White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/192367/

Version: Published Version

Article:
Hughes, Nathan, Flockton, Josephine and Cairns, Paul Antony (2023) Growing Together: An Analysis of Measurement Transparency Across 15 Years of Player Motivation Questionnaires. International Journal of Human-Computer Studies. 102940. ISSN: 1071-5819

https://doi.org/10.1016/j.ijhcs.2022.102940

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Int. J. Human–Computer Studies 169 (2023) 102940

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Growing together: An analysis of measurement transparency across 15 years of player motivation questionnaires

Nathan G.J. Hughes [a,*], Josephine R. Flockton [b], Paul Cairns [a]

[a] Department of Computer Science, University of York, York, YO10 5GH, United Kingdom
[b] Department of Psychology, University of York, York, YO10 5GH, United Kingdom

ARTICLE INFO

Keywords:
Questionnaires
Player traits
Motivations
Content analysis

ABSTRACT

There are many questionnaires to assess player motivation, originating from a diverse range of disciplines. Each discipline differs in their usage and reporting of questionnaires, but there has been no attempt to standardise their application. No standard approach leads to a lack of transparency in usage reporting, which affects the ability of the field to synthesise. This has made it unclear whether player motivation research is a unified community, or a collection of individuals with a similar goal. Therefore, the current work assesses the transparency of reporting practices of player motivation questionnaires published within the last 15 years. 18 questionnaires were identified via a scoping review, then papers citing these questionnaires were analysed for their transparency of reporting practices (n = 238); first via a content analysis of justifications for use, then followed by an analysis of transparency against eight criteria created for this work. Overall, reporting transparency is lacking, driven by little priority for presenting items alongside text. Many papers use questionnaires because they are theory-based or have measured specific variables in previous works, but explicit justification is rare. The work concludes with a transparency checklist based on the eight criteria used, which authors can use to standardise the field and allow for more cohesive research synthesis.

Coming together is a beginning, staying together is progress, and working together is success - Edward Everett Hale

1. Introduction

The field of games research has greatly matured over the last three decades. There is a particular interest in understanding how players differ from one another, variously referred to as their individual motivations, traits, and preferences. Whilst there are theoretical differences between these concepts, there are overlaps in how they are measured, which has led to a reification of the concepts (see Hughes and Cairns (2020) for a more in-depth discussion). This is because whilst these are different concepts, which serve to emphasise different ways of conceptualising individual differences in players, the practical measurement of them shows a high degree of convergence. Each theory typically employs the use of questionnaires to assess the thoughts of players, meaning whilst the underlying stance of the work may differ, the methodology used is the same. Because of this, research using these theoretical differences is interchangeable, as the items used in measurement have become comparable. Therefore, the distinct differences found when considering these concepts at the abstract level have been removed, and the questionnaires become what is actually being measured. For brevity, the concepts of motivation, preference, traits, and individual differences are referred to as 'player motivations' throughout the current work.

As differences in player motivations are subjective and difficult to measure from gameplay data (though there have been attempts e.g., Melhart et al. (2019)), questionnaires are a common method deployed to capture them. Indeed, questionnaires are so common there are multiple scales available to capture these individual differences, with more being constructed at a consistent rate. This is perhaps unsurprising, as what drives players is both complex and context-specific (Hughes and Cairns, 2021). Because of the wide variety of contexts that digital games offer, it is common to see researchers adapt and change established questionnaires to fit their specific research needs and the games under examination. For example, some

* Corresponding author.
E-mail addresses: nathan.hughes@york.ac.uk (N.G.J. Hughes), jrf521@york.ac.uk (J.R. Flockton), paul.cairns@york.ac.uk (P. Cairns).

https://doi.org/10.1016/j.ijhcs.2022.102940
Received 22 November 2021; Received in revised form 31 July 2022; Accepted 22 September 2022
Available online 3 October 2022
1071-5819/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

researchers wish to study specific games, and so modify items to refer to them (e.g., Davies and Hemingway (2014)). Researchers wishing to study online games may use online gaming motivations measures, but drop items that are too specific (e.g., Teng and Chen (2014) and Comello et al. (2016) who used the original (Yee, 2006) scale that mostly focuses on *World of Warcraft)*.

However, the existence of multiple questionnaires, and the frequency of these questionnaires being altered between uses, presents a risk to the compatibility and coherence of research into player motivation. Besides the concern that the existence of so many questionnaires is in itself a conceptual threat for what 'motivation' actually means (Hughes and Cairns, 2020), the question becomes the following: do player motivation researchers have comparable practices in questionnaire-based research? By extension, can different groups of researchers meaningfully compare and collectively build their knowledge?

The second question relates to the concept of a research community. Wenger argued that certain communities can be considered Communities of Practice (CoP; (Wenger, 1999)). A CoP has 3 features: Domain (an area of interest), Community (a way for practitioners to discuss ideas), and Practice (the focus of the community that drives knowledge). As part of Practice, there is the concept of a Shared Repertoire: an agreed set of methods and tools used by members of the community to achieve the research goals. For player motivation, questionnaires could be considered a shared repertoire as part of the wider practice investigating what drives players to play. Further, previous research has explored these aspects of a Community of Practice within academia (e.g., Rolin (2008), Nersessian (2006)), highlighting the applicability of the theory to research communities.

Whilst all are arguably present within player motivation research (there is a shared interest in player motivation, avenues for discussing this interest such as conferences, and a number of methods used by multiple researchers), the aspect of a shared repertoire is the most relevant to the current work. By assessing the methods used by researchers (in this case, questionnaires), it is possible to assess if player motivation researchers have an established shared repertoire — and by extension are indeed a community, rather than a collection of individuals with a common goal. This is an important distinction, as a research community is more likely to last long-term than a collection of individuals, as having a defined set of methods makes the community more resilient and improves the quality of the research (see Ankeny and Leonelli (2016) for an overview). Therefore, the research question instead becomes: is there an observable and collaborative community of player motivation researchers?

To illustrate this question, consider the following worked example. A researcher is keen to understand player motivation within their research, and wishes to build on previous work on the subject via the use of an established questionnaire. They will face two methodological challenges. First, which questionnaire is the most appropriate based both on their own research needs and the existing work done in the field? Second, how should they use the chosen questionnaire to ensure comparability and compatibility with existing work? The latter challenge is especially important, as even small changes to questionnaires (such as changing the wording of an item or dropping an item that seems irrelevant) can potentially alter participant responses (Cairns, 2019). Doing so can substantially change the interpretation of findings within and between studies purporting to use the 'same' questionnaire.

The researcher in this example may be new to the field, and unaware of the intricacies of how questionnaires are used. It is reasonable to assume they will use existing work as the basis for their own work, which would naturally support the building and growth of the body of knowledge in the field. However, where questionnaire use is not justified (why this questionnaire, and not another?) and their deployment is not transparent, there is a risk that new research is simply not compatible with existing research. Differences in research outcomes may be quirks of differences in method rather than anything more substantive. Without transparency in questionnaire usage, there is no way to distinguish the two. Without trustworthy comparison between research findings, there is no way to build on previous work. With no building on previous work, there cannot be a research community. Therefore, how questionnaires are used is as important to the field as why they are used.

The goal of the current paper is to explore if the field of player motivation research has developed into a research community over the last 15 years, by examining the state of questionnaire-based methods. An extensive analysis of the literature allows us to assess how player motivation questionnaires are currently used and reported, which reflects the community practices currently deployed. From this, we can propose best practice to support community growth and research coherence, by creating an accessible checklist. Therefore, the aim of the current work is threefold:

1. Assess the justifications for player questionnaire usage
2. Assess the transparency of current reporting practices
3. Provide guidance and support to encourage best practice in questionnaire-based player motivation research

2. Background

In this section, the field of measuring player motivation is discussed, along with how this growth has led to the current state of questionnaire design and reporting. Following this, practices are outlined that increase transparency when using questionnaires, which form the basis of the criteria used when evaluating papers in this study.

2.1. How we got here: A brief history of player questionnaires

Following the seminal work of Bartle (1996) that theoretically outlined how players are not a monolith but instead a diverse range of people with differing reasons for play, understanding how to measure these differences has been a focus of many researchers. Whilst questionnaires are an obvious way to measure these differences due to their ease of administration, that does not mean they are easy to design or interpret. Indeed, the non-monolithic and complex nature of why players differ from one another is so hard to capture that no one questionnaire currently available seems apt to measure all the myriad of reasons (Hughes and Cairns, 2020). Because of this, multiple questionnaires have been designed that either capture new dimensions of player differences (such as the introduction of the motivation for 'Continuance Intention' by Wu et al. (2010)), or refine and improve on existing questionnaires (e.g., the trait scale proposed by Tondello et al. (2019) which improved on BrainHex (Nacke et al., 2014)).

This iterative process reflects the organic growth of the field, where the aim is for measurement of player differences to improve with new studies and new designs. The growth is further explained by the fact that, as the nature of games allows for research from a multidisciplinary viewpoint, multiple disciplines have become involved since the work of Bartle (1996). A narrative review of the literature conducted for the current work identified four distinct disciplines involved in the identification and measurement of player motivations. This identification was supported by the scoping review of literature, as detailed in the study below. These can be loosely summarised as four areas:

- **Games Researchers**: inspired strongly by Bartle (1996) and building on the work of Yee (2006), these researchers consider games from the primary standpoint that they are a specific phenomena worth studying. Work here looks at player experiences in specific games (e.g., Billieux et al. (2013)), and attempts to predict behaviour based on player motivations (e.g., Bowman et al. (2012)).

- **Psychologists**: these researchers attempt to apply theoretical and established models of human behaviour to games. This is typically by using Self-Determination Theory (Ryan et al., 2006) and its associated sub-theories, but also includes work on explaining how gaming can affect behaviours in real life (e.g., Mills et al. (2018)).
- **Media Researchers**: by considering games as a new form of media, these researchers apply theories of media use to games. This is typically considered from a Uses & Gratifications perspective (Sherry et al., 2006), where findings are based on how people interact with the medium of games, and what games can tell us about media consumption (e.g., Lee and Schoenstedt (2011)).
- **Education Researchers**: one of the more recent disciplines to study games, those interested in education attempt to understand what about games is fun so this can be used to make learning more effective/engaging (e.g., Tavakkoli et al. (2014)). This also can be done by understanding how games teach players, and applying these methods to real-life teaching (e.g., Monterrat et al. (2017)).

Whilst these disciplines are different in the reasons that drive them to understand games and their players, a technique they typically have in common is that of using questionnaires. Questionnaires, especially those that explore what motivates players, are one of the most common techniques deployed in all of these disciplines. For example, questionnaires have been use to explore escapism (Warmelink et al., 2009), well-being (Goh et al., 2019), and the effects of controller type on the play experience (Birk and Mandryk, 2013), highlighting the range of topics covered across these disciplines. The multidisciplinary nature of the field has been true since its conception, with three main strands creating questionnaires for player motivation in the same year of 2006. Games research began measuring player motivation from Yee (2006), who was inspired by the seminal work of Bartle (1996). At the same time, Media scholars used a Uses & Gratifications theoretical lens to construct a motivation questionnaire, by Sherry et al. (2006). Finally, Psychologists used Social Determination Theory to construct the Player Experience of Need Satisfaction questionnaire (PENS; (Ryan et al., 2006)).

Since then, there has been little cross-communication between questionnaires. Each questionnaire inspires the creation of new questionnaires within the discipline of origin, with little example of inspiration from other disciplines. This can be seen from reading the background literature cited in newer questionnaires; the BrainHex (Nacke et al., 2014) and Trojan (Kahn et al., 2015) motivation questionnaires cite (Yee, 2006) as inspiration, whilst the Gaming Motivation Scale (GAMS; (Lafrenière et al., 2012)) and the Motives for Online Gaming Questionnaire (MOGQ; (Demetrovics et al., 2011)) build on PENS (Ryan et al., 2006). Whilst there has not been as many new questionnaires created by Media scholars, Wu et al. (2010) built on Sherry et al. (2006) to include 'Continuance Intention'. The one noted exception is the Intrinsic Motivations to Gameplay questionnaire (IMG; (Vahlo and Hamari, 2019)), which purposefully references multiple disciplines as inspiration. This is however the exception to the rule, as most questionnaires sit within one discipline of influence. Finally, the discipline of Education is notably newer to enter games, and so does not have the same history. However, it is still a significantly separate collection of works with different goals for measuring player motivation, and includes recent works such as the Video Game Pursuit Scale (VGPu; (Sanchez and Langer, 2020)).

Overall, these disciplines are notably different in their evolution and application of questionnaires — especially in terms of motivation for studying individual differences in players, but also the theories that underpin these differences. The presence of different practices within the field of games research is not surprising; games research is a relatively new topic, but is established enough to have a sense of community where researchers have a shared interest (i.e., individual differences in players).

This building on each other's research resembles Wenger's concept of a Community of Practice (CoP; (Wenger, 1999)). Researchers interested in player motivation have a shared domain (an interest in how players differ from one another), a community where the domain can be discussed (the presence of conferences, journals, workshops, visiting speakers, all contribute to the sharing of ideas), and practice via a shared repertoire, in terms of agreed methods. This last concept is most relevant to the current work, where the use of questionnaires could be considered a shared repertoire that multiple researchers understand to be a way of assessing player motivation. By exploring how researchers use questionnaires, there is an implicit focus on practice, as this is the way in which methods in the field are deployed. Therefore, the current work is an exploration of the extent to which the field has a shared repertoire when deploying questionnaires.

However, the plurality of backgrounds and disciplines studying around and with one another bring with them differing priorities on how to use and report questionnaires. Using questionnaires for different purposes is not inherently wrong, as there are a multitude of motivations for studying players. However, using supposedly the same questionnaires in different *ways* is problematic. By researchers bringing each of their own priorities and motivations for studying differences in players, there is no obvious standard for the use or reporting of questionnaires within the context of games. If there is no standardisation of the practice of using questionnaires – perhaps caused by the multidisciplinary nature of the field – there cannot truly be a 'shared' repertoire. By extension, there cannot be an established Community of Practice, which limits the ability of researchers in the community to learn from one another. This is a problem for the continued growth of the field, as a lack of standardisation allows for questionnaires to be used in different ways, which in turn could lead to differing findings. This would mean findings are less likely to be comparable or meaningfully build on one another.

Furthermore, by only reporting aspects of questionnaire usage that are deemed important to the respective discipline, this obscures aspects that may be important to others (or ought to be important for researchers within the same discipline). Whilst disciplinary practices are abstract and difficult to measure, in the same way that the common beliefs of a society are hard to define (e.g., Kuhn (1987)), there are indications that this may be true. For example, there are notable differences in how students of these disciplines are trained when using statistical approaches. The ACM SIGCHI curriculum for Human–Computer Interaction (Hewett et al., 1992) is relatively sparse on statistical literacy, indicating it may not be a current priority for the field. The document states students should be taught statistics, but does not go into detail on what statistical methods are appropriate. Contrast this with the American Psychological Association (APA) curriculum for undergraduate psychology students, which dedicates more detail to what specific statistical analysis students should be capable of doing (Halonen et al., 2013). These differences in priority can be considered a proxy of community common practice, and indicate that different disciplines approach games research differently.

This also creates a situation where there is yet to be an established common language used by the field when reporting questionnaires, which will inevitably lead to less collaboration or meaningful advancement. To relate this back to questionnaires used in player research, suppose that psychology researchers deem reporting reliability scores is important, as this demonstrates the questionnaire is reliable within the current sample. In contrast, perhaps game researchers place less emphasis on reliability and care more for why a specific questionnaire has been used (as there are numerous available to the field). That is not to say that either discipline does not care about reliability or a justification for use, just that the emphasis is placed differently due to their disciplinary priorities. The effect of these priorities is a reduced chance of non-prioritised aspects being reported, making the work less accessible to other disciplines. By omitting aspects of questionnaire

usage and reporting, other researchers can neither meaningfully understand or build on one another, reducing the chance of substantive research synthesis.

To overcome these consequences in both a new and multidisciplinary field, there is a need for standardisation of the shared repertoire — in this case, questionnaire usage and reporting. The individual actions of researchers, potentially influenced by their disciplines of origin, combine to create the practices of the field of measuring player differences. It is therefore important to know both a justification for questionnaire use and how it was used within the current sample, as both indicate researcher intent and the validity of the approach. These practices form the basis of how research is conducted and reported, and so assessing them provides insight into the current structure of the field. Therefore, being transparent about how questionnaires have been used makes it easier to evaluate their usage, which makes it easier to establish a shared repertoire, and consequently aids research synthesis. To assess the extent to which questionnaires are reported differently across disciplinary influences, it is first important to consider how questionnaires should be reported, to maximise transparency. This is discussed in the next section.

2.2. What is good questionnaire reporting practice?

The goal of method reporting is to be as clear as necessary to allow other researchers to understand the work carried out, but also that they could reproduce the work if needed. To allow this, reporting should aim to be transparent, which is equally true for questionnaire usage. This is because how items are phrased, how they are asked to participants, and how the structure of their respective sub-scales is treated can affect results (Cairns, 2019). Consequently, this means researchers who use the same questionnaire but alter it in different ways from the original risk being unable to truly compare results to other work. Without comparison, there is no way to build on previous work and add to the body of knowledge investigating player motivation.

Therefore, this section outlines previous guidance given for measurement reporting, and builds on this to provide explicit criteria for questionnaires within player research. Eight practices that increase the transparency of questionnaire usage are outlined. On the whole, the more practices included in a paper, the more transparent the reporting of the questionnaire is. This in turn means the work is more reproducible and can be more effectively built on by future work.

2.2.1. Previous guidance for questionnaire reporting

Methodological concerns for scientific writing are not new, though are less explored within player research. Unclear and un-standardised reporting practices have been noted in the medical sciences, understood as being driven by a lack of priority on methodology in favour of results (e.g., Van Calster et al. (2021)). This is not just a problem of authors but a lack of oversight and enforcement from publication venues, as many publication guidelines do not state how questionnaires – or indeed any type of survey – should be reported (as little as 7% of medical journals; (Bennett et al., 2011)). With no clear and unified guidance for how methodologies should be deployed, the medicine discipline risks being undermined, with trust lost in findings being valid or reliable. Whilst the focus of medicinal research is not the same as player research, a lack of clear reporting practices in both leads to the same issues of lost trust.

The push underway in medicine should serve as a warning to those researching games that methodology is not a trivial aspect to be overlooked. Setting a precedent now will protect against future 'crises' that could occur from a lack of standardised methods. In a similar fashion, if psychology had taken notice of the replication crisis (where multiple key studies in the field are repeatedly shown to not be significant; (Maxwell et al., 2015)) earlier within its field, perhaps it would not have been a crisis after all. By learning from the mistakes other fields have made in undervaluing methods, player research may be able to avoid these pitfalls before they manifest.

Whilst less common, there are a few examples of research evaluating questionnaire reporting practices within the field of player research. One notable paper is the recent review of CHIPLAY submissions by Aeschbach et al. (2021). The authors analysed the 24 publications from 2020 that used self-reported measures on how transparent this reporting was, and if justifications for measures were given. This was done by analysing papers on 5 aspects based on the work of Flake and Fried (2020): construct definition, construct operationalisation, measurement selection, measurement modification, and measurement self-development. Of particular interest to the current work are the aspects of measurement selection and modification. The authors found justifications for why a particular measure was selected – measurement selection – were uncommon (16 out of 84 measurements), and within this justifications ranged from short sentences to more in-depth explanations. In contrast measurement modification was common at 38.71%, and 69.05% provided some administration details about the measurement (e.g., number of Likert points used). This indicates reporting practices within CHIPLAY are varied and not always transparent. However, the work by Aeschbach et al. (2021) is limited to one conference in one year, so cannot reflect the field overall. This is especially limiting as the interdisciplinary nature of the field discussed above leads to the possibility that not all researchers will publish in a venue such as CHIPLAY. This means their practices cannot be evaluated, as such an analysis would require a venue-agnostic search approach.

Other guidance for reporting practices tend to be somewhat technique-agnostic, where multiple techniques are described and given advice. This means advice is either less tailored towards the specifics of questionnaires, or only provides brief advice (such as how to report exclusion criteria or removal of data; (Boynton, 2004)). Even guidance that is specific to questionnaires sometimes spends more time discussing how to report results than how to report the administration (e.g., Boynton (2004)). The reporting guidance of Kelley et al. (2003) provides a checklist for survey research, which considers a collection of survey methods that include questionnaires. However, the goal of the guidance was to discuss the entirety of survey reporting practice (from the research aim to the discussion of findings), meaning less time was dedicated on how to report the method used (for the focus of this work, the questionnaire). There are seven main points contained within the guidance for reporting survey research, of which only one is related to reporting the questionnaire itself. Within this point there are three sub-components; reporting a justification for the method, describing the research tool (such as its psychometric properties and providing a reference to the original version), and describing the sample.

However, these sub-components are short and abstract, making them less practically useful for a checklist style approach. Whilst the sub-component asking researchers to describe the research tool is the closest to the aims of the current work, it is still lacking in specificity. What details are needed to describe the tool? What psychometric properties should be reported? How much detail is needed for a method to be considered 'justified'? This lack of specificity, which allows the guidance to apply to a number of survey techniques, comes with the drawback of being practically un-useful for researchers specifically deploying questionnaires — a common problem seen when using any heuristic (e.g., Hermawati and Lawson (2016)).

Therefore, whilst guidance exists for reporting questionnaires, there is a general lack of specific and easy to interpret criteria that cover the smallest of steps to improve transparency. To address this, the current work aims to provide guidance for questionnaire usage that is specific in its practical application, yet useful for any application within player motivation research, via the creation of a transparency checklist. This is built from the common practices of the multiple disciplines present, and is designed to capture the most essential aspects of transparency in reporting. These essential aspects are discussed in the following section.

2.2.2. The 3 aspects of transparent questionnaire reporting

When reporting a questionnaire transparently, it is important to consider the following three aspects:

- **Validity**: The concept that what is being measured via a questionnaire is 'true' and measures what is intended to be measured. It allows readers to evaluate if the chosen measures are sound, and potentially relevant to their own work/future work.
- **Reproducibility**: The ability to understand what has been done in a study allows researchers to rerun the work and look for replication in findings, or run a similarly designed study. The latter allows researchers to deploy the same questionnaire for a similar but different study, with the confidence that the new work accurately builds on the old. In a sense, designing for reproducibility is the act of looking forward, by considering how future works will benefit from the current work.
- **Clarity**: The ease in which readers can understand how a questionnaire has been used, including the simplicity in explanation and the presentation of the information.

This leads to transparency being defined as a combination of the three above aspects. Therefore, transparency is the ability for a reader to understand what a questionnaire has measured and how it was used, which is written in a concise and clear way. Doing this effectively makes it clear how the questionnaire relates to the research aims. For questionnaire reporting, there are a number of ways these three aspects of transparency can be considered. Generally speaking, considering reproducibility and validity naturally increases transparency in writing, as doing so enables readers to clearly see what has been measured, how it has been measured, and why. Therefore, it is important to consider how reproducibility and validity can be captured within questionnaire reporting, in a way that is clear for the reader to follow (clarity).

Each of these aspects can be assessed in several ways. To allow assessment of transparency, the current study devised a number of criteria that fall under these aspects. The following section explains the creation of these transparency criteria, and which aspect they correspond to.

2.2.3. Transparency criteria creation

As discussed in Section 2.2.1, existing guidance for reporting survey research is typically abstract and lacking in focus for the reporting of the questionnaire itself. Further, they have not been designed through the lens of transparency and the three aspects discussed previously (reproducibility, validity, and clarity). Therefore, to assess transparency in the current work, new criteria were created from the previous guidance that were more specific and suitable for a checklist approach.

This was done by extracting criteria specific to measurement reporting from the previous guidance discussed above. These were then broken down into smaller, more actionable guidance, such as defining psychometric properties as relating to face validity and reliability, using the advice given in Cairns (2019) which discusses how to use statistics within HCI. This led to the creation of eight criteria that provide a basis for measuring questionnaire transparency within player motivation research. Whilst there are more considerations that could be made (e.g., sample size, data cleaning steps), these eight criteria capture what could be considered the 'bare minimum' of information that increases transparency of questionnaire reporting. These fall under the aspects of transparency in the following ways:

Validity was assessed on three aspects:

- **Discusses Why Questionnaire Used**: With numerous questionnaires to choose from, why does the reported research use this one in particular? This relates to construct validity, as explaining why something is measured links it back to a theoretical base. A clear justification can also be scrutinised by readers who can decide if they agree with the questionnaire selection. This criterion was inspired by the transparency analysis of Aeschbach et al. (2021) which explored justifications for measurement use.
- **Examples of Items Provided**: Clearly indicating to the reader the types of questions asked relates to face validity, as readers can assess for themselves whether the items reflect the construct they are purporting to measure. By providing this in the text the reader does not need to be an expert in the questionnaire chosen, or seek the information themselves from the original document. This was inspired by the guidance of Kelley et al. (2003) which asked researchers to report the psychometric properties of their surveys.
- **Reliability Checks Run on Sample**: Whilst reliability is a separate concern to validity, reporting the reliability of a questionnaire on a current sample is done to show that results are comparable. Reporting reliability scores therefore shows that the current study sample can be compared to previous work using the same questionnaire. This was also inspired by the guidance of Kelley et al. (2003) for reporting psychometric properties.

For **Reproducibility**, the following five criteria were used:

- **Correct Citation**: Whilst perhaps the simplest step to achieve, it is important to exercise caution when using player questionnaires. There are multiple versions available for some, with varying item structures that correspond to different factors and concepts, represented by the different versions of the questionnaires. Reporting the correct version allows for replication as readers will know what items have been used.
- **Number of Likert Points Reported**: The number of points used for items can vary between questionnaires, so reporting the one used in the current study helps to clarify the setup of the experiment. Whilst altering the number of points does not typically alter the results (Cairns, 2019), it is still good practice to be clear of the procedures involved. This criteria was inspired by the analysis of Aeschbach et al. (2021) who measured the number of Likert points within the analysis of administration details. Other aspects of Likerts are also useful for readers to know, such as the anchors used, which can further improve transparency. However, the current criteria are intended to be the 'bare minimum' of inclusion for transparency, and so only one aspect of Likerts was considered for simplicity.
- **Can Tell if Items are Dropped**: Removing items from a sub-scale is bad practice if results are to be compared to previous work. This is because removing items alters the structure of the sub-scale, to the point it is possible the sub-scale no longer measures what it intends to (Cairns (2019)). Therefore, researchers should make clear if they have removed items, so readers know to be cautious when comparing results to other work or if replicating findings. This criteria was inspired by the analysis of Aeschbach et al. (2021), who brought attention to how measurements are modified.
- **Can Tell if Items are Reworded**: Similarly to above, altering items by changing the wording can also damage the structure of a sub-scale (Cairns, 2019). The extent to which this is true depends on how much items are altered, and the context in which it is done. For example, are the items changed from 'game' to 'games', or are phrases rewritten? Is the questionnaire used in a practical setting to assess how motivation affects responses to a game, or is the questionnaire itself being validated in a new setting/sample? These examples highlight that researchers have multiple reasons to use questionnaires, and so to accommodate this rewording should be reported as transparently as possible. Overall, it is advisable to edit items as little as possible, though it is sometimes necessary and somewhat unavoidable (e.g., to fit a specific gaming context). Researchers should be clear if items have been reworded so readers know this may reduce the ability to compare results to other works, and so they know what items to ask if replicating the findings.

- **Items Available With the Text**: This final consideration is the easiest solution to including many of the above considerations. Allowing readers to see the items used for themselves provides the most transparency, and the easiest way to replicate the work as they can lift the items from the text. Due to word limits it is not always possible to include items within the text, but it is possible to include these as an appendix or on an external website such as the Open Science Framework (OSF). This criterion was inspired by the Open Science movement (Foster and Deardorff, 2017) which aims to increase transparency by making data and materials available with papers.

Clarity as a consideration is not so easily defined or measured. Therefore, it is not included as a criteria, but is offered more as a reminder to researchers that how information is displayed to the reader impacts on the transparency of the work. A paper could contain all 8 criteria stated above, but if this is done obscurely this reduces the chance the reader will be able to find them in the paper when searching.

Therefore, including these details when reporting questionnaire use allows for considerations of both validity and reproducibility, which in turns increases transparency. Because of this, the present study uses these criteria to create a checklist that can assess the level of transparency for questionnaire reporting. However, it is important to note these criteria do not reflect the quality of a paper. A highly transparent paper can still be flawed in other ways, even within the criteria assessed (e.g., items are clearly reworded, but the impact this has on the results is not mentioned). The criteria used in this study therefore only consider transparency, and do not assess how this influences the quality of the work being reported.

3. Method

3.1. Aim

To date, there has been no scoping review of the usage and reporting of player motivation questionnaires, that also acknowledges the varied disciplinary background of researchers. Because of this, there has been no evaluation of how established the research community is in terms of a shared repertoire, and consequently no establishment of a standardised practice on which all researchers can build on. Now that the field is over 15 years old, it is important to analyse how questionnaires are used so that this common practice can be observed and standardised. This standardisation will ensure researchers using the same questionnaires are doing so in comparable ways, allowing the field to confidently grow together.

Therefore, the aim of the current study is to analyse current reporting practices when using player motivation questionnaires, to explore the maturity and cohesiveness of the field overall. This is done through a multidisciplinary lens, where both why specific questionnaires are used and how transparently this reporting is made are assessed.

3.2. Questionnaire collection

To collect the number of questionnaires available that measure player motivation, a literature review was conducted. A scoping review approach was used, as this can observe research conduct as well as help identify knowledge gaps in the literature (Munn et al., 2018).

The following search string was used in a variety of search engines (such as the ACM digital library and Google Scholar): [“Video Game” OR “Game”] AND [“Player Trait” OR “Motivation” OR “Preference” OR “Player Type” OR “Typology”] AND [“Questionnaire” OR “Inventory” OR “Instrument”] AND [“Factor Analysis”] AND [“Develop”]. This resulted in a starting sample of 368 papers that purported to measure an aspect of player motivation. Data cleaning steps are summarised in Table 1.

Table 1
The steps taken to achieve the final sample of questionnaire papers analysed.

Search step	Number of papers
Starting sample	368
Specific to games	194
Specific to motivation/preference/trait	155
No addiction/problematic gaming	119
No gamification/serious gaming	103
No unidimensionality	67
General sample	54
Peer Reviewed publications	43
General game application	22
Used at least once	**18**

Despite the specificity of the search term, many results were not relevant to games/digital games, so these were excluded to leave 194 papers. As this work was looking at player motivation (including preferences and traits), measures were excluded that studied player experience, engagement, immersion, flow and enjoyment. These questionnaires relate to what players feel and do within games, and are therefore situational. Motivations on the other hand are supposed inherent properties of a player that drives them towards certain activities. As the current research is focused on the individual differences players bring into games, rather than what they experience within them, these situational questionnaires were not included. This is with the exception of the Player Experience of Need Satisfaction, as it has been used as a motivation questionnaire in previous work (e.g., Lee et al. (2017)), and is one of the most commonly used. Therefore, removing these questionnaires left 155 papers.

Next, measures relating to addiction/problematic gaming were excluded, as these relate to effects of gaming rather than reasons to game. For example, many of these questionnaires ask players how excessive gaming affects their well-being (e.g., the Ten-Item Internet Gaming Disorder Test asks players to reflect on how gaming affects their work and relationships with others; (Király et al., 2019)). This makes them outside of the scope of the current work, leaving 119 papers.

Questionnaires relating to gamification/serious games were then also removed, as the motivations of interest are those for gaming rather than learning. Many questionnaires in this area ask players to evaluate gamified/learning environments, rather than ask them what they are motivated to do in these games. For example, the questionnaire by Zurita Ortega et al. (2020) was excluded as items were aimed at measuring the user’s evaluation of the game (e.g., “Indicate the degree to which it has been easy for you to learn to play”). Of note is that this cleaning step did not remove all work from the discipline of Education — questionnaires were included that could be applied to learning but whose main focus were understanding player motivations (such as the Video Game Pursuit Scale e.g., “I lose track of time when I play video games”; (Sanchez and Langer, 2020)). This left 103 papers.

Only questionnaires that considered multidimensional aspects of players were considered, as the current work aimed to understand general player motivations (i.e., questionnaires that measure a multitude of motivations). Unidimensional scales are typically very specific (such as griefing behaviours in online multiplayer games; (Ladanyi and Doyle-Portillo, 2017)), and so difficult to compare their usage to one another. Therefore papers that only measure one difference (e.g., curiosity) were removed, leaving 67 papers. Measures specifically designed for children were then excluded so that all questionnaires were designed for adults (leaving 61 papers), and those with a specific sample such as ‘Chilean Millennials’ and female gamers were also excluded (leaving 54 papers). Doing so left a selection of questionnaires designed for what could be considered a ‘general’ gaming population, where general was based on author intent (i.e., the creators intended for the questionnaire to be used on a non-specific gaming population). For example, Vermeulen et al. (2017) includes items such as ‘If I don’t do well in a game, it might be viewed as stereotypic of my gender’, which, whilst related

Table 2
The questionnaires used in the current work, with their discipline of origin and acronym used throughout.

Questionnaire	Discipline of origin	Acronym
A Framework and Taxonomy of Videogame Playing Preferences	Games	FTP
Beyond the "Core-Gamer": Genre Preferences and Gratifications in Computer Games	Media	Core
An Instrument for Measuring Individual Motives for Playing Digital Games	Psychology	DeGrove
The Gaming Motivation Scale (GAMS)	Psychology	GAMS
Empirical Taxonomies of Gameplay Enjoyment: Personality and Video Game Preference	Education	GEM
Falling in Love with Online Games: The Uses and Gratifications Perspective	Media	Wu
Five-Factor Inventory of Intrinsic Motivations to Gameplay (IMG)	Hybrid	IMG
The Gaming Attitudes, Motives, and Experiences Scales (GAMES)	Psychology	GAMES
The Electronic Gaming Motives Questionnaire	Psychology	EGMQ
BrainHex: a Neurobiological Gamer Typology Survey	Games	BrainHex
The Demographics, Motivations, and Derived Experiences of Users of MMORPGs	Games	Yee
The Metacognitions about Online Gaming Scale	Psychology	MOGS
The Motivational Pull of Video Games: A Self-Determination Theory Approach	Psychology	PENS
The Trojan Player Typology	Games	Trojan
Gameplay Activity Inventory (GAIN) for Modeling Player Profiles	Games	GAIN
Video Game Pursuit (VGPu) Scale	Education	VGPu
Video Game Uses and Gratifications as Predictors of Use and Game Preference	Media	Sherry
The Motives for Online Gaming Questionnaire (MOGQ)	Psychology	MOGQ

to motivation, is specific to the female gamer context. As the authors of these questionnaires wished to study specific contexts, it was considered an unfair comparison to include them in this study. In doing so, this could bias the selection towards specific countries of collection such as the United States. However, many of the questionnaires in the final sample collected participants online, widening the pool of potential participants outside of the author's home country.

Only those papers accepted as peer reviewed publications were included (i.e., no preprints or theses). Questionnaires had to take the form of Likert point responses (as opposed to a binary choice or open ended questions), as Likert scales are considered one of the most reliable ways to measure self-reported traits (Likert, 1932; Maurer and Pierce, 1998). These scales must also have been validated in some way within the study, typically via a factor analysis. Removing these aspects left 43 papers. Finally, only measures that considered games as a unified entity were included, removing papers relating to specific game genres such as sport video games (e.g., Kim and Ross (2006)). This left a sample of 22 questionnaires that represents statistically validated, multidimensional questionnaires that look at general player motivations of adults across potentially all types of digital games.

However, at the time of data collection only 18 of these questionnaires included citations that actively used the cited questionnaire. Therefore, the final sample is highlighted in Table 2, which also indicates the discipline that influenced the work. Discipline was established by assessing the motivations for creation as discussed in Section 2.1. For example, Psychology questionnaires typically reference psychological theory (such as Self Determination Theory), whilst Media questionnaires cite media theory (such as Uses & Gratifications). Education questionnaires were motivated to understand how game motivations can be applied to education settings, whilst Games questionnaires wished to understand players for the sake of understanding game behaviour. One questionnaire was designed to specifically join together both Games and Psychology research – the Inventory of Intrinsic Motivations to Gameplay (IMG) – and so was considered a hybrid.

From this base set of collected questionnaires, the following section discusses how the citations (and by extension their uses) of these questionnaires were collected.

3.3. Citation collection

In total, the 18 questionnaires had been cited 6970 times on Google Scholar at the time of collection. All citations for each questionnaire were analysed except for Yee and PENS, as they could not practically be fully analysed due to their large citation counts (1727[1] and 2656 respectively at time of collection). However, as these are the two most highly cited works in the field, it was still deemed necessary to include them. Therefore, the first 1000 results from Google Scholar were analysed for each. Results from these two questionnaires consequently do not reflect the entire range of uses, but provide a large sample (roughly 39% of PENS and 59% of Yee 2006a) that can indicate trends in general use.

A further consideration for collection was that some questionnaires have multiple versions. For example, BrainHex includes a paper for preliminary findings and the final model ((Nacke et al., 2011) and (Nacke et al., 2014)). As items did not change between papers, the citations for both papers were analysed but treated as the same BrainHex questionnaire. Other questionnaires have revised versions, such as the Gameplay Enjoyment Model (GEM; (Quick et al., 2012)), but most notably the work of Yee. For GEM, the citations for both versions were collected, but as the revised version had no active uses at the time of collection these were treated as one questionnaire. In the case of Yee, the 2006a paper was used to collect citations. This is because it is the original version, though less cited than the 2006b version (but moreso than 2012). As described above, the high volume count for Yee across its many versions made full data collection impractical, so the first 1000 citations of the 2006a paper were analysed. This means results involving Yee are limited to works citing this version, which will not represent the full range of Yee uses.

Overall 4587 papers were analysed, where only those that actively, empirically used one of the questionnaires were included. Empirical use required a paper to deploy the questionnaire to a sample, rather than use the theory behind them to inform experimental design. Questionnaires were not always used in their entirety (e.g., selected sub-scales were lifted), and sometimes items from questionnaires were combined with others to create new scales. The latter was done either by re-using the items with minimal rewording, or were 'inspired' by the items and therefore not actively used. Therefore, papers that deployed the items in new sub-scales with minimal rewording were included, along with those using specified sub-scales. Furthermore, inclusion criteria were as follows: The work is published (no preprints), and the work is in a paper format (no theses or book chapters). This resulted in 270 papers in total.

A list of the papers analysed can be found at https://osf.io/zetbw/, which also assigns the number used to refer to it in this work. This is done to remove the intention of 'naming and shaming' any one paper, as this work aims to provide an overview and general suggestions for the field to improve. Assigning numbers to papers has been done for this reason before, such as Aeschbach et al. (2021). Therefore, the following results refers to papers in the format of "Paper 1".

[1] For the original 2006a paper.

3.4. Data analysis

3.4.1. Reasons for use

A conceptual content analysis was performed on the papers by collecting any sentences that referenced the questionnaire used, and coding any aspects of the questionnaire that have been reported contained within these sentences (Krippendorff, 2018). These references were found by searching papers for both the citation of the questionnaire used, and for its common name if relevant (e.g., BrainHex). Sentences containing the reference were selected, as well as the surrounding sentences to provide further context. Typically this resulted in sentences being extracted from literature review and methodology sections, though sometimes also discussions. The text extracted and used to code the reasons for use can be found at the following OSF link (https://osf.io/zetbw/).

The sentences per paper were then analysed for what about the questionnaire was mentioned in an iterative fashion. Papers could be given multiple codes based on how extensive the explanation for usage was. The first pass relied on the wordings used within the papers (such as 'replicating past work' and 'comprehensive model'), which resulted in 37 codes. These initial codes were then collected and organised into more cohesive and standardised names, resulting in 15 codes which were then reapplied to the data. As 15 codes is large for qualitative coding, these codes were also assigned into one of four meta-codes, described in the following Results section. These meta-codes provide a summary overview of the reasons for use, whilst preserving the nuance of the codes contained within.

After assigning the codes to the papers, a separate code referring to how the aspects of the questionnaire were referenced was given; explicitly (e.g., "we used this questionnaire because…") or implicitly (i.e., features of the questionnaire detailed but never explicitly linked to why it was chosen e.g., 'reliable', 'comprehensive'). Papers that only stated what the questionnaire measured (e.g., 'motivation') were classed as providing no reason.

This distinction is important for two reasons. Firstly, when no explicit sentence is included, aspects mentioned for a questionnaire may not be justifications but rather statements of fact. For example, does stating that the original Yee questionnaire had a substantially large sample (e.g., Paper 216) mean this was the reason it was chosen, or is this mentioned only to provide context? Secondly, it is common practice when writing literature reviews sections to reference multiple questionnaires (e.g., Papers 21 and 9). This outlines the scope of the field, but highlights a problem when only one questionnaire is subsequently used without explicit reasoning for why — what about *this* questionnaire, and not the others discussed, made it the best option? Whilst highlighting specific aspects of a questionnaire does not conclusively mean this was the reason it was chosen, it does indicate what researchers value in questionnaires, reflecting their priorities. Therefore, both implicit and explicit reasons for questionnaire use were recorded.

3.4.2. Transparency scores

These papers were evaluated on the eight criteria described in Section 2.2.2:

1. The **correct** questionnaire is cited (including the correct version)
2. The paper discusses **why** this specific questionnaire was chosen (moreso than simply stating what it measures e.g., 'motivation', but rather why this *specific* questionnaire as opposed to another)
3. The number of **Likert** points used is reported (either with the text or as part of a table)
4. The paper includes **examples** of items from sub-scales deployed (rather than describing the sub-scale)
5. It is possible to know if items were **dropped** (e.g., the number of items used are reported, or items are available with the text).
6. It is possible to know if items were **reworded** (e.g., an explicit statement is given, or items are available for inspection with the text). This includes all instances of rewording, from small changes such as 'this game' to 'games' to full rewrites
7. **Reliability** analyses were computed and reported (typically via Cronbach's alpha)
8. The items are **available** with the text (including the appendix or an external website such as the Open Science Framework (OSF))

The presence of each resulted in a score of 1, for a maximum total of 8. Therefore, higher scores indicate higher transparency in reporting questionnaire use. Alongside analysing transparency scores, a content analysis was run on the reasons listed for why questionnaires were used. To do this, any reference to the questionnaire or citation to the paper was collected from the paper and analysed.

3.4.3. Assigning paper discipline

Whilst it is easy to observe the originating discipline of a created questionnaire using information such as the theories cited behind its conception, it is less clear how to assign disciplines to papers citing this work. It would not be feasible to assign a discipline to each paper, as the unclear boundaries of community participation make this information unattainable without asking each author for their self-identified discipline.

To overcome this, the idea of discipline influence is used. Influence refers to the idea that when people use something, they can adopt some of its wider properties secondhand. When citing work from a specific discipline, a number of assumptions and biases of that discipline can be naturally inherited, such as how a questionnaire 'should' be reported. This is in many ways how science operates; those that come after copy and adapt from those that came before by adopting their beliefs (Kuhn, 1970).

In the context of using player motivation questionnaires, researchers will be exposed to the practices of the discipline in which the questionnaire originated. It is therefore possible that each discipline identified within the player motivation questionnaires (Games, Psychology, Media, and Education) has their own set of lenses, which draw researchers to prioritise certain reporting practices over others. This is not a bad thing in itself, as science is achieved through incremental work where lenses are naturally inherited. However, the likely existence of multiple lenses makes it more difficult for different disciplines to access the work of others. To achieve a true community of practice, such lenses should be acknowledged and understood.

Therefore, the current work performs a further exploratory examination to assess the presence of disciplinary influence within the data. Whilst it is not possible to know a paper author's discipline from the current data, it is possible to see which questionnaire has been used, and so what discipline they may have been influenced by. To assess discipline influence, papers were categorised depending on what questionnaire was cited; for example, a paper citing PENS was classified as influenced by Psychology, whilst a paper citing Yee was classified as a Games influence. Therefore, these categories potentially reflect the lens of the originating discipline, and so may indicate how the multiple disciplines prioritise aspects of reporting. By extension, this may indicate how the multidisciplinary nature of player motivation research has influenced the community of practice surrounding questionnaire use.

4. Results

In total, 270 papers were collected that used one of the 18 questionnaires studied, of which 57 (21%) created a new scale. The breakdown of number of uses per questionnaire is displayed in Table 3.

In total there were 306 uses of questionnaires, higher than the 270 papers studied as multiple questionnaires were sometimes used (32 papers). To keep score comparisons fair, these 32 papers are not considered in the following analyses. In these cases, it was common

Table 3
A summary of the number of uses for each questionnaire studied, in comparison to their original citations.

Questionnaire	Total uses	Original citations	Percentage
PENS (Ryan et al., 2006)	75	2656 (1000)	8%
Yee (Yee, 2006)	72	1727 (1000)	7%
Sherry (Sherry et al., 2006)	35	948	4%
MOGQ (Demetrovics et al., 2011)	30	209	14%
Wu (Wu et al., 2010)	29	380	8%
BrainHex (Nacke et al., 2014)	19	357	5%
GAMS (Lafrenière et al., 2012)	17	141	12%
GAMES (Hilgard et al., 2013)	6	117	5%
Trojan (Kahn et al., 2015)	5	103	5%
DeGrove (De Grove et al., 2016)	5	55	9%
MOGS (Spada and Caselli, 2017)	3	30	10%
IMG (Vahlo and Hamari, 2019)	2	11	18%
Core (Scharkow et al., 2015)	2	82	2%
GEM (Quick et al., 2012)	2	9	22%
EGMQ (Myrseth et al., 2017)	1	4	25%
VGPu (Sanchez and Langer, 2020)	1	4	25%
GAIN (Vahlo et al., 2018)	1	14	7%
FTP (Tondello et al., 2017)	1	36	3%

for a paper to produce different scores depending on the questionnaire usage being studied. For example, one questionnaire could score highly with an in-depth explanation of its use, whilst the second questionnaire is only provided a sentence. This means the overall score of the paper would be damaged, making it difficult to fairly compare to works with only one usage (the majority of the sample). Therefore, 238 papers make up the final analysis.

The most commonly used by count is Yee and PENS with 75 and 72 uses respectively, followed by Sherry, MOGQ and Wu at 35-29 uses each. This is followed by BrainHex and GAMS with uses between 19 and 17, with the remaining questionnaires being used less that 10 times each. Therefore, 53% uses are accounted for solely by Yee and PENS, with 91% accounted for in the top seven most used (those with more than 10 uses each). This indicates a clear preference for certain questionnaires over others, though this may in part be explained by the year of publication. Indeed, the top three most used are also the oldest, with all three published in 2006.[2] The number of uses for each questionnaire is low compared to the number of overall citations, indicating that most work that references questionnaires does not seek to use them empirically. It can be seen however that the more citations a paper has, the more likely it is to have been empirically used at least once.

The following sections explore the reasons given by researchers for their choice of questionnaire, followed by an analysis of transparency scores.

4.1. Reasons for use

A content analysis was run on the reported reasons for using questionnaires, where a total of 413 aspects were mentioned across 238 papers, for a mean of 2.74 reasons mentioned per paper (median = 2). Of these, 44 papers gave explicit reasons for questionnaire use (e.g., Papers 55 and 134), leaving 125 with implicit reasons and 69 with no reason outside of what the questionnaire is used to measure (e.g., "we used this questionnaire to measure player motivation"; Papers 51 and 14). As many questionnaires could be used to measure motivation, this reasoning does not support any one specific questionnaire being used. Overall then, 72% of papers stated a reason for why a specific questionnaire was used.

Within the papers stating reasons for use (beyond what it assesses), 15 types of reasons were identified. These are presented in Table 4, along with a description of the category and the type of reason it is classified as. Fig. 1 shows the distribution of each reason. Following is an overview of each reason with examples to illustrate the categories.

The most common reason given for choosing a specific questionnaire, **Based on X**, was that it was based on a theory of relevance (65 papers; 16%). For example, the most common underlying theory referenced was Self Determination Theory (41 out of 65 papers), which frequently made reference to the questionnaire being based on this theory; e.g., "Based on SDT and other relevant theories (e.g., presence), Przybylski and colleagues developed the Player Experience of Need Satisfaction (PENS) measure" (Paper 159). Uses and Gratifications Theory from Media was also mentioned (8 papers; e.g., "Students' preference for competition was assessed before the serious game started by using a selfreport questionnaire adapted from the uses and gratifications scale by Sherry et al. (2006)" (Paper 94)), as well as neurobiology in reference to BrainHex (9 papers; e.g., "we have also collected [...] their player profile according to the BrainHex model [18], which is based on neurological research related to gameplay" (Paper 19)).

The second most common reason, **Used to Measure X**, refers to instances when the questionnaire has been used to measure a specific variable of interest (62 papers; 15%). This could be by referring to what the original publication assessed to establish criterion validity (e.g., Paper 4; "Consistent with this assumption, Lafrenière et al. (2012) observed that a stronger intrinsic motivation toward gaming was associated with perceiving higher needs satisfaction during gaming"), or by referencing what others have used the questionnaire to measure (e.g., Paper 161; "[PENS] has led to a better understanding of how vitality is maintained or enhanced [5], how personality interacts with need satisfaction [9], and what compels people to play as opposed to why people choose to play [2]"). Therefore, this reason considers what constructs the questionnaire can be associated to, and what correlations have been found previously.

Thirdly, papers would make reference to questionnaires that were **Valid** (59 papers; 14%). This could be in reference to the psychometric properties of the questionnaire/how the questionnaire was originally validated (17 papers e.g., Paper 7 — "This measure has demonstrated good reliability (i.e., internal consistency) and validity (i.e., relating to need satisfaction and gaming frequency as expected)), or the questionnaire having been validated by others (15 papers, such as Paper 40 — "Research indicates that the GAMS has adequate levels of validity and reliability [44]".). This may also relate to who the questionnaire has been validated with (7 papers), such as Paper 31 — "all participants completed the Digital Games Motivation Scale (DGMS), an internationally validated questionnaire used to assess different motivations for playing games [20, 21]". However, it was also common for papers to simply state the questionnaire was valid without further elaboration (20 papers, such as Paper 5: "Recently, Lafrenière, et al. (2012) developed a gaming motivation scale, a valid assessment of gaming motivation"). This makes it difficult to assess what the authors mean by valid in the context of their usage, as it could refer to the validation performed by the original authors, others, or the sample used.

38 papers made reference to a questionnaire being **Mature** in some way, where researchers described the questionnaire as being comprehensive or complete in a way that suggests it is more appropriate than other measures (9% of papers). For example, the term 'complete' was used to distinguish BrainHex from other questionnaires — "We chose the BrainHex model to classify our students regarding their gameplay style because it is one of the most complete works in the field" (Paper 19). This could also be referred to as the questionnaire being established, and therefore mature in its existence; "The measurement was adapted from well-established constructs in the literature" (Paper 223). Further, some papers argued the chosen questionnaire covers the 'full' range of motivations to play (e.g., Paper 66: "[MOGQ] is a 27-item self-report measure used to assess the full range of motives for online gaming").

31 papers highlighted the chosen questionnaire was **Reliable** (8% of papers). Similar to Valid, reliability was not always specified as to what

[2] Sherry also has a version released in 2003, which is sometimes cited in the papers studied. However, this version is not openly available online nor is as widely cited.

Table 4
The 15 reasons given for using a specific questionnaire, in order of their frequency. Raw frequency is shown in Fig. 1.

Reason	Type of reason	Definition: The questionnaire...
Based on X	Structural	Has a theoretical underpinning
Used to Measure X	Community practice	Has been used to measure a construct
Valid	Soundness	Has been demonstrated as valid
Mature	Community practice	Is comprehensive, has existed for a length of time, or built from previous works
Reliable	Soundness	Has been demonstrated as reliable
Structure	Structural	Has specific qualities that make it appropriate for use e.g., non-genre specific
Data Approach	Structural	Has been based on a specific type of data
Sample size	Structural	Has been developed from a large sample size
Used in previous work	Community practice	Has been used by others in the past
Popular	Community practice	Has been frequently used by others
Replication	Replication	Structure is being tested in a new setting
Original sample	Structural	Has an appropriate original sample
Recent	Community practice	Has recently been developed
Ease of use	Structural	Is easy to access and use
First	Community practice	Is the first to measure a specific construct

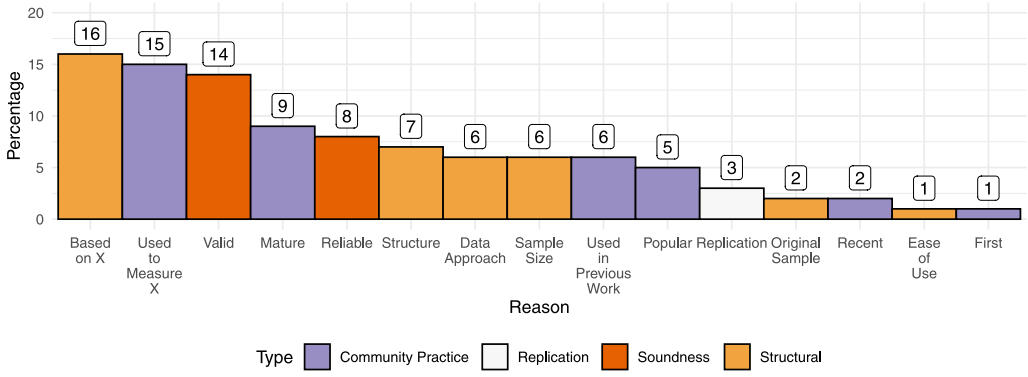

Fig. 1. A histogram of the reasons given for using a specific questionnaire.

exactly this meant. For example, it was commonly alpha scores (19 papers), but 3 papers only referred to 'internal consistency' — "The scale has previously shown good reliability in terms of internal consistency" (Paper 61). Further, 9 papers did not specify what reliability referred to (e.g., Paper 137: "The PENS scale used in previous studies (Przybylski et al. 2009, Tamborini et al. 2010) revealed a good reliability").

29 papers make reference to the **Structure** of the original questionnaire, typically an aspect that made it uniquely fitting for the current setting (7% of papers). For example, this could relate to being specific to a type of genre (e.g., Paper 69; "Demetrovics and colleagues [15] were able to identify seven primary motivational factors in gaming behavior applicable to all types of online games"), or could be non-genre specific (Paper 192; "The PENS scale has been utilized to assess players' in-game needs satisfaction in various game genres"). This could also refer to the questionnaire having a number of sub-scales which provide further nuance to results (e.g., Paper 164; "We selected them on the basis that they offer multiple subscales designed to assess different components of player experience").

26 papers mentioned the **Data Approach** of the original questionnaire, such as what data and data analyses were used to develop the questionnaire (6% of papers). This was frequently a mention of factor analysis (12 papers), but equally could also be a previous data collection step such as focus groups or interviews to generate items (12 papers) — "This instrument is based on a number of recurring motives reported in interviews"; (Paper 83). Other papers either mentioned a combination of these two (4 papers), or discussed the empirical nature of the item creation (6 papers).

25 papers referenced the **Sample Size** of the original questionnaire (6% of papers), especially if it was notably large; "the BrainHex model was selected because of its large number of respondents" (Paper 21). It was often unclear what specifically about the large sample was appealing to the authors, as the statement was provided in isolation (e.g., Paper 209; "Yee, 2006a, Yee, 2006b looked at gamer motivations by surveying a sample of 3000 online gamers"). However, sometimes papers used original sample size as a proxy for the validity/reliability of the questionnaire; "The PENS subscales were created for research by Ryan et al. (2006) and further validated in two rounds of confirmatory factor analysis using survey data from 2,000 regular video game players" (Paper 170).

23 papers discussed how the questionnaire had been **Used in Previous Work** (6% of papers). This reason is subtly different to 'Used to Measure X', which focuses on what specific constructs have been related to the questionnaire. In contrast, this reason identifies the questionnaire being used by others in and of itself as a reason for the current use (e.g., Paper 149; "participants were asked to fill in the Player Experience and Needs Satisfaction (PENS) Questionnaire [27] as this has previously been applied in games research [3]"). This sometimes highlighted that, because the questionnaire is already in use, it is therefore valid to use in the current setting (e.g., Paper 150; "To assess player experience, we used validated instruments that have been used to measure player experience before [26]"). Therefore, whilst these two reasons are similar (what the questionnaire has been used to measure versus the questionnaire having been used by others), they are still separable reasons — indeed, the reasons only occurred in the same paper in 7 cases (e.g., Paper 126).

19 papers stated the chosen questionnaire was **Popular** for its use within the field (5% of papers). This typically meant the questionnaire had been used frequently before by others, implying this indicates it is widely accepted by the field; "The Game Motivation Scale, developed by Yee (2006a), has been the most popular measure employed in game-related research for nearly a decade" (Paper 268). This is therefore slightly different to the previous reason 'Used in Previous Work'. Popular refers specifically to the *frequency* of use, whereas papers in the previous reason only showed the work has been used *at all*. As such, only four papers contained both Used in Previous Work and Popular as reasons for use, making them distinct reasons.

Table 5
The top five most commons reasons given for questionnaire use by discipline of influence.

Games (79 papers)		Psychology (115 papers)		Media (40 papers)	
Mature	14%	Based on X	23%	Based on X	22%
Valid	13%	Used to Measure X	19%	Used to Measure X	20%
Sample Size	12%	Valid	15%	Valid	17%
Used to Measure X	10%	Reliable	9%	Data Approach	15%
Structure	9%	Mature	6%	Mature	7%

12 papers were performing a **Replication** of one of the original questionnaires, and so were testing it in a new sample/setting (3% of papers). A variety of questionnaires have undergone replication studies, however PENS was the most common (5 papers).

9 papers made reference to the **Original Sample** used when creating the questionnaire (2% of papers). This was usually to highlight the sample was diverse ("We selected this model because these authors have recently validated it with videogame players of multiple genres across cultures such as Japan, Canada, and Finland".; Paper 49), unlike the previous reason 'Sample Size' which was concerned only with the number of people involved.

8 papers mentioned a questionnaire was **Recent** in its creation (2% of papers), where 6 referenced BrainHex and the remaining two were GAMS and GAMES. This is almost contradictory to some of the more common reasons, which cite the maturity or age of the work as an indication that the work is well accepted by the field. In contrast, papers using this reason did so as a recent questionnaire meant it was more 'advanced', so would be based on the most amount of prior work; "Although the BrainHex survey still has to be improved, it seems to be yet the most advanced player type survey" (Paper 24).

5 papers referenced the questionnaire for its **Ease of Use** (1% of papers). Four of these were in relation to BrainHex and the last was for MOGS, but all referred to the ability to access and administer the questionnaire online (e.g., Paper 19: "an online questionnaire for this model is available, which makes it of easy access and easy administration"). Therefore, this ease of use refers both to the researchers and the participants.

Finally, the least common reason for usage was that the questionnaire was the **First** of its kind. This could be first to use a specific theory (e.g., Paper 74: "Spada and Caselli (2017) developed and validated the first self-report measure designed to assess metacognitions about online gaming"), or could be the first to study the area empirically — "Yee conducted the first empirical studies aimed at identifying the various motivations of online game players" (Paper 153).

When comparing the types of reasons overall, many relate to the Structure of the questionnaire, such as how it was built and what it measures. Indeed, many of the most commonly stated reasons relate to how the questionnaire was conceptualised (38% of reasons). Other reasons relate to how it has been used in the past and how often, indicating a reasoning based on Community Practices. These are as equally common, accounting for 37% of stated reasons. A final large collection of reasons relates to demonstrated proof the questionnaires are fit for purpose (the Soundness of the questionnaire), in that they are valid or reliable. These accounted for 22% of reasons.

In summary, the top four most common aspects (Used to Measure X, Valid, Based on X, and Reliable) explain more than 50% of reasons. This indicates the most common reasons for questionnaire use are based around what the application of a questionnaire has been in the past, the theory that it is built on, and demonstrating it has sound structural properties.

4.1.1. Comparison of discipline influence

To further understand the multidisciplinary nature of player motivation questionnaire usage, the following section assesses how the discipline of influence (i.e., the discipline of the originating questionnaire) may affect justifications for use. When comparing the most common reasons for questionnaire use by discipline of influence, there is further evidence of differing priorities. This is highlighted in Table 5, which shows the top 5 most common reasons for use by influential discipline. Note Education is not included here as there were only 3 uses in total, making the analysis not robust enough to draw conclusions from.

The influence of the Psychology and Media disciplines share many similarities, only differing on a priority for Reliability for the former and Data Approach for the latter. In contrast, those influenced by the Games discipline show a marked difference in priority compared to Psychology and Media, though do still overlap with a focus on Mature, Valid, and Used to Measure X reasons.

Overall, questionnaire justification contains a multitude of reasons, where the most common are: what it has been used to measure, its validity, the theory it is based on, and the reliability. Within the discipline influences analysed there is some divergence in priority, particular between Games and those of Psychology and Media. Now that why questionnaires are used has been discussed, the scores for transparency in questionnaire usage are explored.

4.2. Score analysis

The distribution of transparency scores for each questionnaire are presented in Fig. 2, where their respective disciplinary influence is highlighted. The mean transparency score for each questionnaire varies between 3 and 6.50 out of 8 (with an average of 4.97), indicating notable inconsistencies in reporting practices. Even amongst those used more than 10 times (the top 7 most used), scores range from 3.72 of BrainHex to 5.89 of Wu (average score of 5.07).

The breakdown of each criteria is explained in Section 4.3, but it is first interesting to explore what differences in overall score can inform about the current state of the field. For example, there is the assumption that reporting practices should improve over time as a field matures. As publications move increasingly online, there is theoretically more space available for discussing why questionnaires are used as well as their structural characteristics. Therefore, it is expected transparency scores should improve over time accordingly. This was explored by plotting the scores achieved against the year of publication, shown in Fig. 3.

As can be seen, there is no improvement in transparency over time. This could be driven by differences in word limits between conferences and journals — conferences not allowing space for item publication or clear questionnaire choice justification could negatively impact transparency scores. The difference between journal and conference publication scores are therefore shown in Fig. 4. Whilst journal transparency has remained mostly neutral at an average of 5.43 out of 8, conference transparency has shown a general slight decrease, save for a notable improvement in the most recent year of 2020 (3.84 out of 8). Only seven papers scored 8 out of 8 for transparency, and all of them were published in journals. In contrast, of the nine papers that scored 1 out of 8, six were published in conferences.

Therefore, the lack of improvement in transparency over time may be partially explained by publication venue, and, by extension, word limits. However, scores have remained relatively stable for journals over time. This means even with more lenient word limits, it is unlikely for currently published work in journals to score 8 out of 8, indicating it is not publishing restrictions alone driving the lacking transparency criteria. Consequently, the field of player motivation cannot be assumed to be improving naturally over time. Now that the general score has been explored, trends in each of the 8 criteria are discussed.

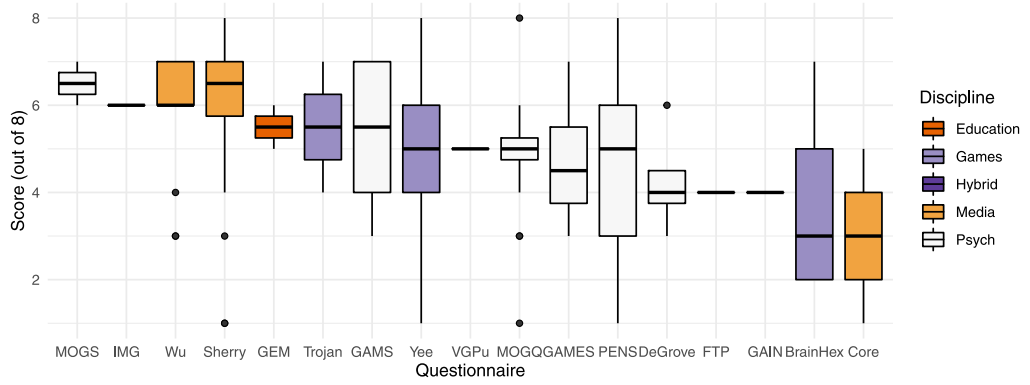

Fig. 2. The distribution of transparency scores for all 18 questionnaires analysed.

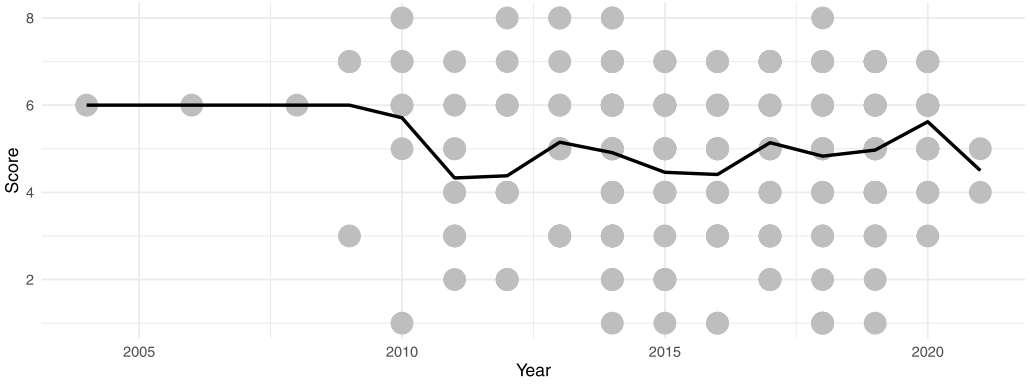

Fig. 3. A scatterplot of transparency scores over time. The average score for each year is also plotted.

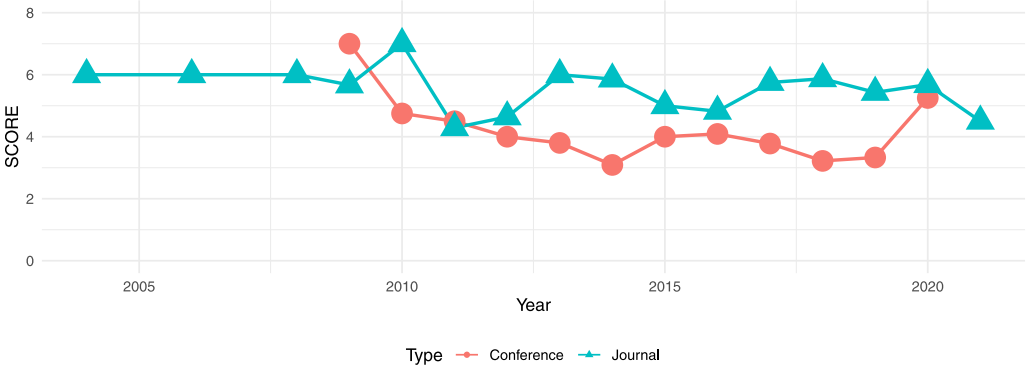

Fig. 4. The average score for journal vs conference papers plotted over time.

4.3. Transparency of reporting

The average percentage for each criteria analysed for the 238 papers is shown in Fig. 5. Whilst most papers correctly cited the questionnaire (and specific version), reported the number of Likert points used, and reported the number of items used, many did not provide examples of items, or present the items within the paper or on an external website. By extension, this made it difficult to know if items were reworded from the original scales. Each criteria is now explored in turn, with specific attention given to the top 7 most cited questionnaires (those used more than 10 times), as these reflect questionnaire usage more reliably. Further observations relating to these criteria are also explored where relevant (such as comparisons between journals and conferences, and the ways in which certain papers reworded items), to provide context to the scores given.

4.3.1. Correct citation

Whilst citations were typically cited correctly overall, there are exceptions. 14 papers did not correctly cite the questionnaire they used, of which 13 were from the top seven most used. Eight of these were of Yee (giving a score of 86% of Yee papers with correct citations), most likely caused by the high number of versions available published in similar years. For example, as there are two versions published in 2006, incorrect citations commonly referenced the wrong one (e.g., Papers 209 and 261). These incorrect citations are only possible to observe if items are available with the text in some way (either by having the items available, or in the example items given when introducing the questionnaire). Therefore, it is possible that more incorrect citations occurred, but due to a lack of transparency in reporting items used, this cannot be known. In the case of papers using Yee, there are 10 papers that did not provide items, examples of items, or the number of

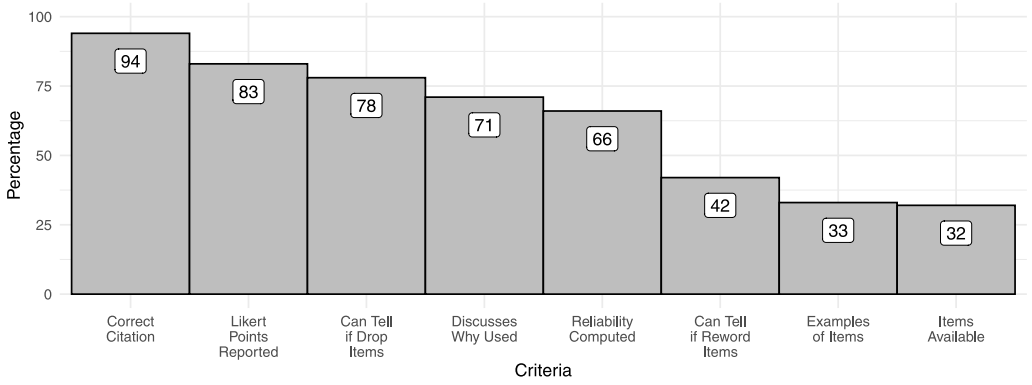

Fig. 5. A histogram of the average percentage of papers complying with each of the eight transparency criteria.

items used (e.g., Papers 202 and 207). This information could be used to infer which version was used by comparing the number of items used and the wording of them. Their absence makes it impossible to know if the correct version is indeed cited (for the purposes of this work they were assumed to be correct).

Overall, almost all incorrect citations were due to referencing the wrong version of the same questionnaire. This is especially problematic in the case of Yee, where the items and scales between versions varies significantly. However, one extreme example was of a paper that cited the 2012 version of Yee et al. (2012) but in fact used items from Trojan, even though the original Trojan paper was not cited in the work (Paper 206). Therefore, care should be taken when referencing questionnaires to reduce the risk of incorrect citations.

4.3.2. Likert points reported

83% of papers provided the number of Likert point used. Within the top 7 most used questionnaires, BrainHex reported these points the least at 41%. This is in contrast to the most reported of the top 7, which was GAMS at 92%. Most papers reported Likert points clearly within the methodology, though some provided this information within figure captions or within tables (e.g., Paper 31). In other cases, Likert points were reported for one scale, but not others (e.g., Papers 92 and 110). This could mean all questionnaires used the same Likert scale, but this would not always be true, as sometimes different Likert scales are used (e.g., Paper 22). Therefore, the most transparent papers reported the number of Likert points for each questionnaire, or made it clear the same points were used throughout.

4.3.3. Can tell if items dropped

Overall, papers made it easy to tell if items had been dropped from sub-scales, and therefore making them different from the original questionnaire. Of the top seven most used, the ability to tell if items were dropped ranged from 96% for MOGQ to 65% for PENS. However, in some instances a questionnaire is reported to use the same number of items as the original questionnaire, but upon inspection of the items made available with the text, an item has in fact been dropped. For example, in Paper 6 the GAMS questionnaire is used — the questionnaire is reported as 17 items, when the original is 18. In Paper 12, the Seeker sub-scale from the BrainHex questionnaire is used, but only two of its three items. This is not discussed in the text, and is only observed as the items are presented in a table, highlighting the importance of this criteria for transparency. By extension, this also means papers that reported the number of items used, but do not provide the items themselves (90 papers), are also at risk of inaccurate reporting. This discrepancy would therefore not be observable.

Furthermore, there are instances of confusion over the number of items a questionnaire contains. This mostly concerns the BrainHex questionnaire, where items were reported either as 21 items (e.g., Papers 12 and 18) or 28 (Papers 16 and 25). The online administration of BrainHex contains non-Likert items at the end of the survey, which are used to calculate player trait scores; one for each sub-type, leading to 7 additional items. This means papers may be using two different version of BrainHex – one with the added non-Likert items, and one without – whilst still citing the same questionnaire. This makes it challenging to compare findings, and can be easily overlooked as the citation of the questionnaire remains the same for both versions.

Of the papers that did provide information on the number of items, 62 dropped items (26%). This was done for a variety of reasons, such as a specific item was deemed to not fit the current research aims (e.g., Paper 35 wished to only look at items "associated with what could be termed as violent") or the particular game type being studied (e.g., Paper 250 wished to study all online games, not just *World of Warcraft* which is the main focus of Yee, whilst Paper 235 sought the opposite), as well as time constraints (Paper 229). However, there are also papers that state no reason at all (e.g., Papers 83 and 199). Furthermore, in 52 papers it was not possible to tell if items had been dropped, as the number used was not reported and items were not available. This means the number of papers that dropped items ranges between 26% to a possible 48%.

Dropping items in the top 7 most used was mostly by those using Wu (72%), as it was common for a number of Wu items to be taken and made into new scales (e.g., Papers 181–188). Whilst this is a more common practice in Media papers, Sherry did this to a lesser extent (32% of papers), so it is not as universal. In contrast, those using the MOGQ rarely dropped items (8%).

Overall, most papers made it clear the number of items used, and by extension if items had been dropped. However, some papers do not report this, making it unknown if the scale is used exactly the same as the original or not. As some papers report a questionnaire usage the same as the original, yet still drop items (e.g., Papers 6 and 12), this reporting should be done carefully and explicitly.

4.3.4. Discussed why this questionnaire

The previous Section 4.1 discusses the reasons for why specific questionnaires were used, therefore this section focuses on the prevalence of reasons given. 72% of papers provided a reason for use; within the top 7 most used questionnaires this ranged from 50% for Wu to 94% for BrainHex. Furthermore, reasons given for BrainHex were the most likely to be explicit (76% of reasons, such as Papers 1 and 17), whilst those using Wu or Sherry gave no explicit reasons (e.g., Papers 112 and 94). Indeed, besides BrainHex none of the top 7 most used gave explicit reasons more than 25% of the time. Therefore, whilst many papers give a reason for questionnaire use, this is rarely done explicitly save in the case of BrainHex.

4.3.5. Reliability computed

Two thirds of papers reported the reliability of the questionnaire within their sample, usually calculated as an Alpha score, though sometimes Omega or Composite Reliability. Of the top seven most used, reliability was most commonly calculated for Wu (94%), with BrainHex calculated the least (29%). This indicates a wide range of priorities for computing reliability. Many papers that did not compute reliability did comment on the original reliability from the questionnaire used (e.g., Paper 61). Reporting the original reliability score is useful to signal that a questionnaire has a robust structure, but it does not confirm whether the questionnaire has performed reliably with the current sample. This made it more difficult to assess if reliability had been computed on the current sample, and so know if the questionnaire remains reliable across different studies.

4.3.6. Can tell if items reworded

It was possible to know if items had been reworded in less than half of papers (42%), mostly driven by the lack of items available with the work (as discussed later). This criteria has a higher percentage than Items Available as sometimes it was possible to conclude items had been reworded by studying any example items given (e.g., Papers 89 and 106). Of these papers where it was possible to identify rewording, only 14 did not reword items in any way. This means item rewording was common (at least 36%), ranging from small tweaks to make narrative sense (such as 'game' to 'games' or 'this game' e.g., Papers 9 and 56), to full rewriting of items (e.g., Paper 24).

Furthermore, similarly to the criteria of Items Dropped, some papers stated no rewording took place, but upon item inspection this was found to be untrue (e.g., Paper 138 made slight alterations to the PENS items). This opens up the same problem of a potential for even more papers to have rewritten items, without the ability to see this due to a lack of item availability. Therefore, the true rate of rewording could range from as low as 36% to as high as 94%. A lack of transparency in item reporting makes this impossible to establish, highlighting again the importance of item availability.

4.3.7. Examples of items reported

The second least common criteria was the reporting of example items, at 33%. Of the top seven most used, this was most commonly done by MOGQ with 54%, and the lowest was BrainHex at 6%. This makes examples of items uncommon for all questionnaires. Instead of reporting example items, many papers would describe the sub-scale in question (e.g., Papers 137 and 92), or would simply state the names of the sub-scales (Papers 136 and 6). However, this is not the same as showing an item; examples of items provide face validity, where readers can see exactly the nature of items being asked. Describing a sub-scale on the other hand informs the reader on what the underlying construct is purporting to measure, an interpretation that a reader has no way to confirm or disagree with. Whilst word counts may be a concern for including example items, this does not seem to have affected their prevalence; 36% of conference and 31% of journal papers included examples of items, making them comparable.

4.3.8. Items available

The least common criteria met was the availability of items, at 32%. This reflects the overall lack of transparency across papers using questionnaires, as the inability to see items limits the ability to see exactly what has been asked of participants (and by extension if items have been dropped or reworded). Of the top seven most used, items were most commonly available in Wu at 72%, whilst in comparison only 4% of MOGQ uses did so. This highlights a significant variation in the priority for reporting items. The overall lack of availability may be slightly driven by the publication venue; 34% of journals had items available compared to 26% of conferences.

Of those that did report items, these were typically done by including an appendix (e.g., Papers 215 and 223), presenting the items in a table (Papers 28 and 81), or linking to an Open Science Framework (OSF) document (Paper 10). However, there were instances of papers that referred the reader to the items in an appendix, where no such appendix was present (e.g., Paper 256). Therefore, care should be taken so that supplementary materials are attached to papers in a clear and accessible way.

4.4. Comparison of discipline influence

Similarly to Section 4.1.1, how the originating discipline of the questionnaire cited affects reporting practices is now explored. Fig. 6 shows the distribution of scores for the top seven most cited questionnaires, colour coded with their respective discipline of influence. Whilst those influenced by Media have a consistently high level of transparency, Games was less so, mostly due to the low transparency in reporting of BrainHex noted previously.

Fig. 7 shows the distribution of criteria for each discipline of influence. There is clear evidence that the reason why Media questionnaires are reported more transparently is because of the availability of items (which allows item rewording to be seen more easily). Beyond this, Media questionnaires scored higher than the average for all criteria except for Discussing Why a questionnaire has been used, as discussed in Section 4.3.

When looking at the influence of the Games discipline, there is a noted lack of reporting Likert points, and to a lesser extent a lack of examples of items, computing reliability, and citing the correct questionnaire. In contrast, Games questionnaires were slightly more likely than average to provide items, and by extension make it clear if items had been reworded.

Finally, Psychology questionnaires were the least likely to provide items and by extension show if items had been reworded. This may be driven by the inability of authors to publish the propriety PENS questionnaire. There were however the most likely to provide Likert points, examples of items, and discuss why a specific questionnaire was used.

4.5. Summary

Overall, there is a wide distribution in how papers both justify questionnaire usage and how transparently this usage is reported. There is a focus on the structural soundness of questionnaires as well as the theory that built them, though those influenced by Games put more emphasis on a questionnaire being reflective of previous work than those of Psychology and Media. In terms of transparency there is also a wide distribution of scores, with questionnaires from Media scoring the highest on average. This is driven by their likelihood of including items with the text, a criteria lacking in Games and Psychology papers. Furthermore, papers published in journals had higher transparency scores than those in conferences, and overall transparency scores have not improved over time (against common perception; (Aeschbach et al., 2021)).

5. Discussion

The aim of the current study is to analyse current reporting practices when using player motivation questionnaires, to explore the maturity and cohesiveness of the field overall. This was done by assessing current reporting practices for player motivation questionnaires, where the differing priorities between the multiple disciplines present in the field were highlighted to explain these practices. Following a literature review, 18 questionnaires were identified that measure player motivations, and had been used at least once since their publication. Works that used these questionnaires were collected from the citations of these 18 questionnaires, and were analysed on eight aspects of transparent reporting practice.

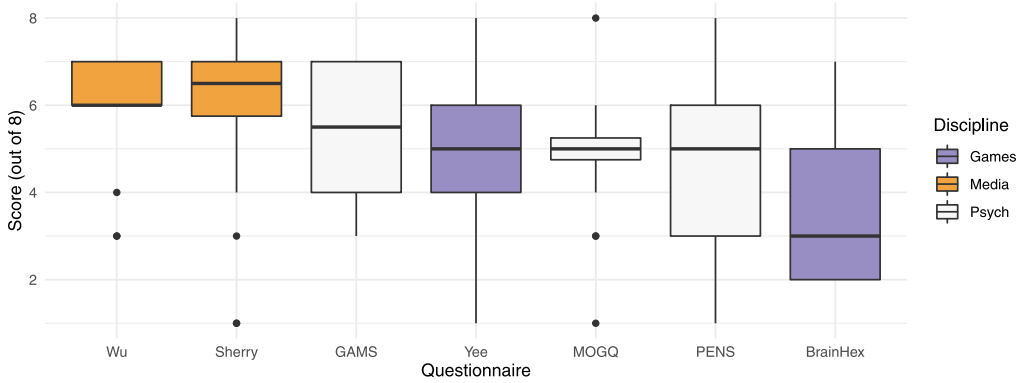

Fig. 6. The distribution of transparency scores for the top seven most used questionnaires analysed. Colours indicate the discipline of influence.

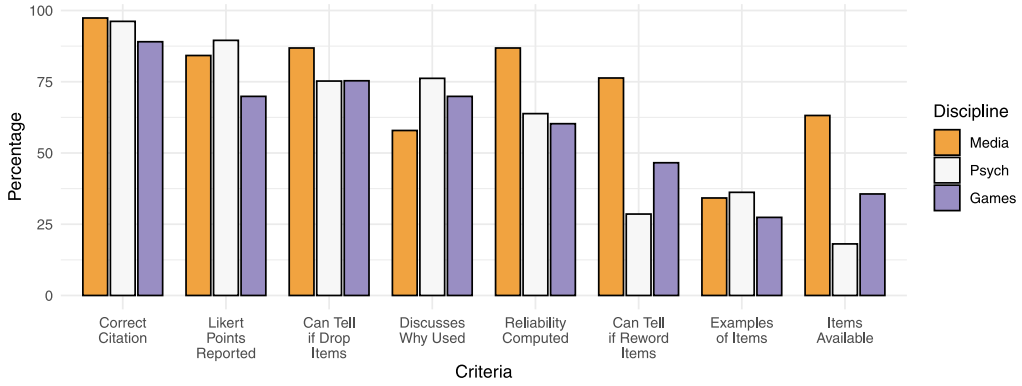

Fig. 7. The percentage of papers complying with each criteria by discipline of influence.

5.1. Questionnaire selection justifications

To assess justifications for use, the reasons given for why specific questionnaires were used was assessed. The presence of at least 18 questionnaires that measure player motivation makes this justification necessary, as it both indicates researcher intent and helps readers evaluate construct validity. This justification should ideally be as explicit as possible, such as by using the phrase "we used X because...". Few papers stated an explicit reason for why one questionnaire was used and not another from the multiple options available, at 19% of papers. In contrast, 52% provided implicit reasons (i.e., factors typically associated with 'good' questionnaire design – reliable, large sample, diverse sample – were reported but not linked to their use in the specific paper), and 29% of papers stated no reason at all. The most common reason for use was that the questionnaire was based on a specific theory. This is followed closely by what the questionnaire has been used to measure by previous works, and that the questionnaire is demonstrably valid in some way.

When comparing disciplines of influence, Psychology and Media share many similarities in priority, only differing on a priority for Reliability for the former and Data Approach for the latter. This is understandable, as the discipline of Psychology has a strong focus on statistics, whereas how questionnaires are created is important for Media where many papers make their own custom scales. Furthermore, as both Psychology and Media are theoretically-driven disciplines, it makes sense why they both prioritise theory and usage. In contrast, those influenced by Games were notably different in priority, potentially highlighting a lack of theory focus in the Games discipline in favour of how questionnaires are constructed and what they measure.

Therefore, it is currently difficult for researchers to evaluate the content validity of a chosen questionnaire in the field. By extension this makes it harder to understand when and how other researchers are using them, especially as there are different reporting priorities found within the disciplines of influence. This reduces the chance of papers coherently building on one another, highlighting the importance of accounting for multidisciplinary work that exists within the field.

5.2. Reporting transparency

The second aspect of the current analysis involved assessing how transparently questionnaires are reported. Transparency is important because questionnaires can be used and altered in a variety of ways, where some of these changes may make comparisons to other work incompatible. Without transparent reporting, there is no way to know how altered questionnaires are to one another, and so the field cannot be confident that new work builds on previous findings. To measure transparency, this paper explored the extent to which 8 criteria are met within currently published work; the more transparent the paper, the higher the score out of 8. These criteria covered validity and reproducibility arguments, which are aspects that lead to transparent reporting.

Overall, the average score for transparency was 4.94 out of 8, indicating that generally papers report around half of the practices. The most common practice was citing the correct version of the questionnaire at 94% of papers, and the least common was providing items available at 32%. The lack of items available replicates previous work, which found that less than a third of questionnaires in the field of Information Technology and Information Systems were provided (Van Biljon, 2014). The rate of modified items was also comparable to the work of Aeschbach et al. (2021), who found measurements were modified 38.71% of the time (versus 42% in the current work).

When considering questionnaires that have been used more than 10 times each (the top seven most used), papers influenced by the discipline of Media were reported the most transparently. Media papers

had the highest percentage in each criteria, apart from discussing why a questionnaire was used, examples of items, and reporting Likert points (where it was second highest). The least transparently reported discipline of influence was Games, with the lowest percentage in 5 of the 8 criteria. This may indicate those influenced by Games, and by extension the Games discipline itself, put lesser priority for reporting statistical features of a questionnaire, especially in comparison to Psychology. There is clear room for improvement in all of the 8 criteria assessed, though the largest improvement would be the inclusion of items within or associated to the text. This naturally increases the likelihood of being able to see if items have been dropped or reworded, which is why Media questionnaire reporting is particularly transparent (with 64% of uses including items).

One explanation for this trend in transparency is the restrictions placed on conferences for word limit. It is reasonable to assume that a stricter word limit and the typical lack of appendices would mean some criteria would be left out to save space. Indeed, when comparing the venue of publication (journal or conference) there was a clear difference in average scores: journal papers scored an average of 5.43 and conferences 3.84. Journals contained items 41% of the time, versus 11% for conferences. Therefore, to improve transparency, conferences should allow for items to be included with the text. As these can be provided digitally this would not be to a detriment of page limits within a conference, but would greatly increase the ability of readers to assess how a questionnaire has been used.

Another explanation for the trend is the influence of time. It could be assumed that newer papers are more transparent, as reporting practices become more widespread and observed by others (as suggested by Aeschbach et al. (2021)). However, there was no observable improvement in transparency over time, and the latest year studied does not show a marked increase in transparency. There was also no difference in transparency between conferences and journals over time, indicating transparency does not seem to be 'self correcting' itself naturally. This would suggest that, in the absence of formal education in this area, reporting practices are essentially learned from the practices prevalent in the discipline. There is no natural drift to better practices over time, and so the hope for this paper is that it will explicitly inform what are considered better practices for future researchers to build on.

5.3. The lenses of specific disciplinary influences

A further aim of the current work was to uncover the influence of having multiple disciplines within the field of player motivation. This was done by considering how the discipline of the questionnaire used may influence the reporting practices of those using it, and as a consequence make the lens of the discipline itself observable.

The lens of **Games** places value on reporting why specific questionnaires have been used, potentially due to an awareness of the number of questionnaires available. However, there is a lack of value placed in reporting the structural aspects of the questionnaires, such as the number of Likert points, computing reliability, and providing examples of items.

The **Psychology** lens places emphasis on reporting the structural aspects of questionnaires, including the number of Likert points, the number of items used, and providing examples of items. However, there is less focus on providing items available with the text, which reduces the ability to judge if items have been reworded. This could be biased by the PENS scale not being freely available which reduces the ability to report items, however the MOGQ questionnaire which is published in its original paper has one of the lowest rates of items being reported (4%).

The **Media** lens provided the highest level of transparency, driven mostly by the high percentage of items available within papers. This naturally led to a higher score in the ability to identify reworded items. A further strength in Media is computing reliability of questionnaires, however a noted lack of focus on explaining why a specific questionnaire has been used, and even within this many only stated Media theory as the sole reason when more than one questionnaire is available based on this theory.

As **Education** research is the newest discipline to the field, there are few uses of these questionnaires, where within this many citations are from the original authors. This means there is little data to analyse and so inferences are minimal. However, as a general summary there is a focus on reporting the structural characteristics of questionnaires (e.g., Likert points, the number of items used, computing reliability), but less focus is on providing items with the text, or discussing why a questionnaire has been used.

Therefore, whilst there are overlaps in priorities for transparency reporting, there are notable differences, potentially driven by the differing disciplinary lenses. When reading papers influenced by specific disciplines, it is worthwhile acknowledging the lenses that come with the work so it can be used in context, as well as know what certain readers may expect to be present. By seeing these lenses, it is possible to understand what is likely to be reported, and where the potential blindspots are.

For example, if a Psychologist or a person influenced by Psychology used to reporting reliability analyses comes across a paper influenced by Games, they may be confused to see a lack of alpha scores. This may make it difficult for the Psychology influencee to engage with the work, or know how to incorporate it into their understanding. Conversely, a Games influencee may be confused by the lack of explanation given to why a specific questionnaire has been chosen, and a Media influencee may expect to see the items available with the text. To allow multiple disciplines to engage and collaborate within the field of player motivation, it is important to respect these lenses of priority, and aim to benefit a wide variety of researchers by including all eight of the criteria discussed in the current work when deploying questionnaires. By reporting the priorities of all disciplines in future work, more of the field can engage with the methods deployed, leading to a more cohesive field.

5.4. A note on reflexivity

Care has been taken to highlight that, whilst the nature of this work is to show discrepancies in reporting practices, this is not to 'name and shame' any particular authors. Indeed, the citations for each paper discussed in the results has been assigned a number to be looked up, to avoid drawing attention within the main body of work. These discrepancies in transparency also do not reflect on the authors themselves, or is meant to make comments about their abilities to produce good science. There are several reasons why an author may overlook a transparency criteria that does not include intentional malfeasance.

As an example, the authors of this work were also included in this dataset twice, and scored 1 and 6 respectively, demonstrating a wide variation even within authors. Upon reflection, the score of 1 was found to be caused by a lack of thought towards the inclusion of PENS as a control variable, as it was not the main focus of the study. Therefore this is not due to intentional obscurity, but rather not considering how the inclusion of PENS can vary so significantly, even if there is an assumption that readers will know how it was used. Future authors can take this reflection as evidence that even those with good intentions may overlook aspects that do not seem important at the time.

5.5. Reflections on the field

As transparency is a multi-faceted concept, it is important to explore whether certain aspects of transparency have been adopted by the field and not others. The current work broke these down into three concepts; validity (what has been measured, and will it measure what it intends to), reproducibility (can other researchers accurately reproduce

some or all of the use of measures in their own work), and clarity, where the first two made up the 8 criteria used to score papers. When considering validity (the criteria of discussing why a questionnaire is used, providing examples of items, and running reliability checks), these were less likely to be reported than reproducibility criteria. This indicates a focus on reporting the 'what' of questionnaire usage rather than the 'why'. Furthermore, this apparent disconnect between validity and reproducibility criteria is also present within the works analysed, which can reduce the overall clarity of the work. Whilst the former is typically found in the background/literature review section, the latter is contained in the method. This decoupling of the two aspects of transparency reduces the clarity of the work, as it can be difficult to connect why a questionnaire has been chosen to how its use has been reported.

As an example, consider Paper 17. The BrainHex questionnaire is justified at length in the background, but in the methods there is no longer a mention of these reasons for use. To connect why and how a questionnaire is used, the reader must switch between sections. This reduces clarity and coherency, leading to less transparency as the reader does not intuitively know which section to look at to find specific information. To improve this, researchers could re-couple these arguments under the method section. That is not to say that researchers cannot discuss the questionnaire in the background, but more the reasons for use in the specific study should be placed in the method section to maximise clarity.

Therefore, the current work has highlighted that current reporting practices for player motivation questionnaires are neither wholly transparent nor fully justified. This lack of a shared repertoire means past and future work cannot reliably build on one another, reducing the likelihood of the field of player motivation becoming a coherent research community. This lack of a shared repertoire has led to widespread variation in how questionnaires are deployed and reported, which has decreased transparency and trust in the work that has been done. The criteria analysed in this paper are not complex or demanding of researcher time; second guessing whether a questionnaire is correctly cited should not ideally be a concern for a reader, yet incorrect citations were found in 6% of papers. It could be argued these oversights are small, and that the majority of authors remember to include these criteria in some capacity as they are simple and could be argued to be 'common sense'. However, Gawande (2011) explains how forgetting the simplest actions is easy to do, even (perhaps especially) for experts in their work.

The field is not old enough or standardised enough to allow for the level of trust where each questionnaire usage can be taken at face value. This is demonstrated by a number of instances where authors have incorrectly reported an aspect of questionnaire usage: a paper saying it uses a questionnaire the 'same as the original authors', and yet dropping items in the process; a paper reporting to use one version of a scale yet using another; a paper saying the items are available in the appendix when no such appendix exists. These instances highlight that, beneath the transparency scores obtained in this work, there is a deeper concern that authors may have made errors. These errors were only captured as items were made available for scrutiny. This raises the following question: what about the majority of papers that did not provide items? It is impossible to know the upper boundary of errors made without this information, meaning the field could contain a far greater amount of errors that are not detectable, further damaging trust in reported findings.

Because of the concerns over transparency and trust, the criteria highlighted in this work have to be explicitly discussed until they become second nature. Standardising the use of questionnaires will allow trust to be rebuilt, and so the next section explores how best to share these criteria with authors.

Table 6
A checklist of transparency for reporting player motivation questionnaires.

Transparency criteria	
Validity	
Is there discussion of why this specific questionnaire was chosen?	☐
Are there examples of items from scales deployed?	☐
Were reliability analyses computed?	☐
Reproducibility	
Is the correct questionnaire cited (including correct version)?	☐
Are the number of Likert points used reported?	☐
Is it possible to know if items were dropped?	☐
Is it possible to know if items were reworded, even if slight?	☐
Are the items available with the text?	☐

5.6. Checklists: a light in the dark

This study used eight criteria described above as evaluation points, to assess the transparency of current reporting practice of player motivation questionnaires. These can form a 'checklist' where researchers can use it when writing method sections to ensure these aspects are clearly described. This approach has been used before (as described in Section 2.2.1), though typically in a more abstract and high-level way (e.g., Kelley et al. (2003)). A checklist approach has a variety of benefits; they are simple to follow, easy to evaluate, and improve performance of routine actions (which, arguably, writing methodology sections become for researchers). For example, Gawande (2011) discusses how checklists in intensive care units reduced ten-day line infection rates from 11 to 0%, which involved doctors being reminded by nurses to follow standardised checklist items they were aware of but infrequently forgot to do. Whilst reporting player research is not as life threatening, the same principles can be applied to improve the robustness of the field in similar ways. A paper accurately and transparently reporting questionnaire usage allows other researchers to clearly understand the work, as well as replicate the same approach if desired. They can also see where they wish to differ from previous use, with a strengthened ability to acknowledge this divergence and explicitly explain their reasons. Therefore, a checklist based on the 8 criteria measured in this study is proposed in Table 6, ordered under the sub-headings of Validity and Reproducibility. Whilst Clarity was introduced as the third component of transparency, is it not included here due to its more subjective nature, making it less useful as part of a checklist. Researchers are however still encouraged to consider Clarity when reporting questionnaires.

The checklist is unlikely to bring large benefit to individual research teams, but transparent reporting brought about via a standardised checklist approach could allow future researchers the ability to confidently build on this work. Therefore, standardisation of questionnaire reporting is for the benefit of the field overall, and should be made as easily accessible to all researchers as possible.

In summary, each discipline of influence has demonstrated strengths in certain aspects of the 8 criteria analysed. These should be acknowledged and built upon to create a robust and transparent standard for questionnaire reporting, building on the strength of each discipline. This allows the field to come together, improving the standards for the betterment of all.

5.7. Limitations & future work

There are a number of limitations to the present study. As mentioned previously, the transparency criteria used does not indicate if a paper is good or without flaws. It is possible to have a highly transparent paper that is still lacking in some way; for example, a paper saying the questionnaire was chosen as it is popular is a justification, but the extent to which this is a good justification is subjective. Therefore, the current criteria can only assess whether a paper is transparent, and cannot indicate quality.

A further limitation is that the criteria devised for this study are not an exhaustive list of elements that affect transparency. They represent a combination of previous guidance on reporting practices, whilst reducing this guidance to the smallest and simplest unit of measurement. For example, reducing 'psychometric properties' to reliability scores and Likert points are smaller and more actionable, making them suitable for a checklist approach. This means there are further ways to measure transparency that may give the same papers differing scores, but the current criteria can still provide a useful baseline for researchers to use in their own work.

The scoping review conducted has a number of limitations. Firstly, not all questionnaires were considered, such as those from gamification and problematic gaming. This helped to narrow the focus of the work, but presents large volumes of citations excluded that are peripherally related to player motivation. The analysis conducted in this study could be repeated on these papers to assess their transparency and compare it to the questionnaires included here. Secondly the review was not an exhaustive list, and so citations could have been missed. For example, not all papers were accessible (either due to lack of access or not being indexed via Google Scholar), and not all citations for the questionnaires from Yee and PENS were assessed due to their high volume. Therefore, the results may not reflect all uses, but do provide a large enough sample to assess general trends within the field. A further limitation of the scoping review is the chance for human error in the collection of transparency criteria, as they could be missed. Papers that did not present questionnaire usage clearly, such as by putting aspects of questionnaire usage into figure headings or separating justifications from the methods section, are at an increased risk of this being true. This was mitigated by the use of two researchers for data collection, but some papers are still likely to have been missed.

The content analysis approach also has limitations. The codes generated are influenced by the researcher creating them, and so may not have been generated if the analysis was done by a different person. The justifications for questionnaire use is a more subjective interpretation than the transparency scores, and are only generated from the text in the paper. If authors were to be asked directly about the uses, further codes are likely to exist, providing an opportunity for future work.

As the separation of influential disciplines was created based on the discipline of the questionnaire being used, this also has limitations. This analysis sought to understand how disciplines may influence the transparency reporting practices of those that use it, but transparency is also likely influenced by the discipline of the authors themselves. As this was not possible to learn from the data, self-identification of discipline could not be assessed, and so may be a confounding factor. The assessment of influential discipline should therefore be viewed as an exploratory analysis of the effect of the questionnaire being cited on resulting papers, and not a reflection on the authors of citing papers.

Finally, the current work only assesses one method to observe a shared repertoire, which is the use of questionnaires. There are other commonly used methods in the field, such as qualitative approaches, which may also reflect the community of practice. Future work could examine how other methods have been used across the disciplines of influence, such as content or thematic analyses, to explore whether the field is more synthesised for other shared repertoires.

Overall, to address these limitations future work could survey authors of papers for their self-identified discipline, widen the review criteria to other motivation topics and methodological approaches, and include more citations for Yee and PENS.

6. Conclusion

Understanding the reasons why players differ in their reasons to play is a complex issue. Questionnaires are one way to assess these motivations, but the field of player research has no standardised guidance for their use. Exacerbated by the presence of multiple disciplines researching player differences, differing priorities for how questionnaires should be used and reported are prolific. This has led to the creation of multiple questionnaires, with no agreed process for reporting their usage in a transparent way.

The present work analysed the transparency of current work using player motivation questionnaires, and found a noted lack of questionnaire usage being reported in a clear way. A majority of papers did not include the items used in their work, obscuring the ability to see if items have been reworded or dropped. Many papers also did not report the structural properties of the items (such as the number of Likert points and the reliability of sub-scales in the current sample). Furthermore, whilst citing the correct version of a questionnaire was the most common transparency criteria to be met, the fact it was not present 100% of the time (94%) is cause for concern. When comparing discipline of influence, Media had higher transparency scores, driven by the higher rates of including questionnaire items in the work that could be inspected. Alongside this, an analysis of reasons for specific questionnaire use were analysed. Whilst there is a wide distribution in how papers justify questionnaire usage, there is a focus on their structural soundness, as well as the theory that built them. When comparing discipline of influence, Games put more emphasis on a questionnaire being reflective of previous work than that of Psychology and Media.

Overall, there is a lack of standardisation in questionnaire reporting, which, combined with a subsequent lack of transparency, has made it difficult to assess whether studies using the 'same' questionnaire are indeed applying them in the same way. This has complicated research synthesis, and cast doubt on whether the advancement of knowledge in the field is sound. Now that the field of player motivation research has matured, it is time to establish a shared repertoire for how questionnaires should be used, to allow results to be confidently built atop one another and for the field to truly become a research community. Other fields that have overlooked methodological issues have faced replication crises and the risk of results being invalidated/mistrusted. By learning from these fields and standardising questionnaire reporting now, player motivation research may be able to avoid this outcome.

The present work provides a checklist for researchers in the field to use to standardise questionnaire usage. This will aid future work in avoiding the pitfalls of opaque reporting practices, and allow all disciplines to access the work conducted by others. Doing so will allow the field, with its multiple disciplines and research interests, to grow together, instead of in parallel.

CRediT authorship contribution statement

Nathan G.J. Hughes: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualisation. **Josephine R. Flockton:** Investigation, Writing – original draft, Writing – review & editing. **Paul Cairns:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the EPSRC, United Kingdom Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/L015846/1].

References

Aeschbach, L.F., Perrig, S.A., Weder, L., Opwis, K., Brühlmann, F., 2021. Transparency in measurement reporting: A systematic literature review of CHI PLAY. Proc. ACM Hum.-Comput. Interact. 5 (CHI PLAY), 1–21.

Ankeny, R.A., Leonelli, S., 2016. Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. Stud. His. Philos. Sci. A 60, 18–28.

Bartle, R., 1996. Hearts, clubs, diamonds, spades: Players who suit MUDs. J. MUD Res. 1 (1), 19.

Bennett, C., Khangura, S., Brehaut, J.C., Graham, I.D., Moher, D., Potter, B.K., M. Grimshaw, J., 2011. Reporting guidelines for survey research: An analysis of published guidance and reporting practices. PLoS Med. 8 (8), e1001069.

Billieux, J., Van der Linden, M., Achab, S., Khazaal, Y., Paraskevopoulos, L., Zullino, D., Thorens, G., 2013. Why do you play World of Warcraft? An in-depth exploration of self-reported motivations to play online and in-game behaviours in the virtual world of Azeroth. Comput. Hum. Behav. 29 (1), 103–109.

Birk, M., Mandryk, R.L., 2013. Control your game-self: Effects of controller type on enjoyment, motivation, and personality in game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 685–694.

Bowman, N.D., Schultheiss, D., Schumann, C., 2012. "I'm attached, and i'm a good guy/gal!": How character attachment influences pro-and anti-social motivations to play massively multiplayer online role-playing games. Cyberpsychol. Behav. Soc. Netw. 15 (3), 169–174.

Boynton, P.M., 2004. Administering, analysing, and reporting your questionnaire. Bmj 328 (7452), 1372–1375.

Cairns, P., 2019. Doing Better Statistics in Human–Computer Interaction. Cambridge University Press.

Comello, M.L.G., Francis, D.B., Marshall, L.H., Puglia, D.R., 2016. Cancer survivors who play recreational computer games: Motivations for playing and associations with beneficial psychological outcomes. Games Health J. 5 (4), 286–292.

Davies, J.J., Hemingway, T.J., 2014. Guitar hero or zero? J. Media Psychol..

De Grove, F., Cauberghe, V., Van Looy, J., 2016. Development and validation of an instrument for measuring individual motives for playing digital games. Media Psychol. 19 (1), 101–125.

Demetrovics, Z., Urbán, R., Nagygyörgy, K., Farkas, J., Zilahy, D., Mervó, B., Reindl, A., Ágoston, C., Kertész, A., Harmath, E., 2011. Why do you play? The development of the motives for online gaming questionnaire (MOGQ). Behav. Res. Methods 43 (3), 814–825.

Flake, J.K., Fried, E.I., 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. Adv. Methods Pract. Psychol. Sci. 3 (4), 456–465.

Foster, E.D., Deardorff, A., 2017. Open Science Framework (OSF). J. Med. Library Assoc.: JMLA 105 (2), 203.

Gawande, A., 2011. The checklist manifesto: How to get things right. J. Nurs. Regul. 1 (4), 64.

Goh, C., Jones, C., Copello, A., 2019. A further test of the impact of online gaming on psychological wellbeing and the role of play motivations and problematic use. Psychiatr. Q. 90 (4), 747–760.

Halonen, J.S., Buskist, W., Dunn, D., Freeman, J., Hill, G., Enns, C., et al., 2013. APA guidelines for the undergraduate psychology major (version 2.0). Washington, DC: APA.

Hermawati, S., Lawson, G., 2016. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? Applied Ergon. 56, 34–51.

Hewett, T.T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., Verplank, W., 1992. ACM SIGCHI Curricula for Human–Computer Interaction. ACM.

Hilgard, J., Engelhardt, C.R., Bartholow, B.D., 2013. Individual differences in motives, preferences, and pathology in video games: The gaming attitudes, motives, and experiences scales (GAMES). Front. Psychol. 4, 608.

Hughes, N., Cairns, P., 2020. Practically invalid - A statistical analysis of player trait questionnaires. http://dx.doi.org/10.31219/osf.io/kehmu, OSF Preprints, URL: https://osf.io/kehmu.

Hughes, N.G., Cairns, P., 2021. Opening the world of contextually-specific player experiences. Entertain. Comput. 37, 100401.

Kahn, A.S., Shen, C., Lu, L., Ratan, R.A., Coary, S., Hou, J., Meng, J., Osborn, J., Williams, D., 2015. The trojan player typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations. Comput. Hum. Behav. 49, 354–361.

Kelley, K., Clark, B., Brown, V., Sitzia, J., 2003. Good practice in the conduct and reporting of survey research. Int. J. Qual. Health Care 15 (3), 261–266.

Kim, Y., Ross, S.D., 2006. An exploration of motives in sport video gaming. Int. J. Sports Market. Sponsorship.

Király, O., Bőthe, B., Ramos-Diaz, J., Rahimi-Movaghar, A., Lukavska, K., Hrabec, O., Miovsky, M., Billieux, J., Deleuze, J., Nuyens, F., et al., 2019. Ten-item Internet Gaming Disorder Test (IGDT-10): Measurement invariance and cross-cultural validation across seven language-based samples. Psychol. Addict. Behav. 33 (1), 91.

Krippendorff, K., 2018. Content Analysis: An Introduction to Its Methodology. Sage publications.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, Vol. 111. Chicago University of Chicago Press.

Kuhn, T.S., 1987. What are Scientific Revolutions. MIT, Cambridge.

Ladanyi, J., Doyle-Portillo, S., 2017. The development and validation of the Grief Play Scale (GPS) in MMORPGs. Pers. Individ. Differ. 114, 125–133.

Lafrenière, M.-A.K., Verner-Filion, J., Vallerand, R.J., 2012. Development and validation of the gaming motivation scale (GAMS). Personal. Individ. Differ. 53 (7), 827–831.

Lee, C., Lee, K., Lee, D., 2017. Mobile healthcare applications and gamification for sustained health maintenance. Sustainability 9 (5), 772.

Lee, D., Schoenstedt, L.J., 2011. Comparison of esports and traditional sports consumption motives. ICHPER-SD J. Res. 6 (2), 39–44.

Likert, R., 1932. A Technique for the Measurement of Attitudes (Archives of Psychology, No: 140), Vol. 7, no. 3. Columbia University, New York City.

Maurer, T.J., Pierce, H.R., 1998. A comparison of likert scale and traditional measures of self-efficacy.. J. Appl. Psychol. 83 (2), 324.

Maxwell, S.E., Lau, M.Y., Howard, G.S., 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am. Psychol. 70 (6), 487.

Melhart, D., Azadvar, A., Canossa, A., Liapis, A., Yannakakis, G.N., 2019. Your gameplay says it all: Modelling motivation in Tom Clancy's the division. In: 2019 IEEE Conference on Games. CoG, IEEE, pp. 1–8.

Mills, D.J., Milyavskaya, M., Heath, N.L., Derevensky, J.L., 2018. Gaming motivation and problematic video gaming: The role of needs frustration. Eur. J. Soc. Psychol. 48 (4), 551–559.

Monterrat, B., Lavoué, É., George, S., 2017. Adaptation of gaming features for motivating learners. Simul. Gaming 48 (5), 625–656.

Munn, Z., Peters, M.D., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E., 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med. Res. Methodol. 18 (1), 1–7.

Myrseth, H., Notelaers, G., Strand, L.A., Borud, E.K., Olsen, O.K., 2017. Introduction of a new instrument to measure motivation for gaming: The electronic gaming motives questionnaire. Addiction 112 (9), 1658–1668.

Nacke, L.E., Bateman, C., Mandryk, R.L., 2011. BrainHex: Preliminary results from a neurobiological gamer typology survey. In: International Conference on Entertainment Computing. Springer, pp. 288–293.

Nacke, L.E., Bateman, C., Mandryk, R.L., 2014. BrainHex: A neurobiological gamer typology survey. Entertain. Comput. 5 (1), 55–62.

Nersessian, N.J., 2006. The cognitive-cultural systems of the research laboratory. Organ. Stud. 27 (1), 125–145.

Quick, J.M., Atkinson, R.K., Lin, L., 2012. Empirical taxonomies of gameplay enjoyment: Personality and video game preference. Int. J. Game-Based Learn. (IJGBL) 2 (3), 11–31.

Rolin, K., 2008. Science as collective knowledge. Cogn. Syst. Res. 9 (1–2), 115–124.

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

Sanchez, D.R., Langer, M., 2020. Video game pursuit (VGPu) scale development: Designing and validating a scale with implications for game-based learning and assessment. Simul. Gaming 51 (1), 55–86.

Scharkow, M., Festl, R., Vogelgesang, J., Quandt, T., 2015. Beyond the "core-gamer": Genre preferences and gratifications in computer games. Comput. Hum. Behav. 44, 293–298.

Sherry, J.L., Lucas, K., Greenberg, B.S., Lachlan, K., 2006. Video game uses and gratifications as predictors of use and game preference. Playing Video Games: Motives, Responses, and Consequences 24 (1), 213–224.

Spada, M.M., Caselli, G., 2017. The metacognitions about online gaming scale: Development and psychometric properties. Addict. Behav. 64, 281–286.

Tavakkoli, A., Loffredo, D., Ward Sr, M., 2014. Insights from massively multiplayer online role playing games to enhance gamification in education. J. Systemics Cybern. Inform. 12 (4), 66–78.

Teng, C.-I., Chen, W.-W., 2014. Team participation and online gamer loyalty. Electron. Commer. Res. Appl. 13 (1), 24–31.

Tondello, G.F., Arrambide, K., Ribeiro, G., Cen, A.J.-l., Nacke, L.E., 2019. "I don't fit into a single type": A trait model and scale of game playing preferences. In: IFIP Conference on Human–Computer Interaction. Springer, pp. 375–395.

Tondello, G.F., Wehbe, R.R., Orji, R., Ribeiro, G., Nacke, L.E., 2017. A framework and taxonomy of videogame playing preferences. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play. pp. 329–340.

Vahlo, J., Hamari, J., 2019. Five-factor inventory of intrinsic motivations to gameplay (IMG). In: Proceedings of the 52nd Hawaii International Conference on System Sciences. Hawaii, USA, 2019, HICSS, pp. 2476–2485.

Vahlo, J., Smed, J., Koponen, A., 2018. Validating gameplay activity inventory (GAIN) for modeling player profiles. User Model. User-Adapted Interact. 28 (4), 425–453.

Van Biljon, J., 2014. Questioning the questionnaire: Expediency of reviewing and publication versus adequate description and methodological justification. In: 8th European Conference on IS Management and Evaluation. pp. 262–270.

Van Calster, B., Wynants, L., Riley, R.D., van Smeden, M., Collins, G.S., 2021. Methodology over metrics: Current scientific standards are a disservice to patients and society. J. Clin. Epidemiol..

Vermeulen, L., Van Bauwel, S., Van Looy, J., 2017. Tracing female gamer identity. an empirical study into gender and stereotype threat perceptions. Computers in Human Behavior 71, 90–98.

Warmelink, H., Harteveld, C., Mayer, I., 2009. Press enter or escape to play-deconstructing escapism in multiplayer gaming. In: DiGRA Conference. pp. 1–11.

Wenger, E., 1999. Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press.

Wu, J.-H., Wang, S.-C., Tsai, H.-H., 2010. Falling in love with online games: The uses and gratifications perspective. Comput. Hum. Behav. 26 (6), 1862–1871.

Yee, N., 2006. Motivations for play in online games. CyberPsychol. Behav. 9 (6), 772–775.

Yee, N., Ducheneaut, N., Nelson, L., 2012. Online gaming motivations scale: development and validation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2803–2806.

Zurita Ortega, F., Medina Medina, N., Gutierrez Vela, F.L., Chacon Cuberos, R., 2020. Validation and psychometric properties of the gameplay-scale for educative video games in Spanish children. Sustainability 12 (6), 2283.