Entertainment Computing 47 (2023) 100591

Contents lists available at ScienceDirect

Entertainment Computing

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

Theories, methodologies, and effects of affect-adaptive games: A systematic review

Maximilian Croissant [a,*], Guy Schofield [b], Cade McCall [c]

[a] University of York, Department of Computer Science, United Kingdom
[b] University of York, Department of Theatre, Film, Television and Interactive Media, United Kingdom
[c] University of York, Department of Psychology, United Kingdom

ARTICLE INFO

Keywords:
Adaptation
Emotion
Video games
Affective computing

ABSTRACT

Affect-adaptive games gained in popularity over the last years in human computer interactions studies, promising potential benefits for player experience, performance, and even health. It is however not yet clear how affective games are being evaluated, what the precise effects are, and how they are based on emotion theoretical concepts that are still not universally agreed upon. This systematic review investigated these questions by analysing relevant high-quality evaluation studies of the effect of affect-adaptive video games on various outcomes in regards to their effects, theoretical assumptions, and methodologies. Out of 3,930 papers, 26 studies were included based on preregistered inclusion and exclusion criteria. A high variance regarding theoretical assumptions and methodological approaches was observed, as well as an overall poor methodological rigour, leading to the conclusion that more work is needed in constructing better methodological standards for game evaluation studies and theoretical considerations when developing and testing affect-adaptive video games.

1. Introduction

Over the last decades, video games have risen in popularity and established themselves as one of the most commercially successful forms of entertainment, yielding over 160 billion dollars in revenue in 2020 [1]. Beyond the commercial interest, research suggests that video games offer benefits to mental health (e.g. [2,3]), via treatments or trainings (e.g. [4,5]) or by offering positive affective experiences (e.g. [6,7]).

To further the development of games promoting such benefits, both industry and research have attempted to develop and evaluate new ways of improving the effects associated with video games. One of the most popular areas of improvement is game adaptation based on individual player data. Adaptation has been seen as one of the main ways to increase accessibility for all kinds of players, align player enjoyment across populations, and introduce new gameplay opportunities [8,9]. For example, dynamic difficulty adjustment (DDA) has been presented as an effective way to narrow player skill gaps and therefore raise overall game enjoyment by adapting game materials to models of player skills [10].

Other approaches describe affective adaptation, which is the adaptation of game material to emotional or affective data from the player. Rooted in the discipline of affective computing, originally described by Picard [11], affect-adaptive games try to directly elicit emotional responses by mapping various game characteristics to potential emotional responses. This process is seen as fundamentally based on psychological groundwork and often relates to concepts introduced in psychological research. For example, based on the model of optimal experiences by Csikszentmihalyi and Csikszentmihalyi [12] and Chen [13] described the flow model in video games as a form of balance between individual abilities of players and challenge to maximize enjoyment and minimize frustration or boredom. Other popular adapted theories include self-determination theory (SDT), describing the motivational pull of video games based on basic psychological needs [14,15]. In general, because games offer emotional benefits and motivational appeal, directly measuring and adapting the affective relationship between games and players is often seen as an ideal way to address interindividual differences in experiencing games and therefore maximizing the potential of beneficial outcomes in aspects such as enjoyment, health, education, or well-being [16].

However, as of yet there are still many open questions relating to affect-adaptive video games. Prior reviews have found promising effects for affect-adaptive games [17,18] and reported an extensive overview about findings and methods specifically for physiology-based games [19]. A similar understanding about the reported effects of

* Corresponding author.
E-mail address: mc2230@york.ac.uk (M. Croissant).

https://doi.org/10.1016/j.entcom.2023.100591
Received 22 May 2023; Accepted 20 June 2023
Available online 4 July 2023

affect-adaptive games would be very valuable in assessing the benefits and risks involved in the design and development process. It is currently not clear how affect-adaptive games perform against control conditions, what outcomes (such as health benefits or player enjoyment) are being investigated, and how large reported effect sizes are. Furthermore, it is not clear in which ways affective games are grounded in psychological theories. While emotion as a concept has been examined through many established theoretical and empirical works, there are still many fundamental conceptual uncertainties present in modern psychological research (see for example [20] for a current overview). It is still not clear how emotions are structurally represented (e.g. dimensionally vs. distinct), what underlying mechanisms explain their elicitation (e.g. socially constructed vs. innate), which components participate in emotion development and expression, and how emotions can be measured. Because of these conceptual issues, applications of emotion theories are at risk of blindly relying on assumptions that are currently still under debate — or that may already be outdated. It is therefore important to examine the theoretical and mechanistic assumptions underlying the design and testing of affective-adaptive games. Finally, the quality of provided evidence for the effect of affect-adaptive games in terms of their methodological approaches is not yet clear and may further provide important data to evaluate the true potential of affective games.

To summarize, while many promising effects and underlying psychological models are discussed in current research, we are lacking systematic evidence to evaluate the success of affective adaptation in enhancing the positive effects video games have on players. Additionally, there is a lack of research indicating how well theoretical and methodological approaches are applied in affect-adaptive games research. This study therefore proposes a systematic review approach to fill these gaps and provide more insight about the current state-of-the-art in affect-adaptive research.

2. Background

2.1. Emotion theory

In affective computing, the terms “emotion” and “affect” are often used interchangeably, but in psychology affect is generally seen as an umbrella term, describing multiple possible affective states that can be differentiated through certain features [21–23]. For emotions specifically these features include a fixed (often short) duration with varying emotion-specific onset and offset periods of a mostly high intensity [24], which differentiates them from other affective states, such as mood, stance, attitude, or affective traits. In affective games, emotions are often the main affective variable of interest, but the overall game experience has also considerable relations to other affective states, such as mood or stance. In this review, we include all studies concerned with an adaptation to affective data, which might in theory include measures of overall mood, but also includes all measures related to emotions.

But even looking at the psychological concept of emotion and its theoretical components, more uncertainties arise. In 2010, Izard [25] conducted a survey study with 35 highly acclaimed scientists in the field of emotion research, asking six questions about the definition, functions, and underlying mechanisms of emotions. They found considerable disagreements in almost all answers, with only a 25% agreement in basic definitions of emotions, and even more disagreement in their views of emotion function, emotion elicitation, and the relationships between emotion, cognition, and action. In 2022, most of these disagreements are still not resolved [20].

Contentious aspects of emotion theoretical approaches are concerned with basic assumptions such as the underlying structure of emotions:

Dimensional emotion models originate from work mapping emotions onto a pleasantness-unpleasantness scale [26], which was expanded by the works of Russell [27] who popularized the circumplex model of affect, which added a dimension for activation or arousal, providing more depth for emotion classification and even guidance to assign emotions based on values in these dimensions. To this day, many subjective measures, such as the Positive and Negative Affect Scales (PANAS; Watson et al. [28]) reflect these dimension and are used to measure affective states in an experimental or diagnostic setting. Other scales, such as the self-assessment manikin (SAM; Bradley and Lang [29]) even add another dimension (Dominance) to differentiate between emotions. The influential cone model by Plutchik [30] was built upon the circumplex model and used assumed mappings between individual emotional states and these dimension to describe their relationship to each other.

In contrast to dimensional models are distinct emotion perspectives, which are based on assumptions of specific and distinct emotion expressions and action motivations. Modern descriptions of discrete emotions has been popularized by the works of Tomkins [31], or Izard [32], which resulted in the development of general emotion categories, such as joy, sadness, anger, fear, or disgust. The main implication of this uncertainty are the potential problems in describing and measuring emotions. For example, measures of peripheral physiology focusing on the autonomic nervous system (ANS) have been found to inconsistently reflect distinct emotional states in a meta-analysis by Cacioppo et al. [33]. Rather, such measures (for example heart rate) can be used to infer dimensional emotional information, most notably arousal, but also to some degree valence [33]. Behavioural measurements, however, such as those of facial or body behaviour may convey valence information [34], but also may have significant specificity for discrete emotional states (e.g. [35]). Additionally, while emotion terms (and therefore subjective ratings) can be quite intuitively mapped onto one or more dimensions (for example in the circumplex model [36]), there are a number of findings that support the notion of emotion-specific properties, such as the unique involvement of the insula in disgust processing [37]. Quite often, the underlying structure of emotion is assumed based on the possibilities dictated by measurement instruments, making both dimensional and discrete views prevalent and arguably equally important [38]. It is however crucial to acknowledge that neither one nor the other approach can currently be considered as the true underlying structure of emotion and a joined theoretical approach of both perspectives would need clear and universally agreed upon criteria that do not yet exist [39].

More points of contention include the question if emotions are innate and universal (“basic”) categories or the result of social constructions. Although modern theories agree that both biological and sociocultural factors play a role in the development and expression of emotions, there are still fundamentally different views regarding the importance and roles of those factors. Following the logic made famous by Ekman [40], researchers arguing for the existence of basic (or universal) emotions build their theories on findings supporting cross-culture emotion expressions, especially in facial expressions [41,42], and neurophysiological data examining affective processes related to “old”, evolution-shaped systems in the mammalian brain [43,44]. In this view, emotions are basically hardwired and especially on an unconscious (or “deep”) level universal, while cultural influences begin to play a role on a conscious, second-order level [44,45]. The constructivist perspective argue for emotions as sociocultural constructions that do not emerge from emotion-specific brain patterns, but that the brain provides mechanisms for affective learning, leading to the construction of emotions within cultural and social contexts [46,47]. Again, arguments are being made for both perspectives, although they interpret the nature of emotion differently. Basic emotion theories often explain the functions of emotions through an evolutionary lens: Anger and fear lead to approach and avoidance respectively and fulfill therefore different roles in behaviour motivation, dictated by the biological development of humans [48]. Constructivist views on the other hand see learning and sense-making as the evolutionary advantage of emotion, which enables action tendencies, communication, and social influence

within experienced interactions [49]. In applied contexts such as affective games, the expectation of the effect of emotion-eliciting material (e.g. expected player behaviour or relationship between game material and emotional reaction) can change drastically depending on the theoretical approach, particularly its assumptions about universality.

Beyond the sources of uncertainty in emotion theory, there are currently many aspects that are generally agreed upon. Most modern theories argue for the importance of cognitive appraisal of a triggering situation in emotion elicitation [20], individual differences in emotion experience and expression, as well as context-dependent differences in emotion experience and expression. Furthermore, emotions involve multiple components (such as a behavioural component, physiological component, and subjective feeling component) that interact in various ways and have important implications for measurement. An important implication of this is that there is currently no clear mapping between emotional states and specific state indicators (i.e. measures of componential expressions). In other words, even well-established measures of affective physiological data such as heart rate (HR), heart rate variability (HRV), or electrodermal activity (EDA) can only measure specific aspects of emotion components (such as physiological arousal) that can indicate only some emotions in some circumstances [34].

Overall, there are many potential theoretical obstacles when it comes to emotion theory application that need consideration when assessing affect-adaptive game studies. How studies approach these conceptual problems and even assist in solving some of these questions by validating measures or elicitation mechanisms in game contexts will also need examination.

2.2. Affective gaming

Affective computing has been a prominent topic within human–computer interaction (HCI) research, exploring the measurement of and reaction to user emotions by a computer system [11]. It therefore represents an interactional relationship that has the potential to provide optimal emotional experiences by taking the current affective state of the user into account. In an effort to bring affective computing research to games, Hudlicka [50] described principles and current issues of the three main components of affective games: emotion sensing and recognition, computational models of emotion, and emotion expression or adaptation. Building on this, she outlined requirements for an ideal emotion engine that could accurately measure and interpret emotional data from the player and feed it into a model, as well as create realistic emotional behaviours for NPCs [6]. Affective design in general is therefore mainly concerned with addressing these requirements and developing solutions within three affective tasks:

1. **Emotion Sensing**: Lux et al. [51] identified 76 studies that use biofeedback devices as an affective measurement, ranging from measures of cardiovascular activity to electrodermal activity, body movement, or respiration. For games specifically, common measurements include physical measures like body movement; physiological measures like skin conductance, heart rate, muscle movement, or brain waves; and observation measures like facial or vocal expression [52]. Currently, there is no universally accurate instrument to measure emotions and recognition methods depend on emotion model assumptions, individual differences, and context. Furthermore, measurements are often seen as invasive, expensive, and unpractical [53].
2. **Computational Emotion Modelling**: Models of emotions are most commonly researched in artificial intelligence game studies with the main aim being the development of realistic affective game agents. Hamdy and King [54] collected requirements to develop emotional agents and provided an overview of computational emotion models. They pointed out that models often have to simplify the complex nature of emotions and are also quite costly and difficult to develop. Similarly Hudlicka [55] found that models often do not address detailed implications of psychological theories. They concluded that in order to fit with modern, complex theories of emotion, believable and realistic agents need to address theoretical uncertainties first, meaning that more systematic and integrative research is necessary. In a systematic review by Wang et al. [56], current practices in emotion modelling for affective computing and their implications were described, uncovering methodological difficulties present in the field.
3. **Adaptation**: Finally, research considering emotion adaptation focuses either on affect-based changes in agents or the game world [57]. Agents are again used to express emotions based on the underlying model and showcase mostly "believable" emotional behaviour, while the game world is specifically designed to reinforce a target emotion. For example, adaptive difficulty has been used to limit frustration [58], and adaptive camera movement has been used to augment a game's narrative [59].

In order to facilitate research addressing these tasks, Yannakakis and Paiva [57] provided descriptions of three game system modules: an emotion detection module (a module to measure and model player emotions), an adaptation module (a module to adapt the game world to these player emotions), and an elicitation module (a module to elicit target player emotions). These modules are embedded into a shared high-level concept of the emotional interaction between player and game, known as the affective feedback loop [57,60]. The closed nature of the loop is emphasized, as the ongoing adaptation of the game system to the changes in players' emotions is argued to be a unique characteristic of games compared to other mediums [50] and is also generally believed to facilitate emotional benefits, such as health benefits, more accessible games, new gameplay opportunities, and higher enjoyment.

Bontchev [17] analysed 14 video games that integrated affect-based adaptation techniques. They found that affective-adaptive games generally were effective in achieving goal-oriented changes (e.g. more enjoyment while playing). However, because of often incomplete internal models of affective player behaviour, they conclude that there is much more work to do to achieve a complete and realistic system for affect adaptation in video games. In a systematic review Robinson et al. [19] analysed 162 biofeedback games and found effects not only for player engagement, but also for treatment in health related affective games. However, they also note that many physiological game studies show insufficient critical reflection, both in terms of how technological limits are reported, and how rigorous evaluation is executed. As of yet, there is no systematic review that focuses on high-quality studies evaluating modern affect-adaptive games to analyse the adaptation effectiveness, studies' implementations of emotion theoretical assumptions, and their methodological approach in a comprehensive manner.

2.3. The current research

Affect-adaptive games have been a topic of interest in human–computer interaction research for many years now, as they promise a variety of benefits to players, ranging from increased enjoyment to mental health benefits [16,58]. However, there is a lack of comparative studies to investigate how well affect-adaptive games achieve these effects in the published literature. Since emotion research is also still struggling with fundamental theoretical definitions, there is also a question of how robustly affective game studies apply and test psychological theories or if they depend on theoretical assumptions that are still being debated or are even outdated. Finally, there is a need to compare affect-adaptive game studies in terms of methodological rigour to assess how well the reported effect can be generalized.

To our knowledge, this is the first systematic review that analyses affect-adaptive video game studies in terms of (a) the effect of adaptation; (b) the theoretical assumptions regarding emotions; and (c) the

quality of the evidence regarding evaluation of such studies. This study tries to address these gaps by systematically analysing the available research body of affect-adaptive video games to answer the following questions:

1. **RQ1: What evidence is there for effectiveness of game adaptation to player emotions?**
   (a) How many studies evaluate the effect of affect-adaptation within a video game?
   (b) What dependent variable is used to indicate adaptation success?
   (c) What empirical evidence is reported as part of the evaluation?
2. **RQ2: What emotion theoretical assumptions are being applied to build affective adaptation?**
   (a) How are target emotions defined? What theories are used?
   (b) What measures are used to indicate affective states and how are they tested?
   (c) What material is used to elicit emotions and how are they tested?
3. **RQ3: How are affect-adaptive games being evaluated?**
   (a) What sample characteristics are provided?
   (b) What control condition is used for the evaluation?
   (c) What are characteristics of the methodology?

3. Methods

This review follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al. [61]). All studies that empirically evaluated an affect-adaptive video game by comparing the affect-adaptive game to a control condition were considered for inclusion. A protocol for the study was preregistered on the Open Science Framework (OSF; 2022) before data screening commenced, but after the initial database searches, which were conducted first to assess the scope and feasibility of the study.

3.1. Data collection

Electronic databases were searched on April 8th 2022. Databases that are relevant for information technology, health, and social sciences were chosen, which includes: ACM Digital Library (n = 561), IEEE Explore (n = 824), Science Direct (n = 53), and Scopus (n = 2490). Additional studies (n = 2) were identified through reference lists of relevant studies [17,19], as well as through searches of the search terms on Google Scholar. The Google Scholar searches were conducted once with and without the "intitle:" operator and limited to the first 20 result pages. The database searches returned a total of 3930 papers.

3.2. Search terms

Search terms were chosen based on three necessary study characteristics, namely (a) it had to include a video game, (b) it had to include some kind of adaptation, (c) this adaptation was based on emotion. The string for the first characteristic was based on common practice in similar studies (e.g. [19]) and included GAME* OR GAMING. The search string for the second characteristic was based on game adaptation literature and used synonyms for adaptation processes and included ADAPT* OR MODUL* OR ADJUST*. Lastly, the string based on the third characteristic was based on affective computing studies and terms used for emotions or emotional components, namely: AFFECT* OR EMOTION* OR VALENCE OR AROUSAL OR EXPERIENC*.

3.3. Inclusion criteria

This review aims to investigate the reported effect of affective-adaptive games and how these effects are being empirically evaluated. Therefore, it focuses on high quality comparative studies, leading to the following inclusion criteria:

1. Peer-reviewed papers (including conference papers)
2. Full-length papers
3. Available in English or German
4. Test an adaptive video game based on affective information
5. Evaluates the adaptation effects empirically against a control condition

3.4. Exclusion criteria

Following the reasoning to provide insight about high-quality work in the field, studies were excluded if they showed one of the following characteristics:

1. Do not include an empirical study (i.e. reviews, study protocols, 'work-in-progress')
2. Evaluate only through qualitative or descriptive means
3. Do not compare to a control condition that is not affect-adaptive
4. Evaluate only based on case studies (defined as $N < 5$)

It is important to note that all non affect-adaptive control conditions where included in the study, including performance-adaptive or non-adaptive games that were tested in a within-design. Additionally, it was not a necessary criterion to include randomized condition assignment, as for example in quasi-experimental designs. Evaluations therefore did not need to consist of randomized controlled trials (RCTs) to be included.

3.5. Data analysis

The initial search returned 3930 papers, 755 of which have been identified as duplicates and were removed. Title and abstracts were screened by the principal investigator and papers that demonstrated a clear mismatch to any of the relevant research questions (e.g. papers that do not involve video games or HCI in general) were excluded, leading to the removal of another 2965 papers. The 210 remaining papers were assessed by reading the full texts of the papers and coded in regards to the inclusion and exclusion criteria. Out of these papers, 32 were excluded for not involving an empirical study, 7 were excluded for not involving a video game, 38 were excluded for not involving an adaptation, 55 were excluded for not basing the adaptation on affective data, 36 were excluded for not evaluating the effects of affective-adaptation empirically, 14 were excluded for not involving a control condition within the evaluation, and finally 2 were excluded for only evaluating through a case study. The final set of papers consisted of 26 studies that were further analysed within this review. A full representation of the process as proposed by PRISMA guidelines [61] can be viewed in Fig. 1.

3.6. Coding

In order to answer our three research questions individually, specific aspects of the full sample were coded under predefined conditions for each question.

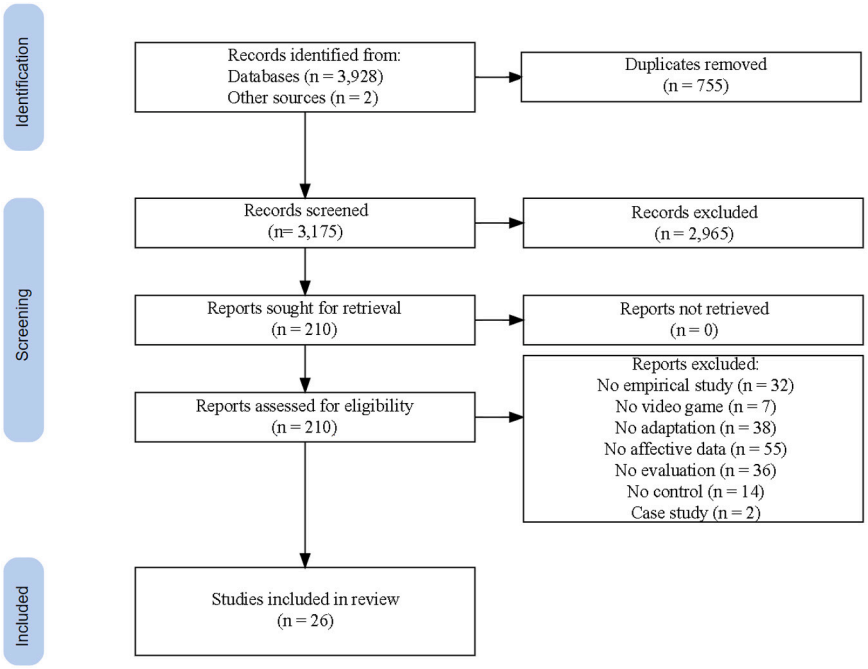

Fig. 1. Flow diagram of data extraction process.

3.6.1. RQ1: Evidence synthesis

Each study was coded by publication year and within each study, each adaptive game was coded by genre (as described in the paper itself). For each game that was tested, the specified outcome variable was coded, including the measurement instrument. The study design was coded based on the control being a within or between condition and the number and nature of all tested conditions were specified. Finally, the effect was coded as positive, mixed/neutral, or negative (experimental condition compared to control condition), and if possible the reported effect size was included and coded as small, medium, or high effects, based on interpretation guidelines as reported by Fritz et al. [63]. For studies where no effect size was included, but sufficient data was provided, effect sizes were calculated and interpreted as described by Fritz et al. [63].

3.6.2. RQ2: Theoretical assumptions

For each paper, the affective state of interest (i.e. source of adaptation) was coded based on the theorized underlying structure (dimensional vs. distinct) and the reported labels of the measured emotional states. Furthermore, the specific measures used to detect the emotional state were recorded. Together these details were gathered in order to examine how affect was measured across studies. Each paper's efforts to validate individual measurement instruments (e.g. through comparison with self-report scales) were also recorded. Tests were either direct (i.e. related to subjective measures of the target emotion), indirect (related to other indication of target emotions), or absent.

Additionally, it was coded what game material was adapted to affective information and whether these game materials were tested in their ability to elicit a target emotion in order to inform the adaptation design. Game materials that were adapted of each game were listed and summarized where appropriate (e.g. "difficulty" for all individual gameplay changes that were made to increase challenge). Test were again either directly (impact of materials was related to subjective measures of target emotion), indirectly (impact of materials was related to other indications of target emotions), or not tested.

3.6.3. RQ3: Methodological approach

Finally, for each evaluated game, the methodological approach was coded, including sample information (N, percentage of male participants, mean age) and the used statistical test. An estimate of achieved statistical power was calculated post-hoc for each study based on the study design, sample size, and an assumed medium effect size (0.5 standard deviations [SD]). Rather than providing an estimate of "achieved" power, this was done because such an estimate completely depends on the observed effect and can therefore be misleading, as it is not theory-based, nor a good indicator of methodological validity [64]. Additionally, many studies did not provide sufficient information to calculate the observed effect size, which would limit the ability to compare all studies. To provide more insights about each statistical power, target effect sizes (ES) were calculated, representing the detectable effect sizes for a study, assuming a power (a priori) of at least 0.8. The target effect size therefore represents the necessary differences between groups in SD to achieve a power of 0.8 or higher.

Furthermore, risk of bias (RoB) was assessed using the Cochrane Collaboration's tool [65]. Risk of bias was assessed based on objective criteria regarding multiple domains: (a) selection bias (i.e. whether or not participants' allocation was concealed and randomized), (b) performance bias (i.e. whether participants were aware of the intervention and if this could affect outcomes), (c) attrition bias (i.e. how much missing data regarding the outcome was reported and how that could influence analyses); (d) detection bias (i.e. whether or not clear and appropriate measures for the outcome were reported and whether deviations arose through data collection strategies); and (e) reporting bias (i.e. whether or not all results from all measurements and analyses were reported). An overall RoB was judged based on the following criteria:

1. Low risk: The study presents a low risk of bias for all domains
2. Some concerns: The study presents some concerns in at least one domain, but no high risk for any domain
3. High risk: The study presents a high risk in at least one domain

A detailed overview of all domains and criteria was provided by Higgins et al. [65].

4. Results

4.1. RQ1: Effectiveness of affective adaptation

26 studies were included in the analysis. A description of study aims, methods, and conclusions can be found in Table 5. An overview

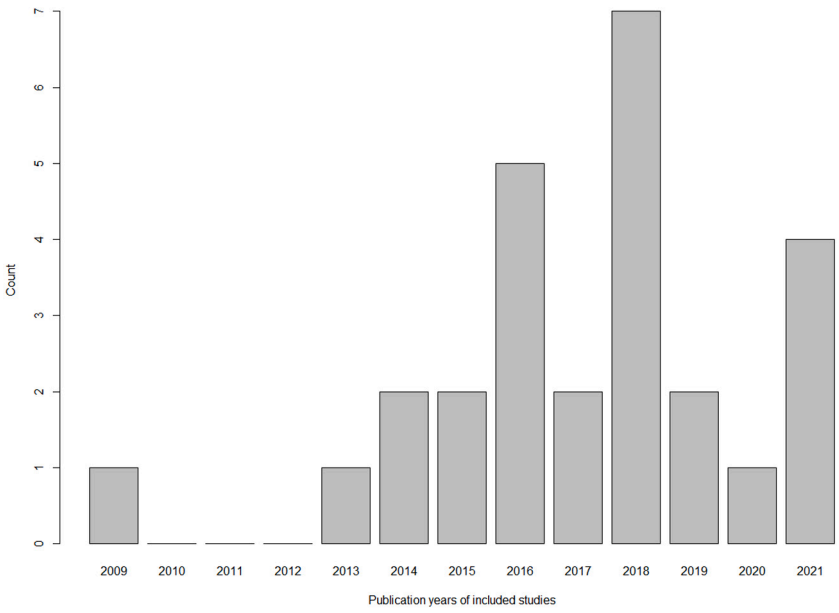

Fig. 2. Counts of included studies by year.

of publication years can be seen in Fig. 2. Over half of all studies (n = 14) were published in 2018 or later. 69% (n = 18) tested an affect-adaptive game against one or more control conditions in a repeated measure design and 31% (n = 8) used a group comparison. 86% of studies (n = 22) used randomized subject assignment, 8% (n = 2) did not use randomization for subject assignment, and 8% (n = 2) did not report sampling procedures.

In these 26 studies, 27 affect-adaptive games have been described. A overview of the games' genres can be seen in Table 1.

To evaluate adaptation effectiveness, 18 different outcome variables were used, assessed through 15 different instruments (see Table 2). The outcome variables can be summarized within three broad categories. The most used outcome category (n = 16) is player experience, which includes outcome variables such as enjoyment, engagement, immersion, aesthetics, dynamics, competence, character believability, fun, flow, and general player experience. 46% of studies (n = 12) measured player experience through a previously validated self-assessment instrument, such as the Game Experience Questionnaire (GEQ; IJsselsteijn et al. [66]; n = 3), the Intrinsic Motivation Inventory (IMI; Markland and Hardy [67]; n = 3), the Flow Experience Measure (FEM; Sung et al. [68]; n = 1), the Player Experience of Need Satisfaction (PENS; Ryan et al. [14]; n = 1), the Immersive Experiences Questionnaire (IEQ; Jennett et al. [69]; n = 1), the Player Experience Inventory (PXI; Vanden Abeele et al. [70]; n = 1), User Response to Interactive Storytelling tool (URTIS; Vermeulen et al. [71]; n = 1), and the Character Believability Questionnaire (CBQ; Gomes et al. [72]; n = 1). Additionally, 27% of studies (n = 7) constructed own scales to assess player experience.

Another category includes affective variables (n = 8), such as arousal, stress, valence, excitement, and anxiety. These were measured mostly through physiological data, including heart rate (HR; n = 1), heart rate variability (HRV; n = 3), electrodermal activity (EDA; n = 3), electroencephalography (EEG; n = 1). Some studies measured the affective outcome through facial expression recognition (FER; n = 1) or voice analysis (n = 1), and finally some through subjective self-assessment tools such as the Self-Assessment Manikin (SAM; n = 1), the Mood Adjective Checklist (UMACL; n = 1), or an own scale (n = 2).

The third category consists of performance metrics (n = 9), either in-game performances (n = 8) or learning performance metrics (n = 1).

While a variety of outcome variables were used, most studies reported a positive effect direction (i.e. increase in affect-adaptive condition compared to control). 65% of studies (n = 17) reported statistically significant positive effects, of which 2 can be considered small, 4 can be considered medium, 6 can be considered large, and the remaining 5 were not reported with sufficient data to calculate effect sizes. Only 4% of studies (n = 1) reported a significant negative effect. 42% of studies (n = 11) reported non-significant effects for at least some of their outcome variables. As shown in Table 2, these effects were either reported as "No effect", "positive n.s.", or "negative n.s." within the studies, with the last two signalling a descriptive trend of the data, but no statistically significant effect.

Table 1
List of genres for adapted games analysed in this review.

Genre	No of studies	% of sample
Action (3D)	5	18
Arcade	3	11
Education	2	7
Horror	4	15
Interactive Story	1	4
Platformer (2D)	4	15
Shooter (3D)	4	15
Training	4	15

4.2. RQ2: Emotions in affect-adaptive games

All presented games aimed at improving a predefined outcome variable through adapting game material to emotional states. They included therefore means to measure the affective states, and some emotion-eliciting material that was the aim of adaptation. An overview of emotion-theoretical assumptions, emotion measures, eliciting materials and whether measures and material was tested in their ability to reflect target emotional states can be seen in Table 3.

42% of studies (n = 11) considered emotions as distinct states, while 50% (n = 13) considered emotions as instances along a dimension. The remaining 8% (n = 2) explicitly defined and measured both distinct and dimensional affective variables. Adaptations were based on a wide variety of affective triggers that were often based on the means of measurement (e.g. arousal cut offs with dimensional measures, classified fear with distinct measures). The number of states that were measured ranges between 1 and 11.

Affective measures were used to indicate states by specific emotion component expressions. The most widely used form of measurement

Table 2
List of included studies, outcome variables, outcome assessment instrument, effect direction (non-significant effects marked with n.s.), and observed effect size if sufficient information was provided.

Authors	Outcome	Instrument	Control	Effect	Effect size
Akbar et al. [73]	Player experience	GEQ (2013)	Non-adaptive	Positive	N/A
Al Osman et al. [74]	Stress reduction	Physiology (HRV)	Non-adaptive	Positive	Medium
Alves et al. [75]	Flow, Performance	GEQ (2009)	Performance adaptation	Negative	Small
Andrew and Chowanda [77]	Valence decrease, Arousal increase	Facial expression	Non-adaptive	Positive	Medium
Blom et al. [78]	Preference	Single item	Non-adaptive	Positive	N/A
Bontchev and Vassileva [79]	Effectiveness, Efficiency, Difficulty	In-game-performance	Non-adaptive	Positive	N/A
Bontchev and Georgieva [80]	Effectiveness, Efficiency, Difficulty	In-game-performance	Non-adaptive	Positive	Medium
Darzi et al. [81]	Player experience	IMI (1997), FEM (2015)	Manual, Random, Performance, Personality adaptation	No effect	N/A
Ewing et al. [82]	Enjoyment, Immersion	UMACL (1990), IEQ (2008)	Manual adaptation	No effect	N/A
Frommel et al. [84]	Perceived competence, Aesthetics, Dynamics	PXI (2016), IMI (1989)	Increasing difficulty, Fixed difficulty	Positive	Large
Hernandez et al. [86]	User experience	URTIS (2010)	Non-adaptive	No effect	N/A
Ibáñez et al. [87]	Presence	SUS (1994)	Non-adaptive	Positive	N/A
Jalbert and Rank [89]	NPC rapport	3-item questionnaire	Non-adaptive	No effect	N/A
Lara-Alvarez et al. [90]	Affective state, performance	Voice analysis, In-game-performance	Non-adaptive	Positive, no effect	Medium
Liu et al. [91]	Player experience, Performance, Anxiety	Single items (9-point likert)	Performance adaptation	Positive	N/A
Moniaga et al. [92]	Challenge and experience	IEQ (2008)	Non-adaptive	Positive	N/A
Negini et al. [93]	Arousal, Player experience	EDA, IMI (1997) & PENS (2006)	Non-adaptive	Positive, no effect	Large
Nogueira et al. [94]	Player experience	GEQ (2013)	Non-adaptive	Positive	Large
Parnandi et al. [95]	Physiological arousal, Performance	Physiology (HRV, EDA), In-game-data	Non-adaptive, Deep breathing task	Positive, No effect	N/A
Parnandi and Gutierrez-Osuna [96]	Physiological arousal, Performance	Physiology (HRV, EDA, BR), In-game-data	Non-adaptive, Deep breathing task	No effect	Large
Rodriguez-Guerrero et al. [97]	Valence, Arousal, Dominance	Physiology (HR, EDA), Subjective (SAM (1994), own scale)	Non-adaptive	Positive (n.s.)	Large
Rosa et al. [98]	Performance, Flow	In-game data, Single items	Performance adaptation, Non-adaptive	No effect	N/A
Salah et al. [99]	Learning, Engagement	Not specified	Non-adaptive	Positive	Small, large
Stein et al. [100]	Long term excitement, Enjoyment	Physiology (EEG), Single item	Non-adaptive	Positive	Small
Tjokrosetio and Chowanda [101]	Character believability	CBQ (2013)	Non-adaptive	Positive	N/A
Vachiratamporn et al. [102]	Fear, Fun, Difficulty	5-point scale	Non-adaptive	Negative (n.s.)	Medium

were physiological measures, used by 62% of studies (n = 16) and were conducted through HR readings (n = 6), HRV readings (n = 4), EDA (n = 10), EEG (n = 4), electromyography (EMG; n = 3), or breathing rate (n = 2). 42% of studies (n = 11) considered observational data of behaviours, such as facial expressions (n = 8), voice analysis (n = 1), gesture analysis (n = 1), or in-game choices (n = 1). Finally, 8% of studies (n = 2) measured subjective feeling in-game as a mean to adapt gameplay through self-report ratings. 62% of studies (n = 16) did not explicitly test how well the used measure indicated target emotional states, meaning that these studies relied on either previously tested or untested theoretical assumptions regarding how well a measure could differentiate between predefined states based on a predefined underlying structure. 12% (n = 3) indirectly tested the measure, by validating it through other means than subjective emotion self-assessment (e.g. comparing physiological measures, or testing reliability of differentiating between game materials emotion-eliciting). 26% of studies (n = 7) tested testing the measure within a certain game context directly, by associating it with the self-reported target emotion in an experimental context.

Table 3
List of included studies, definition of underlying emotion structure, emotional state labels, emotion measure, whether the measure was tested in the study, the in-game adapted material used for emotion elicitation, and whether the effect of this material on emotion elicitation was tested in the study.

Authors	Structure	State labels	Measure	Measure validated	Adapted material	Material tested
Akbar et al. [73]	Distinct	Anger, Frustration, Smile, Relaxation	Facial expression	Not tested	Difficulty	Not tested
Al Osman et al. [74]	Dimensional	Stress	Physiology (HRV)	Not tested	Visual feedback	Indirectly
Alves et al. [75]	Distinct	Anxiety, Boredom, Engagement, Frustration	Physiology (HR, EEG)	Indirectly	Difficulty	Directly
Andrew and Chowanda [77]	Dimensional	Valence, Arousal	Facial expression	Not tested	Difficulty	Not tested
Blom et al. [78]	Distinct	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise	Facial expression	Indirectly	Difficulty	Directly
Bontchev and Vassileva [79]	Distinct and dimensional	[Anger, Disgust, Fear, Happiness, Sadness, Surprise] and [Arousal]	EDA and facial expression	Not tested	Difficulty, Lighting	Not tested
Bontchev and Georgieva [80]	Distinct and dimensional	[Anger, Disgust, Fear, Happiness, Sadness, Surprise] and [Arousal]	EDA and facial expression	Not tested	Difficulty	Not tested
Darzi et al. [81]	Dimensional	Preference	Physiology (EDA, EEG, HR, HRV)	Directly	Difficulty	Directly
Ewing et al. [82]	Distinct	Boredom, Engagement	EEG	Directly	Cognitive demand	Directly
Frommel et al. [84]	Dimensional	Boredom, Frustration	Questionnaires	Directly	Difficulty	Indirectly
Hernandez et al. [86]	Distinct	Distress, Fear, Hope, Joy	Behaviour	Not tested	Narrative trajectory	Not tested
Ibáñez et al. [87]	Distinct	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Gestures	Indirectly	Music	Not tested
Jalbert and Rank [89]	Distinct	Alarmed, Angry, Bored, Content, Depressed, Happy, Miserable, Neutral, Tired	Physiology (EDA, EMG)	Not tested	NPC Dialogue	Not tested
Lara-Alvarez et al. [90]	Dimensional	Valence, Arousal	Voice analysis	Directly	Difficulty, Sound	Partly directly
Liu et al. [91]	Dimensional	Anxiety	Physiology (HR, EMG, EDA)	Directly	Difficulty	Not tested
Moniaga et al. [92]	Distinct	Anger, Frustration, Joy	Facial expression	Not tested	Difficulty	Not tested
Negini et al. [93]	Dimensional	Excitement	Physiology (EDA)	Not tested	Difficulty	Not tested
Nogueira et al. [94]	Dimensional	Arousal, Valence	Physiology (EDA, EMG, HR, HRV)	Not tested	Character representation, Difficulty	Indirectly
Parnandi et al. [95]	Dimensional	Arousal	Physiology (BR)	Not tested	Difficulty	Not tested
Parnandi and Gutierrez-Osuna [96]	Dimensional	Arousal	Physiology (EDA, HRV, BR)	Not tested	Difficulty	Not tested
Rodriguez-Guerrero et al. [97]	Dimensional	Arousal, Dominance, Valence	Physiology (HR, EDA)	Directly	Difficulty	Directly
Rosa et al. [98]	Dimensional	Boredom, Frustration	Physiology (EDA)	Not tested	Difficulty	Directly
Salah et al. [99]	Distinct	Boredom, Frustration, Relaxation	SAM (1994)	Not tested	Difficulty, Aesthetics	Not tested
Stein et al. [100]	Dimensional	Excitement, Frustration	Physiology (EEG)	Not tested	Difficulty	Indirectly
Tjokrosetio and Chowanda [101]	Distinct	Anger, Disgust, Fear, Joy, Neutral, Sadness	Facial expression	Not tested	NPC behaviour	Not tested
Vachiratamporn et al. [102]	Distinct	Anxiety, Fear, Neutral, Suspense	Physiology (HR)	Directly	Enemy position	Not tested

Table 4
Included studies, sample size (N), reported demographics (% male and mean age), statistical test, estimated power assuming a medium effect (0.5 SD), target effect size (ES), and risk of bias (RoB; + refers to low risk; +/− refers to some concerns; − refers to high risk of bias).

Authors	N	% Male	M Age	Test	Power	Target ES	RoB
Akbar et al. [73]	60	68	N/A	t-Test	0.85	0.5	+
Al Osman et al. [74]	12	58	33.92	MANOVA	0.20	1.3	+/−
Alves et al. [75]	21	76	22.43	t-test	0.39	0.71	−
Andrew and Chowanda [77]	31	N/A	N/A	Wilcoxon Rank	0.84	0.5	−
Blom et al. [78]	25	80	N/A	Z-test	0.43	1.0	+/−
Bontchev and Vassileva [79]	30	60	N/A	t-test	0.85	0.5	+
Bontchev and Georgieva [80]	30	60	31.87	t-test	0.85	0.5	+
Darzi et al. [81]	50	74	25.1	t-test	0.23	1.1	+
Ewing et al. [82]	10	40	N/A	t-test	0.29	0.8	+
Frommel et al. [84]	66	73	30	ANOVA	0.99	0.3	+
Hernandez et al. [86]	294	50	19	MANOVA	0.99	0.2	+
Ibáñez et al. [87]	22	67	29.09	t-test	0.3	1.2	+
Jalbert and Rank [89]	16	63	N/A	t-test	0.24	1.4	+
Lara-Alvarez et al. [90]	40	N/A	N/A	t-test	0.93	0.4	+
Liu et al. [91]	9	47	N/A	ANOVA	0.26	1.1	+/−
Moniaga et al. [92]	32	N/A	N/A	Wilcoxon Rank	0.85	0.5	+
Negini et al. [93]	16	94	N/A	ANOVA	0.61	0.7	+
Nogueira et al. [94]	24	67	22.5	MANOVA	0.41	0.8	+
Parnandi et al. [95]	9	78	N/A	Not specified	0.08	2.8	+/−
Parnandi and Gutierrez-Osuna [96]	16	94	N/A	ANOVA	0.12	1.6	+
Rodriguez-Guerrero et al. [97]	11	73	30.5	Not specified	0.19	1.7	+/−
Rosa et al. [98]	36	61	N/A	Friedman test	0.82	0.5	−
Salah et al. [99]	30	67	19	t-test	0.38	1.0	+/−
Stein et al. [100]	24	92	25.59	ANOVA	0.74	1.0	+
Tjokrosetio and Chowanda [101]	52	86	N/A	Z-Test	0.97	0.4	+/−
Vachiratamporn et al. [102]	12	92	25.42	Not specified	0.49	0.8	+

The emotion-eliciting game material that was the source of adaptation was mostly focused on challenge aspects of games. 77% of games (n = 20) manipulated game material with the aim to change a game's difficulty in order to evoke a range of emotions. 19% of studies (n = 5) manipulated the game's aesthetics (through the visuals or audio) as a way to evoke emotions. 15% of studies (n = 4) manipulated non-playable characters (NPCs) or story progressions to reflect affective data of players, and 4% of studies (n = 1) manipulated non-challenge related in-game events to evoke fear. Again, a majority of studies (58%, n = 15) did not test the effects of game material manipulation on the target emotion through self-report measures. 16% of studies (n = 4) tested the emotional elicitation effect of material through indirect measures (such as physiology), and 26% of studies (n = 7) tested the effects of the adapted game material on target emotions directly through self-report measures.

4.3. RQ3: Methodologies

The sample size (n) ranged from 9 to 294 (M = 37.62), with a Median sample size across studies of 24 participants. None of the studies justified the sample size on statistical power assumptions. 96% of studies (n = 25) provided information about demographic details, such as mean age (n = 18), age range (n = 21), gender distribution (n = 24), or game experience (n = 15). Statistical power assuming a medium (0.5 SD) effect size ranged from 0.08 to 0.99 (M = 0.55; Md = 0.46). The target effect size detectable with the study design ranged from 0.2 SD to 2.8 SD (M = 0.89; Md = 0.8). 8% of studies (n = 2) were able to detect a small effect size (up to 0.3 SD), 42% of studies (n = 11) were able to detect a medium effect size (up to 0.6 SD), 58% of studies (n = 15) were able to detect a large effect size (up to 0.9 SD), and 88% of studies (n = 23) were able to detect a very large effect size (up to 1.5 SD). An effect size of up to 2 SD was detectable by 96% of studies (n = 25) and one study was under powered for lower effect sized than 2.8 SD (see Table 4).

62% of studies (n = 16) were found to have a low risk of bias (RoB), i.e. no bias concerns in the observed domains. 27% of studies (n = 7) showed some concerns for risk of bias, and 11% of studies (n = 3) showed domains with a high risk of bias. Al Osman et al. [74] compared a biofeedback game against the same game with hidden feedback. They also introduced participants to the game aim and relaxation strategies through meditation in the biofeedback condition only. These conditions were therefore visible to participants and could impact the outcome, leading to some concerns in the domain of performance bias, even though the sampling was reportedly counterbalanced. Alves et al. [75] reported inconsistent empirical results (i.e. different effect sizes for the same effect), which was judged a high risk for reporting bias. Andrew and Chowanda [77] used strategies of unconcealed randomization, did not report group comparisons for all outcome measures, and proposed some conflicting operationalizations of similar measures (such as negative valence through FER and positive affect through self-report as desired outcomes), leading to a high risk in the domains of detection and reporting bias and some concerns in selection bias. Blom et al. [78] reported multiple outcome variables (preference, challenge, immersion, frustration) through constructed self-assessment questions, but only report descriptive differences for challenge, immersion, and frustration without a statistical test to test these differences, which indicates some risk in reporting bias. Liu et al. [91] provided a clear methodology, but conducted some additional analyses and created variables not previously justified, indicating some potential risk for reporting bias. Parnandi et al. [95] missed some important information in the process description (such as randomization, blinding, or how knowledge of different interventions [such as affective game vs. deep breathing exercise] was controlled in its potential to affect outcome). It is not clear if all outcomes are sufficiently reported, as a statistical test for group comparisons was not provided for all outcome variables, indicating some concerns for reporting bias. Rodriguez-Guerrero et al. [97] were not able to randomize participants across conditions, as experimental data was compared to a previously conducted experiment. Additionally, they provide very limited reports of group differences for all outcome variables, indicating some concern for selection bias and reporting bias. Rosa et al. [98] did not provide a clear analysis plan (including number and types of outcome variables and statistical tests), leading to some potential replication issues and a high risk for reporting bias. Potential order effect due to missing counterbalance was not discussed, indicating some concerns for selection bias. Salah et al. [99] reported extremely large effects (>5 SD group difference), without sufficient indications on the potential nature of these effects. Measures such as "learning effect" were also not clearly defined, indicating some

concerns regarding performance and detection bias. Finally, Tjokrosetio and Chowanda [101] described an unconcealed randomization process, leading to some concerns in selection bias. Additionally, the outcome variable was tested by participants watching specific gameplay videos without playing the tested games, leading to an unclear evaluation of adaptation, as it was not described how emotion adaptation contributed to changes in outcome variables.

5. General discussion

This study aimed at investigating the impact of affect-adaptive games on various possible outcome variables through a systematic review of high-quality evaluation studies of the field. To broaden our understanding of the nature of these studies, both theoretical assumptions regarding emotion research, and methodological concerns were examined. 26 studies were identified that evaluate affect-adaptive games against a form of control condition in an empirical context and their contents were summarized.

5.1. The effects of adaptation

To judge whether or not emotional game adaptation can be seen as effective, there are many variables that need consideration. In the initial search, many studies were identified that describe methods to achieve affect-adaptive adaptation, but many did not focus on evaluation (n = 36), some did evaluate but either without a control condition or only using case studies (n = 16). Still, the empirical evaluation of affective games against controls has been a topic with increasing interest, as 26 studies were identified, most of them published after 2017. These studies test a range of different adaptation mechanisms in different genres of games, with different strategies to measure and model emotions, and even different outcomes of interest.

The most investigated outcome related to at least some domains of player experience. (PX) As a concept, player experience suffers from the lack of a clear conceptualization and measuring standard, which was mirrored by the abundance of different instruments to measure PX aspects. Only recently, efforts have been made to test and improve validity and reliability concerns. For example, Denisova et al. [103] tested the underlying structure of the IEQ, GEQ, and PENS and found considerable similarities, which make a clear distinction between tested PX domains difficult. Similarly, Johnson et al. [104] tested the factor structure of the GEQ and PENS and found they were only partially replicable. Aspects of concepts like immersion and flow show considerable overlaps, leading to further doubts about how many and which domains PX consists of Michailidis et al. [105], influencing the value of PX as a precise and valid research outcome and therefore as a useful development concept. Because integrative and comparable research becomes more and more important to evaluate effects, our findings support the notion of the need of more unified concepts and instruments, especially in terms of game evaluations.

Still, using PX as a broad overall category of interest, mostly positive effects of affect-adaptive games have been reported. For example, Akbar et al. [73] provided empirical evidence for PX improvements through DDA using facial expression recognition for both a 2D platformer and 3D shooter and similar results were reported by Moniaga et al. [92] for a 3D Hack and Slash game. Frommel et al. [84] used in-game self-reported emotions to adapt difficulty in a 2D platformer, leading to large effects. Nogueira et al. [94] extensively tested multiple versions of affective adaptation through physiological data in a survival horror game and identified many large PX domain improvements compared to a non-adaptive game. Ibáñez et al. [87] showed improved presence for a virtual reality horror game with fear-adaptive music against the same game with generic music.

There were some non-significant effects reported, which could indicate mixed results regarding effect of adaptation. Many of these however also indicate small sample sizes and a low statistical power, making it difficult to draw inferences. Darzi et al. [81] for example found no effect on multiple PX domains, but included many conditions, which led to a power of under 0.8 for any effect smaller than 1.1 SD. A similar picture can be seen in the study from [82], who did not find an effect against manual adaptation of difficulty, or Negini et al. [93], who found no effect for PX reports, both showed a generally low power. Jalbert and Rank [89] tested rapport with affect-adaptive NPCs, but also was severely underpowered for any effect smaller than 1.4 SD. While this does not necessarily mean that negative or non-effects are always based on power, it is very difficult to interpret results that are not sufficiently powered to uncover a range of effect sizes. The study by Hernandez et al. [86] provides an exception; they had a large sample size, but still found no effect of emotion-adaptation on PX. In this particular study, emotion was measured through in-game behaviour and classified based on designer-constructed rules, which introduces a range of validity concerns regarding whether or not the affect-adaptive game could truly be considered affect-adaptive (as this was not tested using any validated emotion measure).

One important aspect to note is that while issues in statistical power become immediately apparent in studies with non-conclusive results, there are also issues in studies reporting significant results. Because the observation of significant results with a small sample size means that the observed effect is quite large, a high post-hoc power can be misleading and should not be interpreted as strength of evidence [64]. In fact, most studies in this review only achieve a sufficient power with large (0.8 SD) or very large (>1 SD) effects. Even if these are found, issues in generalizability due to the small sample size should be considered. Salah et al. [99] conducted a study with a low sample size and found an extremely large adaptation effect for engagement (>5 SD). While it can be argued that there is no need for large samples if the theorized effects are large enough to be observable, a small sample is also less likely to represent a given population [64]. Extremely large effects for small samples might lead to unreliable interpretations as the same effect might not hold true for a general population. Studies of Al Osman et al. [74], Blom et al. [78] and Liu et al. [91] have similar issues and report large positive effects in at least some of the observed outcome variables with a low sample size. Statistical power was not explicitly discussed as a factor to justify sample size in any of the examined studies, and neither was accuracy. It is important to note that accuracy (i.e. width of confidence interval) can be seen as a considerable concern with most of the studies (given the median sample size of 24), making even significant effects potentially unrepresentative [106]. Additional concerns regarding generalizability and replicability was the inconsistent reporting of basic demographic data and descriptive statistics.

Studies that focused on affective outcomes reported positive to mixed results. Lara-Alvarez et al. [90] provided evidence for successful improvements in experiences of pleasant-high affective states in an affect-adaptive learning game using pre-validated voice analysis. Stein et al. [100] used an EEG-adaptive version of a 3D shooter and showed higher long-term excitement values compared to the control version of the game. Parnandi et al. [95] and Parnandi and Gutierrez-Osuna [96] showed mostly no differences, comparing a relaxation training game to a non-adaptive game and a deep breathing task condition with a very low sample size, leading to a general conclusion that affective games have promise in their ability to manipulate emotions through context (e.g. the ability to create stressful situations), which cannot be done with regular relaxation exercises, but the proper design and development of affective games need further work to provide consistent results. Rodriguez-Guerrero et al. [97] tested an affective against a non-affective neurohabilitation game with a low sample size and found inconclusive results, indicating complex affective relationships between game materials, player data, and outcomes. Vachiratamporn et al. [102] tested the effects of a fear-adaptive horror game in terms of emotional reactions, which remained non-significant, possibly based on a very low statistical power.

Studies that focused on the effects on performance [79,80] reported positive effects for shooting, puzzle, and exploration tasks in a 3D game for an adaptive game compared to a non-adaptive game, using physiological and face recognition information. In these particular studies, it is argued that the combination of relevant information (in this case affective information and playing style classification) to personalize experiences could lead to the largest effect. The authors conclude that there are still many unknown variables and interactions when it comes to affective adaptation, but the initial promising data points towards the potential of further research, especially research that reduces cost and obtrusiveness of affective recognition and modelling. While much more research exists, the theoretical and methodological differences make clear comparisons and effect interpretations impossible and can only lead to the conclusion that fundamental standards must be applied to better research affect-adaptation. While the reported effects of affective games seem ultimately promising, it may be too soon to fully evaluate them, given these barriers.

5.2. The role of emotion

All studies described a game that adapts its material to affective information, which is either continuously or intermediately measured. In general, affective states of interest can be considered emotional, i.e. states with a relatively short duration and high intensity. While the elicitation of certain moods (e.g. in horror games) was a particular aim, all studies measured and adapted to data relating to emotional reactions, either measuring physiological responses (through HR, HRV, EDA, EMG, or EEG), behavioural responses (through FER, gesture analysis, or voice analysis), or subjective feelings (through self-reports).

Mostly depending on the measurement instruments, the inferred emotional states are either considered as dimensional or distinct constructs with states of interest that are considered useful for a particular game adaptation. For example, some studies [95,96] focus on emotional arousal, measured through physiological arousal in an effort to create games for relaxation training. Others [73,78–80,101] use facial expression analysis to measure distinct emotional states, such as fear, joy, anger, or sadness. One of the main concerns when it comes to emotion measures is the inability of a single instrument to accurately reflect the complex nature of an emotion in its entirety [46]. Inferences made from one or multiple measures are also subject to different sources of variation, such as dispositional differences and current context [34,107]. That means that the validity of the emotion recognition system is highly dependent on the following factors: The measurement instrument, the emotion conceptualization, the given context, and individual differences. As the study by Rodriguez-Guerrero et al. [97] shows, even well-established affective qualities (in this case dimensions of valence, arousal, and dominance), measured through a combination of instruments (such as HR and EDA), can lead to poor accuracy. Still, the majority of studies (n = 16) did not explicitly test how well a certain measure predicted the target state and, instead, built the emotion recognition system on theoretical assumptions. While some of the assumptions have considerable representation in the literature, such as the association between physiological and emotional arousal [53], others are highly contested. For example, there is no clear consensus on which true emotional states are represented well through facial expressions [108]. Researchers (e.g. [73,92]) may therefore interpret potentially non-distinct facial expressions (such as smile and smirk) as distinct emotional states. Another contested point is how and if distinct emotional states could be mapped to affective dimensions (e.g. [77]), as dimensional and distinct theoretical frameworks of emotions often have vastly different theoretical bases [20]. Finally, the exact relationship between a physiological measure and an affective state is not clear for every individual and context [34], so the relationship is hard to interpret without concrete mappings that some of the studies did not provide [79,80,89,100]. While basing decisions on contested assumptions can be in some cases useful, especially in providing more insight into fundamental psychology research, without explicit validity testing, there is a risk in unknowingly misinterpreting ambiguous data. In the study by Ibáñez et al. [87], the gesture-based emotion recognition was tested indirectly by classifying participants who encountered a predefined "emotion-inducing" room within the game world and accuracy was only sufficient to distinguish between participants who visited the fear room or participants who visited any other room. Given a specific game and audience, such an approach could provide a way to adapt between two affective states, although it is unclear if these states truly represent the targeted fear vs. no fear states. Alves et al. [75] combined measures indicating fear and frustration into a combined emotional state to increase accuracy, although the theoretical and practical implications of such a state are not discussed.

The explicit (and direct) testing of measures given a game context and player base has in some of the analysed studies been used to improve emotion recognition strategies: Ewing et al. [82] described a 2-step process, first establishing relationship between measures and target emotions and then designing adaptations. Frommel et al. [84] measured the feeling of a target emotion through self-reports, which directly reflected the base of potential adaptation. Liu et al. [91] based their study specifically on anxiety and established methods to accurately predict anxiety in a preceding experiment. To ensure theoretically valid mappings, the relationship between a proposed emotion model and a given measured emotion component not only supports valid predictions, but also provides the opportunity to focus on any emotional state that might be of interest for game design, including complex emotions like shame or pride. In this sense, designers are not limited to measuring concepts with more established physiological correlates (such as emotional arousal), especially given the influences of context and individual differences that justify testing in any case.

Still, emotion recognition is only a part of the adaptation process. A game is only truly adaptive if it changes in a way that elicits a target emotion, which closes the feedback-loop [60]. Again, most investigated studies make theoretical assumptions regarding such an elicitation process. Most notably, many studies propose affective difficulty adjustment based on the flow model [12,13], which proposes the existence of an optimal experience (lying between dimensions of boredom and frustration) when challenge and skill of a game are balanced. As a consequence, many of the studies chose to adapt difficulty aspects of games (such as health, enemy behaviour, platform size, game speed, etc.) to achieve an optimal experience. But not only is flow a conceptually ambiguous construct in psychology [109], the precise relationship between skill, challenge, and flow is unknown [110]. Furthermore, it is hard to assess whether or not a given adaptation was successful, if fundamental and untested assumptions must be made (e.g. smaller platforms lead to challenge, which leads to frustration). Again, as all emotional reactions, elicitation has been found to be dependent on individual and contextual factors, both in perspectives that argue for basic, innate emotions [45], and in perspectives that argue for constructed emotions [46]. This is not only true for ambiguous concepts such as flow, but all emotions. Hernandez et al. [86] based their adaptation purely on the designer's ability to infer emotional states from made choices, which might have led to the observed lack of adaptation effects. Ibáñez et al. [87] assumed specific relationships between game elements and six basic emotions based on Ekman and Keltner [41] (e.g. light and flowers for joy, insects and slime for disgust) and used these assumptions to train emotion classifiers.

Again, the explicit (and direct) testing of eliciting material given a game context and player may be necessary to avoid unclear mappings between game materials and emotions. For example, Darzi et al. [81] tested the ability of different game characteristics to elicit the targeted emotional changes in a preceding test. Such a process could provide similar benefits as testing the relationship between emotions and measurement instruments. Moreover, if the relationship between the target emotional state and the game material is clear, adaptation can be based on very specific, pre-defined rules that are not based on potentially

Table 5
Complete list of studies included in the review with summaries of aims, methods, and conclusions.

Authors	Aims/Objectives	Methods	Results and conclusions
Akbar et al. [73]	To develop and evaluate a game balancing system based on facial expression recognition with the aim to enhance player experience.	Two groups of 30 participants (68% male) played either a 2D platformer or 3D shooter in two conditions: Affect-adaptive vs. non-adaptive. Participants answered a subsequent player experience questionnaire [66].	The adaptive versions of both genres showed significant improvements in experience domains such as immersion, flow, challenge, and positive affect. No differences were found negative affect and only for the 2D platformer for competence.
Al Osman et al. [74]	To prove practicality of ubiquitous biofeedback serious games by developing and evaluating a physiology-based stress management game.	Exp 1: 15 participants (60% male; mean age 33.47 years) played a biofeedback game with stressful/relaxing tasks to test if physiological stress corresponded to game adaptation. Exp 2: 12 participants (58% male; mean age 33.92) played an adaptive vs. non-adaptive version for 5 days each and answered a post-treatment questionnaire.	The game demonstrated a good reflection of physiological stress as presented in experiment 1. Experiment 2 showed that the adaptive version provided a better mental stress reduction over five days. Limitations are discussed in terms of generalizability.
Alves et al. [75]	To develop a mental state-adaptive FPS and evaluate the adaptation against performance-based adaptation in terms of enjoyment and scoring.	21 participants (76% male); age range 19–27; efforts made to validate affective measurement and emotion elicitation through the game before evaluation; participants played affect- and performance-adapted game with HR and EEG sensors and answered adapted GEQ [76] questions (5 items).	Performance-based adaptation resulted in significantly higher flow-experience scores and significantly better performance compared to affect-adaptation. Discussion states small sample size and lack of generalizability as possible reasons, as well as the limited number of predicted affective states.
Andrew and Chowanda [77]	To apply emotion-based difficulty adjustment based on facial recognition to a horror game to improve player satisfaction.	31 participants (unspecified demographics) played two versions of a survival horror game: One with difficulty adjustment based on facial expressions and one without. Evaluation was based on number of observed positive and low-arousal emotions.	The adaptive game provided significantly less observed positive valence-emotions and low-arousal emotions, which is argued to show a successful fear experience. Descriptive and qualitative data was provided to show good player satisfaction for the adaptive game.
Blom et al. [78]	To develop an online game difficulty personality system based on Facial Expression Analysis (FEA) and evaluate it within a popular platformer.	Exp 1: 38 participants (47% male); mean age 35.1; participants played through three versions of an Arcade game with FEA sensors to evaluate prediction of perceived difficulty; Exp 2: 10 (without head pose analysis) and 25 (with head pose analysis) (80% male); participants played a static and personalized version of a 2D platformer and rated their preference.	Perceived difficulty was measurable through FEA, which provided the possibility to create an heuristic online personalization system that was preferred by players when used in a 2D platformer, compared to a static game version. Similar results were found for a modelling approach that includes head pose analysis.
Bontchev and Vassileva [79]	To clarify how affect-based game adaptation can improve implicit recognition of playing styles and performance within a 3D puzzle game.	30 participants (60% male); mean age 31; participants played a 3D puzzle game with and without affective-adaptation controls in a randomized order and answered a post-game questionnaire indicating playing styles and adaptation enjoyment.	Recognition of playing styles yielded a good accuracy within a game combining affective and performance adaptation. The adaptive version of the game showed higher performance and good enjoyment ratings. Limitations are discussed in terms of generalizability.
Bontchev and Georgieva [80]	To propose and test a linear regression-based model to recognize player styles and test it within an affect-adaptive game.	Exp 1: 34 participants (53% male); mean age 26.85; participants played through an adaptive and non-adaptive VR puzzle game and answered playing style questionnaires [111,112]. Exp 2: 30 participants (60% male); mean age 31.87; same setup to validate playing-style recognition.	Playing style recognition through affect-related and gameplay data was achieved with an accuracy between 73% and 84% and adaptation based on affective data led to improvement in effectiveness, efficiency, and difficulty of a puzzle game. Combination of affective adaptation and playing style-adaptation is recommended.
Darzi et al. [81]	To compare five difficulty adjustment methods in a video game, including manual, random, performance-based, personality-performance-based and physiology-personality-performance-based.	50 participants (74% male; mean age 25.1) played one of five game versions which adapts difficulty: (a) manually, (b) randomly, (c) performance-based, (d) personality-performance-based, (e) physiology-personality-performance-based. Experience was measured through Intrinsic Motivation Inventory [67] and Flow Experience Measure [68].	Physiology-based affective adaptation did not lead to an improvement in game experience, compared to any other group. Physiology-based adaptation may show promising results in validation studies but do not guarantee user experience-improvements, even if all affective relationships are tested in a preceding open loop study.

(*continued on next page*)

Table 5 (*continued*).

Authors	Aims/Objectives	Methods	Results and conclusions
Ewing et al. [82]	To develop and validate a psychophysiological model between a player and a game and apply it to an affect-adaptive game.	Exp 1: 20 participants (45% male; age range 19–36) played Tetris with EEG sensors equipped, followed by subjective questionnaires; Exp 2: 10 participants (40% male) played 3 affect-model adaptive game versions and a manual-adaptive version for difficulty adjustment and answered affective and player experience questionnaires [69,83]	The presented 2-step process to associate physiological data to psychological construct resulted in valid predictions for cognitive demand and effort using EEG measures. The evaluated adapted game showed no improvement in most of the used experience measures. Results are discussed in their utility of a conceptual process model to develop theory-based affective games.
Frommel et al. [84]	To propose an approach of emotion-based difficulty adjustment using self-report measures and evaluate it empirically.	66 participants (73% male; mean age 30) played a 2D platformer with emotion-adaptive difficulty, increasing difficulty, and fixed difficulty. Differences are reported in terms of the Intrinsic Motivation Inventory [67] and the Player Experience Inventory [70].	The emotion-adaptive game shows increased player experience ratings compared to both control groups. Additionally, in-game dialogue-based subjective emotion measures showed a high accuracy. Limitations are discussed in terms of more possible comparison groups (such as performance-based).
Hernandez et al. [86]	To implement and evaluate an AI experience manager to keep players on a predefined emotion trajectory within a narrative video game.	Exp 1: 294 participants (50% male; mean age 19) played either a game managed by the PACE AI experience manager or by a random model and rated their experience scores. Exp 2: 39 participants (41% male; mean age 20); same setup as Exp 1, but with a preceding calibrating task [71].	Experiment 1 showed no statistical difference between groups, possibly based on a missing calibration as a form of reference. As a consequence, Experiment 2 introduced a calibration task, but again there were no significant differences observed, leading to inconclusive results.
Ibáñez et al. [87]	To test if gestures can be used to recognize emotional states and adapt music to these states using a VR game.	22 participants (67% male); mean age 29.09; participants; participants played either an adaptive (gesture-based fear recognition to change music) or non-adaptive VR game and rated presence via the SUS [88].	Head gesture was found sufficient to detect fear, but no other emotional state. Fear-adaptive music in a VR world was shown to increase perceived presence of players compared to a control game. Limitations are reported in terms of the system's responsiveness.
Jalbert and Rank [89]	To assess the usefulness of physiological data to increase rapport with NPCs in an action RPG.	16 participants (63% male; age range 18–34) were assigned to either a adaptive or non-adaptive game version utilizing EMG and EDA data to change NPC behaviour. Rapport was measured with 3 items on a 9-point likert scale.	Evaluation showed no difference in rapport ratings between adaptive and control group, but qualitative questions indicate enjoyment of the adaptation.
Lara-Alvarez et al. [90]	To propose and evaluate an educational game with affective induction through a fuzzy system analysing performance and emotional states.	40 participants (unspecified demographics) played both an educational game with linear and with affective difficulty adjustment based on voice recordings. In-game performance and emotional reactions are used as evaluation.	A previously tested emotion classified showed medium to high accuracy for valence and arousal dimensions. The adaptive game showed significant improvements in experience of pleasant-high states and reduction in unpleasant-low states. Adapting both difficulty and aesthetics was considered a promising approach.
Liu et al. [91]	To design and implement an affect-based difficulty adjustment system based on anxiety measures and evaluate its effect.	Exp 1: 15 participants (47% male; age range 18–34) played six sessions of Pacman over two months while physiological data (HR, EMG, EDA) and subjective reports of anxiety were assessed to create an emotion model. Exp 2: 9 participants (unspecified demographics) played both a performance-based and anxiety-based adapted game and answered questions about their anxiety, enjoyment, challenge, and perceived performance.	Anxiety was accurately predicted (88%) with the created emotion model (Regression Tree) through a combination of physiological measures. Significant improvements for enjoyment, challenge, and perceived performance was reported for the affect-adaptive game compared to the performance-adaptive, with no significant difference for reported anxiety.
Moniaga et al. [92]	To test is facial expression recognition can be used to dynamically balance a game and enhance the experience.	32 participants (unspecified demographics) played both a facial expression-adaptive and non-adaptive game and answered the Immersive Experience Questionnaire (20 items; Jennett et al. [69])	After a initial survey, a simple Hack and Slash game was designed with dynamic balancing based on facial expression recognition. The adaptive version showed improvements for the challenge and player experience domain in the follow-up questionnaire.

(*continued on next page*)

Table 5 (*continued*).

Authors	Aims/Objectives	Methods	Results and conclusions
Negini et al. [93]	To create and evaluate an affective game engine to test how player abilities, enemy design, and environment influences performance and effect.	16 participants (34% male; age range 18–32) played through four game conditions (control, player adapted, NPC adapted, environment adapted) with EDA-based adaptation. Dependent measures included skin conductance response, game performance, and player experience [14,67].	Results show that the adapted versions of the game were more physiologically arousing, indicating successful arousal-adaptation. Results on player experience scales reveal no effect of adaptation. NPC-based adaptation was reported to be especially ineffective as enjoyment-reduction was observed. Limitations are discussed in terms of generalizability.
Nogueira et al. [94]	To develop and test a procedural horror game that adapts to affective physiological states.	24 participants (67% male; mean age 22.5) tested three versions of a horror game with physiological measures (EDA, EMG, HR, HRV): A symbiotic adaptive version (in-game character mirrors player affects), a affective difficulty adjustment, and a non-adaptive version. Participants then answered the Game Experience Questionnaire [66].	The adaptive game versions showed improved ratings on the domains of immersion, tension, positive affect, and negative affect compared to the non-adaptive game. It was also reported that game adaptation was successful in shifting player experiences. Further analyses, including qualitative data, provides evidence for the interindividual differences in emotional experiences and elicitation effects of emotion-adaptive materials.
Parnandi et al. [95]	To develop and evaluate an adaptive biofeedback game that teaches relaxation skills by monitoring players' breathing rates.	9 participants (78% male; age range 22–33) performed a Stroop colour test, and then played either a biofeedback relaxation game, performed deep breathing, or played a traditional game. Physiological data was assessed through HRV and EDA during a follow-up stress-inducing task.	The adaptive game was reported to show good skill transfer in terms of relaxation training and showed significant improvement in terms of physiological arousal compared to the other groups. The main benefit of the adaptive game is reported to be the ability to create stressful situations while training relaxation skills.
Parnandi and Gutierrez-Osuna [96]	To present an evaluate an adaptive biofeedback game for teaching self-regulation of stress.	25 participants (60% male; age range 19–33) tested an emotion-adaptive game using three modalities (EDA, HRV, breathing rate) against a deep breathing treatment and a non-adaptive game after a baseline breathing phase. Performance and physiological data was assessed in a follow-up stress-inducing task.	There were mixed results reported, indicating positive but non-significant improvement in breathing rate and performance for the adapted game versions versus the control, and significant physiological arousal improvement for the breathing rate-adaptive game versus the control game. Results are discussed in terms of the potential of games to manipulate arousal-inducing material to train relaxation skills, but more studies seem necessary.
Rodriguez-Guerrero et al. [97]	To present a biocooperative game control architecture for haptic assistance and difficulty adaptation through physiological affective data.	Exp 1: 6 participants (83% male; mean age 30.5) played a VR rehabilitation game while their physiology (HR, Skin temperature, EDA) and subjective experience were measured. Exp 2: 11 participants (73% male; mean age 30.5) played an affect-adaptive game and physiological, subjective, and performance data was compared to a previous study.	The preceding open-loop experiment showed generally poor correlations between subjective experience and physiological data, providing more evidence about their complex relationship. A mix of multiple physiological measures was used to adapt the game in Experiment 2, leading to improved but non-significant valence and dominance scores, compared to the control game. Mapping between subjective and objective data, as well as game data remains inconclusive.
Rosa et al. [98]	To compare three approaches for difficulty adjustment (affective, performance-based, combined) in their ability to promote flow and test game characteristics as a mean of successful adaptation.	Exp 1: 20 participants (age range 18–24) played a 2D platformer on different difficulty levels and rated the perceived difficulty. Exp 2: 36 participants (61% men; age range 18–25) tested a affective (EDA-based), performance-based, combined, and control version of a 2D platformer while their performance and preference was measured. Exp 3: 155 participants (81% male; age range 15–65) tested the same game without adjustment, with adjustment through platform size, through jump height, and a combined version and then rated their experience.	Experiment 1 provided insights about what difficulty adjustments were successful in a 2D platformer, leading to the manipulation of platform size and jump height, which was tested in Experiment 3 against no modification with no effect on player experience. Adaptation test showed improvement in performance for the difficulty adjustment models, compared to the control, but no effect for player experience ratings.

(*continued on next page*)

Table 5 (*continued*).

Authors	Aims/Objectives	Methods	Results and conclusions
Salah et al. [99]	To develop educational games with affect-adaptive difficulty and interfaces and evaluate its effects in terms of learning gain and player engagement against a non-adaptive version.	30 participants (67% male); mean age 19 years; participants played either subjective feeling-adaptive and non-adaptive game with pre-test and post-test learning and engagement questionnaires.	Adaptive-game group showed a significantly higher learning increase, as well as higher engagement. Adaptive version showed balance between skill and challenge and adaptation in time limit and background music showed most promise.
Stein et al. [100]	To evaluate the practicality of dynamic difficulty adjustment through EEG-measured excitement in a 3D shooter.	Exp 1: 8 participants (87% male; age range 22–28) played a 3D shooter with EEG sensors attached to test game-physiology relationship; Exp 2: 24 participants (92% male; age range 20–29) played through 4 versions of the game (learning, EEG-adaptive, fixed-interval based on mean EEG trigger time, non-adaptive) and answered a subjective experience questionnaire.	The first experiment showed good correlations between game events and EEG data. The second experiment showed higher excitement with the EEG-adaptive games compared to the other groups. The fixed-interval version was rated higher than the non-adaptive version in terms of enjoyment.
Tjokrosetio and Chowanda [101]	To improve NPC believability using facial expression recognition and to provide an empirical validation of the results.	52 participants (86% male; mean age 25.59) watched gameplay videos of an adapted vs. non-adapted NPC character in a 3D action RPG and evaluated the adaptation via the Character Believability Questionnaire and an interview.	The adapted version showed a higher emotional range compared to the non-adaptive version and scored higher in all character believability domains, except predictability.
Vachiratamporn et al. [102]	To develop an affective survival horror game based on a previously developed state prediction model and evaluate its effects against a non-adaptive version.	12 participants (92% male); age range 22–36; pre-validated emotion classification; participants played adaptive and non-adapted game with HR and facial expression sensors and answered 5-point scale items regarding fear, fun, and difficulty.	Non-adaptive version showed higher ratings compared to adaptive version in fun, fear, and difficulty, although these differences remained non-significant. Results are discussed in terms of limited generalizability and possible problematic elicitation methods.

contentious assumptions and address concerns of interindividual and intraindividual differences in emotion processing.

Overall, the nature of emotions given modern theories, including emotion component expressions, and the implications of theoretical perspectives are not thoroughly addressed in almost all studies, leading to potential theoretical uncertainties, influencing the observed results.

5.3. Limitations

Overall, the analysed studies differ in many ways, which makes a clear analysis of the effect of emotion-adaptation not yet possible. In fact, the differences in methodological approaches, lack of effort to ensure generalizability, and lack of effort to reduce risk of bias add to the already present problems of comparability. Meta-analytical strategies, which are seen as one of the best way to aggregate scientific knowledge [113], are difficult to conduct, not only because of differences in approaches and theoretical perspectives (such as outcome variables of interest or emotion models), but also because of differences in methodologies that should be universally prevalent, such as shared and precise PX conceptualizations, appropriate measures, well-constructed and powered experiments, and the sufficient reporting of data. This study therefore was not able to conduct a meta-analysis and limits itself, as a consequence to the data at hand, to descriptive evaluations of some of the theoretical and methodological inconsistencies and trends in reported effects.

6. Conclusion

This review provides aggregated evidence regarding the effects, evaluation methods, and theoretical assumptions of affect-adaptive video games. Not only were mixed effects observed in the investigated studies, a large variance in methodological approach and theoretical justifications was observed, leading to many open questions regarding affective games. This systematic review adds to the body of evidence uncovering gaps in research and practice when it comes to games that adapt to player emotions.

Many of the described studies describe their main contribution as the development and exploration of technological solutions regarding emotion recognition and adaptation and not in the evaluation of affect-adaptive games. This review specifically analysed the evaluations in terms of emotion-theoretical assumptions, methodologies, and findings. From such a perspective, it is clear that more work is required to draw certain conclusions regarding the three main aspects of affective gaming as defined by Hudlicka [6], i.e. emotion sensing, modelling, and adaptation. It may not be the case that technological barriers limit the amount of conclusive data in the field, but theoretical and methodological barriers. The research standards regarding adaptation evaluation shared between studies are limited, especially in regards to generalizability. Ambiguous constructs, measured through instruments with unknown reliability are often used as outcome variables to evaluate adaptive games and there is a large variance between studies that makes it apparent that the meaning of these constructs (especially relating to player experience) is lacking consensus. Similarly, emotion-theoretical details are insufficiently integrated into the research process, leading to potentially erroneous practices in regards to applying emotion theories. The strongest support for the potential of affect-adaptive games in enhancing player experiences, performance, or health lies in studies that specifically test their affective assumptions in terms of: (a) measures of affective data and their relationship to the target emotion; and (b) adapted game materials used to elicit emotions and their relationship to the target emotion. Following such a process gives game designers and researchers the opportunity to gather more information, address concerns regarding influences of individual differences and context on emotional reactions, and avoid making assumptions based on contentious emotion-theoretical perspectives. As many of the described studies show, affect adaptation can be considered as promising if the design and evaluation process is robust. Overall, the currently available body of studies suffer however from theoretical and methodological inconsistencies and a lack of applied research standards. Future studies in the field need to tackle these problems by applying rigorous methods to empirically test their effects and any debated theoretical assumption they make (for example how well the game design choices elicit the target emotion and how well the measures can assess it). As it stands, the current body of evidence cannot be used to draw fixed conclusions about the effects of affect-adaptive video games, but should rather be used to guide and motivate

future research that could bring us closer to the proposed benefits discussed in the successful studies.

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.

Data availability

All data is available through the used databases and within the analyzed research articles.

Acknowledgements

This work was supported through the EPSRC Centre for Doctoral Training in Intelligent Games and Game Intelligence (IGGI) of the Engineering and Physical Sciences Research Council (EP/S022325/1).

Appendix

See Table 5.

References

[1] T. Wijman, Three Billion Players by 2023: Engagement and Revenues Continue to Thrive Across the Global Games Market, Newzoo, 2020, AccessedJune.
[2] M.D. Griffiths, D.J. Kuss, A.B.O. de Gortari, Videogames as therapy: an updated selective review of the medical and psychological literature, Int. J. Priv. Health Inf. Manag. (IJPHIM) 5 (2) (2017) 71–96.
[3] R. Pine, T. Fleming, S. McCallum, K. Sutcliffe, The effects of casual videogames on anxiety, depression, stress, and low mood: A systematic review, Games Health J. (2020).
[4] E.A. Holmes, E.L. James, E.J. Kilford, C. Deeprose, Key steps in developing a cognitive vaccine against traumatic flashbacks: Visuospatial Tetris versus verbal Pub Quiz, PLoS One 5 (11) (2010) e13706.
[5] L. Iyadurai, S.E. Blackwell, R. Meiser-Stedman, P.C. Watson, M.B. Bonsall, J.R. Geddes, A.C. Nobre, E.A. Holmes, Preventing intrusive memories after trauma via a brief intervention involving Tetris computer game play in the emergency department: a proof-of-concept randomized controlled trial, Mol. Psychiatry 23 (3) (2018) 674–682.
[6] E. Hudlicka, Affective game engines: motivation and requirements, in: Proceedings of the 4th International Conference on Foundations of Digital Games, 2009, pp. 299–306.
[7] G.N. Yannakakis, J. Togelius, Experience-driven procedural content generation, IEEE Trans. Affect. Comput. 2 (3) (2011) 147–161.
[8] K.M. Gilleade, A. Dix, Using frustration in the design of adaptive videogames, in: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2004, pp. 228–232.
[9] G.N. Yannakakis, H.P. Martínez, A. Jhala, Towards affective camera control in games, User Model. User-Adapt. Interact. 20 (4) (2010) 313–340.
[10] R. Hunicke, The case for dynamic difficulty adjustment in games, in: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2005, pp. 429–433.
[11] R.W. Picard, Affective Computing, MIT Press, 2000.
[12] M. Csikszentmihalyi, I.S. Csikszentmihalyi, Optimal Experience: Psychological Studies of Flow in Consciousness, Cambridge University Press, 1992.
[13] J. Chen, Flow in games (and everything else), Commun. ACM 50 (4) (2007) 31–34.
[14] R.M. Ryan, C.S. Rigby, A. Przybylski, The motivational pull of video games: A self-determination theory approach, Motiv. Emot. 30 (4) (2006) 344–360.
[15] E.L. Deci, R.M. Ryan, Motivation, personality, and development within embedded social contexts: An overview of self-determination theory, 2012.
[16] R. Lopes, R. Bidarra, Adaptivity challenges in games and simulations: a survey, IEEE Trans. Comput. Intell. AI Games 3 (2) (2011) 85–99.
[17] B. Bontchev, Adaptation in affective video games: A literature review, Cybern. Inf. Technol. 16 (3) (2016) 3–34.
[18] C. Schrader, J. Brich, J. Frommel, V. Riemer, K. Rogers, Rising to the challenge: An emotion-driven approach toward adaptive serious games, in: Serious Games and Edutainment Applications, Springer, 2017, pp. 3–28.
[19] R. Robinson, K. Wiley, A. Rezaeivahdati, M. Klarkowski, R.L. Mandryk, " Let's get physiological, physiological!" a systematic review of affective gaming, in: Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 2020, pp. 132–147.
[20] K.R. Scherer, Theory convergence in emotion science is timely and realistic, Cogn. Emot. 36 (2) (2022) 154–170.
[21] N.H. Frijda, Moods, emotion episodes, and emotions, 1993.
[22] N.H. Frijda, Varieties of affect: Emotions and episodes, moods, and sentiments, 1994.
[23] D. Zillmann, Theory of affective dynamics: Emotions and moods, 2003.
[24] K.R. Scherer, et al., Psychological models of emotion, in: The Neuropsychology of Emotion, Vol. 137, 2000, pp. 137–162.
[25] C.E. Izard, The many meanings/aspects of emotion: Definitions, functions, activation, and regulation, Emot. Rev. 2 (4) (2010) 363–370.
[26] W.M. Wundt, Principles of Physiological Psychology, Vol. 1, Sonnenschein, 1904.
[27] J.A. Russell, A circumplex model of affect, J. Personal. Soc. Psychol. 39 (6) (1980) 1161.
[28] D. Watson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: the PANAS scales, J. Personal. Soc. Psychol. 54 (6) (1988) 1063.
[29] M.M. Bradley, P.J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1) (1994) 49–59.
[30] R. Plutchik, The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, Am. Sci. 89 (4) (2001) 344–350.
[31] S.S. Tomkins, Affect theory, in: Approaches To Emotion, Vol. 163, Hillsdale, NJ, 1984.
[32] C.E. Izard, Basic emotions, relations among emotions, and emotion-cognition relations, 1992.
[33] J.T. Cacioppo, G.G. Berntson, J.T. Larsen, K.M. Poehlmann, T.A. Ito, et al., The psychophysiology of emotion, in: Handbook of Emotions, Vol. 2, 2000, pp. 173–191.
[34] I.B. Mauss, M.D. Robinson, Measures of emotion: A review, Cogn. Emot. 23 (2) (2009) 209–237.
[35] E.L. Rosenberg, P. Ekman, Coherence between expressive and experiential systems in emotion, Cogn. Emot. 8 (3) (1994) 201–229.
[36] J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev. Psychopathol. 17 (3) (2005) 715.
[37] B. Chakrabarti, E. Bullmore, S. Baron-Cohen, Empathizing with basic emotions: common and discrete neural substrates, Soc. Neurosci. 1 (3–4) (2006) 364–384.
[38] E. Harmon-Jones, C. Harmon-Jones, E. Summerell, On the importance of both dimensional and discrete models of emotion, Behav. Sci. 7 (4) (2017) 66.
[39] S. Hamann, Mapping discrete and dimensional emotions onto the brain: controversies and consensus, Trends in Cognitive Sciences 16 (9) (2012) 458–466.
[40] P. Ekman, What emotion categories or dimensions can observers judge from facial behavior? in: Emotions in the Human Face, Cambridge University Press, 1982, pp. 39–55.
[41] P. Ekman, D. Keltner, Universal facial expressions of emotion, in: P. Segerstrale U, P. Molnar (Eds.), Nonverbal Communication: Where Nature Meets Culture, 1997, pp. 27–46.
[42] W. Sato, S. Hyniewska, K. Minemoto, S. Yoshikawa, Facial expressions of basic emotions in Japanese laypeople, Front. Psychol. 10 (2019) 259.
[43] J. LeDoux, The Emotional Brain: The Mysterious Underpinnings of Emotional Life, Simon and Schuster, 1998.
[44] C.E. Izard, Emotion theory and research: Highlights, unanswered questions, and emerging issues, Annu. Rev. Psychol. 60 (2009) 1–25.
[45] D. Keltner, J.L. Tracy, D. Sauter, A. Cowen, What basic emotion theory really says for the twenty-first century study of emotion, J. Nonverbal Behav. 43 (2) (2019) 195–201.
[46] L.F. Barrett, How Emotions are Made: The Secret Life of the Brain, Houghton Mifflin Harcourt, 2017.
[47] K.A. Lindquist, T.D. Wager, H. Kober, E. Bliss-Moreau, L.F. Barrett, The brain basis of emotion: a meta-analytic review, Behav. Brain Sci. 35 (3) (2012) 121.
[48] J.E. LeDoux, Emotion circuits in the brain, Annu. Rev. Neurosci. 23 (1) (2000) 155–184.
[49] L.F. Barrett, Emotions are real, Emotion 12 (3) (2012) 413.
[50] E. Hudlicka, Affective computing for game design, in: Proceedings of the 4th Intl. North American Conference on Intelligent Games and Simulation, McGill University Montreal, Canada, 2008, pp. 5–12.
[51] E. Lux, M.T.P. Adam, V. Dorner, S. Helming, M.T. Knierim, C. Weinhardt, Live biofeedback as a user interface design element: A review of the literature, Commun. Assoc. Inf. Syst. 43 (1) (2018) 18.
[52] Y.Y. Ng, C.W. Khong, A review of affective user-centered design for video games, in: 2014 3rd International Conference on User Science and Engineering (I-User), 2014, pp. 79–84.
[53] A. Dzedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods, Sensors 20 (3) (2020) 592.
[54] S. Hamdy, D. King, Affect and believability in game characters–a review of the use of affective computing in games, in: Proceedings of the 18th Annual Conference on Simulation and AI in Computer Games. EUROSIS, 2017.

[55] E. Hudlicka, Guidelines for designing computational models of emotions, Int. J. Synth. Emot. (IJSE) 2 (1) (2011) 26–79.

[56] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, Inf. Fusion (2022).

[57] G.N. Yannakakis, A. Paiva, Emotion in games, in: Handbook on Affective Computing, Vol. 2014, Oxford University Press, 2014, pp. 459–471.

[58] K.M. Gilleade, A. Dix, Using frustration in the design of adaptive videogames, in: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2004, pp. 228–232.

[59] A. Picardi, P. Burelli, G.N. Yannakakis, Modelling virtual camera behaviour through player gaze, in: Proceedings of the 6th International Conference on Foundations of Digital Games, 2011, pp. 107–114.

[60] P. Sundström, Exploring the Affective Loop (Ph.D. thesis), 2005.

[61] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P. Group*, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Int. Med. 151 (4) (2009) 264–269.

[62] OSF, Affective theories, methodology, and evaluation of emotion-adaptive games: A systematic review, 2022, URL: https://osf.io/qep2u.

[63] C.O. Fritz, P.E. Morris, J.J. Richler, Effect size estimates: current use, calculations, and interpretation, J. Exp. Psychol. [Gen.] 141 (1) (2012) 2.

[64] T. Baguley, Understanding statistical power in the context of applied research, Applied Ergon. 35 (2) (2004) 73–80.

[65] J.P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M.J. Page, V.A. Welch, Cochrane Handbook for Systematic Reviews of Interventions, John Wiley & Sons, 2019.

[66] W.A. IJsselsteijn, Y.A. De Kort, K. Poels, The game experience questionnaire, 2013.

[67] D. Markland, L. Hardy, On the factorial and construct validity of the Intrinsic Motivation Inventory: Conceptual and operational concerns, Res. Q. Exerc. Sport 68 (1) (1997) 20–32.

[68] H.-Y. Sung, G.-J. Hwang, Y.-F. Yen, Development of a contextual decision-making game for improving students' learning performance in a health education course, Comput. Educ. 82 (2015) 179–190.

[69] C. Jennett, A.L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, A. Walton, Measuring and defining the experience of immersion in games, Int. J. Hum.-Comput. Stud. 66 (9) (2008) 641–661.

[70] V. Vanden Abeele, L.E. Nacke, E.D. Mekler, D. Johnson, Design and preliminary validation of the player experience inventory, in: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, 2016, pp. 335–341.

[71] I.E. Vermeulen, C. Roth, P. Vorderer, C. Klimmt, Measuring user responses to interactive stories: Towards a standardized assessment tool, in: Joint International Conference on Interactive Digital Storytelling, Springer, 2010, pp. 38–43.

[72] P. Gomes, A. Paiva, C. Martinho, A. Jhala, Metrics for character believability in interactive narrative, in: International Conference on Interactive Digital Storytelling, Springer, 2013, pp. 223–228.

[73] M.T. Akbar, M.N. Ilmi, I.V. Rumayar, J. Moniaga, T.-K. Chen, A. Chowanda, Enhancing game experience with facial expression recognition as dynamic balancing, Procedia Comput. Sci. 157 (2019) 388–395.

[74] H. Al Osman, H. Dong, A. El Saddik, Ubiquitous biofeedback serious game for stress management, IEEE Access 4 (2016) 1274–1286.

[75] T. Alves, S. Gama, F.S. Melo, Flow adaptation in serious games for health, in: 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH), IEEE, 2018, pp. 1–8.

[76] J.H. Brockmyer, C.M. Fox, K.A. Curtiss, E. McBroom, K.M. Burkhart, J.N. Pidruzny, The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing, J. Exp. Soc. Psychol. 45 (4) (2009) 624–634.

[77] A.N.T. Andrew, A. Chowanda, Dynamic difficulty adjustment with facial expression recognition for improving player satisfaction in a survival horror game, ICIC Express Lett. 14 (2020).

[78] P.M. Blom, S. Bakkes, P. Spronck, Modeling and adjusting in-game difficulty based on facial expression analysis, Entertain. Comput. 31 (2019) 100307.

[79] B. Bontchev, D. Vassileva, Affect-based adaptation of an applied video game for educational purposes, in: Interactive Technology and Smart Education, Emerald Publishing Limited, 2017.

[80] B. Bontchev, O. Georgieva, Playing style recognition through an adaptive video game, Comput. Hum. Behav. 82 (2018) 136–147.

[81] A. Darzi, S.M. McCrea, D. Novak, et al., User experience with dynamic difficulty adjustment methods for an affective exergame: Comparative laboratory-based study, JMIR Serious Games 9 (2) (2021) e25771.

[82] K.C. Ewing, S.H. Fairclough, K. Gilleade, Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop, Front. Hum. Neurosci. 10 (2016) 223.

[83] G. Matthews, D.M. Jones, A.G. Chamberlain, Refining the measurement of mood: The UWIST mood adjective checklist, Br. J. Psychol. 81 (1) (1990) 17–42.

[84] J. Frommel, F. Fischbach, K. Rogers, M. Weber, Emotion-based dynamic difficulty adjustment using parameterized difficulty and self-reports of emotion, in: Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, 2018, pp. 163–171.

[85] E. McAuley, T. Duncan, V.V. Tammen, Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis, Res. Q. Exerc. Sport 60 (1) (1989) 48–58.

[86] S.P. Hernandez, V. Bulitko, M. Spetch, Keeping the player on an emotional trajectory in interactive storytelling, in: Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference, 2015.

[87] M.L. Ibáñez, M. Miranda, N. Alvarez, F. Peinado, Using gestural emotions recognised through a neural network as input for an adaptive music system in virtual reality, Entertain. Comput. 38 (2021) 100404.

[88] M. Slater, M. Usoh, A. Steed, Depth of presence in virtual environments, in: Presence: Teleoperators & Virtual Environments, Vol. 3, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 1994, pp. 130–144.

[89] J. Jalbert, S. Rank, Exit 53: physiological data for improving non-player character interaction, in: International Conference on Interactive Digital Storytelling, Springer, 2016, pp. 25–36.

[90] C. Lara-Alvarez, H. Mitre-Hernandez, J.J. Flores, H. Pérez-Espinosa, Induction of emotional states in educational video games through a fuzzy control system, IEEE Trans. Affect. Comput. 12 (1) (2018) 66–77.

[91] C. Liu, P. Agrawal, N. Sarkar, S. Chen, Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback, Int. J. Hum.-Comput. Interact. 25 (6) (2009) 506–529.

[92] J.V. Moniaga, A. Chowanda, A. Prima, M.D.T. Rizqi, et al., Facial expression recognition as dynamic game balancing system, Procedia Comput. Sci. 135 (2018) 361–368.

[93] F. Negini, R.L. Mandryk, K.G. Stanley, Using affective state to adapt characters, NPCs, and the environment in a first-person shooter game, in: 2014 IEEE Games Media Entertainment, IEEE, 2014, pp. 1–8.

[94] P.A. Nogueira, V. Torres, R. Rodrigues, E. Oliveira, L.E. Nacke, Vanishing scares: biofeedback modulation of affective player experiences in a procedural horror game, J. Multimodal User Interfaces 10 (1) (2016) 31–62.

[95] A. Parnandi, B. Ahmed, E. Shipp, R. Gutierrez-Osuna, Chill-Out: Relaxation training through respiratory biofeedback in a mobile casual game, in: International Conference on Mobile Computing, Applications, and Services, Springer, 2013, pp. 252–260.

[96] A. Parnandi, R. Gutierrez-Osuna, Physiological modalities for relaxation skill transfer in biofeedback games, IEEE J. Biomed. Health Inf. 21 (2) (2015) 361–371.

[97] C. Rodriguez-Guerrero, K. Knaepen, J.C. Fraile-Marinero, J. Perez-Turiel, V. Gonzalez-de Garibay, D. Lefeber, Improving challenge/skill ratio in a multimodal interface by simultaneously adapting game difficulty and haptic assistance through psychophysiological and performance feedback, Front. Neurosci. 11 (2017) 242.

[98] M.P. Rosa, E.A.d. Santos, I.L. de Moraes, M.M. Sarmet, C.D. Castanho, R.P. Jacobi, et al., Dynamic difficulty adjustment using performance and affective data in a platform game, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 367–386.

[99] J. Salah, Y. Abdelrahman, A. Dakrouni, S. Abdennadher, Judged by the cover: Investigating the effect of adaptive game interface on the learning experience, in: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, 2018, pp. 215–225.

[100] A. Stein, Y. Yotam, R. Puzis, G. Shani, M. Taieb-Maimon, EEG-triggered dynamic difficulty adjustment for multiplayer games, Entertain. Comput. 25 (2018) 14–25.

[101] A.N. Tjokrosetio, A. Chowanda, Character believability enhancement using facial expression recognition to improve the players immersive experience, ICIC Express Lett. 15 (2021) 1235–1242.

[102] V. Vachiratamporn, K. Moriyama, K.-i. Fukui, M. Numao, An implementation of affective adaptation in survival horror games, in: 2014 IEEE Conference on Computational Intelligence and Games, IEEE, 2014, pp. 1–8.

[103] A. Denisova, A.I. Nordin, P. Cairns, The convergence of player experience questionnaires, in: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play, 2016, pp. 33–37.

[104] D. Johnson, M.J. Gardner, R. Perry, Validation of two game experience scales: the player experience of need satisfaction (PENS) and game experience questionnaire (GEQ), Int. J. Hum.-Comput. Stud. 118 (2018) 38–46.

[105] L. Michailidis, E. Balaguer-Ballester, X. He, Flow and immersion in video games: The aftermath of a conceptual challenge, Front. Psychol. 9 (2018) 1682.

[106] S.E. Maxwell, K. Kelley, J.R. Rausch, Sample size planning for statistical power and accuracy in parameter estimation, Annu. Rev. Psychol. 59 (2008) 537–563.

[107] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126.

[108] M. Gendron, D. Roberson, J.M. van der Vyver, L.F. Barrett, Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture, Emotion 14 (2) (2014) 251.

[109] K. Jalife, C. Harteveld, C. Holmgård, From flow to fuse: A cognitive perspective, Proc. ACM Hum.-Comput. Interact. 5 (CHI PLAY) (2021) 1–30.
[110] C.J. Fong, D.J. Zaleski, J.K. Leach, The challenge–skill balance and antecedents of flow: A meta-analytic investigation, J. Posit. Psychol. 10 (5) (2015) 425–446.
[111] P. Honey, A. Mumford, et al., The Manual of Learning Styles, Vol. 3, Peter Honey Maidenhead, 1992.
[112] A. Aleksieva-Petrova, M. Petrov, B. Bontchev, Game and learmer ontology model, in: Int. Scientific Conf. Computer Science'2011, 2011, pp. 1–2.
[113] M. Borenstein, L.V. Hedges, J.P. Higgins, H.R. Rothstein, Introduction to Meta-Analysis, John Wiley & Sons, 2021.