkümen et al. 2016), all of which helped inform what this experience may be comprised of in its constituent parts.

Where the literature does not already articulate a useful or appropriate concept of experience, an alternative is to generate an account of the concept based on qualitative research. Grounded theory is well suited to this task (Charmaz 2014) as it aims to develop a theoretical account of phenomena that faithfully represents the experiences and accounts of people. Thus, starting from a recognition of some phenomenon of interest, a grounded theory study sets out to get people's account of that phenomenon and to try to discuss what brings it about. We have used this approach successfully to try to bring clarity to notions of immersion (Brown and Cairns 2004), user experience (Calvillo-Gamez et al. 2015) and time perception (Nordin 2014) in games. Similarly, others have gone from a very general notion that players have experiences when they play games and used focus groups to distinguish and refine the general concept into specific aspects of player experience (Poels et al. 2007).

Regardless of the theoretical basis for the concept to be measured, such theories are always prey to the processes that generated them. Despite the desire of grounded

theory to theoretically sample across people and experiences in order to test the scope and range of an emerging theory (Charmaz 2014), there are both practical constraints on how far the boundaries of a theory can be developed and implicit constraints from the researcher's own interests and biases. It is considered good practice for the researcher to be reflective of how they have influenced the theoretical development but this cannot remove such influences from any resulting theory. Indeed, some biases may be beyond the ability of a researcher to either identify or articulate.

Similarly, with theories based on existing literature, all such knowledge is situated in the studies conducted and the researchers who conducted them. Costikyan (2013) is drawing on his own experiences as a game developer and player of games to identify the sources of uncertainty. No matter how extensive his experience, it will only be with a fraction of all the possible digital games that have been developed and only one perspective on those games. Of course, that his views resonate and are meaningful to other players and researchers of games gives support to his ideas. But it is always hard to see what has been omitted.

In some sense, as long as there is some grounding of the concepts in the actual experiences of people, then there is some legitimacy to the development of those concepts. If we are unable to draw a line under collecting descriptions and data regarding experiences, researchers could wait forever for an exhaustive account of a concept like uncertainty. If you wish to start going deeper then you have to start somewhere. This is not just true of user experience but even physical concepts such as temperature. Emerging theoretical concepts start from a basic understanding of our own senses (Chang 2004). For example, temperature emerges from the basic touch sensation that some things feel warm and some things feel cold. With time, research, false avenues and new theories, it becomes possible to extend the reach of such concepts beyond what could ever be sensed by us directly. So now it makes sense for physicists to make meaningful statements about absolute zero or the surface of the sun. Similarly in HCI, we are setting out to understand the concepts of user experience but we are long way from the rich theoretical accounts like the kinetic theory of gases. However, we are trying to move beyond the basic intuitive sensations to more general accounts of user experience, no matter how constrained by context and individual differences. In time, we will refine, challenge and even discard some ideas about those experiences and their composition, avoiding the temptation to supplant what has come before, and instead building a broad, nuanced and ultimately more useful understanding of a concept.

5.4 Generating Items

Once there is a concrete articulation of the concept to be measured, the next step is to begin to generate items that relate to that concept. The principle of using multiple items in a questionnaire is that the concept itself is subjective and so cannot accurately be directly expressed by people. Instead, each item is intended to tap into one

specific and distinct part of the subjective experience so that, cumulatively, the items together build up the specific and, more importantly, quantifiable account of the experience.

Each item must therefore provide a statement that captures some aspect of the experience against which participants are able to rate their agreement. For instance, with uncertainty in digital games, it was clear that a sense of being lost in a game was an important source and experience of uncertainty. However, this was not necessarily lost in the sense of navigation but in the sense of not knowing what to do. Thus, in developing the uncertainty questionnaire, it made sense to consider items related to lostness. Of course, lostness is only one facet of uncertainty but that players could talk about this gives something concrete to ask about the internal and hidden experience of what it is to feel uncertainty in a game.

In generating initial items for the pool, the items can explore the range of possible wordings and consider both positive and negative phrasings. For example:

- I often felt lost
- I always knew where I was going
- I always had a plan
- I was going round in circles
- I didn't know what to do next

All of these are potential items though more than one is probably not needed as it is only one facet amongst many of what people describe as uncertainty. Selecting which item to use may be done based on the closeness of fit to how people express their experiences or even down to the preferences of the researcher. To guard against choosing too early, it is a good idea to maintain two or three likely candidate phrasings and these can be trialled with participants.

Wording is also important to avoid common, known traps and problems, For example, bipartite questions like "I found this website interesting and enjoyable" make it ambiguous whether people found the website interesting or enjoyable or both. Though often associated, enjoyment and interest are not the same thing. Also, care needs to be taken to avoid questions that do not make sense in some contexts. For example, "The first person perspective drew me in to the game" only applies if the game does in fact have a first person perspective on a virtual world. Extensive resources exist to guide researchers such as Oppenheim (2000) and Müller et al. (2014).

No matter how much care researchers might take, the wording of items can show strong cultural biases. One personality questionnaire that we have used previously in our research had the item "I am a spendthrift." Whilst it is a perfectly reasonable statement, "spendthrift" is not a commonly used word and many non-native English speakers had real trouble with this item as they simply did not know the word. In fact, many native English speakers also had trouble as they had never seen or used the word enough to be sure of its meaning. It may be the case here that the questionnaire had aged badly from a time when spending and how you spent your money was thought about and talked about more. Just as questionnaires may be of their time, they can also be of their place with colloquialisms like "my cup of tea" or

"curve ball." These may be very clear expressions to the researcher and any reviewer that the researcher knows, but they place the questionnaire firmly in a cultural context.

Another form of cultural contextualisation seems to arise from what researchers think people will be responding to. To be specific, in player experience research, there is often in the mind of the researchers a prototypical or even stereotypical idea of what it is to play a game. Such an idea might be that playing a game is sitting down at a gaming console and spending two hours exploring alien worlds in Mass Effect, or it might be stopping for 10 minutes during the day for a quick burst of Candy Crush on an smartphone. The researcher will try to be broad in imagining such prototypes and evaluating items against relevance in these contexts. But all imaginings are necessarily limited. It is not possible to envisage all the possible games, current and future, that players might play and so mentally check each item against them. Even defining game genre is a challenge (Clarke et al. 2015). Thus, to some extent all researchers are guided by their mental prototypes of the technologies that people use. This limits the reach of the questionnaire but without specific ways to articulate the prototypes considered, it is impossible to really acknowledge what those limits are.

As the generation of items progresses and items need refining, sometimes experts are used to review the items for relevance to the intended underlying concept. Though this will help to broaden and challenge the cultural and prototype biases of the researcher, it cannot overcome them, particularly when the experts are chosen from the researcher's colleagues (as they typically are).

5.5 Participants

One way to validate items early on is to ask potential questionnaire respondents to try out the items. This can be done with the large, relatively unrefined item pool where there might be items with overlapping content or where different wordings are used for the same ideas. This allows the participants to give their view of what it is like to do the questionnaire: it is a form of usability test on the items and is sometimes done with only a few participants. This can lead to removing items, rephrasing others or even suggest new items which the researcher did not think of. Later, once items have been selected and refined down to a plausible questionnaire with the right balance of length and conceptual content, the questionnaire is administered usually in a survey with a large, representative group of participants.

Regardless of at what stage participants are involved in the process and how many times participants are involved, as with any quantitative study, there is always the issue of who a sample of participants are. Though statisticians often talk about the distinction between sample and population, there is typically no meaningful population that can be identified. The sample is typically drawn from a pool: students at a university; people who subscribe to a particular forum; passers-by on the day of the field trial. With good demographics, it is possible to characterise to some

extent the diversity of participants but there is no way to know in what sense any particular set of participants are either typical (and if so, typical of what) or idiosyncratic.

Information from participants is often used in questionnaire development to remove items that do not function well, whether this is a result of specific feedback from participants or through statistical analysis. In statistical analysis, the reasons for considering the item weak might be:

- It is often omitted by participants
- It shows little variation, for example, everyone strongly agrees with it so it adds little insight into the concept
- It shows no coherence with the other items

In many discussions of questionnaire development, these reasons are considered good indications that the item is weak. For example, in developing the uncertainty questionnaire we had an item "I found myself going round in circles." We felt that this was a very good characterisation of the experience of uncertainty. There is a sense of doing something but ending, unintentionally, back at the same point. This suggested to us a lack of progress, not knowing what to do or not knowing why something happened when the player did do something. This item however did not load well in our factor analysis suggesting it lacked coherence with the other items or at least less coherence than others. Thus, we eliminated it.

However, it is worth examining this assumption a little further. If a researcher, along with expert reviewers and early trial participants have proposed an item, on what basis is it then considered weak as a result of running with a group of participants? It could be that for these participants, they simply did see themselves as going in circles. Or maybe not enough of them played games where going in circles was a possibility.

This also relates to the notion of prototypes in developing the questionnaire. When a group of participants respond to a user experience questionnaire, they are either bringing to mind or have just engaged in a particular experience. Naturally, this set of experiences goes beyond the prototypical experiences imagined by the researcher. However, these experiences are still specific and concrete to given contexts and the range of contexts is necessarily limited. This is in part influenced by the ways in which participants are recruited. If participants are found through a particular discussion forum about games, they are likely either to be engaged with a particular sort of game or to have particular attitudes to playing games that make them want to engage in that forum.

As with all statistical methods, it is not possible to know for sure whether it is the participants that are somehow not typical or whether the items are indeed not suitable. However, unlike other contexts, such as experiments, where the variation of participants is accounted for by statistical methods, it is not possible to use statistical methods to decide whether it is the participants or the items to blame. Until a sound operationalisation of a concept has been established, for instance through a questionnaire, it is not possible to know how relevant items are to the concept and therefore to account for their variation with statistical methods. And given the

subtleties and nuances of language, though there is not an infinite set of plausible items to include in a questionnaire, it is effectively unbounded within the scope of the questionnaire development process. Just as with participants, the pool of potentially relevant items is only represented by the sample of particular items that we happen to gather together.

5.6 Factor Analysis

The core step in validation of a questionnaire is to do factor analysis. For this, a version of the questionnaire, let's call this Version 1.0, is administered to a large number of people. Version 1.0 is not necessarily expected to be the final version of the questionnaire. It may well include too many items but the previous processes are not able to decide between them. For instance, the Version 1.0 of the uncertainty questionnaire contained 65 items, which we knew was too long for a practical instrument for use in player experience research. The hope is that factor analysis will both highlight items that are not useful to respondents as well as give an indication of which items, in a statistical sense, work better than others.

The purpose of factor analysis is, in essence, to reverse the process of item generation, where a complex concept is broken down into items that each partially reflect the concept, and try to find the commonality between different items that might reflect the hidden concept that underpins them. There are many good books and resources on how to do factor analysis, for example Kline (1994, 1998) and Hair et al. (1998), that go into both the mathematics and the practicalities of doing factor analysis on questionnaires. The purpose here is to give some insight into how meaning arises from these processes. Such books will also give guidance as to what actually is a "large" number of participants.

Typically, when a concept is being captured for the first time by a new questionnaire, exploratory factor analysis is undertaken. This is usually Principal Component Analysis but it may also be a factor analysis approach like Principal Axis Factoring. In my experience, these only give slightly different results. What these methods do give is a way of grouping items in such a way that items from the same group strongly correlate with each other but only weakly with items from the other group. Each group then forms a factor.

However, this is usually not as easy or clear a step as one might hope. While some items will clearly group, some items cross-load, that is, they correlate well with items from two or more distinct factors. Also, though there may be a set of distinct factors, it can be hard to collectively interpret the items in the factors as a unified, meaningful concept. Additionally, some factors naturally correlate with each other because they are all, after all, meant to relate to the same underlying concept.

The role of the researcher is to navigate the challenges of deciding on useful factors with the statistical tools of factor analysis, in particular choosing the number of

factors in a solution and judging what constitutes an item belonging to a factor. The result is a set of factors that underpin the concept in hand.

In our first attempt to analysed the uncertainty questionnaire data using Principal Component Analysis, we found four distinct factors.

- Disorientation
- Exploration
- Prospect
- Randomness

Though the factor analysis was done using the recommended best practice, there is a puzzle at the heart of this. How could a questionnaire intended to measure the single concept of uncertainty result in four distinct factors? These factors correlate together but not strongly so is there one concept of uncertainty that players experience or four? Interestingly, the factors did not divide along the same lines as Costikyan (2013)'s analysis of sources of uncertainty suggesting that different sources may not lead to distinct experiences of uncertainty.

Bifactor analysis was developed in the early days of questionnaire design but was neglected until relatively recently (Reise 2012). Whereas traditional factor analysis posits that data can be represented by distinct factors that may correlate, bifactor analysis assumes a single underlying factor, often called g, that accounts for all common correlation between the factors and then specific distinct variation due to each distinct factor.

The second and more careful analysis of the uncertainty questionnaire was conducted with this model in mind. Our first application of this method deliberately looked only for a single factor solution. Almost all of the 65 items in Version 1.0, loaded well on a single factor. The argument is that this single factor is capturing the underlying notion of uncertainty. Further analysis suggested five distinct factors (Power et al. to appear):

- Uncertainty in Decision Making
- Uncertainty in Action
- Uncertainty in Problem Solving
- External uncertainty
- Exploration

Not all items loaded well on these five distinct factors suggesting that while some items are relevant to uncertainty they are not central enough to form into factors. As there were still a lot of items, we selected items that loaded strongly in each factor and therefore might be understood to be core to the concept represented by the factor. We then applied a bifactor analysis to this 24-item Version 2.0 of the questionnaire.

What we found was that all of the items of Version 2.0 loaded to some extent on the single underlying factor, g, but that External Uncertainty and Exploration both showed strong distinct loadings. Our interpretation is that the first three items are core to the internal sense of uncertainty of what players feel uncertain about but which relies on them to resolve. External uncertainty is due to things outside of their

control: behaviour of other players, hidden information, chance, randomness, or even just perceived randomness. It of course relates to internal uncertainty but as it is perceived to arise from outside the player's control, it is also clearly distinct. Finally Exploration is a strategy to resolve uncertainty. Feeling uncertain leads to the need for exploration but is otherwise unrelated to the other factors, but might relate to how External Uncertainty can become internal uncertainty within the game space.

Even within this model, though internal uncertainty emerges strongly, there is still room for some people to feel that uncertainty in different ways, say from not being able to solve problems or from not knowing what action to take. What bifactor analysis suggests is a broadly unified concept that is nuanced more or less strongly by the different factors according to the players, their contexts and what the games mean to them. It may be this nuancing that led to four factors in our preliminary analysis because internal uncertainty there factored only into disorientation and prospect rather than the three factors we later found.

What should also be noted here is the role of the researcher in developing this model. We arrived at this description of uncertainty in games iteratively and only stopped when we felt we had a good description. The numbers of factors, the choice of what constituted a high factor loading and our choice of items for Version 2.0 were in no way determined algorithmically. The hope, of course, is that though this solution may be idiosyncratic it is nonetheless a meaningful representation of uncertainty and one that would agree with other such measures developed by other researchers. The problem is that once such a measure is in place, the inclination of others to develop similar measures is greatly reduced.

5.7 Analysing Data

Once a suitable instrument for an experience has been developed, it can then be deployed in studies, experiments, surveys and so on. In this way, it is possible to begin to both quantify and manipulate the experiences that people have in different contexts. Experiments will explicitly manipulate the context of interaction and use statistical testing to see the effect on experience, for example, altering the degree of what is visible to see the effect on players' feelings of uncertainty (Kumari et al. 2017). Surveys or other in-the-wild studies enable researchers to build a picture of how people experience a particular concept. Through exploratory statistical analysis, such studies reveal correlations and associations between different aspects of the players and their experiences, for example whether winning or losing in a game influences their sense of social presence (Hudson and Cairns 2016).

Statistical tests are used to do these analyses and typically with questionnaire data, the default is to use the classic parametric statistics like a t-test and ANOVA. Historically, these tests are believed to be robust to deviations from their assumptions and likely to lead to sound analysis. For instance, one such belief is that a t-test still gives sound results even when underlying distributions are not at all

normal provided the sample sizes being compared are equal (Sawilowsky and Blair 1992). However, more recently, such beliefs have both been challenged and also rendered unnecessary thanks to new tests that are genuinely more robust and rely less on inappropriate assumptions (Wilcox 2017). These more modern tests however do require the researcher to be more careful in looking at data.

It is a much larger topic to explore the full range of the implications of modern robust statistics for measuring user experiences (Cairns 2018). However, here it is worth considering something very relevant to a third wave approach which is consideration of the individual and their experiences. Typically statistical tests work with averages, that is, some measure of a sample of participants that aggregates across all of the participants such as the mean or median. However, this not only downplays the importance of the individual participant but also considers individuals as a relatively uniform group whose experiences are in some sense the same.

There are of course good reasons not to put too much weight on individual data points about user experience. Measures of experience are likely to be quite inaccurate partly because a questionnaire is at best measuring facets of a hidden experience and partly because of people's interpretation of the questionnaire. It is only on aggregate over a series of measures that quantification of experience becomes meaningful.

However, where individual participants' data do meaningfully stand out and with implications for analysis are when a measurement outlying. It is possible that any outlying measure is just highly inaccurate but at the same time, the reasons for such inaccuracy must be considered.

Modern statistics has robust tools for identifying outliers of individuals from a sample. One of the most effective is in a boxplot. The box of the boxplot represents the interquartile range of a sample of data, the middle 50% of data points. A point is declared an outlier if it is a fixed proportion of the box's size away from the box. This slightly complicated decision procedure arises so that the outliers themselves are unlikely to influence the decision of what constitutes an outlier. This robust decision procedure is built into most statistical packages that can draw boxplots with the result that it is easy to identify outliers as points singled out on the boxplot. Traditionally, outliers were nuisances. Theses single points can strongly influence the results of parametric tests and so mislead the interpretation of the "average" behaviour. Thus, outliers are often omitted from the analysis (Bakker and Wicherts 2014). However, it is not clear that that is justified. If a person has an unusual or outlying experience, that may still be an important aspect of the technology or interaction under consideration. Yes, it could also unduly influence the statistical analysis but it is also worthy of consideration in its own right.

Thus, the detection of outliers should be a reason to pause and think about the possible causes of outlying values. There are typically four possible reasons why they might occur (Osborne 2010):

1. Data entry error
2. Mischievous participants
3. Bad study design

4. True representation of a participant

Of these, only the first is easily solved. An outlying value may occur because of a miskeying or slip when entering data ready for analysis. In which case, the outlying value has no relevance to any analysis. However, if an outlying value can be tracked back to a participant who had not engaged in a study properly then there is a more serious problem. It is not enough to discount that one outlying participant's data because the same behaviour that led to an outlying value may also be influencing other participants and their measurements as well. This can particularly be a problem when a study has been run online and the researcher has not been on hand to observe participants' behaviours during a study. Checks would need to be made to see if other participants also behaved the same way and then all of the affected data removed from any analysis. More serious still is that the study design itself led to outlying values, for instance, by a failure in the questionnaire software (or even the wetware) to deliver all the questions correctly to all participants.

In these last two cases, outliers, far from being nuisance values, are important indicators of potential problems in the research. They require investigation and the causes of the problems need to be tracked down and if possible eliminated.

When all possible mishaps have been discounted, then the only conclusion can be that some people simply produce outlying values. Some people have different experiences than others and in some cases, sufficiently different to be considered extreme or outlying. But that does not make them illegitimate values to be discounted. Indeed, taking seriously the challenge of third wave thinking, such values may arise from a different type of person in a different situation and could signal important variations in people's experiences. In almost all cases, then, outliers are food for thought either about the nature of the research or the nature of the experiences being researched.

In research where questionnaires are used to measure user experience, it is not at all clear that there has ever been any systematic consideration of outlying measurements. In more traditional usability contexts, we have noticed that there can be persistent outliers in any particular usability task (Schiller and Cairns 2008). For example, we have seen that the time it takes people to navigate an online website always seems to produce one outlying person who takes an unusually long time. However, it would be mistaken to attribute that unusual time as atypical. It seems to be a persistent feature of either that sort of study or of people generally. We have a new project underway to explore this more systematically but up until now it has remained unexplored by others despite the prevalence of usability tests in both the research and practice of UX.

It would be useful to try to look systematically across large bodies of data to see if, similarly, outlying measurements are a feature of studies into experience. Are there individuals or even sets of people for whom experiences simply do not lie in the typical range of given questionnaire? Or more challengingly, perhaps we should be trying to seek out outliers and so highlight the limitations and situatedness of our measures.

5.8 Limitations and Opportunities

From the above account of questionnaire development, there are two immediate implications. First, the epistemologically situated nature of any questionnaire has to be acknowledged and accordingly, any numbers produced that measure user experience are not absolute but represent a particular understanding of what that experience could be. Perhaps this is not news to many HCI researchers but it brings to the fore what is often overlooked or ignored: just because we have hard quantities we do not necessarily have hard concepts. Secondly, it may seem that given the bias and context of any questionnaire, it is pointless to pretend that any questionnaire captures anything meaningful about the general nature of human experience. Our questionnaires cannot capture anything that reflects wider and useful truths about humans and our relationships to technology. In which case, we would be entitled to have an existential crisis about our research careers!

Stepping back though, having a scientific theory which is known to be false is not a problem. History is on our side because almost every scientific theory to this point has been proven to be wrong both in the specifics of its predictions and indeed the underlying "reality" that it represents. Phlogiston, Newton's theory of gravity, Faraday's lines of magnetic flux and so on have all fallen by the wayside as essentially wrong theories. Even now, the two most precise theories in modern physics, general relativity and quantum field theory are known to be wrong because they are fundamentally incompatible. Based on this, we have no right to expect our theories, based on questionnaires or not, to be correct. This is the principle of pessimistic induction.

It is pessimistic but it is also liberating. The judge of the value of a scientific theory is not some reference to an elusive underlying truth but rather whether it is useful: can we make testable predictions? does it help answer questions? does it drive inquiry? None of these criteria assert the truth or correctness of a theory but they suggest that in some sense we are making progress in developing new knowledge.

With this regard, we argue that limited and situated though questionnaires may be for measuring user experience, they help us to make progress. We have not had much opportunity to use the uncertainty questionnaire to see how progressive it is (though our early forays with student projects are encouraging). Considering instead the older immersion questionnaire, the IEQ (Jennett et al. 2008), there has been a lot of work using this questionnaire. Understandably, and as you might hope, it has helped to validate some expected results: more interesting games are more immersive (Jennett et al. 2008); playing under time pressure increases engagement (Cairns et al. 2014). Rather than informing user experience, this adds weight that the IEQ is measuring something relevant to how immersed a person feels in a game. We have also learned less expected things: players are more immersed if they think they are playing against humans (Cairns et al. 2013); players are more immersed if they believe the game is adapting to them (Denisova and Cairns 2015).

The fact the IEQ and other questionnaires do help us to make progress is encouraging. Despite all the limitations and narrowness of the epistemological grounding of the questionnaire, it seems to capture something that starts to reveal new things, things we did not previously realise and things that might be useful and important. Our feeling is that perhaps this is because in some way, people are not so different. Yes, we value different things. Different contexts, different relationships, different times of our life, all give different meanings to the experiences we have with technology and with each other. However, those experiences seem to have some commonality despite all this variety. We cannot know for sure if your experience of uncertainty in games is the same as ours but it is perhaps an act of faith in human nature that in some sense, while not the same, it is very like ours.

This is not to say we can neglect that for some their experience are very different. Many will agree on the experience that is called "red" but we know there are colour-blind people for whom the experience of red is fundamentally different and blind people for whom it is not even meaningful. Modern statistics suggest that we should take seriously both the difference in nuance of experiences between individuals and the differences that lead to radically different experiences.

With these considerations, measuring experiences is not the inherent contradiction in terms that it might at first seem. A successor science view of measuring experiences, or indeed anything, must require us to question the situated nature of the knowledge we produce. Measuring uncertainty allows us to make progress in certain ways but we should not stop at this particular measure but constantly reach and extend both our concept and our measurements to investigate new people, new games and new situations of play, and beyond into other interactive systems. In this sense, a successor science view of measurement accords well with Wright and McCarthy (2004)'s view of experiences as becoming: our measures must be in a process of becoming as well.

The uncertainty questionnaire is in its early days but we are already thinking about how it is relevant to different sorts of people from our mainstream players who helped us to develop it. In particular, it seems that disabled players may experience uncertainty when playing games that is both intrinsic to the game but also driven by the technology that enables them to play. We are therefore looking to validate the questionnaire specifically with gamers with disabilities. We believe that at some level they will experience the same pleasures and frustrations that other players feel, even if they come from different sources, and that this uncertainty questionnaire will allow us to make progress for their experiences as well. However, if it does not, then we will need a new way to understand and measure their experiences.

In the third wave of HCI, we must recognise that there is not just one experience or one way to measure it.

References

Bakker M, Wicherts JM (2014) Outlier removal, sum scores, and the inflation of the type i error rate in independent samples t tests: the power of alternatives and recommendations. Psychol Methods 19(3):409

Baumer EP, Silberman M (2011) When the implication is not to design (technology). In: Proceedings of the SIGCHI Conference on human factors in computing systems, ACM, pp 2271–2274

Bødker S (2006) When second wave HCI meets third wave challenges. In: Proceedings of the 4th Nordic conference on human-computer interaction: changing roles, ACM, pp 1–8

Brockmyer JH, Fox CM, Curtiss KA, McBroom E, Burkhart KM, Pidruzny JN (2009) The development of the game engagement questionnaire: a measure of engagement in video game-playing. J Exp Soc Psychol 45(4):624–634

Brown E, Cairns P (2004) A grounded investigation of game immersion. In: CHI'04 extended abstracts on human factors in computing systems, ACM, pp 1297–1300

Cairns P (2016) Engagement in digital games. In: O'Brien H, Cairns P (eds) Why engagement matters. Springer, Cham, pp 81–104

Cairns P (2018) Being less wrong: essays on statistical methods in HCI. Cambridge University Press, Cambridge

Cairns P, Cox AL, Day M, Martin H, Perryman T (2013) Who but not where: the effect of social play on immersion in digital games. Int J Hum Comput Stud 71(11):1069–1077

Cairns P, Cox A, Nordin AI (2014) Immersion in digital games: review of gaming experience research. In: Handbook of digital games. Wiley, Hoboken, pp 339–361

Calvillo-Gamez EH, Cairns P, Cox AL (2015) Assessing the core elements of the gaming experience. In: Bernhaupt R (ed) Game user experience evaluation. Springer, Cham, pp 37–62

Chang H (2004) Inventing temperature: measurement and scientific progress. Oxford University Press, Oxford

Charmaz K (2014) Constructing grounded theory. Sage, Thousand Oaks

Chen J (2007) Flow in games (and everything else). Commun ACM 50(4):31–34

Clarke RI, Lee JH, Clark N (2015) Why video game genres fail: a classificatory analysis. Games Cult 12:445–465

Costikyan G (2013) Uncertainty in games. MIT Press, Cambridge, MA

Cox AL, Cairns PA, Walton A, Lee S (2008) Tlk or txt? using voice input for SMS composition. Pers Ubiquit Comput 12(8):567–588

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, ACM, pp 23–33

Denisova A, Guckelsberger C, Zendle D (2017) Challenge in digital games: towards developing a measurement tool. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, ACM, pp 2511–2519

Dowell J, Long J (1998) Conception of the cognitive engineering design problem. Ergonomics 41(2):126–139

Fox CR, Ülkümen G (2011) Distinguishing two dimensions of uncertainty. In: Brun W, Keren G, Kirkebøen G, Montgomery H (eds) Perspectives on thinking, judging, and decision making. Universitetsforlaget, Oslo, pp 21–35

Gould SJ (1996) The mismeasure of man. WW Norton, London

Hacking I (1983) Representing and intervening: introductory topics in the philosophy of natural science. Cambridge University Press, Cambridge

Hair JF, Anderson RE, Tatham RL, Black WC (1998) Multivariate data analysis, 5th edn. Prentice Hall, Upper Saddle River

Harrison S, Tatar D, Sengers P (2007) The three paradigms of HCI. In: Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA, pp 1–18

Harrison S, Sengers P, Tatar D (2011) Making epistemological trouble: third- paradigm HCI as successor science. Interact Comput 23(5):385–392
Hassenzahl M (2004) The interplay of beauty, goodness, and usability in interactive products. Hum Comput Interact 19(4):319–349
Hudson M, Cairns P (2014) Measuring social presence in team based digital games. In: Riva G, Waterworth J, Murray D (eds) Interacting with presence. de Gruyter, Warsaw, pp 83–101
Hudson M, Cairns P (2016) The effects of winning and losing on social presence in team-based digital games. Comput Hum Behav 60:1–12
Jennett C, Cox AL, Cairns P, Dhoparee S, Epps A, Tijs T, Walton A (2008) Measuring and defining the experience of immersion in games. Int J Hum Comput Stud 66(9):641–661
Kahneman D, Tversky A (1982) Variants of uncertainty. Cognition 11(2):143–157
Kline P (1994) An easy guide to factor analysis. Routledge, London
Kline P (1998) The new psychometrics: science, psychology and measurement. Routledge, London
Kuhlthau CC, Heinström J, Todd RJ (2008) The 'information search process' revisited: is the model still useful. Inf Res 13(4):13–14
Kumari S, Power C, Cairns P (2017) Investigating uncertainty in digital games and its impact on player immersion. In: Extended abstracts publication of the annual symposium on computer-human interaction in play, ACM, CHI PLAY '17 Extended abstracts, pp 503–509
Kuutti K, Bannon LJ (2014) The turn to practice in HCI: towards a research agenda. In: Proceedings of the 32nd annual ACM conference on human factors in computing systems, ACM, pp 3543–3552
Lazzaro N (2009) Why we play: affect and the fun of games, entertainment interfaces and interactive products. In: Sears A, Jacko JA (eds) Human-computer interaction: designing for diverse users and domains. CRC Press, Boca Raton, pp 155–176
Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 22(140):55
Long J, Dowell J (1989) Conceptions of the discipline of HCI: craft, applied science, and engineering. In: People and Computers V: proceedings of the fifth conference of the British Computer Society, Cambridge University Press, vol 5, p 9
Müller H, Sedley A, Ferrall-Nunge E (2014) Survey research in HCI. In: Ways of knowing in HCI. Springer, New York, pp 229–266
Nordin A (2014) Immersion and players' time perception in digital games. PhD thesis, University of York
O'Brien HL, Toms EG (2010) The development and evaluation of a survey to measure user engagement. J Am Soc Inf Sci Technol 61(1):50–69
Oppenheim AN (2000) Questionnaire design, interviewing and attitude measurement. Bloomsbury Publishing, London
Osborne JW (2010) Data cleaning basics: best practices in dealing with extreme scores. Newborn Infant Nurs Rev 10(1):37–43
Poels K, De Kort Y, Ijsselsteijn W (2007) It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology. In: Proceedings of the 2007 conference on future play, ACM, pp 83–89
Power C, Denisova A, Papaioannou T, Cairns P (2017) Measuring uncertainty in games: Design and preliminary validation. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, ACM, pp 2839–2845
Power C, Cairns P, Denisova A, Papaioannou T (to appear) Player uncertainty in games: measuring when players gets stuck. Under review
Pugh J, Power C (2015) Swimming the channels: an analysis of online archival reference enquiries. In: Abascal J, Barbosa S, Fetter M, Gross T, Palanque P, Winckler M (eds) Human-computer interaction. Springer, Cham, pp 99–115
Reeves S (2015) Human-computer interaction as science. In: Proceedings of the fifth decennial aarhus conference on critical alternatives, Aarhus University Press, pp 73–84
Reise SP (2012) The rediscovery of bifactor measurement models. Multivar Behav Res 47(5):667–696

Ryu YS, Smith-Jackson TL (2006) Reliability and validity of the mobile phone usability questionnaire (mpuq). J Usability Stud 2(1):39–53
Salen K, Zimmerman E (2004) Rules of play: game design fundamentals. MIT Press, Cambridge, MA
Sawilowsky SS, Blair RC (1992) A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychol Bull 111(2):352
Schiller J, Cairns P (2008) There's always one!: modelling outlying user performance. In: CHI'08 extended abstracts on human factors in computing systems, ACM, pp 3513–3518
Ülkümen G, Fox CR, Malle BF (2016) Two dimensions of subjective uncertainty: clues from natural language. J Exp Psychol Gen 145(10):1280–1297
Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 1079–1088
Wilcox RR (2017) Introduction to robust estimation and hypothesis testing, 4th edn. Academic Press, London
Witmer BG, Singer MJ (1998) Measuring presence in virtual environments: a presence questionnaire. Presence Teleop Virt 7(3):225–240
Wright P, McCarthy J (2004) Technology as experience. MIT Press, Cambridge, MA
Wright P, McCarthy J, Meekison L (2003) Making sense of experience. In: Blythe MA, Overbeeke K, Monk AF, Wright PC (eds) Funology. Kluwer Academic Publishers, London, pp 43–53