A Bechdel Test for Computer Games?

Mio Firkins
Heriot-Watt University
Edinburgh, Scotland
EH14 4AS
mio.firkins@gmail.com

Ruth Aylett
Heriot-Watt University
Edinburgh, Scotland
EH14 4AS
r.s.aylett@hw.ac.uk

ABSTRACT

This paper reports work addressing the research question “could a Bechdel-like test be an indication of how women are represented in video games?” through developing such a test for video games: the *Indicative Representation of Women In Games (IRWiG)* test. We describe its development process: a multidisciplinary approach combining a literature review with the development and application of ontologies representing the constructional elements of games that relate to portrayal of women. The IRWiG test was evaluated through a public evaluation survey and an internally-conducted games analysis. The test proposes four criteria for analysing a female character: her character development, appearance, abilities and relevant stereotypes, and the skippability of content in which she is active. An overall agreement rate of 74% was found between between users’ opinions of how a woman is represented in a game, and the application of the IRWiG test.

Keywords

Games analysis, sociology, gender studies, women’s studies, game design, character design, character development, narrative

INTRODUCTION

Women have historically been excluded from fair and equal representation in media, contributed to in-part by pervasive negative stereotypes (Sharda 2014). The Bechdel-Wallace test, or “Bechdel test” (Bechdel 1985), has highlighted the need for better representation of women in film media, however no equivalent exists for video games, where the problem of stereotypical and problematic representation is widespread.

Because the Bechdel test does not address the interactive nature of video games, the existing test cannot be directly transferred. However, there is potential for the formulation of more robust Bechdel-like tests that could be applied to interactive media such as video games. Designing a Bechdel-equivalent test requires a multidisciplinary approach, combining elements of gaming research and sociology, particularly gender studies. Evaluating the original Bechdel test and its contemporaries can provide a set of test design principles to use and avoid for the IRWiG test, and establish how representation as a whole is assessed in video games.

Proceedings of DiGRA 2024

BACKGROUND

The Bechdel test

First published in a comic strip in 1985 by Alison Bechdel, the Bechdel-Wallace test, generally referred to as the Bechdel test, creates a discussion about the representation of women in film media by offering three criteria a film must pass:
1. Are there two women in a film?
2. Do these two women have a conversation with each other?
3. Is this conversation not about a man?

The test gained greater reach than expected (Bechdel 2015), becoming widely used and contributing to the broader discussion of underrepresentation of women in films (Koivunen et al. 2014).

Despite this success, the Bechdel test has limitations, largely relating to its simplicity, limited scope, and Boolean threshold. One approach to improving the original Bechdel test is to create multiple tests that span a range of aspects in a film, from characters in a game to a game's production, as used in The Next Bechdel (Hickey et al. 2017). This consists of 12 tests that can be broken-down into 4 categories: crew, non-white women, protagonists, and supporting cast.

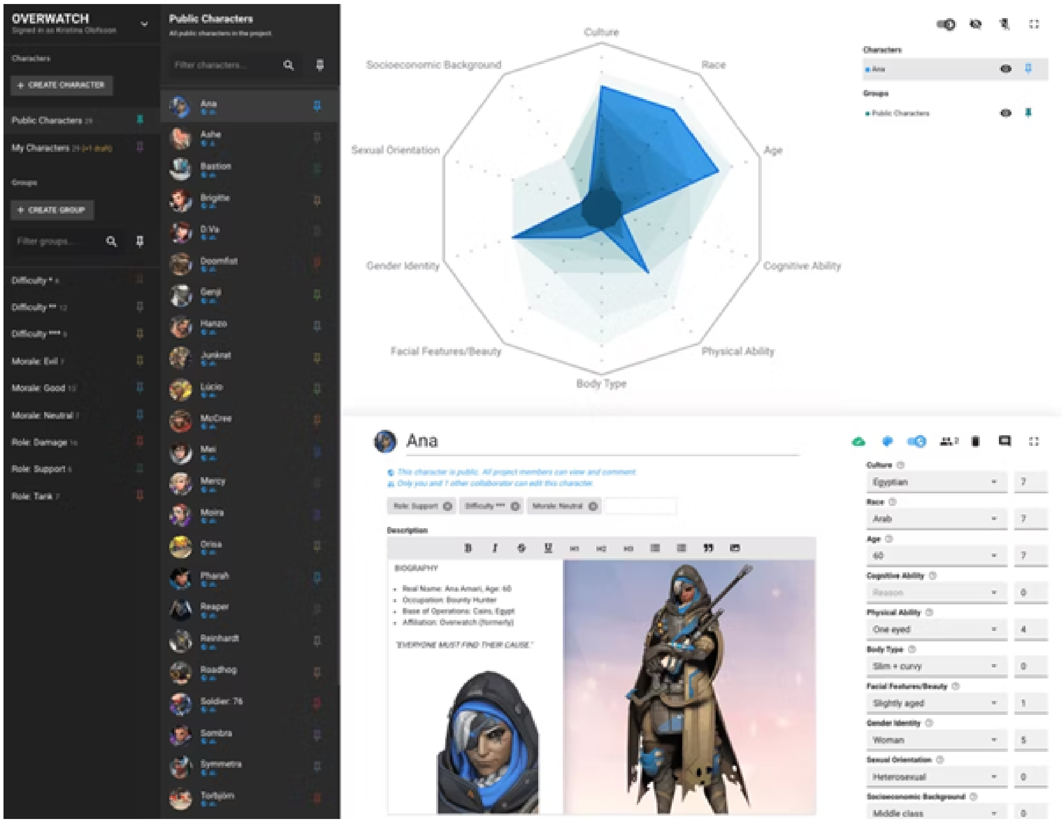
The Next Bechdel culminates with the "Feldman Score", which combines elements of each of these categories. Films pass by scoring above a pre-specified value rather than passing a set of criteria. However, by not requiring a minimum per-category score, a film can pass overall, despite failing scores in some categories, which could be rectified by breaking down these elements of the Feldman Score into separate sections, further emphasising the multidimensionality of representation. One should also note that each test within The Next Bechdel was designed by different people with different backgrounds and specialisms, with some requiring in-depth data collection such as statistics for film crew, and/or the proportion of women in crowded establishing shots. This may reduce applicability.

While Hickey et al. state that "We need more than one test", they do not discuss the practicalities of applying multiple tests. Whether classing a film as having good representation depends on passing a set number of tests out of the total, or passing in each category is left ambiguous but is an important distinction. In their own findings, the highest number of passes relate to the tests on film protagonists, and the fewest to the crew-related tests.

Representation in gaming

Bechdel-like tests in film is a widely researched topic (O'Meara 2016; Bouchat 2019). Our work suggests that the same cannot be said for video games, with few proposed methods of evaluating representation when considering Esposito's definition of a video game as ”a game which we play thanks to an audiovisual apparatus and which can be based on a story" (Esposito 2005). In this case, an audiovisual apparatus would be an electronic computation device with accompanying input and output peripherals. One outlier method that has been suggested by industry lies in the "Diversity Space Tool", announced by Blizzard Entertainment in May of 2022 (Alt 2022). This was designed to support the creation of non-stereotypical characters in their games, but was met with widespread ridicule and criticism (Francis 2022).

Using characters from the already-released title Overwatch (Blizzard Entertainment, 2016) to demonstrate the tool, it was seemingly aimed at Blizzard's future titles such as Overwatch 2 (Blizzard Entertainment, 2022) and Diablo IV (Blizzard Entertainment, 2023), see Figure 1.



Ana

This character is public. All project members can view and comment.
Only you and 1 other collaborator can edit this character.

Role: Support · Difficulty *** · Morale: Neutral

Description

BIOGRAPHY

- Real Name: Ana Amari, Age: 60
- Occupation: Bounty Hunter
- Base of Operations: Cairo, Egypt
- Affiliation: Overwatch (formerly)

"EVERYONE MUST FIND THEIR CAUSE."

Culture	Egyptian	7
Race	Arab	7
Age	60	7
Cognitive Ability	Reason	0
Physical Ability	One eyed	4
Body Type	Slim + curvy	0
Facial Features/Beauty	Slightly aged	1
Gender Identity	Woman	5
Sexual Orientation	Heterosexual	0
Socioeconomic Background	Middle class	0

Figure 1: The Diversity Space Tool: screenshot from Blizzard's PR announcement prior to removal (Alt 2022; Cerafica 2022).

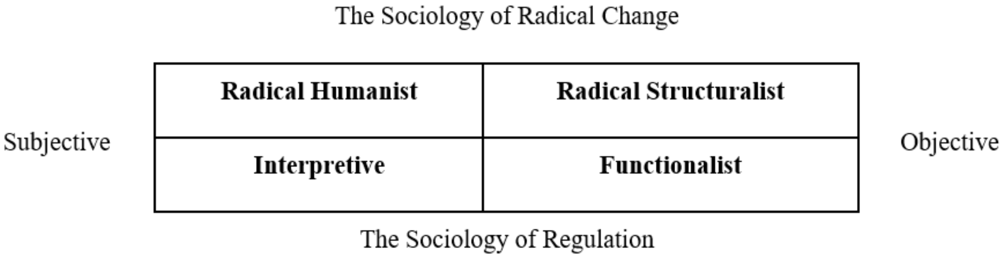
The tool took various character attributes such as age, race, sexuality, physical ability, and scored them on a 0–10 scale. Scores were plotted on a radar chart to display the character's overall "diversity". One point of concern in this framework is tokenism. While Blizzard argues "the Diversity Space Tool can clearly delineate between token characters and true representation", turning a character into numbers on a scale is tokenism: "a series of broad categories to plonk characters in simply to […], tick boxes. That's what tokenism is" (Sterling 2022). Though it is not that their developers avoided trying to include true diversity in their projects, with one developer having stated when the tool was revealed: "I swear our own company tries so hard to slaughter any good will the actual devs who make the game have built" (spellissa 2022).

The Diversity Space Tool has a "Gender Identity" variable on its radar chart output, implying some gender identities are more diverse than others. The cisgender woman Ana is given a Gender Identity score of 5, halfway along the gender diversity scale. Whether a cisgender man would score 0, or a transgender woman score the same as cisgender women, or higher, is unclear. Non-binary identities add further confusion. Many other examples could be conceived, showing the problematic nature of pitting different aspects against each other in an attempt to find the most diverse character (Tassi 2022).

The backlash to the King Diversity Tool demonstrates that tests that quantitatively and comparatively assess diversity are difficult to design correctly. A Bechdel-equivalent test should avoid numerically comparing types of people or trying to set up "objective" tier lists of minorities. While our work focuses on representation of women, rather than other aspects of diversity, these are pitfalls to be avoided by anyone seeking to test or improve diversity.

Objectivity vs subjectivity in games analysis

Methods of games analysis have deep roots in social theory, which itself has been debated more widely for at least the last 40 years. In Sociological Paradigms and Organisational Analysis, Burrell and Gibson (1979) proposed 4 primary paradigms of thought in a 2x2 matrix: Functionalist, Interpretive, Radical Humanist, and Radical Structuralist, see Figure 2.



Radical Humanist	**Radical Structuralist**
Interpretive	**Functionalist**

Figure 2: Four paradigms for the analysis of social theory (Burrell 1993)

For our work, the most relevant axis is objective vs subjective, as this can be applied to contexts outside of social theory (the authors note, without going into detail, that these paradigms can be utilised in wider contexts). Burrell & Gibson argues that each paradigm is mutually exclusive, that how one goes about analysis defines which paradigm they belong to, and that changing paradigm is both difficult and unlikely. This raises issues for a multifaceted approach, as "perspectives and subsidiary theories relevant to intersubjective phenomena cannot be accommodated" (Louis et al. 1983). Despite this, Burrell & Gibson's paradigms have gained widespread adoption in related disciplines (Louis et al. 1983; Poonamallee 2009).

The field of games research and the scientific analysis of games is much newer, having gained momentum as recently as 2001 (Aarseth 2001), and so its research methodology is still maturing. Debates around choice of methodologies have paralleled such discussions within social theory, with similar conceptual categories being addressed (Copier 2003). However, discussions in games research lay more emphasis on how best to incorporate other fields of study and which fields are most relevant. Despite this, there is general agreement that creating strict inclusion/exclusion criteria for games research will only undermine it, and that allowing for fluid methods of analysis works best for the medium. This flexibility is elaborated on by Carr et al., (2004) who used early games-research methodology on several role-playing games of the time, later succinctly stating games researchers should be "researchers who play, players who research."

Since this is a games research project, albeit focusing on gender representation, these suggest that researchers should have first-hand video game experience. When selecting which games to test against the current Bechdel test and the later IRWiG test, selection should be based on participants' direct experience. It is also important to gather a range of game types based on criteria such as release dates and genres, which is also relevant to the later survey created for evaluating the IRWiG test. Many game studies have focussed on specific genres (Carr et al. 2004; Fortim and Moura Grando 2013), not appropriate for the creation of a widely-applicable test.

RESEARCH METHODOLOGY

Principles of Creating and Evaluating the IRWiG Test

The Bechdel-like IRWiG test relies heavily on the results of prior work conducted by the authors (Firkins 2023), which identified six elements that heavily-relate to how women are represented in video games: Abilities & Stereotypes, Plot & Themes, Character Appearance, Dialogue, Skippability of Content (being whether content can be missed or skipped by a player while still reaching the end of a game's main narrative or end credits e.g. side quests or optional content), and Character Development. The first five elements were identified by both the paper authors and by participants of a public survey (n=52), with Character Development being identified by a subset of respondents as relating to all top-level elements of games. The creation of the IRWiG test was also informed by background literature around pre-existing tests for films (O'Meara 2016; Bouchat 2019; Hickey et al. 2017). Like The Next Bechdel, which split up multiple topics into smaller tests, we decided to test ontological components individually, in turn making selecting and iterating each criterion in the IRWiG test easier and faster. Errors we aimed to avoid included creating a test that tokenises characters or allocates types of people different point values to imply that different minorities are a "better" representation than others. Criteria in the IRWiG test should avoid having numeric scores other than a binary pass/fail. While the Feldman Score does have varying final scores, it is composed of criteria that have a single point value for pass or fail. Moreover it does not rank films with (e.g.) 7 points higher than those with 5, meaning only the threshold matters. The IRWiG test should mirror this binary pass/fail system.

When developing an evaluation methodology, attempts to mitigate selection bias were addressed via pre-selecting a set of games to be tested that varied in sales numbers, genres, release dates, and development studio sizes. Games in this list are limited to ones that the primary author has direct experience with, to leverage personal insight on how representation is handled in a given title, and to avoid differences in perspective that might result from a lack of direct experience (Aylett and Louchart 2007). This list was used as part of a larger public evaluation via survey, conducted to partially mitigate the lack of insight from field experts (Zhao and Li 2009), and to simulate a widespread use of the IRWiG test akin to the Bechdel test in media discussions. This also applies the principle of Carr et al.'s (2004) conclusions that different people working together only benefits the research being conducted. While Carr et al. discussed this in the context of researchers of differing methodologies, applying a test to video games is more understandable to laypersons. The general public will also be able to give feedback on the IRWiG test, therefore the high barrier of entry that applies in research-heavy, specialised topics does not exist. This allows for identifying potential points of failure in the test were it to receive overall negative feedback in an initial evaluation.

Due to the exploratory nature of this research, factors such as cultural and language differences, historical contexts of games' release dates or in-game settings, or how trends in the video game landscape affect game development decisions, were not considered as part of this paper. While they play a large role in how games are made and perceived, it was decided that those topics would increase the scope of this investigation too much, and would be a better fit as their own subsequent topics of research or further study.

Creating the IRWiG Test

As stated above, the six ontological elements of the IRWiG test were those assessed as significantly impacting how women are represented in games in an earlier study (Firkins 2023). Each characteristic is discussed below and evaluated for its inclusion in the IRWiG test.

"Character Development"

A woman character having a development arc was a commonly identified element of the survey-based ontology in the prior study (Firkins 2023), both in terms of appearing under all three top-level elements and frequency of inclusion by participants. Character development can occur within many circumstances, so placing too many restrictions or additional clauses on this criterion could produce a less-effective test. However, it was identified in the same prior study that the most overwhelmingly frequent response (n = 12) to what character traits correlate to a "strong" woman was independence. Adding a qualification that she needs to be able to face and solve some challenge without relying on someone else (irrespective of gender identity), adds specificity without creating additional restrictions. A proposed wording for this criterion was therefore: "Does she undergo character development, or has struggles/obstacles she overcomes without wholly relying on another character?"

"Abilities & Stereotypes"

Abilities and stereotypes were separate ontology elements, but can be closely tied by linking gameplay abilities to both character and gameplay stereotypes. Stereotypes, tropes, and archetypes aren't inherently bad, as they help people to heuristically sort information into less overwhelming collections (Mather et al. 1999); the issue arises when stereotypes become what defines a character, and said character only exists as a set of stereotypes. In order to push for characters to not be wholly defined by basic stereotypes, while still allowing for their use in a limited capacity, this criterion was worded as "Does her characterisation/gameplay role extend beyond one or two widespread stereotypes?"

"Character Appearance"

A criterion based around a character's appearance could be formulated without nuance by asking "is the character sexualised", but this would ignore both the long-standing debate in games and film, and that sexualisation is not an immediate indication of bad representation (Lerum and Dworkin 2009).

However, the significant role of context is double-edged; a believable justification of a character is often what makes the difference between what the general audience typically considers good or bad sexualisation. To account for different sources of justification, the criterion was formulated as: “Is her outfit practical for her role, and if not, is it justified narratively, thematically, or by her characterisation?” This wording still has some issues with specific examples of women in games, like Quiet in Metal Gear V: The Phantom Pain (Konami, 2015), where sexualisation that is technically justified by the game can be viewed as over-the-top fanservice for the player, or can be up to player interpretation such as the titular character in the Bayonetta games (PlatinumGames, 2009, 2014, 2022). However, creating clauses for further specific cases like these would both result in diminishing returns, and reduce the widespread applicability of this criterion.

“Skippability of Content”

Interactivity sets video games apart from other screen-reliant forms of entertainment like films or television (Grodal 2000), and in many games this is accentuated by containing optional content the player can choose to participate in or not. In some games such as Xenoblade Chronicles 3 (Monolithsoft, 2022), main characters have optional side quests during which they undergo character development or have important personal moments, which players may miss out on. This optionality can entirely alter how players perceive a character by the end of a game, and there is a risk that all meaningful narrative content in which women have active roles might be placed into optional or hard-to-find content. It is arguably important that this kind of narrative content sits within the main campaign. This is not to say that all narrative content with women playing an active role should be non-optional – this may be impractical – but it is important that it is not all relegated to sidelines or optional questlines. The criterion was therefore formulated as: “Does she play an active role in the main story and isn’t hidden to only optional/side content?”

“Dialogue”

Dialogue is the major basis for the original Bechdel test, and for a large percentage of films this is reasonably applicable, though note it cannot always be fairly applied to silent films, documentaries, or those with small casts. The same issues apply to video games, if not more so, with many titles featuring fourth-wall-breaking dialogue at the player or a silent stand-in for the player instead of in-world conversations, internal monologues, or including no spoken dialogue at all. Due to the number of games that would unfairly fail a Bechdel-equivalent test with dialogue-focussed criteria and to allow for a larger variety of games to be tested, this criterion was excluded from the IRWiG test.

“Plot & Themes”

A character’s relevance to the plot in a game was implicitly tested by the two criteria on Character Development and Skippability of Content, and a separate criterion relating specifically to plot and themes might overcomplicate the IRWiG test. We therefore chose not to include a criterion on plot and thematic relevance.

The IRWiG test

The backbone of the IRWiG test was a combination of background research and a comparison of ontologies to find a list of common characteristics that were found to be relevant to how women are portrayed in games, and criteria around these characteristics were designed and culminated in the following test:

A game passes if a woman in it meets these criteria:
- Does she undergo character development, or has struggles/obstacles she overcomes without wholly relying on another character?
- Is her outfit and physique practical for her role, and if not, is it justified narratively, thematically, or by her characterisation?
- Does she play an active role in the main story and isn't hidden to only optional/side content?
- Does her characterisation/gameplay role extend beyond one or two widespread stereotypes?

To pass the IRWiG test, a game must meet all of these criteria, and a character being tested does not need to be the main character or a playable character.

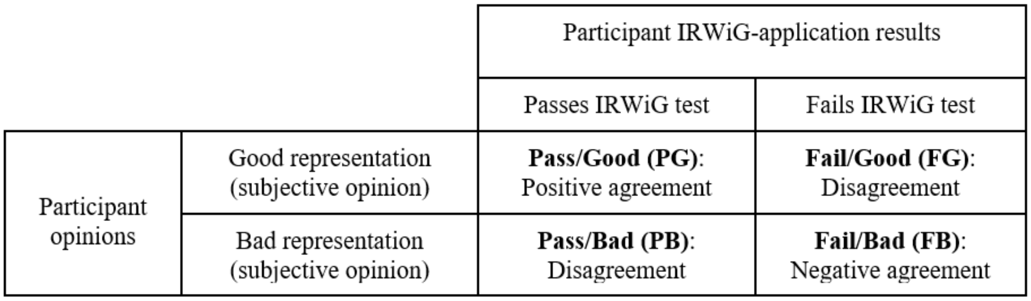
Designing an evaluation strategy

The finalised IRWiG test required testing in conditions similar to the use of the original Bechdel test in the real world. The research question "could a Bechdel-like test be an indication of how women are represented in video games?", was addressed via a public survey to examine the agreement between respondent opinions on representation in a set of games versus application of the IRWiG test to the same games. For each game tested, respondents could answer any one of five options, four of which correspond to a value in Figure 3, plus an additional ”Haven't played/unsure” option. The options ”Passes test; has good representation” and ”Fails test; lacks good representation” show agreement, while ”Fails test; has good representation” and ”Passes test; lacks good representation” show disagreement.

To avoid participants being overwhelmed with too large groupings of questions, games were broken down into sets of 6 games. A total of 60 games were tested by each participant, over 10 genres. Open-ended questions for participants to comment were added for every genre, where the responses were analysed by summarising responses into a set of shared keywords without altering response meanings, and patterns in keyword frequencies identified. Games being tested were chosen based on the games list tested internally, and from a survey conducted in prior work (Firkins 2023); games that were mentioned several times as having good representation were added, and from these 2 lists combined, the game genres were identified. Gaps in these genres that did not already have 6 games in them were filled in by games with high sales numbers (Valve 2005, 2023; Nintendo, 2019, n.d), high player counts on the PC platform Steam (SteamDB 2013-2023), or high review counts on Backloggd.com (*Top popularity* 2019-2023).

A final section in the survey was a general evaluation of the IRWiG test itself; participants were asked about each criterion's effectiveness in determining whether a game had an example of good female representation, and if they felt that there were any major aspects it failed to consider.

The survey underwent one round of piloting with five preselected participants, and all suggestions were minor edits. One pilot participant suggested changing the genre classification for some games, and two pilot participants had no complaints at all. The two other pilot respondents suggested small changes to the instructions to make points clearer, such as setting some words in bold. When asked whether too many genres were being tested, two pilots stated no, and two others suggested removal of one or two, but noted that they did not feel the current total was excessive. As no major concerns were stated with the number of genres, none were removed. This also kept the evaluation of the IRWiG test as a genre-agnostic test, much like the Bechdel test. The survey was also approved by the Maths and Computer Science Ethics Board at Heriot Watt University (Edinburgh, UK), and participant data was anonymised for GDPR-compliance. Following the pilot study, participants were recruited via a Microsoft Forms link posted in several online communities (video game-centric Discord servers, academic email chains, and relevant Reddit communities), and offline via posters displayed around the Heriot Watt University campus. To discuss how well the IRWiG test performed, answers were counted for each game and genre of games, and frequency statistics was conducted (excluding “haven’t played/unsure”). From the sets of responses given, correlations were conducted on how participants scored games via the IRWiG test against their own views on how each game represented women. For the test to be effective, the desired outcome was a high proportion of positive and negative agreement out of all responses.



		Participant IRWiG-application results	
		Passes IRWiG test	Fails IRWiG test
Participant opinions	Good representation (subjective opinion)	**Pass/Good (PG)**: Positive agreement	**Fail/Good (FG)**: Disagreement
	Bad representation (subjective opinion)	**Pass/Bad (PB)**: Disagreement	**Fail/Bad (FB)**: Negative agreement

Figure 3: Table of evaluation survey response meanings

When being asked for feedback, respondents were first given a 5-point Likert scale with which to measure how useful they found each of the four criteria in determining if a game possessed good or bad representation of women. They were then given the opportunity to provide open feedback.

SURVEY RESULTS

This survey ran for three weeks in early 2023. In total there were 63 participants; 30 (48%) identifying as men, 20 (32%) as women, and the remaining 13 as non-binary or other identities. Most participants were from North America or Europe (92%), with five exceptions; three of whom did not disclose their location.

Few participants identified as irregular game players, possibly due to the requirement for participants to have some experience with games; >40% of participants play 10+ games/year.

To discuss how well the IRWiG performed, each of the four non-empty answers were counted for each game and genre of games, and frequency statistics was conducted on them (occurrences of “haven’t played/unsure” were removed and not counted). From the sets of responses given, it was possible to find correlations in how participants used the IRWiG test against their own views on how games represent women. For the test to be effective, the

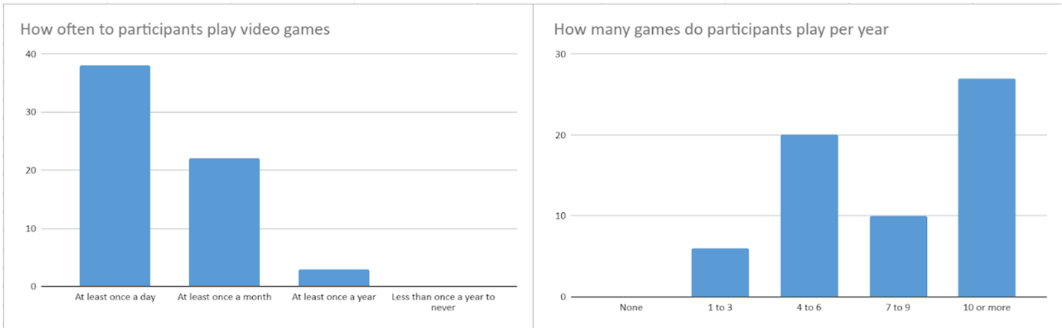


Figure 4: Participants' responses to the questions "How often do you play video games" and "How many games do you play per year".

desired outcome was a high proportion of positive and negative agreement out of all four responses, see Figures 5 and 6.

Responses to open-comment questions were summarised and collated via the same methods to similar questions in previous work (Firkins 2023). This generated a frequency table of response themes and general views in responses without altering meaning as much as possible.

IRWiG test results

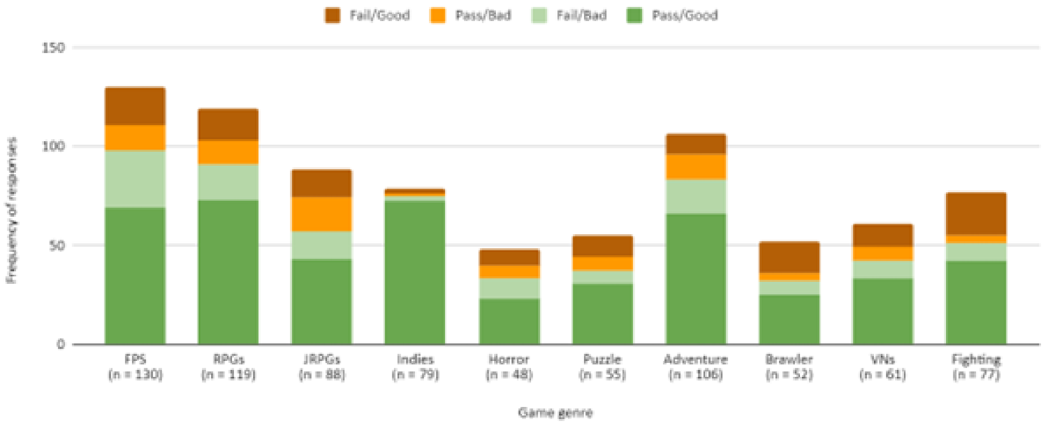


Figure 5: Frequencies of participant responses when using the IRWiG test, collated by genre (excluding "Haven't played/unsure").

Excluding "Haven't played/unsure", 815 responses were collected; 477 pass/good, 85 pass/bad, 131 fail/good, and 122 fail/bad. The rate of games passing the IRWiG test ranged from 56–95% by genre, and the rate of games being considered to have had good representation of women between 63–92%. There was a total agreement rate between participants' IRWiG test results and their opinions how well games represent women of 74%. All individual genres had agreement rates of >60%, with the Brawler genre scoring lowest (62%), and the Indie genre scoring highest (95%).

It should be noted that due to the exploratory nature of this research, and the limited num-

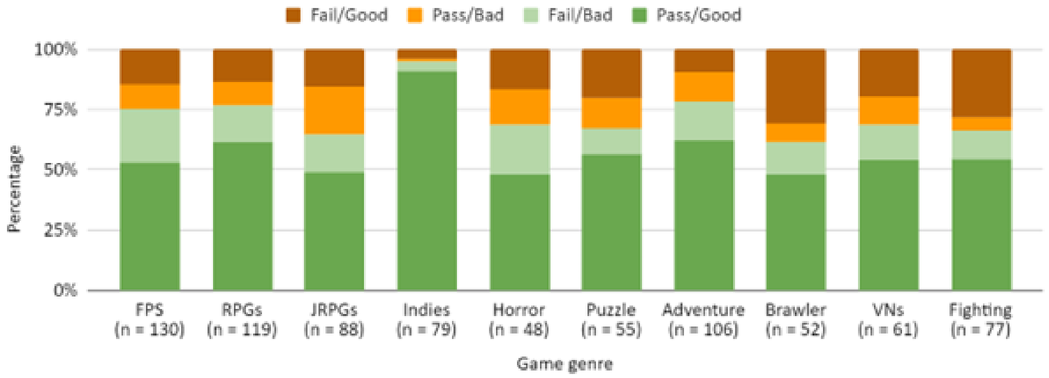


Figure 6: Percentages of each IRWiG response rate (excluding "Haven't played/unsure"). Agreement comprises combined percentages of Pass/Good and Fail/Bad).

ber of responses, that these results are not statistically significant. When splitting the results by respondent gender, there were slight discrepancies in the agreement rate between men, women, and non-binary/other identities, but generally were within 20% of each other. Women participants reported lower agreement rates for horror games, and higher rates for puzzle games vs the other respondents, while non-binary/non-disclosed gender identity participants reported lower agreement for visual novel games. No other major differences were observed by participant demographics.

When the list of games played by the project author was used to evaluate the IRWiG test, the results were broadly the same between when the IRWiG test was used by participants and by the project author. Out of all games tested both internally and by survey participants n = 19/25 (76%) games had the same IRWiG test result for both parties. The remaining six games where there was a disagreement between the survey participants and the project author were investigated, and the IRWiG test criteria that were causing this disagreement were isolated:
• Xenoblade Chronicles (Monolithsoft, 2010): participants were uncertain whether women played an important-enough role in the main narrative
• The Legend of Zelda: Twilight Princess (Nintendo, 2006): there was ambiguity in the criterion relating to stereotypes, and how a character extends beyond them so as to pass
• The Witcher 3: Wild Hunt (CD Projekt Red and Sapkowski, Andrzej, 2015): the primary author believed women did not play a substantial role in the main narrative, but survey respondents disagreed
• The Talos Principle (Croteam, 2014): the games' philosophical themes of "what makes a human" and the non-gendered self-insert player character caused confusion when participants applied the IRWiG test
• Genshin Impact (Hoyoverse, 2020) and NieR: Automata (PlatinumGames, 2017): women with weakly-justified sexualised clothing throughout both titles, failed the appearance-related IRWiG criterion when they would otherwise pass the test

For participant feedback on the IRWiG test criteria, all received useful/very useful responses overall. The Appearance metric had the lowest usefulness score, with comments discussing the complexity of questions, including arguments for both sides. In particular, the criterion on tropes potentially requires more caveats or greater specificity, with some responses considering it too vague, or believing it ignored the heuristic and cultural role tropes can play. It was believed that this would be partially mitigated in the IRWiG test by specifying a character should "extend beyond one or two widespread stereotypes", allowing for tropes to still factor into a character, but this potentially increased the ambiguity of the criterion.

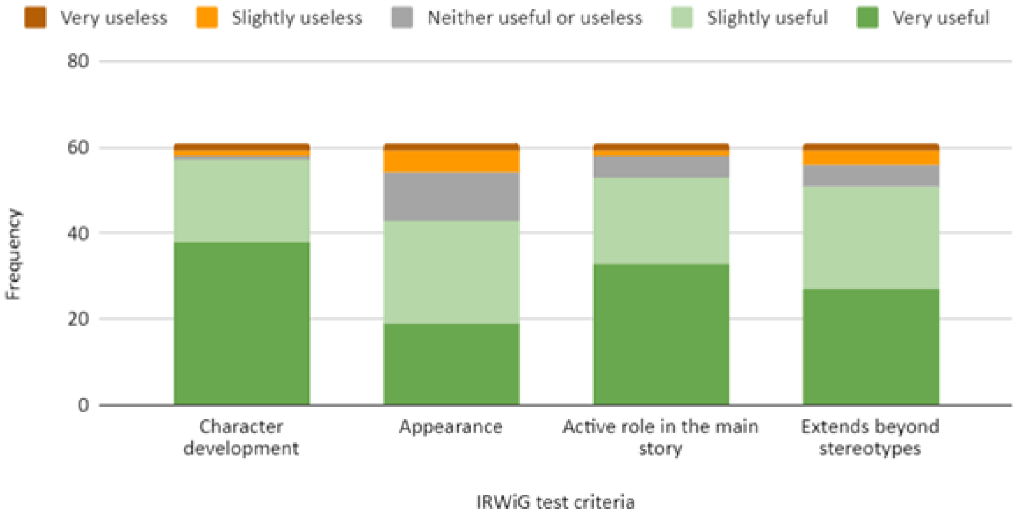


Figure 7: Participant responses to how effective each IRWiG test criteria was.

As for potential new criteria to test games against, some that were suggested by participants include:

- Ideas or themes that the game as a whole conveys
- Comparing writing quality between characters of all genders
- A character's place in the wider in-game world
- Testing character physique rather than their outfit
- Bechdel-like conversation criteria

Discussion of results

The above results, and an overall agreement rate of 74% between participant views and their application of the IRWiG test, suggest that an indication of how well women are represented in video games could be assessed via a Bechdel-like test. Results of the IRWiG test were broadly in-line with participants' views on representation of women across every genre, regardless of experience with video games. This indicates that the IRWiG test is not reliant on game experience, and may be effectively understood and applied by both high intensity gamers, and those with far less experience with the medium, potentially because the criteria of the IRWiG test are not reliant on gaming-specific knowledge, but draw on general media principles also seen in film and TV. People with little gaming experience are able to use general media literacy to apply the IRWiG test to games. While one criterion in the test is more specific to gameplay, some can also be applied to characterisation outside of gameplay, as well as stereotypes in other media.

Potential reasons for the disparities amongst genders in the Horror and Puzzle genres were unclear. In both genres, participants identifying as men answered nearly as much as the combined number of women and non-binary/those not providing an identity. Small sample size issues are a possible factor, as differences were often driven by just two or three responses. A further possible reason for the Horror genre is women being more cognizant of the tropes that pervade the genre in both films and games (Spittle 2011), but this can only be confirmed through further research.

The Indie genre had the highest agreement rate between participants and their application of the IRWiG test. It also contained both the highest pass rate with the IRWiG test and was the genre considered to have the best representation of women. As mentioned by some respondents in the open-ended question for Indie games, games in the genre often provided more room for open exploration of different themes and ideas that larger-budget AAA games are only rarely permitted (Ruffino 2013). This freedom is utilised in varying ways, from more heavily featuring women (Perreault et al. 2022) to employing broader ranges of art styles (Alvarez 2016). The combination of these can allow Indie games more leeway to tell stories with characters who are women that do not overly rely on others to overcome challenges. However the choice of games used in the survey could have skewed results unintentionally, as the majority of the Indie games tested had playable women characters.

Of the 79 responses within the Indie genre, only 4 (5%) indicated a the game in the genre had poor representation. This raises the question of whether there was an accidental bias in the survey, or whether the genre may be an outlier within the industry. It is also worth noting that some of the highest-performing and critically acclaimed games in the genre feature female lead characters, indicating that consumers actively seek out this style of game. This could have potentially biased the dataset being used away from games in this genre with poorer representation. The Indie genre in particular would need further testing to either eliminate or confirm some of these potential issues, but indicative findings and the freedom for Indie game development suggest the genre has better representation of women than others; this is supported by other studies (Perreault et al. 2022).

The genre with the lowest agreement rate between participant opinions and the application of the IRWiG test was the brawler genre, with 62% agreement. There was an individual IRWiG application pass rate of 56%, and 79% of respondents considered games in this genre to have a good representation of women. With a substantial proportion of participants stating that games fail the IRWiG test despite having good representation, this shows the test is challenged by this particular genre. Brawler games take many forms but typically focus on precise 3rd-person combat gameplay over a strong narrative. While some games chosen for the survey in this genre do possess campaigns and riveting stories, they can still be secondary to the focus on gameplay that the genre is known for. Games in this genre often feature multiple playable characters, differentiated via stereotypes, to help them stand out from each other, and highlight their style of gameplay. The combination of story as low priority and stereotyped characters may lead to more games in this genre failing the IRWiG test. That participants consider brawler games to have had better representation than the test results show indicates that this may require further investigation.

Participant feedback on the IRWiG test

When evaluating the test criterion on Appearance, it was expected that it would have the highest discordance, as debates around whether character sexualisation can be examples of good representation are contentious within video game culture (Lindner et al. 2020). No participants argued against appearance in video games being important when giving open-ended feedback, but rather noted that the outfit design typically found in some genres let characters down, notably in the JRPG genre, or that appearance is particularly subjective. The discussion on genre-specific tendencies in character and outfit designs warrants further research, taking international and cultural diversity into account.

When discussing character designs and sexualisation of women in games, a high-profile example is Bayonetta from the game series of the same name (PlatinumGames, 2009, 2014, 2022). This title was deliberately chosen for inclusion due to high sales and frequent mentions in these debates. Of the 17 participants who had played the game, 16 believed it to have good representation of women, albeit 6 of those feeling it failed the IRWiG test, with one "unsure" participant noting "Bayonetta is a really tough one". Discussions on the line between women in games being "hypersexualised and problematic" and "powerful women taking control of their sexuality" have dominated characters like Bayonetta (Harper 2015), and highlighted how difficult subjectively testing how well games represent women based on appearance can be. It was unlikely that this subjectivity could be incorporated succinctly into a Bechdel-equivalent test, so potential future work could evaluate how an updated IRWiG test performs with vs without Appearance. However, given that Appearance was deemed an important element in the survey-based ontology, removing it entirely may be problematic.

In open feedback on the IRWiG test, several participants noted how they believed it to work well for RPGs and JRPGs, although they found issues with excluding optional and side content, particularly for RPGs with a heavy focus on side material such as The Witcher 3: The Wild Hunt (CD Projekt Red and Sapkowski, Andrzej, 2015) or the Legend of Zelda: Breath of the Wild (Nintendo, 2017). Participants were also unsure on how to evaluate player-created playable characters, such as Shepard in the Mass Effect trilogy (Bioware, 2007, 2010, 2012), or self-insert characters such as Chell in Portal 2 (Valve, 2011), and others noting that a game passing or failing sometimes depends on choice of character's gender, an example given by some participants being Dishonored 2 (Arkane Studios, 2016), where one solely plays as the man Corvo Attano or woman Emily Kaldwin for the game's entire duration.

It was also noted that some games have no specific narrative, limiting compatibility with one of the IRWiG test criterion, namely of a woman playing an active role in the narrative. Multiplayer-centric games like Overwatch (Blizzard Entertainment, 2016) and League of Legends (Riot Games, 2009) have large character rosters, with external media that flesh out their respective game worlds and characters, but with no in-game main story it becomes a challenge to test how well they handle women. Overwatch had the largest fail/good rate of any game being tested, potentially indicating that criteria that have a reliance on narrative should be modified to account for this problem.

There were mixed opinions on whether the IRWiG test was too complex or needed more criteria, with multiple responses arguing for either side. Some felt it was too hard or complex for casual use, and that remembering all the criteria was difficult and would limit real-world applicability, while others believed it needed more criteria, or that already-existing criteria needed further caveats to account for more games or characters. It was pointed out that some games feature villainous women who do not undergo character development or overcome obstacles, but instead are themselves obstacles for the protagonists to overcome . In some instances these villains are still examples of good representation of women, as is also seen in comic books (Sereni 2020). Expanding the related criterion and adding more caveats to include this character archetype was suggested to help alleviate this.

Personal interpretation is inherently a large element of the Bechdel test (Kravina 2022), such as what meets the threshold for a conversation, and this ambiguity is well recognised (Jang et al. 2019). Personal interpretation was an expected aspect of the IRWiG test, and may be difficult to entirely eliminate. The results and feedback for the six aforementioned titles which had disagreement between the survey participants and the project author, particularly the Talos Principle due to its nongendered player avatar, point towards personal interpretation already being a factor. This aspect is something that can be alleviated by ensuring criteria wording is clear and concise, but would remain an inherent part of a test for this purpose, which should be kept in mind, particularly for games designed to have multiple narrative interpretations or open games with skippable content.

Some participants proposed creating multiple tests for different genres, which each contain different criteria. This could warrant future research, however there would be difficulties in designing how to test multi-genre games, as well as how to define those genres, which is a large enough discussion in the gaming industry and community to be out of scope here (Clearwater 2011). Any further research that follows this route would either have to be careful to avoid turning the debate into a mere discussion of genre semantics, or tackle it head-on as related work to build on.

The IRWiG test evaluation methodology also allowed for gaining insights into the potential social impact it might have on a small scale, and the discussions it might prompt. After the survey period had ended, members in one online community had long discussions about the IRWiG test. These debates around the effectiveness of the test highlighted potential games outside of the survey that would pass or fail, as well as discussing reasons participants had for passing or failing certain games. The general consensus within these discussions was that discussion participants believed the test needed iterating on, but that it was a good basis for having conversations around how women are represented in video games.

CONCLUSIONS

The purpose of this project was to create a Bechdel-equivalent test for use in video games, and evaluate it under the research question ”could a Bechdel-like test be an indication of how women are represented in video games?” The IRWiG test was built on background research covering the multidisciplinary nature of this research topic, and ontologies created by both the primary author and by the general public via an online survey. The test was then assessed in an evaluation survey in which public participants used it on a predetermined list of games, as well as internally by the primary author.

Results found a correlation between opinions of how women are represented in the medium, and opinions of whether games pass the IRWiG test. While rates of agreement varied by genre, a notable and evident relation was observed. Feedback on the IRWiG test stated that the themes the individual criteria attempted to evaluate were important to the topic, but the specific wording needed addressing to raise the agreement rate. Through using respondent Likert scaling it was identified that the lowest-received criterion was in relation to character appearance, though all were given overall positive feedback. However, gaps in the IRWiG test did lead to games intrinsically failing due to their core design and purpose, suggesting further development is required.

The methods used to analyse the survey data proved sufficient to answer the research question, and form a potential basis for future work. While using survey questions that mimicked confusion matrices still allowed for viable data that was transferable to a different type of analysis, a different approach might allow for more granular discussion and evaluations.

This project does not aim to definitively answer whether any game has good or bad representation, and no findings should be misconstrued as such. The work presented is indicative and exploratory. The Bechdel test is itself not a definitive method of evaluation for whether movies have good representation of women, and neither would any similar test for video games be so. The debates and discourse around gender and women in video games are lengthy and heated, and it was not the intent of this paper to add to such discourse in any harmful way. Bechdel-like tests form a potential starting point for dialogue on how women are portrayed in games, as participants in this project's surveys have noted, as has work evaluating the Bechdel test itself (Classroom 2018), but should not be used to gauge how good or bad a game (or film) might be. There is however potential for this research to be noticed and gain profile in the wider gaming community. The primary potential implication would be its effect on game development, which could either be driven by the public using the IRWiG test on games similar to websites such as The Bechdel Movie Test (*The Bechdel Movie Test* 2008-2022), or from use in-house during development, like the King Diversity Tool (Alt 2022). However, it is paramount to stress the indicative nature of these results.

FUTURE WORK

Overall, this project indicates that a Bechdel-like test such as the IRWiG test might provide useful indicative information regarding how women are represented in video games, and could support healthy and beneficial discussions of this topic. Further work on this topic may yield valuable results for the industry and wider gaming community.

While this project adequately addressed the research question, it also raised more questions that potentially warrant future investigation. In the continuation of creating Bechdel-like tests for video games, revising the IRWiG test as a single test for multiple genres and creating new separate tests for individual genres are both possible research paths. One could introduce the concept to game development studios, who might use Bechdel-like tests proactively for game development. Potential differences in how culture impacts perspectives of representation of women in video games was also identified as worth investigating, but was not possible in this project due to scope and the limited demographics of survey respondents. This work has very much focused on western cultures and their games. Video games are both played and developed globally: this can and should be addressed in future work.

During this work, gaps in literature were also identified that would be worth filling. Many of these gaps were the result of already-existing literature lacking follow-ups in recent years. The last publicly available research investigating how video games portray and advertise to women was published in 2016, and the ratios of women protagonists in games last addressed in 2013. Women respondents in the initial survey stated that they felt increasingly welcomed in the video game community and that more games were now being made for them. This, and the rise of critically acclaimed games prominently featuring women, were at time of writing only anecdotal observations, and their relationships and trends warrant further, formal investigation.

BIBLIOGRAPHY

Aarseth, Espen. 2001. *Computer Game Studies, Year One.* Web Page, 30/10/2022. http://www.gamestudies.org/0101/editorial.html.

Alt, Eric. 2022. *King's Diversity Space Tool.* Web Page, December. https://www.activisionblizzard.com/newsroom/2022/05/king-diversity-space-tool.

Alvarez, Gonzalo. 2016. *Pencils, Paints, or Pixels?: How Aesthetic Choices of Indie Games Affect Interactive Experience.* Generic.

Arkane Studios. 2016. *Dishonored 2.* [PC] Microsoft.

Aylett, Ruth, and Sandy Louchart. 2007. "Being there: Participants and spectators in interactive narrative." In *International Conference on Virtual Storytelling,* 117–128. Springer.

Bechdel, Alison. 1985. *The rule.,* 2, https://lithub.com/read-the-1985-comic-strip-that-inspired-the-bechdel-test/.

———. 2015. *Lesbian Cartoonist Alison Bechdel Countered Dad's Secrecy By Being Out And Open.* Interview. https://www.npr.org/2015/08/17/432569415/lesbian-cartoonist-alison-bechdel-countered-dads-secrecy-by-being-out-and-open.

Bioware. 2007. *Mass Effect.* [PC] Electronic Arts, Microsoft.

———. 2010. *Mass Effect 2.* [PC] Electronic Arts.

———. 2012. *Mass Effect 3.* [PC] Electronic Arts.

Blizzard Entertainment. 2023. *Diablo IV.* [PC] Blizzard Entertainment.

———. 2016. *Overwatch.* [PC] Blizzard Entertainment.

———. 2022. *Overwatch 2.* [PC] Blizzard Entertainment.

Bouchat, Kathryn Gray. 2019. "Testing the Bechdel Test."

Burrell, Gibson. 1993. *Sociological paradigms and organisational analysis : elements of the sociology of corporate life / by Gibson Burrell and Gareth Morgan.* Farnham: Farnham: Ashgate. Book.

Carr, Diane, Gareth Schott, Andrew Burn, and David Buckingham. 2004. “Doing game studies: A multi-method approach to the study of textuality, interactivity and narrative space.” *Media International Australia* 110 (1): 19–30. ISSN: 1329-878X.

CD Projekt Red and Sapkowski, Andrzej. 2015. *The Witcher 3: Wild Hunt.* [PC] CD Projekt Red.

Cerafica, Annakatrina. 2022. *The Activision Blizzard Diversity Space Tool Controversy Explained.* Electronic Article. https://gamerant.com/the-activision-blizzard-diversity-space-tool-controversy-explained/.

Classroom, Pop Culture. 2018. *The Bechdel Test: Why It’s Important For Students.* Electronic Article, August. https://popcultureclassroom.org/2018/08/22/the-bechdel-test-why-its-important-for-students/.

Clearwater, David. 2011. “What defines video game genre? Thinking about genre study after the great divide.” *Loading...* 5 (8). ISSN: 1923-2691.

Copier, Marinka. 2003. “The other game researcher: participating in and watching the construction of boundaries in game studies.” In *DiGRA Conference,* 404–419.

Croteam. 2014. *The Talos Principle.* [PC] Devolver Digital.

Esposito, Nicolas. 2005. *A Short and Simple Definition of What a Videogame Is.* Conference Proceedings.

Firkins, Mio. 2023. *A Bechdel Test for computer games? Final-year undergraduate dissertation.*

Fortim, Ivelise, and Carolina de Moura Grando. 2013. “Attention whore! Perception of female players who identify themselves as women in the communities of MMOs.” In *DiGRA Conference.*

Francis, Bryant. 2022. *Activision Blizzard’s new ”diversity space tool” gets frosty reception from devs.* Electronic Article. https://www.gamedeveloper.com/culture/activision-blizzard-s-new-diversity-space-tool-gets-frosty-reception-from-devs.

Grodal, Torben. 2000. “Video games and the pleasures of control.” *Media entertainment: The psychology of its appeal,* 197–213.

Harper, Todd. 2015. “Beyond Bayonetta’s Barbie Body.” *AoIR Selected Papers of Internet Research,* ISSN: 2162-3317.

Hickey, Walt, Ella Koeze, Rachael Dottle, and Gus Wezerek. 2017. *The Next Bechdel.* Web Page. https://projects.fivethirtyeight.com/next-bechdel/.

Hoyoverse. 2020. *Genshin Impact.* [PC] Hoyoverse.

Jang, Ji Yoon, Sangyoon Lee, and Byungjoo Lee. 2019. “Quantification of gender representation bias in commercial films based on image analysis.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–29. ISSN: 2573-0142.

Koivunen, Anu, Ingrid Ryberg, and Laura Horak. 2014. “Swedish cinema’s use of the Bechdel test is a provocation that works.” *The Guardian.*

Konami. 2015. *Metal Gear Solid V: The Phantom Pain.* [PC] Konami.

Kravina, Carolina. 2022. *Is The Bechdel Test still Relevant?* Electronic Article, July. https://raindance.org/is-the-bechdel-test-still-relevant/.

Lerum, Kari, and Shari L Dworkin. 2009. "“Bad girls rule”: An interdisciplinary feminist commentary on the report of the APA task force on the sexualization of girls." *Journal of sex research* 46 (4): 250–263. ISSN: 0022-4499.

Lindner, Danielle, Melissa Trible, Ilana Pilato, and Christopher J Ferguson. 2020. "Examining the effects of exposure to a sexualized female video game protagonist on women's body image." *Psychology of Popular Media* 9 (4): 553. ISSN: 2689-6575.

Louis, Meryl Reis, Gibson Burrell, and Gareth Morgan. 1983. *Sociological Paradigms and Organizational Analysis.* Generic.

Mather, Mara, Marcia K Johnson, and Doreen M De Leonardis. 1999. "Stereotype reliance in source monitoring: Age differences and neuropsychological test correlates." *Cognitive Neuropsychology* 16 (3-5): 437–458. ISSN: 0264-3294.

Monolithsoft. 2010. *Xenoblade Chronicles.* [Nintendo Wii] Nintendo.

———. 2022. *Xenoblade Chronicles 3.* [Nintendo Switch] Nintendo.

Nintendo. 2019, n.d. *Investor Relations Information.* Web Page, 2/2/2023. https://www.nintendo.co.jp/ir/en/.

———. 2017. *The Legend of Zelda: Breath of the Wild.* [Nintendo Switch] Nintendo.

———. 2006. *The Legend of Zelda: Twilight Princess.* [Nintendo Wii] Nintendo.

O'Meara, Jennifer. 2016. "What "The Bechdel Test" doesn't tell us: examining women's verbal and vocal (dis) empowerment in cinema." *Feminist media studies* 16 (6): 1120–1123. ISSN: 1468-0777.

Perreault, Mildred F, Gregory Perreault, and Andrea Suarez. 2022. "What Does it Mean to be a Female Character in "Indie" Game Storytelling? Narrative Framing and Humanization in Independently Developed Video Games." *Games and Culture* 17 (2): 244–261. ISSN: 1555-4120.

PlatinumGames. 2009. *Bayonetta.* [PC] Sega.

———. 2014. *Bayonetta 2.* [Nintendo Wii U] Nintendo.

———. 2022. *Bayonetta 3.* [Nintendo Switch] Nintendo.

———. 2017. *NieR: Automata.* [PC] Square Enix.

Poonamallee, Latha. 2009. "Building grounded theory in action research through the interplay of subjective ontology and objective epistemology." *Action Research* 7 (1): 69–83. ISSN: 1476-7503.

Riot Games. 2009. *League of Legends.* [PC] Riot Games.

Ruffino, Paolo. 2013. "Narratives of independent production in video game culture." *Loading...* 7 (11). ISSN: 1923-2691.

Sereni, Eleonora. 2020. “” When I’m Bad, I’m Better”:: from early Villainesses to contemporary antiheroines in superhero comics.”

Sharda, Adhikari. 2014. “Media and gender stereotyping: The need for media literacy.” *International Research Journal of Social Sciences* 3 (8): 43–49. ISSN: 2319-3565.

spellissa, skelly. 2022. *God I swear our own company tries so hard to slaughter any good will the actual devs who make the game have built Overwatch doesn’t even use this creepy distopian chart, our writers have eyes. The artists: have eyes. Producers, directors, etc, as far as I know also all have eyes.* Social Media. The in-document citation is done by hand bc endnote is being weird, 14/5/2022. https://twitter.com/_mlktea/status/1525507447548366848.

Spittle, Steve. 2011. ““Did This Game Scare You? Because it Sure as Hell Scared Me!”FEAR, the Abject and the Uncanny.” *Games and Culture* 6 (4): 312–326. ISSN: 1555-4120.

SteamDB. 2013-2023. *Most played games.* https://steamdb.info/charts/.

Sterling, Jim. 2022. *Activision’s ”Diversity Tool” Is F*cking Awful (The Jimquisition).* Social Media. https://www.youtube.com/watch?v=g46-gV7TfB4&t=218.

Tassi, Paul. 2022. *Activision Blizzard Heavily Edits King’s Diversity Generator Blog, Removes All ‘Ranking’ Images.* Electronic Article. https://www.forbes.com/sites/paultassi/2022/05/14/activision-blizzard-heavily-edits-kings-diversity-generator-blog-removes-all-ranking-images/?sh=5f37a9b93de0.

The Bechdel Movie Test. 2008-2022. https://bechdeltest.com/.

Top popularity. 2019-2023. https://www.backloggd.com/games/lib/played/.

Valve. 2011. *Portal 2.* [PC] Valve.

———. 2005, 2023. *Top Sellers.* Web Page, 2/2/2023, 14/3/2023. https://store.steampowered.com/search/?filter=topsellers.

Zhao, Lili, and Chunping Li. 2009. “Ontology based opinion mining for movie reviews.” In *International Conference on Knowledge Science, Engineering and Management,* 204–214. Springer.