On Choosing Games

And What Counts as a "Good" Game

Katrin Becker, Ph.D. (ABD)
Graduate Division of Educational Research,
University of Calgary,
2500 University Drive NW, Calgary, Alberta T2N 1N4
Email: becker@minkhollow.ca
1-403-932-6322

Dr. James R. Parker
Digital Media Lab
University of Calgary,
2500 University Drive NW, Calgary, Alberta T2N 1N4
Email: jparker@ucalgary.ca
1-403-220-6784

Abstract

This chapter will discuss the growing importance of applying considered rationales to which games are chosen for study, whether it be for ethnography, classroom use, or anything else.

A brief overview of how games are currently bring chosen for study is presented through a meta analysis of studies with games that were published between 2003 and 2006 in order to demonstrate that most published games studies do not include a supported rationale for the games chosen.

The chapter will then present various ways that game choices can be justified, and propose and explain a data fusion technique that can be applied to game reviews and other lists in order to facilitate representative and defensible game choices.

On Choosing Games And What Counts as a "Good" Game

Introduction

Why is it important to justify the choice of game being used as an example in a scholarly article or for the purposes of study? In the early days of games studies there seemed little call for careful scrutiny of one's game choices. We studied what we had handy and wrote about the games we were already playing. However, if we want to make the case that the game in question is *good* by some measure (however we decide to define "good") then we really should have some evidence to back this up. When a single game or a small number of games are chosen as the subject(s) of study they form part of the bounded system that is the case being examined, and also forms part of what makes the case of special interest (Stake, 1995). If we are proposing the use of a game in the classroom or the study of some specific game to learn something applicable to our agenda whether that agenda is to examine the educational potential of the game or to learn something else about the game that may inform other instruction, then as academics we have a responsibility to explain why *that* game is suitable for our purpose.

One reason for putting thoughtful effort into justifying the choice of a game used in a study is that it helps to make the study itself more credible. This has implications for the increased acceptance of game studies academically as well as for helping to improve relations between academia and the games industry. In a recent article offering suggestions for how the Academy could build stronger ties with the Games Industry, John Hopson argues that we should "(u)se examples from bestsellers. A good example from a popular game is more effective than a great example from something they've never heard of. Industry people often suffer from an 'if-they're-so-smart-, why-ain't-they-rich' attitude towards smaller titles. Even if the small title is a perfect example of how the theory works, they're going to be less likely to listen if they haven't heard of the game ahead of time. Commercial success is one way of making sure that the audience will respect your examples, but you can also use titles that are well known or critically acclaimed but which weren't necessarily huge blockbusters. It's also important to keep your examples as current as possible, because many industry folks will see a three-year-old example as ancient history" (Hopson, 2006).

Critical and commercial success are key recognizable and accepted (albeit subjective) measures of a game's popularity, and that popularity in turn gives some indication of that game's perceived quality as judged by players, developers, and game critics. When it comes to resources that are primarily creative or artistic in nature, subjective measures are often the only ones we have. In sports for example, such as sprinting, determining who the fastest sprinter is can be done quite objectively – it is a matter of comparing competition times and the runner with the fastest time wins but no such objective measure exists for most creative endeavors, and since games are creative designs we can only produce subjective measures. To further compound the problem, lists of 'top games' tend to be quite unstable and change not only from year to year as new titles gain recognition, but sometimes from day to day as in review sites where players can contribute. One consequence of this is that no single list can reasonably be used to support claims about a particular game's qualities. One solution is to combine multiple lists into one comprehensive one. By combining multiple lists, we can increase our confidence in the qualifications of games that end up on top. However, the challenge in combining measures from these various sources is that the criteria used to produce lists of 'good' games are often so divergent that they cannot be compared or combined directly. Categories and scores vary, the methodology used to rate and rank the games varies, even the contributors vary – in some cases they are paid professional critics, in other cases association members or even the public at large contributes votes and reviews. The data fusion technique described in this chapter offers a solution to this problem that is both verifiable, and repeatable. Combining a number of different measures to come up with a single measure ensures that games that end up at the top of the final list qualify as successful by more than one measure and have been assessed by more than one source. Using a systematic approach to ranking games results in a list with which most (industry, gamers, and critics) could agree.

Why Do We Study Games?

Game Studies continues to develop as a discipline just as digital games continue to evolve. While there remains an interest in examinations of specific games for various purposes, as the number and sophistication of titles released in a given year continues to rise, it is time to begin looking more closely at how we are choosing the games we study, the criteria we use for those studies, and how we support our claims about the suitability of the game for our purposes.

Although commercial success as demonstrated by sales figures is an important measure of success, it is not the only one, and may not be the most important one for any particular study. Often, studies of individual games are conducted with the hopes of being able to generalize at least some of the conclusions to other games and/or other players. Given the number and variety of games with no cleanly defined delineations of genre, can it be assumed that it is possible to examine one game and make generalizations to other games? How should these generalizations be qualified or limited? Games are no longer trivial, nor frivolous so this is not a straightforward question. There were approximately 2500 game titles released in 2005. With so many titles released in one year it becomes harder and harder to justify choosing a game based on personal preferences. Claims that a particular game meets certain criteria critical to the analysis should be supported by something beyond the author's say-so. There will have to be some way of providing evidence supporting the claims we make about the qualities of the game that we have determined are necessary to our study. As studies on, with, and of games become more accepted and common in mainstream educational research, and as the number of games to choose from continues to grow it will also become more important to justify the choices of subjects. This has not been common practice to date.

How Do Researchers Choose Games for Study?

Digital games have been around for about 40 years now (Williams, 2006) and game studies as a recognizable discipline has been around for about ten years (Wolf, 2001). While there remains an interest in examinations of specific games for various purposes such as Kurt Squire's work with *Civilization* (Meier, 2001; Squire, 2003), as the number and sophistication of titles released in a given year continues to rise, it becomes important to look more seriously at how we are choosing the games we study, the criteria we use for those studies, how we support our claims about the suitability of the game for our purposes, and how generalizations to other games should be limited or qualified.

Since the question of how games are selected by researchers has not previously been examined the author conducted a qualitative meta-analysis (Delgado-Rodríguez, 2001) of what methods researchers reported using in choosing games for study. Papers and reports published primarily between 2003 and 2006 were examined with the goal of determining the reporting frequency of explanations of game subject choices. While it was not known if selection criteria were applied to the choices of games that did not get reported in the studies, it was *not* the goal of this analysis to offer a critique of the choices themselves, simply to examine how they were made. Note that a lack of explanation in the publication does not prove a lack of consideration for the study. It is certainly possible that carefully considered reasons motivated the game choices in many of the studies presented here, but that these were simply not included in the publication. The worthiness of the choice that was made is also not being examined here, and indeed many well-known game scholars are included in the list of papers examined. In many cases there would be little controversy over the claim that the chosen game has the specified characteristics. In some cases there would also be no dispute that the particular type of game is a suitable choice (and perhaps even the most suitable choice) for the study as reported. Many of the reports have contributed to the body of knowledge in games studies in important and significant ways.

A distinction was made in the meta-analysis between the description of the game (including gameplay and any noteworthy features of the game) and a rationale for the choice of the game. Virtually all papers examined offer a description of the game(s) used. Fewer (37%) explained why this game meets the need of the study, and fewer still (15%) supported that explanation with citations. It is suspected that many game choices were, at least in part opportunistic, as the researchers had access to or were already playing this game. Only one researcher actually stated that they were already playing the game as their explanation for choosing it. In other cases, the researcher

states that they play the game but it is not made clear whether the study began before or after that individual began to play that game, nor how much influence the author's own game playing preferences had on the choice. Comments such as, "I've been playing this game for years" places game studies in a somewhat unique position as both casual and avid gamers draw on their own playing experiences to inform their studies. This kind of connection places many game studies in the realm of what Glesne has called "Backyard Research" which can make separating researcher roles from pre-existing ones complicated and difficult (Glesne, 1999).

Generalizability of game studies is one issue that can be addressed by more rigorous justification of game choices. In a longitudinal study of violence in an online videogame, Williams and Skorik raised questions about the generalizability of games which have implications far beyond their own study. "The online database www.allgame.com lists descriptions of more than 38,000 different games across 100 platforms. To collapse this wide variety of content into a variable labeled "game play" is the equivalent of assuming that all television, radio, or motion picture use is the same (Williams & Skoric, 2005). As Dill and Dill have noted, 'This is akin to lumping films like *The Little Mermaid* with *Pulp Fiction*, and expecting this combined 'movie viewing' variable to predict increases in aggressive behavior' (1998, p. 423). One interpretation of this statement is that we are not currently paying sufficient attention to the great variety of games available. Such a large number of games means that we cannot assume that one game is as suitable as any other for the purposes of study (i.e. we cannot collapse all adventure games into one category for the purposes of study). Studying ONE game does not necessarily allow us to generalize our findings. While a suggestion to force all games researchers to use some sort of "scientific" approach to choosing games is clearly unreasonable, paying closer attention to how we choose games can certainly help address legitimate questions about a game's fitness for purpose in the context of a study. It may not be necessary to explain why someone has chosen Shakespeare or Chaucer to study, but games have not yet attained the level of acceptance that classic literature has and we should still be explaining our decisions. If we choose a game because it is one we personally like, that may be justified, but we still need to address how that makes that game a worthy candidate for study. If we choose a game because it is popular, then we should be able to support that with facts or citations that can stand up to scrutiny.

The meta-analysis conducted by the author included 52 papers that were examined in detail. 91 games were identified comprising 71 distinct titles (some studies used more than one game but numerous studies used the same games such as *World of Warcraft*). Only one paper out of 52 examined reported having applied some systematic technique to identifying games for study. 19% offered no explanation for why they chose the game they did. Several offered the explanation that they were already playing it. In most cases (89%) claims that the game met the criteria described were not supported. Only one study described a rationale for the exclusion of one or more games from study and one other report explained the methodology used to select the game for the study (they allowed the study participants to vote on a game, citing prior research that suggested participant interest was an important factor in the study's success). The results of the meta-analysis indicate that a very small minority of game researchers currently report on the methodology used for the choice of a game in a study, or use examples of excluded games to support their choices. Very few explain how or why their stated game requirements support the goal of the study. While some cite references to support at least some of their claims about why this **kind** of game is needed for this study, almost none cite any references supporting their claim that the chosen game actually meets those requirements. By far the most common attribute supported by other references is the claim about the game's popularity and the most common outside reference is to sales figures.

Towards A Solution

We need to begin supporting our claims about a game's fitness for purpose which "equates quality with the fulfillment of a specification or stated outcomes (Harvey, 2004). If a researcher claims that a particular game is an appropriate choice for a particular study it is also appropriate to address the question, "Says who?" Given the great number of games available, it is no longer sufficient to claim that a particular game meets certain criteria without somehow supporting that claim in a verifiable way. A game that may be suitable for one sort of study may not be at all suitable for another. Even though critical and commercial success are both recognizable and accepted measures of a game's popularity, and popularity in turn gives some indication of that game's perceived quality as judged by players, developers, and game critics these are also highly subjective measures. Combining a number of different measures to come up with a single measure ensures that games that end up at the top of the

final list qualify as successful by more than one measure and have been assessed by more than one source. If, for example the main premise for examining commercial games in a study is to learn how games teach and otherwise support learning by 'studying the masters', there must be some way to convincingly determine that the games from which the final choices are made are of a stature that would qualify them as among the masterpieces.

Measures of Critical Success

Critical success is typically thought to include acclaim by professional critics which can also include winning various recognized awards. Critical success can come from many sources, and three different kinds are included in the approach described here. One is the game developers who actually design and build the games, and that voice is heard through the two primary industry professional organizations: The Academy of Interactive Arts and Science, and The International Game Developer's Association (AIAS and IDGA). Both hold annual award ceremonies where members nominate and vote for games in various categories. Another source of critical acclaim is the gamer-reviewer: someone who does not work for a games company or press agency but plays the games and is willing to contribute publicly available reviews of individual games. The third source is the official press. There are now a great many websites and magazines devoted to gaming and most contain review sections. Most will publish gamer reviews along side those written by paid correspondents. All three sources can be combined to create a more robust assessment than any single source can alone.

Measures of Commercial Success

There are many examples in literature, film and other media of works that have achieved critical acclaim but not commercial success and vice versa. Both are measures of success and with modern acceptance of player-reviews as one form of acclaim, both should be included in a list that claims to include 'good; or 'best' games. Commercial success is typically defined by sales, which in games is measured as units sold in a given year. There appears to be only one source for this data in the US: NPD® [1]http://www.npd.com/corpServlet?nextpage=entertainment-categories_s.html, which is also the source used by the ESA (Entertainment Software Association), and most other press agencies. NPD® is the primary source of video game sales and consumer information.

While it is acknowledged that commercial success is no assurance of quality it is an indication of popularity and inclusion in the top ten or twenty games in any given year is a significant achievement. There are thousands of titles released each year, and nearly 230 million games were sold in the US in 2005 (ESA, 2006). That means that less than 1% (possibly less than .05%) of these games makes it onto this best seller's list.

Combining Data

There is evidence to suggest that word of mouth, game demos, and reviews are all important factors influencing a decision to purchase a game (Dobson, 2006), but there is also evidence to suggest that review scores do not significantly affect game sales (Boyer, 2006). One possible consequence of this is that since there is no statistically significant relationship between sales and reviews, both values should be included in the selection process as neither one provides a complete picture alone. As already stated, we also know from literature, theater, and film, that popularity as evidenced through sales does not always match critical acclaim, yet both tell us something about that work's quality.

The sources of the values used to compile various lists are important and some effort should be made to identify the primary data sources whenever possible. For example all of the sales data the authors were able to find could be traced back to a single source: the NPD statistics. This means that there is no point in combining sales lists from multiple sources since they in turn all got their data from the same source – it simply amounts to counting the same data multiple times. This phenomenon should also be remembered when combining other data from multiple sources – we need to check where they got their data from.

1 The NPD Group has redesigned its website since this chapter was written, and data is no longer as accessible as it was. In 2006, the section of PD where game sales data was available was called ***NPD Funworld®***.

Data Fusion Methodologies for Combining Data

Coordinating Decisions Using Many Data Sources

Humans generally make complex decisions by considering many factors. There must be a way for the human mind to assign some kind of value to the reliability of the many sources of information to which it has access and to rank the relevance and significance of these sources for a particular choice that needs to be made. Unfortunately, self-reflection rarely makes available the actual mechanisms, and so researchers are left to use the methods of mathematics and systems analysis to devise algorithms for making decisions based on many data sources.

Many of the methods for decision making in this context come from the field of artificial intelligence, specifically pattern recognition. Consider a computer program that examines an image and looks for faces. One way to do this is to look for elliptical areas of basically a flesh color. Imagine that this works about 80% of the time. Now imagine a program that looks for eyes in an image, and draws a face around them. This may work 70% of the time. However, together, if one uses both of these methods on the same image, it may be possible to raise the success rate to 85-90%. Of course, the combination of sources of game rating information is a vastly different problem from the recognition of faces, but the same basic decision combination methods can be applied to both.

Data Fusion Methods

Data Fusion is the process of combining data from multiple sources into some form of coherent data set. Sources may be of similar types, such as multiple telescopes, or dissimilar types, such as optical, acoustic, radar, and infrared data. An important aspect of data fusion methods is the ability to deal with conflicting data. It is sometimes necessary to create temporary or interim results that the system can improve and change as the available data improves and changes. Much of the literature in this area discusses how to find targets (objects) using sensing devices. On the face of it, this sounds like the sort of solution that is needed for the problem at hand.

There are many sophisticated methods for data fusion in engineering applications, where the different sources of data are discrete sensors. For example, a Bayesian network (Jensen, 2001), also referred to as a belief network, is a statistical model that can represent a set of values or variables and their probabilities. Some researchers use Dempster-Shafer theory (Shafer, 1976) for data fusion, in which we can consider also the confidence we have in the probabilities assigned to the various outcomes as well as the probabilities themselves. Fuzzy logic (Zadeh, 1965) is another method that has been used for data fusion. This involves sets and logic and the idea of a degree of membership. An item is either a member of a set or not, as a general rule, but in a fuzzy set, an object can be a member to a specific degree. Traditional probability can indicate the likelihood of a car being parked in stall 52 or in stall 53. In reality some people park partly in both, and with fuzzy sets this situation can be described as the degree to which the car is in stall 52 and the degree to which it is in 53. This is obviously a better way to represent the real situation.

All of these methods deal well with data that is accurate to known likelihoods and tolerances. They do not, in general, handle unreliable data, decisions, or data from unknown sources with unknown properties. Thus, they will not be especially useful in dealing with the problem of using multiple sources of game success data. (although they would be very good at tracking missiles).

Types of Decisions

In order to devise a mathematical system for merging decisions from multiple sources, it is necessary to find a representation for the various kinds of decision that exist, and to devise a formalism for manipulating these decisions. As it happens this is a very old problem, and has some very old solutions (Farquharson, 1969; Straffin, 1980a; Straffin, 1980b). There are basically three different forms of decision, which are referred to as Type 1, Type 2, and Type 3. As an example on which to create understandable examples, consider the following

problem: there are three candidates for the chairman of a committee (C1, C2, and C3), and five people assigned the task of deciding which one shall be the chairman (P1 .. P5).

A Type 1 decision is a simple choice. For example, P1 may choose C1 to be chairman - that's a Type 1 decision. There are no options, no way to determine how significant the choice was, what the second choice might be, or how close the first and second choices were. The usual way to combine Type 1 choices is to treat them as votes, and usually a simple majority vote determines a winner. This means that whichever of the three candidates receives three or more votes wins (I.E. 3 is more than half of 5, and so is a simple majority). So if the votes are:

P1 chooses C1 P2 chooses C3 P3 chooses C1 P4 chooses C3 P5 chooses C1

The winner is C1, with 3 votes.

A Type 2 decision is a ranking of the set of options, from most to least acceptable. The basic idea is that each decider determines a relative order for the selections; for example, the first choice for chairman is person C2, the second choice is C3, finally the last choice is C1. This example would be output as a list: (C2 C3 C1). It is not necessary to know the degree to which C2 is more popular than C3. There is obviously more information here than in the case of a Type 1 decision, and making use of it should yield a more reliable composite result; that is, a choice that represents a good melding of information from all sources.

Type 2 decisions are what can be expected from game assessments, generally. The 'PC Magazine top 10", for example, would generally list the best ten games of a particular year in order, where the best game of that year would be first in the list. The problem to be addressed is easy to describe: given a collection of 'top 10' lists (where 10 could really be any number at all) from various sources, how can they be merged so that the overall top games can be identified?

The Borda count (also called Borda's method of marks) is an ancient scheme for resolving this kind of situation, in which each alternative is given a number of points depending on where in the ranking it has been placed (Black, 1958; Borda, 1781). A selection is given no points for placing last, one point for placing next to last, and so on, up to R-1 points for placing first in a list with R elements. In other words, the number of points (the weight) given to a selection is the number of classes below it in the ranking. The Borda winner is the selection with the largest count value.

Considering the ongoing example of five people selecting a chairman, let each person rank the candidates first to last. An example would be:

P1	C1, C3, C2
P2	C3, C2, C1
P3	C1, C3, C2
P4	C3, C2, C1
P5	C1, C3, C2

The count for Candidate C1 is 2 (from P1) + 0 (from P2) + 2 (from P3) + 0 (from P4) + 2 (from P5) = 6. However, C1 is not the winner anymore. The Count for C2 is 0+1+0+1+0 = 2, and for C3 is 1+2+1+2+1 = 7, so C3 is the winner. How did that happen? C3 had a higher overall level of support, never finishing last. However, it can be argued that a majority selection should always take precedence of a Borda count (Parker, 1999). This debate has been going on for centuries, and for the purpose of ranking and selecting games for experimentation is probably moot.

In the general case, each potential selection i=1,2,...R receives some number v_{i1} of first place votes, some number v_{i2} of second place votes, and so on. These are combined to give a desirability index (or Borda Count) D_i for each selection. The Borda method for computing Di is (Parker, 2001):

$$D_i = \sum_{j=1}^{R} W_j v_i \qquad \text{(EQ 1)}$$

The multipliers W_j are, in this case, just the number of selections having a lower rank than class j, or simply the value R-j. These values are sums across all lists/deciders to give an overall count, and the selection with the largest overall count is considered to be the best overall choice.

If the lists are different lengths then there is an obvious problem. The first place game in a list of 100 is given a count of 99, whereas the first place in a list of 10 has a value of 9. Is the former instance really 11 times better than the latter? No, probably not. The way this can be handled is to always assume that lists have a standard length, and then convert the scores into those for a list of that length. For example, assume that there are three ranked lists to be combined: one has 10 elements, one as 100, and the last has 20. Assume a standard list length of 100 - using the size of the largest actual ranked list means that it is easier to calculate the new scores. For each element in the list of 10, do the following:

1. Add 1 to the Borda count.
2. Multiply this value by 100/10 = 10.
3. Subtract 1.

This converts the count out of 10 into a count out of 100. In general, assume that the standard list length is L (100 in the above case) and the length of the list being converted is L_c (=10 above). For any Borda ranked element in the L_c list, there is a Borda count for that list - call it B_i. Then the new count for the standardized length L would be:

$$B = \frac{(B_i + 1)L}{L_c} - 1 \qquad \text{(EQ 2)}$$

B is called the normalized count. So, the first place element in a list of 10, having a Borda count of 9, would have a new count of (9+1)*100/10 - 1 = 10*100/10 - 1 = 100-1 = 99. This is what we'd expect. The second place element, having a count of 8, would have a normalized count computer as (8+1)*100/10 - 1 = 90-1 = 89. Each successive element in the list a Borda count that is 10 smaller than the previous. The distance between them is uniform, but is not 1 as in the case when the lists are all the same length. Indeed, the distance between adjacent elements in the new list is given by L/Lc.

Finally, a Type 3 decision has a numerical value associated with it. This value indicates how much confidence can be had in that decision, and so can be thought of as a probability or an acceptability value. It is rare in non-physical situations to have this form of information, and it is very unlikely indeed that game ratings will be associate with numerical values that can be normalized. Sales values, whether in dollars or units, are not easy to scale to a standard range, although raw values can be used if needed. The way to combine values in this case is to average the individual values from the sources.

Using More Than One Voting Strategy

If the use of more than one data source can be merged into a single, more reliable, set of data, then is it possible to use multiple, distinct voting or merging techniques to create a more reliable composite decision? The answer is yes, but to pursue this line of exploration requires a brief foray into some theory behind voting and decisions. Why is this a useful thing to do? It was mentioned in the discussion of the Borda count that there can be disagreements between methods, such as when the majority vote gives one answer and the Borda count gives another. A secondary issue is to deal with ties; the Borda count can result in identical counts for multiple decisions, and there should be a way of breaking these ties. Using votes, Borda counts, and other methods to create a single result is a practical way to solve the problems.

Using a secondary technique to resolve ties is an obvious thing to do. As a matter of policy, the majority criterion is adopted; that is (Straffin, 1980a):

If a majority of ranked lists have an alternative X as their first choice, a decision rule should choose X as the first ranked choice.

This is a weaker version of the Condorcet Winner Criterion (Condorcet, 1785):

If there is an alternative X which could obtain a majority of votes in pair-wise contests against every other alternative, a voting rule should choose X as the winner.

How do these pair-wise contests work? Simply by determining the rank of each choice in the relevant list, and declaring the winner to be the choice of highest rank. Indeed, the Condorcet criterion. Consider the following three rank lists from three different sources:

List 1	List 2	List 3
Metroid Prime	The Legend of Zelda: Ocarina of Time	The Legend of Zelda: Ocarina of Time
The Legend of Zelda: Ocarina of Time	Soul Calibur	Soul Calibur
Goldeneye 007	Metroid Prime	Metroid Prime
Call of Duty 2	Tekken 3	Resident Evil 4
Resident Evil 4	Goldeneye 007	Goldeneye 007

Pairwise contests between *Metroid Prime* and *Goldeneye 007* are all won by *Metroid Prime*: in list 1, *Metroid* is first to *Goldeneye's* third, and in list 2 and 3 *Metroid* is third to *Goldeneye's* fifth. In all cases, *Metroid* places higher than *Goldeneye*, and so *Metroid* is the Condorcet winner.

The Condorcet method is the method of choice, but it unfortunately tends to result in many ties. So, why not combine Condorcet with Borda? The so-called Black (Black, 1958) strategy chooses the winner by the Condorcet criterion if such a winner exists; if not, the Borda winner is chosen. This is appealing in its simplicity, and can be shown to have other important mathematical properties (Parker, 1995). There are other choices for voting strategies, all of which have their own advantages and disadvantages. None have the simplicity and reliability of the Black scheme, in general.

A Specific Example

Although the preceding discussion may give the appearance of being overly complex, the actual calculations and rankings need not be done by hand. One way to do this is by using a spreadsheet or database application. Lists of games and their relative rankings can be gathered for various sources. The amount of information to retain depends on the needs of the study. The author conducted an analysis of several games identified as 'masterpieces' in order to determine how these games helped players learn the things they needed to learn in order to succeed in the game. The rationale for the study was the claim that the best games *already* do a good job of supporting learning (Becker, 2006) and that we, as instructional designers can gain insight into the design of educational games by studying them regardless of their educational value. In a previous analysis conducted by the author a popular but purely entertaining game (*The New Super Mario Bros.*) was compared to an educational game (*Math Blaster, Master the Basics*) and although both were platform games, the commercial game supported the required learning within the game better than the educational game (Becker, 2007). In that example it was not ***what*** was learned that was of value to the study, but rather ***how the learning was facilitated***. The criterion for choosing a game for the masterpiece study was that the game indeed be one of the masterpieces and so it was necessary to justify that designation. Whether or not the game could be considered to be 'educational' in the sense that what was learned had curricular value was not of interest in this case – the goal was to identify how the game helped players learn what they needed to learn to win or get to the end of the game. It was important to include genre, platform, and ESRB (Entertainment Software Rating Board**)** rating information in this study, but these values may not be needed in other studies. The lists were initially compiled using a spreadsheet because of the ease of sorting and inclusion of formulas for calculating normalized counts. The spreadsheet also allowed for fairly simple editing such as switching columns around when raw data lists

differed. Once the raw lists were complete, they were copied to a database application because that allowed for simple production of reports that grouped, counted, and summarized values in various ways. Further, once the basic tables have been defined, they can easily be filled with new data. Once the final list was generated, only the top 100 games were kept. For the purposes of this study only games rated 'E' for everyone or 'T' for teen were considered acceptable as the ultimate goal was to be able to use techniques discovered in the games to help design educational games that would be used in schools. As a result all 'M' (mature) rated games were eliminated – these games often contain violence or other mature themes that would not typically be suitable in a classroom environment. As the author sought to study learning support in top games, all multiplayer games were also 'disqualified' as the goal of this study was to discover mechanisms in the ***game*** that helped players learn rather than how ***other players*** helped players learn. After all other qualifications specific to this study were met, the resultant list still included approximately 50 games. The study called for three games to be chosen, and it was then possible to choose three from this list with about as much confidence as possible that these games are indeed classed among the 'masterpieces'.

Although the specifics of which information to include on these lists will vary from study to study, the only information crucial to all is some way to uniquely identify a particular game, and a consistently calculated numeric value that can be associated with that game that indicates its rank in that particular list. In a different study attempting to discover something about learning communities for example, the number of fan sites, or number of postings to the official website could be among the values used to help rank games for suitability. One of the advantages of this approach is that it becomes possible to develop ranking mechanisms on various criteria and then combine them in a structured way to produce a single list of games best suited to the particular criteria of that study, and to do so in a transparent and verifiable way. Once such a list has been generated, the researcher can then feel free to choose from that list on the basis of personal preference and still be confident that the game meets the stated criteria as well. The one exception to the ability to choose from the final list according to preference might be if the study were attempting to discover something about the use of a game to which players were not attracted. Such a study might yield useful insights in the design of educational games as it can rarely be assumed that all learners will be equally motivated to engage with any particular game. For a study such as that it might help to rank games in a demographic way and then match games with players who do not meet the demographic requirements.

Once the ranking criteria have been developed and applied, the most time consuming part of the collection process is ensuring that games in the combined lists will be recognized by their names (or other identifier) and a certain amount of 'massaging' will be necessary to ensure that all occurrences of *The New Super Mario Bros.* for instance are recognized as the same game across all of the lists. Given that the compiled lists will be focusing on specific criteria such as "best of", the number of individual games named will be a small subset of the total games released, so while tedious, this task is neither difficult nor particularly error-prone when done by hand. Once the individual lists have been created the normalized count for each entry can easily be computed using a formula and then the final list can be compiled by generating a report. Lists should be limited in size although the exact length is somewhat arbitrary. If for example the list length is set to 100, then only games appearing in the top 100 of any list need to be included. If a list is not sorted in its raw form, it can easily be sorted by whatever criteria the researcher has determined to be important. In the case of "good" games, the review rating would be a value used in sorting. As long as the sources and dates of collection are recorded, the final list can be verified by other researchers, thereby lending credibility to the results.

Discussion

Do games have potential for use in educational contexts? Which games are best to use and under what circumstances should these games be incorporated into formal ad informal learning settings? If we design educational games for specific contexts, what exemplars should we use for the design and how will we judge them? We are only beginning to scratch the surface of how to answer these questions along with a great many others, and the methodology described in this chapter is one part of the process towards creating studies that are rigorous, and defensible.

As long as aspects of quality are part of the selection criteria there can be no truly objective method for choosing games likely to yield insights and results as required by scholarly study, attempts can still be made to ensure that the final list from which games are chosen includes those titles that a substantial number of informed individuals such as gamers and industry professionals would agree were games worthy of study. This not only helps to ensure scholarly integrity, but acknowledges industry expertise in a way that can help to foster improved communication between academia and game industry professionals.

The categorization of games by genre is far from clear-cut, yet if the genre of the game is important to a particular study then that classification too should be supported by confirmation in addition to the author's own claims. If a particular game cannot be placed easily in the required category (such as a First Person Shooter, Strategy, or Role-Playing Game), then perhaps a less controversial candidate should be considered. When games are chosen that have particular features, efforts must be made to show evidence that those games do indeed include the required features. If a particular game is identified as the 'best' or a 'good' candidate for a particular study, it is fair to expect justification.

It is unlikely that it will be possible to eliminate subjective measures of fitness when choosing games as the subjects of study, but that should not deter us from using a structured process designed to rank potential candidates so we can have some confidence that the game we use is appropriate. The methodology here allows criteria important to a specific study to be quantified, and provides a way of combining diverse measures to produce a single ranked list.

While acceptance of research with and on digital games continues to grow, it is still not seen as mainstream research and one way to help ensure that our research is given the same scholarly consideration as more established fields is to pay attention to details. One of those details is in research design. Research design must include plans that are lodged in "ideas well grounded in the literature and recognized by audiences (eg. Faculty committees) that read and support proposals for research" (Creswell, 2003, p. 3).

Appendix: Sources of Game Reviews and Other Data

- **The NPD Group** (formerly National Purchase Diary) is a leading provider of consumer and retail market research information. This is the primary source of game sales data in Canada and the U.S. http://www.npd.com/corpServlet?nextpage=entertainment-categories_s.html [sample data: http://www.npd.com/press/releases/press_070119.html]

Professional Industry Organizations

- **Academy of Interactive Arts and Science** (AIAS). Holds annual award ceremonies where members make nominations that are then voted upon in much the same way as the Academy of Motion Picture Arts and Sciences. From their site: "*Interactive Achievement Award recipients are determined by a vote of qualified Academy members. As such, selection as an Interactive Academy award finalist or recipient represents the strongest possible peer recognition. No person may become a voting member of the Academy unless he or she can demonstrate a threshold level of experience and professional credits in the industry. Interactive Academy voting is secret, conducted on-line, and supervised and certified by our partners at eBallot."* http://www.interactive.org
- **International Game Developer's Association** (IGDA): This is the primary professional developer's association, which holds an annual awards ceremony. Any IGDA member in good standing is eligible to nominate a game and vote for finalists. Five finalists are chosen by the advisory board in each category. http://www.gamechoiceawards.com/

Press and Gamer Review Sites

- **Game Critics Awards**: "*Game Critics Awards, an independent group of journalists from 36 leading North American media outlets that cover the videogame industry. Each year the Game Critics Awards present its Best of E3 awards."* http://www.gamecriticsawards.com/

- **Metacritic**: "*Metacritic® compiles reviews from respected critics and publications for film, video/dvd, books, music, television and games. Our unique Metascores® show the critical consensus at a glance by taking a weighted average of critic grades.*" http://www.metacritic.com
- **IGN**: (Independent Game Network) Maintains a Top 100 list, as well as an Editor's Choice list. Both lists focus on recent releases. http://www.ign.com/
- **GameSpot**: A C|NET organization that provides both user and paid reviewer information on games. It maintains a Top Games list that rates games on a 10 point scale. http://www.gamespot.com/
- **Gamespy**: http://archive.gamespy.com/
- **MobyGames**: A Community contributed site that is building a comprehensive list of all computer and videogames. Mobygames maintains Best Of lists which are based on user votes. Scores are listed out of five and include a count of the number of votes that were cast. This list changes in response to user contributions. http://www.mobygames.com

References

Becker, K. (2006). Pedagogy in Commercial Video Games. In D. Gibson, C. Aldrich & M. Prensky (Eds.), ***Games and Simulations in Online Learning: Research and Development Frameworks*****: Idea Group Inc.**

Becker, K. (2007). ***Battle of the Titans: Mario vs. MathBlaster.*** **Paper presented at the 19th Annual World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA), 2007, Vancouver, Canada, June 25- June 29, 2007.**

Black, D. (1958). ***The Theory of Committees and Elections*****: Cambridge University Press.**

Borda, J.-C. (1781). Memoire sur les elections au scrutin. ***Histoire de l'Academie Royale des Sciences*****.**

Boyer, B. (2006). ***Survey: Game Score-to-Sale Theory Again Disproven*****. Retrieved Nov. 14 2006, from http://www.gamasutra.com/php-bin/news_index.php?story=10924.**

Condorcet, M. d. (1785). ***Essai sur l'application de l'analyse a la probabilite des decisions ren¬dues a la pluralite des voix*****. Paris.**

Creswell, J. W. (2003). ***Research design : qualitative, quantitative, and mixed methods approaches*** **(2nd ed.). Thousand Oaks, Calif.: Sage Publications.**

Delgado-Rodríguez, M. (2001). Glossary on meta-analysis. ***Journal of Epidemiology and Community Health, 55*****, 534-536.**

Dobson, J. (2006). Survey: 'Word Of Mouth' Most Important For Game Buyers [Electronic Version]. ***Gamasutra*****, 2006. Retrieved Nov. 14 2006 from http://www.mi6conference.com/Magid_MI6.pdf.**

ESA. (2006). Essential Facts About the Computer and Video Game Industry: 2006 Sales, Demographics, and Usage. Retrieved Jun 25 2006, 2005, from http://www.theesa.com/archives/files/Essential%20Facts%202006.pdf

Farquharson, R. (1969). ***Theory of Voting*****. New Haven: Yale University Press.**

Glesne, C. (1999). ***Becoming qualitative researchers : an introduction*** **(2nd ed.). New York: Longman.**

Harvey, L. (2004). Analytic Quality Glossary [Electronic Version]. ***Quality Research International*****. Retrieved July 13 2007 from http://www.qualityresearchinternational.com/glossary/.**

Hopson, J. (2006). We're Not Listening: An Open Letter to Academic Game Researchers [Electronic Version]. ***Gamasutra*****. Retrieved Nov. 10, 2006 from http://gamasutra.com/features/20061110/hopson_01.shtml.**

Jensen, F. V. (2001). ***Bayesian networks and decision graphs*****. New York: Springer.**

Civilization III **Meier, S. (Designer) [Game] Firaxis (Developer) (2001) [Computer Game] [Windows] Published by Infogrames.**

Parker, J. R. (1995). ***Voting Methods for Multiple Autonomous Agents.*** **Paper presented at the ANZIIS '95, Perth, Australia, Nov. 27, 1995.**

Parker, J. R. (1999). ***Multiple Sensors, Voting Methods, and Target Value Analysis*****. Paper presented at the Signal Processing, Sensor Fusion, and Target Recognition VIII, SPIE Aerosense, Orlando, Florida, April 6-9.**

Parker, J. R. (2001). Rank and Response Combination from Confusion Matrix Data. ***Information Fusion, 2*****(2), 113-120.**

Shafer, G. (1976). ***A mathematical theory of evidence*****. Princeton, N.J.: Princeton University Press.**

Squire, K. (2003). ***Replaying History: Learning World History through playing Civilization III.*** **Unpublished Doctor of Philosophy, Indiana University.**

Stake, R. E. (1995). ***The art of case study research*****. Thousand Oaks: Sage Publications.**

Straffin, P. D. (1980a). ***Topics in the theory of voting*****. Boston: Birkhäuser.**

Straffin, P. D. J. (1980b). ***Topics in the Theory of Voting*****. Boston: Birkhauser.**

Williams, D. (2006). A (Brief) Social History of Video Games. In P. Vorderer & J. Bryant (Eds.), ***Playing Computer Games: Motives, Responses, and Consequences*****. Mahwah, NJ: Lawrence Erlbaum.**

Williams, D., & Skoric, M. (2005). Internet Fantasy Violence: A Test of Aggression in an Online Game. ***Communication Monographs, 72*****(2), 217-233.**

Wolf, M. J. P. (2001). ***The medium of the video game*** **(1st ed.). Austin: University of Texas Press.**

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353.

Key Terms and Their Definitions

Backyard Research
A term used to describe research conducted in an environment in which the researcher already holds another role. An example would be a classroom teacher conducting research within her own classroom.

Borda Count
A well-known methodology for assigning scores to multiple ranked lists that can then be combined to produce a single ranked list that incorporates the results of the other lists.

Data Fusion
Data Fusion is the process of combining data from multiple sources into some form of coherent data set.

ESRB (Entertainment Software Rating Board)
A non-profit, voluntary regulatory body that assigns ratings and enforces advertising policies in interactive entertainment software. This is the body responsible for defining the ratings found on most commercial videogames. The ratings are effected by the level of violence in the game as well as the subject matter. A summary of the ratings is included here, but for a detailed description see the ESRB website: http://www.esrb.org/ratings/index.jsp

- EC (Early Childhood) suitable for ages 3 and older.
- E (Everyone) have content that may be suitable for ages 6 and older. EVERYONE 10+ E10+ (Everyone 10 and older) have content that may be suitable for ages 10 and older.
- T (Teen) have content that may be suitable for ages 13 and older.
- M (Mature) have content that may be suitable for persons ages 17 and older.
- AO (Adults Only) have content that should only be played by persons 18 years and older. RP (Rating Pending) have been submitted to the ESRB and are awaiting final rating. (This symbol appears only in advertising prior to a game's release.)

First Person Shooter
A game played from the first person perspective where the game space is seen from a position slightly behind and over the shoulder of the character being played. The player takes on the role of one of the game characters and the primary mode of game play involves the use of weapons that are used to shoot opponents.

Game Play
The experience of playing a game.

IDGA (International Game Developer's Association)
The premier association for people involved in the game development industry. See more at: http://www.igda.org/

Longitudinal Study
A research study that involves repeated observations over long periods of time, usually including the same items which are often correlated.

Normalized Count
The normalized count is the count in a list divided by the total number of observations. In the method described in this chapter, the normalized count is the score associated with a game that relates to its position in that list. The number is normalized so that the first place game of any list will have the same score, thereby contributing the same weight towards its total. In other words, it makes the first-place game in each list worth the same regardless of the actual length of the list.

Qualitative Meta-Analysis
An analysis of the methods used in a collection of studies.

Role Playing Game
A game usually played from the first-person perspective where the player pretends to be one of the characters in an unfolding story. Roles may be assigned with little flexibility as for example playing James Bond in *Goldeneye 007* or with a great deal of player input such as in *Workd of Warcraft* where players may choose the gender, race, and profession of their character as well as many other variables.

Voting Strategy
A mathematical system for merging decisions about choices among several alternatives that come from multiple sources.