Induction and profiling of strong multi-componential emotions in virtual reality

Meuleman, Ben; Rudrauf, David

Induction and profiling of strong multi-componential emotions in virtual reality

Ben Meuleman, and David Rudrauf

Abstract—Psychological theories of emotion have often defined an emotion as simultaneous changes in several mental and bodily components. In addition, appraisal theories assume that an appraisal component elicits changes in the other emotion components (e.g., motivational, behavioural, experiential). Neither the componential definition of emotion nor appraisal theory have been systematically translated to paradigms for emotion induction, many of which rely on passive emotion induction without a clear theoretical framework. As a result, the observed emotions are often weak. This study explored the potential of virtual reality (VR) to evoke strong emotions in ecologically valid scenarios that fully engaged the mental and bodily components of the participant. Participants played several VR games and reported on their emotions. Multivariate analyses using hierarchical clustering and multilevel linear modelling showed that participants experienced intense, multi-componential emotions in VR. We identified joy and fear clusters of responses, each involving changes in appraisal, motivation, physiology, feeling, and regulation. Appraisal variables were found to be the most predictive for fear and joy intensities, compared to other emotion components, and were found to explain individual differences in VR scenarios, as predicted by appraisal theory. The results advocate upgraded methodologies for the induction and analysis of emotion processes.

Index Terms—Emotion, emotion elicitation, appraisal, virtual reality

1 INTRODUCTION

There is a broad consensus in psychology that an emotion episode involves changes in multiple mental and bodily components simultaneously [1], [2], [3], [4]. Commonly cited components of emotion include a cognitive, a motivational, a physiological, a behavioural, an experiential (feeling), and a regulation component (Figure 1). Their joint pattern of changes defines what is understood as an "emotional reaction", and it is believed that specific qualitative emotional states (e.g., joy, fear, anger) are characterized by a particular combination of changes in these components. A theory of emotion causation that has generally subscribed to the componential definition of emotion is appraisal theory [5], [6]. This theory proposes that emotional reactions to a given situation are driven by cognitive evaluations (i.e., appraisals) about the personal importance of that situation. That is, a person must appraise how a situation affects their personal goals, desires, or beliefs in order to react emotionally to it. From a componential view, appraisal theorists identify the cognitive component as the primary driver of changes in the other emotion components. Following a certain appraisal—or combination of appraisals—it is assumed that the organism will prioritize certain actions (e.g., escaping from an appraised threat) and invest an effort in those actions through a coordination of mental and bodily changes [7], [8], [9], [10]. Appraisal theorists have put forward concrete predictions about how appraisals can generate emotional episodes that are patterned across multiple components, including motivation [11], [12], [13], [14], [15], physiology [16], [17], [18], [19], expression (e.g., [19], [20], [21], [22]), and feeling (e.g., [9], [18], [23], [24]).

Given the explicit causal hypothesis underpinning appraisal theory, one would expect this to be a popular basis for paradigms that induce emotions for scientific study. Surprisingly, this turns out to not be the case. In fact, many conventional paradigms for emotion induction appear to be lacking in proper theoretical justification entirely. This raises the concern that such paradigms may fail to induce strong and/or multi-componential emotions altogether, preventing emotion processes to be studied accurately. Studies that did base induction on appraisal theory

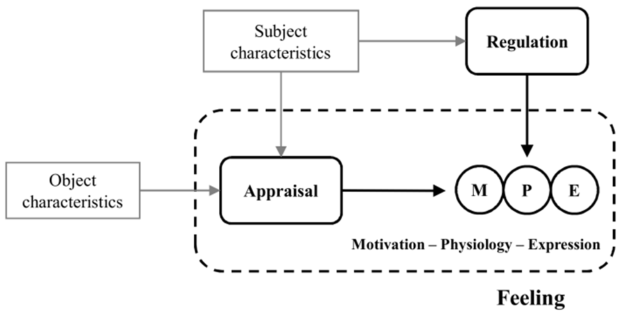

Figure 1. Emotion as a pattern of componential changes, triggered by appraisal, with feeling (dashed box) as an integrated awareness of other component changes. All boxes in black are considered as part of an emotion episode, with appraisal and regulation mediating the link between object/subject differences and emotional responding.

have been mostly conducted within the field of appraisal theory research itself and although these paradigms are faithful to the theoretical framework, in practice the paradigms suffer from other restrictions that also cast doubt on their ability to induce strong emotions. We sought to ad-

- Ben Meuleman is with the Swiss Center for Affective Sciences, Campus Biotech, Chemin des Mines 9, CH-1202 Genève, Switzerland. E-mail: ben.meuleman@unige.ch.
- David Rudrauf is with the Swiss Center for Affective Sciences, Campus Biotech, Chemin des Mines 9, CH-1202 Genève, Switzerland. E-mail: david.rudrauf@unige.ch.

dress these problems in the present study by combining appraisal theory with the medium of virtual reality (VR), showing empirically how VR can induce strong multi-componential emotions with theoretically justifiable scenarios that also allow the testing of hypotheses related to emotion theory. In the next sections, we first discuss the advantages and disadvantages of current emotion induction paradigms in more detail, and present virtual reality (VR) as a potential "best-of-all-worlds" solution to drawbacks in those paradigms. Following this, we outline the goals and hypotheses that we drew from appraisal theory and which informed our study and the use of VR. Finally, data are presented from a study in which participants experienced a number of immersive virtual reality scenarios. Integral to our approach was the use of integrated multivariate analyses, which allowed us to quantify emotional reactions in their intensity, their quality, and their pattern of componential changes.

1.1 Current Induction Paradigms

The study of emotion processes has involved the development of numerous paradigms for the induction of emotion. These paradigms can be broadly categorized according to whether they occur in the field or a laboratory, and whether the induction scenario is passive or active. In practice, passive scenarios (i.e., a person is passively confronted with an emotion-inducing stimulus) occur more often in the lab, whereas active scenarios (i.e., the person actively participates in the emotion-inducing scenario) occur more often in the field. By far the most popular paradigms for emotion induction involve passive emotion induction in the lab and include viewing pictures (e.g., International Affective Picture System (IAPS) database; [25]), viewing emotional facial expressions, watching film clips, listening to music, reading emotional words, recalling a past emotion experience, or imagining an emotion episode (see [26], [27], for reviews). These paradigms have a number of appealing advantages, such as experimental control of stimulus presentation, standardization of viewing and measurement conditions, and elaborate measurement of emotion (e.g., multiple wired devices simultaneously) with minimal risk of interference.

However, passive paradigms for laboratory induction of emotion (e.g., pictures, videos, music) are difficult to connect to the assumptions made by componential emotion definitions and appraisal theory. Rather than presenting the subject with a scenario that can be appraised according to personal relevance and that can be acted upon by a meaningful choice of action, the subject is confronted with an isolated stimulus whose "affective quality" is expected to transfer to the passive observer (e.g., a fearful face induces fear). The underlying causal model appears to be one of "contagion", but this is not a widely held theory of emotion causation. Moreover, the presented stimuli are assumed to possess intrinsic affective qualities, whereas appraisal theories posit that such qualities are irrelevant compared to how the stimulus is appraised. Finally, due to the passive nature of the task, the meaning of important emotion components is rendered ambiguous or inappropriate, such as motivations (e.g., action tendencies) and behaviours. It does not make sense to run away or avoid a picture of a spider on a lab computer, whereas that behaviour might certainly be observed in a real-life encounter with a spider. In other words, there is a lack of ecological validity to presenting a stimulus that has an assumed intrinsic emotional quality, but is otherwise disconnected from its meaning in other components of emotion.

Emotional reactions that are elicited with classic induction paradigms are typically weak, in terms of their felt intensity (e.g., subjective fear intensity), their duration, and their componentiality (e.g., involving changes in all emotion components simultaneously). This drawback need not be problematic when the purpose of the emotion induction ultimately serves another object of study (e.g., inducing an emotional state to examine its impact on some cognitive or social task) but when the object of study is the emotion process itself this aspect is critical. Firstly, according to some emotion theorists, a state should not be considered as "emotional" until changes in all emotion components have been established [28]. Secondly, some emotion theorists have put forward hypotheses that involve all components of emotion simultaneously, such as componential synchronization (e.g., [18], [29], [30], [31]) or experiential integration [3], [32], [33]. Testing these hypotheses is prohibited in emotion induction paradigms that do not engage adequately all mental and bodily subsystems simultaneously.

Outside of passive laboratory paradigms, there have been some attempts at using more ecologically valid induction tasks. In appraisal theory research, studies have experimentally manipulated appraisal criteria in vignettes or video games, and observationally collected emotion data in naturalistic settings (see [28], for a review of task formats). The advantage of such protocols is that they are based directly on the causal model of appraisal theory. Unfortunately, they face some practical restrictions. Experimental paradigms such as vignettes rely on self-report of an imagined rather than an experienced emotion episode. This lacks ecological validity and encourages the subject to access scripts and stereotypes when imagining and reporting on their emotion [34], [35]. Video games involve the subject actively into the emotion-eliciting scenario but still separate the subject from the scenario by a screen and a game character or avatar, making it ambiguous to what extent the player at any time is an observer or participant in the game. Sometimes an emotional response belongs directly to the subject (e.g., a subjective feeling) but sometimes it is expressed through the avatar (e.g., avoidance behaviour). Observational paradigms such as field studies (e.g., [36], [37], [38], [39]) boast ecological validity but often lack the appropriate scientific control for reliable induction of emotion, can be impractical, or limit the comprehensive measurement (e.g., physiological changes).

Ideally, a good paradigm for emotion induction would place the subject in a live scenario that can be manipulated according to their appraised concerns, that engages all mental and bodily subsystems meaningfully, and that can take place within the control of a standard lab environment. An appealing medium that fits to these requirements is virtual reality (VR).

1.2 Virtual Reality Paradigms

VR environments allow researchers the simulation of scenarios that would be difficult to operationalize in the real field for technical (simultaneous multimodal recording of bodily activity) and/or ethical concerns (safety), such as driving simulation (e.g., for studying anger and aggressive behaviour), flight simulation (e.g., for studying phobias), height simulation, and crowd simulation (e.g., for stress induction). Unlike classic emotion induction paradigms, VR actively engages the whole mind and body to respond to the ongoing challenges. Unlike video games, VR does not separate the subject from the virtual world through a computer-interface and a game character—he or she is fully part of and immersed into that world [40]. Due to the aforementioned advantages, VR scenarios are generally expected to have a relatively high ecological and construct validity when compared to other methods for emotion induction.

VR is increasingly applied in affective science to the study (and treatment) of pathological emotion such as phobia (e.g., [41], [42]). In such studies, subjects are typically exposed to a scenario appraised as threatening (e.g., public speaking) with the aim of reducing phobic responses or anxiety by repeated exposure [43], [44], [45], [46], [47], [48], [49]. Other psychological studies that have used VR have attempted to explore the limits of bodily agency and control (e.g., [50], [51], [52]), and the impact of norm violations in virtual social settings [53], [54], [55] , [56], [57]. These studies have provided compelling evidence for the experience of intense and multi-componential emotions. However, the emotional reactions that are being induced in VR are rarely the object of study in and of themselves. Either these reactions are peripheral to the true object of study, or—in the case of VR exposure therapy—the primary goal is in fact to reduce these responses via therapy. No study so far has taken advantage of VR and an explicit emotion theoretical framework to study emotion processes on their own.

1.3 Current Study

The goal of the present study was to explore the potential of VR for the induction of intense and multi-componential emotions. To achieve this, participants in the study experienced a number of commercially available VR games that were chosen according to specific criteria. Following each game, we assessed by self-report to what extent the participant had experienced changes in several emotion components, including appraisal, physiology, motivation, feeling, and regulation. For all components, we measured both their intensity (i.e., numerical strength) and their quality (i.e., items assessing qualitative aspects of that component).

Our hypotheses for this study were derived both from componential definitions of emotion (H1 and H2) and from appraisal theory (H3, H4, and H5). Thus we expected the following:

1. **H1**. Subjects would experience intense responses in more than one component of emotion simultaneously
2. **H2**. These componential responses would cluster into qualitatively differentiated patterns (e.g., a fear pattern, a joy pattern)
3. **H3**. Appraisal variables would explain relatively more variation in qualitative feeling intensities than variables of other components
4. **H4**. Appraisal variables would account for game differences in qualitative feeling intensities
5. **H5**. Appraisal and regulation variables would account for subject differences in qualitative feeling intensities

With respect to H1 and H2, these hypotheses addressed the “strength” of the observed emotional responses, which we defined as intense (subjectively), qualitatively distinct, and multi-componential in its pattern of mental/bodily changes. With respect to H3, due to the strong link that is posited between appraisal criteria and all other components of emotional responding by appraisal theory, it is expected that appraisal should account for most—if not all—variation in that responding [28]. Studies comparing the predictive power of different components are currently scarce, but a cross-cultural study investigating the componential meaning of 24 emotion terms [1] found that the appraisal component alone was capable of discriminating 70% of the 24 terms reported across 35 cultures [58]. For the present study, we expected likewise that the appraisal component would explain more variation in emotional responding than the other components. With respect to H4 and H5, it is expected that individual differences in emotional responding are largely mediated by appraisal and regulation variables (Figure 1). That is, two people each confronted with a different stimulus should nonetheless display the same pattern of emotional reactions, provided that they appraise the stimuli identically and regulate their reactions identically. Appraisal variables represent abstractions of the emotion-eliciting properties of stimuli, and are therefore more generalizing than any specific stimulus property (e.g., form, content), while regulation strategies represent directly how a person chooses to deal with an ongoing emotion, and are therefore more generalizing than specific subject characteristics (e.g., age, gender, personality). Although appraisal is expected to account for both stimulus and subject differences in emotional responding, it should primarily account for the former (H4). Regulation is expected to account exclusively for the latter (H5). For the current study, we wished to establish the generalizability of the emotions induced in VR , that is, as not being determined by idiosyncratic game or subject aspects but by appraisal and regulation.

TABLE 1
VR GAMES

Game name	Code	Description	Dominant emotions	Mean duration (min)
Tilt Brush	BRUSH	Paint three-dimensional doodles with various colourful brushes	Interest	5.7
TheBlu – Whale Encounter	WHALE	View underwater sea life and encounter a whale	Wonder, awe	2.8
Fruit Ninja VR	NINJA	Slash flying fruit with ninja swords to score points	Amusement, joy	5.4
The Lab – Longbow	LONGBOW	Shoot attacking enemies with a bow to defend your castle	Amusement, joy	8.4
The Brookhaven Experiment	ZOMBIES	Shoot attacking zombies to defend your life	Fear, disgust	4.9
Richie's Plank Experience	PLANK	Navigate a plank that suspended from a tall sky-scraper (i.e., virtual height exposure)	Fear, anxiety	2.0
Zero G	ZEROG	Navigate between two space stations in virtual zero gravity	Fear, anxiety	7.1

In order to test these hypotheses, we used a combination of unsupervised and supervised methods of data analysis. The unsupervised analyses were aimed at detecting and profiling multi-componential patterns of emotional responding, whereas the supervised analyses were aimed at testing explicitly the appraisal-related research hypotheses. For the latter, we used multilevel linear modelling with random effects accounting simultaneously for within-subject and within-stimulus correlation.

2 METHOD

2.1 Participants

In total, 53 subjects (27 women) participated in the study, aged between 18 and 41 years old (mean age 28.7). Thirty-three subjects were employees of the campus and participated voluntarily without remuneration. The remaining subjects were students recruited on the campus by word of mouth and received CHF 15 for participating. All subjects were healthy adults with no current or past affective disorders such as phobia. Twenty-seven subjects indicated that they had absolutely no prior experience with VR. The remaining subjects indicated to have a largely limited experience with VR (once or twice).

A session for one subject lasted between 40 to 60 minutes, with the subject playing between 1 and 6 games (on average 3.8 games). This resulted in a total of 202 observations. However, in order to obtain a data set that was balanced across the seven selected VR games, we subsetted the data prior to analysis such that each game counted 22 observations, chosen completely at random. This reduced the total number of observations to 154 but resulted in only one subject to be completely removed.

2.2 VR Games

Seven virtual reality games were selected from the software that is commercially available on Vive's Steam platform (see Table 1). These seven games were selected from a larger pool of games, according to four criteria:

1. **Emotional clarity**: Games had to elicit one—or two at the most—dominant emotions that were connected to the gameplay task. We excluded games with emotionally ambiguous content, or that elicited mixtures of many emotions, or that elicited complex social emotions.
2. **Duration**: Games could not exceed 10 minutes of gameplay. Although we were not able to balance our selection completely by duration, the average duration across all games was relatively short (5 min).
3. **Simple gameplay**: Games could not have complex game controls, cognitively demanding puzzles, physically demanding tasks (e.g., dancing), elaborate storytelling, or extended menus to navigate.
4. **Valence balance**: The selected games had to cover a reasonably balanced spectrum of positive and negative emotions to elicit. Of the seven games, one was somewhat neutral (Tilt Brush), three elicited positive emotions (Whale Encounter, Fruit Ninja, Longbow; interest, amusement joy) and three elicited negative emotions (The Brookhaven Experiment, Richie's Plank Experience, Zero G; anxiety, fear, disgust)

Using these four main criteria for selection resulted in the list of games presented in Table 1. These seven games covered a variety of different gameplay types and styles, including passive versus active, exploration-oriented versus action-oriented, and realistic versus cartoonish. Full information on the gameplay mechanics and the instructions for these games can be found in the Supplementary Material. Only one of these games, The Brookhaven Experiment, has been the subject of an earlier study on emotion [59].

2.3 Questionnaire

In order to measure emotional responses to the selected VR games, we developed a questionnaire consisting of three different psychometric scales, 1) a questionnaire for evaluating componential emotions with conventional Likert items, 2) a circumplex diagram for feeling self-report in qualitative categories (Geneva Emotion Wheel; [60]), and 3) a manikin for topographical self-report of feeling [61], [62].

The componential emotion questionnaire was adapted from the coreGRID questionnaire based on the GRID questionnaire for componential assessment of emotion terms [1]. Eighteen items were drawn directly from the coreGRID: 6 physiology items (e.g., "*To what extent... did your heartbeat get faster?*"), 6 motivation items (e.g., "*To what extent... did you want to explore the environment?*"), and 6 general feeling items e.g., "*To what extent... did you feel good?*"). To these items we added 14 appraisal items (some of which were adapted from the coreGRID; e.g., "*I thought that some events... were more pleasant than expected*"), and 4 regulation items (e.g., "*I tried to control negative emotions by... looking away from intense imagery/situations*"). The full list of items can be found in Table 2. All items were evaluated on a 7-point likert scale range from and asked to which a given statement was applicable, from 1 ("*Not at all*"), over 4 ("*Moderately*") to 7 ("*Very much*").

The circumplex diagram for assessing qualitative categories of feeling consisted of the Geneva Emotion Wheel ([60]; Appendix A, Figure S1]. In this wheel, 20 categories of feeling are arranged in a circle, with intensity bubbles (0 through 5) extending from the center of the wheel to the corresponding emotion category. In addition, the center of the wheel contains an option to check "none", when no particular emotion was felt, and "other", when an emotion was felt not included in the list of 20. When rating an emotion with the GEW, the subject is simply asked to check any categories that applied to the emotion episode and to indicate their felt intensity, with the square marker corresponding to 0 intensity.

The manikin for topographical self-report consisted of a line drawing of a featureless human body (no gender or

TABLE 2
QUESTIONNAIRE ITEMS FOR ASSESSING COMPONENTIAL EMOTIONS

Questionnaire item	Questionnaire item
Physiology: "To what extent..."	**Appraisal**:"I thought that some events..."
… did your heartbeat get faster?	… were pleasant
… did your breathing get faster?	… were unpleasant
… did you get weak in the knees?	… were unexpected
… did your stomach clench up?	… were more pleasant than I had expected
… did your muscles tense up?	… were more unpleasant than I had expected
… did you perspire or have moist hands?	… blocked objectives of the demo/game
	… advanced objectives of the demo/game
Motivation: "To what extent..."	… appeared to be dangerous
… did you want the demo to last or be repeated?	… appeared to require urgent action
… did you want to stop the demo?	… put me in control of the situation
… did you want to vent or curse?	… put the demo/game in control of the situation
… did you want to succeed at the demo?	… were unfair
… did you want to explore the environment?	… were morally inappropriate
… did you want to run away in whatever direction?	… appeared to be unrealistic
… did you want to get totally absorbed in the situation?	
	Regulation: "I tried to control my negative emotions by..."
Feeling: "To what extent..."	… reminding myself that events were virtual/not real
… did you feel good?	… looking away from intense imagery or situations
… did you feel energetic?	… working harder to achieve the demo/game's objectives
… did you feel in control?	… trying to see the humor in the situation
… did you feel immersed?	
… did you feel disoriented?	

race indicated; Appendix A, Figure S2), both from the front and from the back. When rating the location of their feeling with this manikin, the subject is asked to indicate where on the body they had felt their emotion most strongly. Positive feelings are marked by a blue pen, while negative feelings are marked by a red pen. This rating tool was allowed us to map in finer detail the feeling component of emotion, by not just investigating quality and intensity, but also its sensed location in the body (see [61], [62], for a similar approach). This was particularly relevant for some of the fear-eliciting games that were selected for the study (e.g., virtual height challenge), and which can be associated with strongly localized sensations in the legs, knees, or hands.

2.4 Procedure

Participants entered the VR room and were asked to read, agree to, and sign the informed consent form. Afterwards, the experimenter re-emphasized that the study could be stopped by the subject at any moment (e.g., when the subject felt dizzy, motion-sick, or otherwise found the experience emotionally upsetting). Next, the experimenter briefly interviewed the subject about their prior VR experience, and then proceeded to explain how to use the HTC Vive. The subjects were informed that they would be wearing a head-mounted display (HMD) that would provide stereoscopic visual input, and that there would be sound via headphones. Their movements in virtual reality would be tracked continuously with two optical trackers that were shown in opposite corners of each room. It was emphasized that these two trackers would not videotape the subject. When this was clear, the experimenter explained the HTC Vive controllers and their primary buttons.

When this was clear, the participant put on the HTC Vive HMD and adjusted the straps for comfort. They were given the two controllers and the headphones. Standard visual input during this setup consisted of the HTC Vive initial setup space (black grid with a generic mountain backdrop). The subjects were first asked to perform simple tasks such as looking around and inspecting the virtual controllers. Next, they were asked to move forward slowly until a green grid appeared to them. When this happened, the subjects were told to stop moving and it was explained that this grid functioned as the HTC Vive's warning boundary, cautioning to VR users when they reach the edge of the physical play space. The subject was warned to never cross this boundary, lest the trackers would lose the headset or the subject would hit a wall or object. When this was clear, the experimenter verified if they were ready to continue with the first game.

The first game that subjects played was always WHALE. This game was chosen as an introduction due to its relative passivity and simplicity, while nonetheless providing the player with a vivid first impression of virtual reality (the encountered whale). After the subject finished the WHALE game, the controllers were taken back, the headphones removed, and the HMD taken off. The subject was then presented with the emotion questionnaire and asked to rate the emotions they had experienced during the game. When finished, the subject proceeded to the next game, which was chosen randomly from the remaining game pool. This procedure—a game followed by a questionnaire—was repeated until the hour of the session was completed.

At the end of the session, the subject was informally interviewed about their experiences and debriefed about the study goals when inquired. Finally, if remunerated, the subject was paid CHF 15 for their participation and thanked.

2.5 Data Analysis

Data analysis consisted of three parts, 1) descriptive analysis of GEW ratings and feeling items, 2) unsupervised detection of componential emotion patterns among all self-report items, 3) supervised modelling of qualitative GEW feelings with the other componential items as predictors.

For the descriptive analysis, we looked at ratings of the GEW and the five feeling items in order to assess, in general, the intensity of emotion that was induced by VR, as well as the level of immersion. For the unsupervised data analysis, the full data set of self-report items was submitted to a hierarchical clustering analysis, both row- and column-wise. For the column-wise clustering, we sought to detect similarities between self-report items across the participant experiences (e.g., does fear intensity cluster together with physiological arousal items?). This was a general step to evaluate whether there existed clusters containing emotion responses from more than one emotion component simultaneously, rather than clusters containing only responses of one component at a time (e.g., a physiology cluster). For the row-wise clustering, we sought to detect similarities between participants across profiles of emotional reactions (e.g., is there a distinct fear pattern of responses among participants?). For the latter clustering, we wished to investigate not only whether there were multi-componential patterns of emotional reactions (H1 and H2 of our research hypotheses), but whether these patterns were general, rather than clustering only to individual VR games (e.g., a ZEROG cluster). To verify this, we cross-tabulated the optimal cluster solution against the VR games in our data set. Prior to clustering, we partialed out the "subject effect" in the data by extracting residuals from a multivariate regression that used self-report items as responses and the subject factor as predictor. This we did to remove dependencies between observations belonging to the same participant.[1] For the hierarchical clustering, we used an agglomerative nesting algorithm using Ward's method for merging of clusters.

For the supervised data analysis, we modelled standardized intensities of individual emotion categories—obtained with the GEW—as a function of standardized componential items. Associations between componential items and selected GEW intensities were modelled first in a bivariate manner (e.g., hand perspiration predicting fear intensity), and then in a multivariate manner (e.g., more

[1] Note that this procedure only rules out baseline differences in subject responses, not more complex types of individual differences.

than one predictor for fear intensity in the model). The former was aimed at identifying patterns of multi-componential responses to VR scenarios, and thus served as a confirmatory analysis for the clustering part (H2). The latter was aimed at testing our appraisal theory hypotheses (H3–H5) and at finding the best predictive model for each selected GEW intensity. For the bivariate analyses we ran 20 × 53 separate multilevel linear models on the unresidualized data, with selected GEW intensities as separate dependent variables and componential items (see Table 2) as separate independent variables. In order to account for non-independence of repeated responses both within subjects and VR games we included crossed random intercepts for these two effects into the multilevel linear models. Significance of associations was evaluated by a t-test on each regression coefficient, using Satterthwaite's approximation for the degrees of freedom for random effects models (see [63]). Due to the large number of statistical tests, we opted to set the significance level at a conservative value of $\alpha = 0.0001$. This is equivalent to a Bonferroni correction for 500 independent tests, with the actual number of tests for this analysis at 1060.

For the multivariate analyses, we focused only on modelling those GEW intensities that were found to be involved in meaningful data patterns during the clustering part, and were found to be associated with a fully multi-componential[2] pattern of emotional responding in the bivariate supervised modelling. Again using multilevel linear models, we first tested the relative importance of emotion components (H3) by fitting and comparing models for each selected GEW intensity that contained predictor variables of only one emotion component at a time (e.g., fear intensity predicted only by appraisal, only by physiology, etc.). We then compared these models in terms of marginal R^2, conditional R^2, and the Akaike Informartion Criterion (AIC). The R^2 indices differ according to whether proportion of variance explained is conditioned upon the random effects present in the model (conditional R^2), or collapsed across random effects (marginal R^2), which can sometimes yield different results [64], [65]. In addition, because the number of variables per emotion component was unequal, we also considered mean marginal and conditional R^2 per model. AIC, finally, is an approximately unbiased measure of a model's generalization capacity and penalizes models for redundant parameters. Due to the latter property, adjusting manually for the number of variables per component is not necessary for AIC comparison. After fitting and comparing models by emotion component, we conducted a stepwise multilevel model selection for each selected GEW intensity to find the best subset of predictor variables across all components, adding predictive componential items in a forward manner by minimizing AIC. Stepwise model building was terminated when AIC stopped decreasing by a value larger than 2, which is considered a conventional threshold for determining variable relevance [66], [67]. Once again, for all these multilevel models, a random intercept was included to account for both repeated subject and repeated VR game dependence.

For testing whether individual/game differences in emotional responses were explained by differences in appraisal and regulation (H4 and H5), we evaluated the need for the random subject and random VR game intercepts in the fitted appraisal component and regulation component models for each selected GEW intensity. Note that, in a multilevel model, random variables can be considered as latent predictors of individual differences. For example, a random subject intercept allows that repeated measures from the same subject deviate from the population response by a constant value (i.e., each subject has its own intercept). From an explanatory point of view, such variables act as "blind guesses" as to what is causing non-independence among repeated measurements (e.g., within-subject correlation). However, their inclusion should no longer be necessary—or less—when the true causes of the non-independence are included in the model (e.g., appraisal information). Therefore, we expected that a random game intercept would become redundant in a model that included appraisal criteria as predictors, and that a random subject intercept would become redundant in a model that included appraisal criteria and regulation strategies as predictors. We tested redundancy by comparing random effects structures using AIC as a criterion, again considering decreases in AIC smaller than 2 (or negative) to be evidence against adding parameters. We did not use conventional significance testing for this analysis due to the complications of testing whether a random effects parameter is significantly different from zero [see 63].

All analyses were conducted using the R statistical software, version 3.3.3 [68]. For hierarchical clustering, we used the package "cluster" [69]. For multilevel linear modelling, we used the packages "lme4" [70] and "lmerTest" [71]. Model comparisons involving fixed effects used maximum likelihood (ML) estimation for parameter estimates, while model comparisons involving random effects used restricted maximum likelihood (REML) estimation [63].

3 Results

Prior to the analyses, the data were pre-processed. All ratings of the self-report questionnaire were subtracted by 1, such that the "*not at all*" category corresponded to a numerical value of 0. Next, missing values were replaced by 0 values, which accounted for only 2% of all the data. This we did to avoid case-wise deletion of entire observations due to having just 1 missing value. Zero imputation was favoured due to being conservative.

3.1 Descriptives

Average VR immersion was rated 4.84 (on a 0–6 scale), indicating higher than moderate immersion. No significant differences in average immersion were found for the 7 VR games, $F(6,147) = 1.79$, $p = 0.1051$. Average intensity ratings (on a 0–5 scale) for each VR game on the 20 qualitative emotion categories of the GEW are given in Table 3. Across all games, interest, amusement, joy, and

[2] At least one significant association with each emotion component under study.

TABLE 3
AVERAGE INTENSITY RATINGS (0–5 SCALE) FOR GEW EMOTION CATEGORIES BY VR GAME

Emotion	VR game: ZOMBIES	NINJA	PLANK	LONGBOW	WHALE	BRUSH	ZEROG	Average
Interest	**3.8**	**4.3**	**4.0**	**3.7**	**4.6**	**4.3**	**4.4**	**4.2**
Amusement	**3.7**	**4.8**	**3.6**	**4.6**	**3.9**	**4.3**	**3.7**	**4.1**
Pride	2.2	2.2	2.4	2.3	0.8	1.5	1.1	1.8
Joy	2.9	**3.9**	2.8	**3.7**	**3.5**	**3.8**	**3.0**	**3.4**
Pleasure	**3.5**	**4.3**	**3.2**	**4.5**	**4.0**	**4.3**	**3.4**	**3.9**
Contentment	2.1	2.9	2.3	2.4	2.7	**3.0**	**3.0**	2.6
Admiration	1.8	1.5	2.0	1.6	**3.5**	2.6	2.3	2.2
Love	0.2	0.4	0.1	0.4	0.6	0.5	0.6	0.4
Relief	2.6	1.1	2.0	1.0	1.1	1.3	1.0	1.4
Compassion	0.2	0.2	0.0	0.1	0.7	0.2	0.2	0.2
Sadness	0.3	0.2	0.1	0.1	0.0	0.0	0.4	0.2
Guilt	0.1	0.0	0.1	0.3	0.3	0.0	0.3	0.2
Regret	0.5	0.9	0.2	0.7	0.0	0.0	0.6	0.4
Shame	0.3	0.1	0.2	0.4	0.2	0.1	0.5	0.3
Disappointment	1.1	1.7	0.7	1.5	0.2	0.5	2.0	1.1
Fear	**3.0**	0.3	**3.3**	0.7	2.0	0.2	1.5	1.6
Disgust	1.3	0.0	0.0	0.0	0.2	0.1	0.1	0.2
Contempt	0.4	0.4	0.2	0.1	0.3	0.0	0.2	0.2
Hate	1.1	0.2	0.1	0.3	0.0	0.0	0.2	0.3
Anger	1.3	0.6	0.1	1.4	0.0	0.0	1.0	0.6

Values higher than 3 have been highlighted in bold.

pleasure were rated the highest in intensity, pointing to a general experience of VR being pleasant—even when scary. For other emotions, average intensity depended more strongly on the type of game, with high fear intensity for ZOMBIES and PLANK, and high admiration intensity for WHALE. Finally, many negative emotions received low intensity ratings across all VR games, such as guilt, contempt, compassion, and sadness. At least for this selection of VR games, it appeared that these emotions were largely found to be not applicable. In general, however, the VR games elicited always at least one fairly intense emotion, with no participant among the 53 ever choosing the "no emotion" option of the GEW.

3.2 Unsupervised Modelling

First we conducted a hierarchical clustering on the set of questionnaire items (i.e., column-wise clustering of the data). For this we used data that was residualized by removing baseline subject differences (see method). Results of the hierarchical clustering on questionnaire items are depicted as a dendrogram in Figure 2. From this dendrogram, we found visual evidence for four major clusters of items. The bottom cluster we considered an empty cluster (36 items), in that it grouped items with either low variance or many 0 values, such as the "no emotion" and "other emotion" options of the GEW, which were never rated with a non-zero value. The qualitative emotion categories that were found not-applicable to describe experiences to our selection of VR games (see Table 3) also showed up in this cluster. Finally, the empty cluster also contained all of the bodily location items, indicating that these measurements either lacked applicability in the experienced VR games, or were not well differentiated in terms of other emotional responses.

The remaining three clusters did point to patterns of emotional reactions that were multi-componential. From top to bottom (Figure 2), we identified a fear cluster (18 items), a joy cluster (15 items), and a gaming cluster (6 items). The gaming cluster consisted of emotional reactions related to achieving and succeeding game objectives, such as appraisals of urgency, goal advancement, and self-control, an action tendency to vent negative emotion, and regulation strategies that focused on game objectives and reappraisal through humor. We did not consider this cluster as truly emotional, however, since it lacked physiological and feeling items, and did not seem structured around a qualitative emotion category. The remaining two clusters did manifest this type of structure. The fear cluster clearly contained a multi-componential pattern of emotion responses, including appraisals of danger, unpleasantness, unexpectedness, game-control, and worse-than-expected unpleasantness, action tendencies to escape or even stop the VR experience, all reactions related to physiological arousal (e.g., sweating, breathing), regulation strategies for ignoring intense imagery and reappraising the reality of the VR, and finally feelings of disorientation, relief, and fear. Although it would seem that the physiological questionnaire items clustered in one group, stomach queasiness and weak limbs were somewhat distinct from breathing, hear rate, muscle and perspiration changes, with the latter potentially also occurring due to physical exertion. The joy cluster also contained a multi-componential pattern of emotion responses, including appraisals of pleasantness and better-than-expected pleasantness, action tendencies to stay immersed in VR (e.g., absorb, repeat, explore), and numerous positively valenced feelings (e.g., joy, amusement, pleasure).

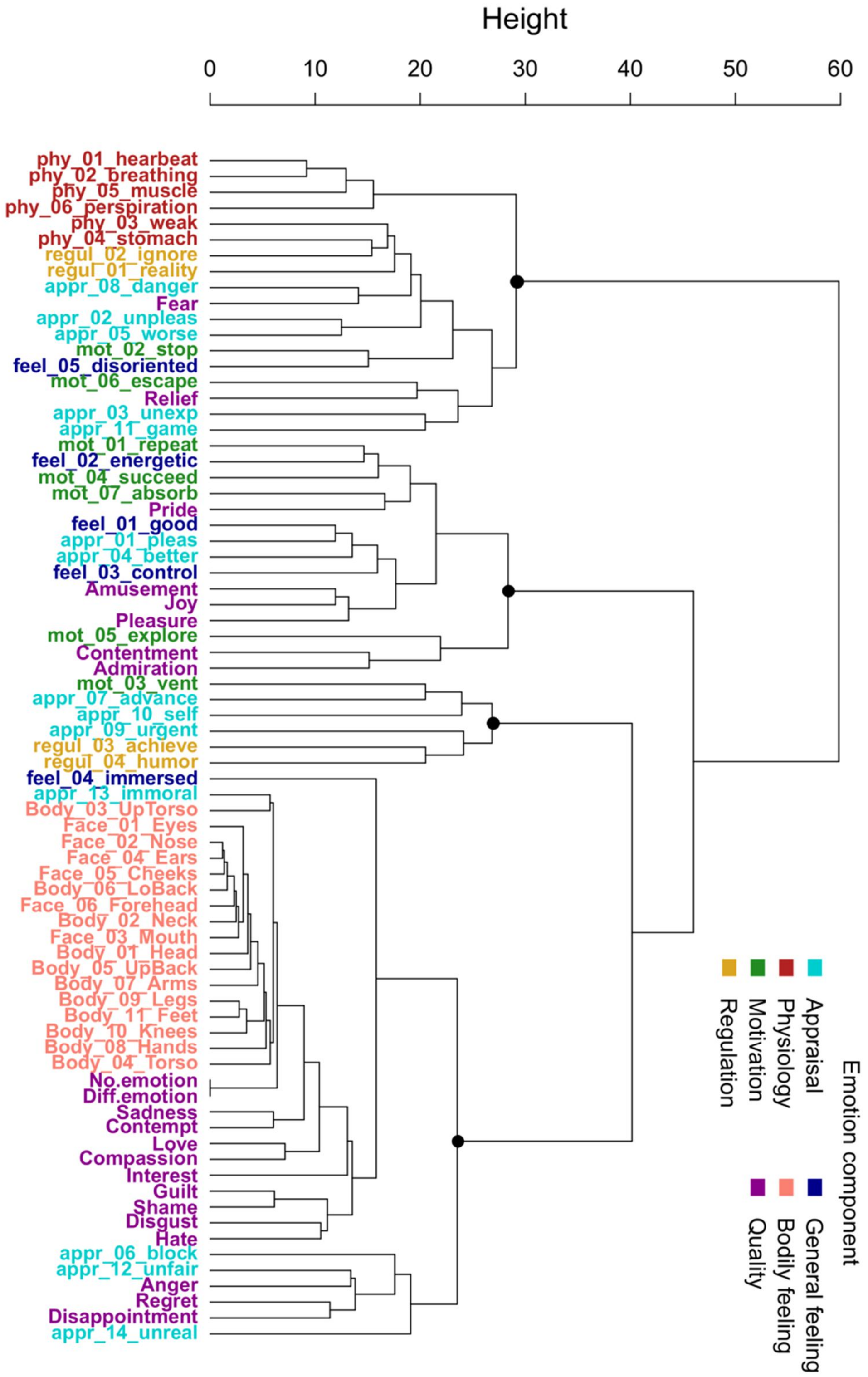

Figure 2. Dendrogram of hierarchical clustering of questionnaire items.

Second we conducted a hierarchical clustering on the set of response profiles per subject (i.e., row-wise clustering), using again the subject-residualized data. The resulting dendrogram pointed visually to four well-differentiated clusters across all VR experiences. In order to investigate whether these clusters were connected to individual VR games or were more general, we cross-tabulated these cluster membership against the VR games (Table 4). This table suggested that emotional reaction patterns generalized across games, with clusters containing subject experiences from more than one VR game together. Experiences with fear-inducing games (ZOMBIES, PLANK, ZEROG) tended to group together strongly in the first and second clusters, while experiences with the more physically oriented games (ZOMBIES, NINJA, LONGBOW) appeared in the third cluster. The fourth cluster primarily captured experiences with the BRUSH game, but also contained a substantial amount of experiences with the WHALE game.

TABLE 4
CROSS-TABULATION OF RESPONSE PATTERN CLUSTERS AGAINST EXPERIENCED VR GAMES

	VR game							
Cluster	ZOMBIES	NINJA	PLANK	LONGBOW	WHALE	BRUSH	ZEROG	*N*
1	4	2	12	1	12	6	18	55
2	9	0	8	0	3	1	2	23
3	9	14	0	18	0	0	1	42
4	0	6	2	3	7	15	1	34

Average VR immersion was rated 4.84 (on a 0–6 scale), indicating higher than moderate immersion.

In order to examine the contents of these four clusters, we averaged emotion ratings of all questionnaire items within each cluster and ranked them from highest to lowest. Figure 3 depicts the five highest and five lowest such ratings for each cluster. These results indicated that—as expected—both clusters 1 and 2 represented fear patterns, although cluster 2 seemed to contain the most intense and negative fear, as indicated, e.g., by high appraisal of danger, high fear intensity, and high unpleasantness. For cluster 1, the patterns suggested a weaker, more benign version of fear, in the sense of anxiety or even thrill-seeking, as indicated by the highest rated item in this cluster, the action tendency to explore. Cluster 3 pointed to a gaming cluster, as indicated by high physiological ratings, focus on self-control and achievement of game objectives, and high appraisal of urgency of the situation. As noted, this cluster captured almost exclusively the competitive game types in our VR game selection (ZOMBIES, NINJA, LONGBOW). Cluster 4 pointed to a general pleasantness or relaxation cluster, high feelings of pleasure and control, versus low ratings of danger, urgency, game focus, and unpleasantness. This cluster captured most of the experiences associated with the BRUSH game, which did not present the subject with game objectives or any challenging stimuli. Instead subjects were free to explore and express their creativity.

The mixture of games in each cluster confirmed that these clusters were more general than just reflecting one specific game experience (Table 4), whereas the emotional ratings in each cluster supported that these clusters captured multi-componential emotion patterns (Figure 3), and were therefore more general than just single-component patterns (e.g., a pure physiology cluster).

3.3 Supervised Modelling

First we conducted a bivariate analysis of association

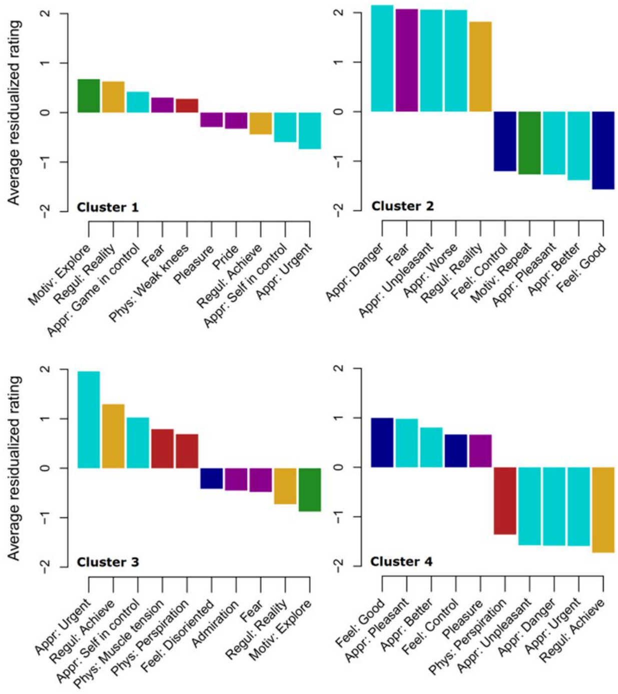

Figure 3. Five highest and five lowest rated questionnaire items per cluster. Note that data represent average residualized responses. For color codes refer to Fig. 3. A larger version of this image can be found in the Supplementary Material, Fig. S3.

between component items and GEW intensities, in order to confirm the patterns revealed by the unsupervised data analysis, and investigate further the componential hypothesis (H2). Since the cluster analysis indicated that the items relating to bodily feeling location were either low in variance or irrelevant, we excluded these items from all subsequent analyses. Results of the bivariate analysis suggested that fear, joy, and pleasure intensity were associated with the most highly "patterned" component information, with 12, 7, and 6 significantly associated items, respectively (Table 5, bivariate results). Moreover, these items were spread across different emotion components, indicating that the patterns were genuinely multi-componential. For the remaining GEW intensities (e.g., relief, guilt, anger), the number of significantly associated component items numbered between 0 and 3, at the most, with no pattern spread across components like fear, joy, or pleasure. If these emotions were felt at all during the VR games, they were not associated with a multi-componential pattern of responses. The full table of coefficients and significance tests for the pairwise analysis is provided in the Supplementary Material, Table S1.

TABLE 5
STANDARDIZED COEFFICIENTS FOR COMPONENTIAL ITEMS IDENTIFIED AS RELEVANT PREDICTORS FOR FEAR AND JOY INTENSITY

Fear			Joy		
Item	Bivariate	Multivariate	Item	Bivariate	Multivariate
Appr: Unpleasant	0.40		Appr: Pleasant	0.47	0.37
Appr: Worse	0.39		Appr: Unpleasant	-0.33	
Appr: Advance		-0.11	Appr: Better	0.43	
Appr: Danger	0.57	0.29	Appr: Advance		0.15
Motiv: Repeat	-0.32	-0.18	Motiv: Repeat	0.39	0.20
Motiv: Stop	0.34		Motiv: Absorb	0.32	
Motiv: Escape	0.38	0.14	Phys: Weak knees	-0.30	
Phys: Heartbeat	0.35	0.14	Feel: Good	0.38	
Phys: Breathing	0.37		Feel: Energetic	0.32	
Phys: Stomach	0.35		Feel: Control	0.33	
Phys: Weak knees	0.29				
Phys: Perspiration	0.29				
Feel: Energetic		0.15			
Feel: Control	-0.27	-0.21			
Regul: Reality	0.46	0.16			
Regul: Humor	0.33	0.17			

Bivariate results show effects significant at α = 0.0001. Multivariate results show selection of effects by forward stepwise AIC minimization.

TABLE 6
FIT INDICES FOR MULTILEVEL MODELS FOR FEAR AND JOY INTENSITY

	Fit index				
Model	R^2_{marg}	R^2_{cond}	*Mean* R^2_{marg}	*Mean* R^2_{cond}	AIC
Fear					
Appraisal (14) only	**0.61**	**0.64**	0.04	0.04	**327**
Motivation (7) only	0.22	0.61	0.03	0.08	341
Physiology (6) only	0.23	0.53	0.03	0.08	354
Feeling (5) only	0.11	0.48	0.02	0.09	368
Regulation (4) only	0.25	0.54	**0.06**	**0.13**	344
Best forward subset (9)	0.62	0.69			280
Joy					
Appraisal (14) only	**0.25**	**0.66**	0.01	0.04	**387**
Motivation (7) only	0.19	0.60	0.02	0.08	394
Physiology (6) only	0.11	0.46	0.01	0.07	412
Feeling (5) only	0.20	0.55	**0.04**	**0.11**	388
Regulation (4) only	0.02	0.43	0.00	0.10	421
Best forward subset (3)	0.26	0.66			367

R^2_{marg} = marginal *R*-squared. R^2_{cond} = conditional *R*-squared. AIC = Akaike Information Criterion. Mean *R*-squared values divide original *R*-squared by the number of variables considered. Optimal values in bold.

Both the unsupervised analysis and the bivariate supervised analysis identified fear, pleasure, and joy as involving a broad pattern of component responses. For subsequent multivariate supervised analyses, we decided to focus exclusively on multilevel modelling of fear and joy intensity. Pleasure intensity was excluded from these analyses due to being extremely highly correlated with the bivariate componential effect pattern of joy (r = 0.97). To test our hypothesis regarding component importance (H3), we fitted multilevel models to fear and joy intensity, using predictor variables of only one emotion component at a time. Results of the analysis are presented in Table 6. For fear intensity, the appraisal model achieved the lowest AIC value among component models, as well as the highest marginal and conditional R^2, with 61.3% and 64.4% proportion of variance explained, respectively. When adjusting R^2 for the number of predictors, however, a slightly different picture emerged, with the regulation model achieving better mean marginal and conditional R^2, 6.4% and 13.5%, respectively. For joy intensity, the appraisal model achieved the lowest AIC value among component models, as well as the highest marginal and conditional R^2, with 25.7% and 66.3% proportion of variance explained, respectively. When adjusting R^2 for the number of predictors, again a different picture emerged, with the feeling model achieving better mean marginal and conditional R^2, 4.0% and 11.1%, respectively.

A best-subset selection across all components by forward stepwise modelling resulted in a model for fear intensity with 9 predictor variables (Table 5, multivariate results). Many of these overlapped with items identified as significant in the bivariate analysis, with appraisal of danger as the most important predictor in both analyses. Significant items from the bivariate analysis that disappeared in the multivariate model likely point to redundant correlates of more predictive items, such as the physiology items. Two items appeared as meaningful in the multivariate analysis (appraisal of goal advancement, feeling energetic) that were not identified in the bivariate analysis, suggesting the possibility of a suppression effect. Note, however, that these two modelling strategies used different criteria to evaluate the relevance of the items (significance testing versus AIC minimization). For joy intensity, a best-subset selection resulted in a model with only 3 predictor variables (Table 5, multivariate results), far fewer than the number of statistically significant bivariate associations. For both analyses, appraisal of pleasantness showed the strongest effect on joy intensity. Interestingly, as with fear intensity, appraisal of goal advancement was relevant only after controlling for other appraisal items.

Finally, in order to test our hypotheses of game and individual differences (H4 and H5) in these data, we evaluated whether the inclusion of random effects for subjects and games improved model fit after adjustment for appraisal and regulation items. To do this, we compared four random effects structures, which were (i) no random effects, (ii) a random intercept for subjects, (iii) a random intercept for games, and (iv) random intercepts for subjects and games. These structures were compared for five models, (1) an intercept-only model, (2) an appraisal-only model, (3) a regulation-only model, and (4) an appraisal-regulation model. We used AIC as a criterion for model comparison, considering differences larger than 2 to be meaningful. Results painted a different picture for the models of fear versus joy intensity (Table 7). For fear, the fixed intercept-only model suggested that random differences between games (corresponding to within-game correlation) were far more substantial than random differences between subjects (corresponding to within-subject correlation), with AIC even increasing when a random subject intercept is added to the fixed intercept-only model. However, the best fitting model for fear is the one that includes both appraisal and regulation variables, and no random effects. This indicates that, contrary to our expectation, appraisal and regulation information primarily account for within-game correlation of fear responses, rather than within-subject correlation of fear responses. For joy, the fixed intercept-only model suggested that random differences between subjects in joy intensity were more substantial than random differences between games. This remained true even when adjusting for relevant appraisal variables. The appraisal-only model generally achieved lower AIC than other fixed effects structures, but still requires the random subject intercept to account for unexplained within-subject correlation.

4 Discussion

In this study we investigated the potential for virtual reality (VR) to induce strong emotions, as expressed by their intensity, their comprehensiveness in affecting multiple mental and bodily components, and their qualitative differentiation in terms of these patterns. Our interest in this medium was guided by componential definitions of emotion and causal models of appraisal theory. Despite the importance of these concepts and models in emotion theory, they have not informed the development of classic paradigms for emotion induction, such as viewing pictures or videos, or listening to music. This has made it difficult—or sometimes impossible—to induce intense and fully multi-componential emotions which, in turn, has hindered studies that wish to understand the emotion process, for example in the manner that its constituents unfold and interact over time. Virtual reality offers an ideal medium for this purpose in that it immerses the person completely into a custom environment and engages that person actively into the ongoing scenario. As the technical and graphical aspects of this technology continue to improve and become more realistic, VR is expected to become a fundamental tool for psychology research.

TABLE 7
Comparison of Random Effects Structures by AIC

Fixed effects	Random effects: None	Subject	Game	Subject + Game
Fear				
Intercept only	440	442	382	383
Appraisal only	324	325	326	327
Regulation only	370	372	344	344
Appraisal + Regulation	**306**	308	308	310
Joy				
Intercept only	440	419	441	416
Appraisal only	424	**385**	426	387
Regulation only	443	420	445	421
Appraisal + Regulation	429	393	431	395

The present results supported our choice to use VR as an optimal medium for operationalizing emotion theoretical concepts for induction. Participants in the study experienced several immersive VR games and reported on their emotional experiences in a questionnaire. Analysis of the data was guided by five major hypotheses related to our expectations about the strength of the induced emotions and the role of the appraisal component in those emotions. Descriptive and cluster analyses supported the intensity and multi-componentiality of the induced emotions (H1 and H2). Although intensities for qualitative feeling categories—from the Geneva Emotion Wheel—were skewed toward positive emotions (e.g., joy, amusement, pleasure), participants also reported intense fear, and no participant reported experiencing no emotions. In general, it appeared that experiencing VR was pleasant for its own sake, even when the presented scenario was challenging or scary. This thrill-seeking aspect appeared to be both related to joy and fear in VR, as indicated by the row-wise clustering of participant experiences (e.g., Figure 3, cluster 1). Multi-componentiality of emotional responding was evident in the column-wise clustering of componential questionnaire items, which revealed a fear and joy cluster, each characterized by a combination of physiological, motivational, appraisal, and feeling items. Although emotion theories often define emotions as consisting of such patterning across mental and bodily subsystems [1], profiling these patterns has not been a systematic aspect of emotion measurement and research, especially when the chosen induction paradigms lacks a clear connection to components such as appraisal or motivation. The cluster analyses that we conducted suggest that a richer characterization of emotion can be obtained by combining a suitable induction method (VR) with comprehensive measurement. Moreover, our results indicated that the extracted clusters reflected general emotion patterns, not specific to either one particular component of emotion (Figure 2) or to one particular game (Table 4).

Results of supervised multilevel modelling supported the patterning found in the clusters (H2), and also supported hypotheses relating to appraisal theory (H3, H4, H5). With respect to the importance of the appraisal component for explaining variation in emotional responding (H3), we found that appraisal variables were the most predictive for fear and joy intensity, in terms of marginal and conditional R^2, as well as AIC. For fear intensity, proportion of variance explained reached as much as 61.3%. When adjusting R^2 for the number of predictor variables per component, appraisal proved to be somewhat less successful compared to the regulation (for fear) and feeling (for joy) components, although appraisal remained the most informative with regard to minimizing AIC. Appraisal variables also emerged as important predictors during the best-subset selection of multilevel modelling, with appraisal of danger and appraisal of pleasantness as the most predictive for fear and joy intensity, respectively (Figure 5). The latter effect might appear to be somewhat trivial, since nothing is explained by stating joy correlates with appraising something as pleasant. However, the model for joy also included appraisal of goal advancement as an important predictor, suggesting that VR is not only enjoyed for its intrinsic fun but that achieving game goals is critical too. In general, the results supported the notion that appraisal variables explain substantial variance in emotional responding [58].

We also investigated the role of appraisal and regulation in explaining inter-game and inter-individual differences in emotional responding. Emotion theories—such as appraisal theory—have proposed that such differences are primarily, if not completely, brought about by how the person appraises the emotion-eliciting stimulus, and how the person chooses to regulate the resulting emotion, rather than any specific stimulus or personality characteristic (Figure 1). Our multilevel approach allowed us to test this assumption by evaluating the need for latent individual difference variables (i.e., random effects) when appraisal and regulation information had already been accounted for (Table 6). For modelling fear intensity, random differences in VR games disappeared when controlling for appraisal and regulation variables but, contrary to expectation, no evidence was found for random differences among subjects in general. For modelling joy intensity, random differences among subjects were and remained substantial, even after controlling for appraisal and regulation variables, but no evidence was found for random differences in games in general. The lack of clarity for these data suggest that, while the multilevel model is a promising approach with regard to testing individual difference hypotheses in emotion, it should be addressed by a more systematic study and experimental design than was utilized for the current research.

Finally, the current results highlighted strength of using a combination of unsupervised and supervised statistical analyses for profiling emotional responding, the use of multilevel models for taking into account repeated measures correlation across multiple levels (subjects and games) and for testing individual difference hypotheses on emotion, and the use of non-inferential criteria such as AIC for deciding on the meaningfulness of data patterns.

A number of limitations to the present study should be acknowledged, firstly the fact that our study was observational. Our chosen set of VR games did not reflect a systematic manipulation of appraisal criteria. Although we advocate appraisal theory as a basis for emotion induction, we have not been able to directly test its causal role in this study. Customized VR programs should be ideally suited to this research, however, and should become a major focus for future appraisal studies. In a manner that is unprecedented, VR enables researchers to devise ecological scenarios with full environment control, without participants' separation by screens and avatars—as in video games—or the requirement to imagine stories—as in vignette paradigms, leading to superior emotion induction. A second limitation is that our study relied entirely on self-report, even for objectively measurable emotion components such as physiology. While objective physiological recordings would be desirable, there are currently technical challenges regarding the integration of wireless physiological measurements and the treatment of motion artefacts to allow participants' free motion in VR. Also, physiological recordings were not a requirement for this study, which focused on the reportable contents of consciousness associated with emotion, and how these contents were related in the overall subjective experience. In addition, an advantage of our questionnaire approach was that all data was measured on commensurable scales, facilitating data analysis. A third limitation is that some part of the induced emotions in this study may have reflected the novelty aspect of experiencing VR, that is, the excitement and thrill of being in VR. Indeed, the data showed high ratings for positive emotions regardless of game content and unsupervised analyses found evidence for a gaming cluster of emotional responses. At present, VR does not enjoy the mainstream familiarity of other immersive and narrative media such as books, films, or music. A majority of participants in this study experienced

VR for the first time. In addition to the pure novelty factor, intrinsic pleasantness of VR was likely generated by the play aspects of some games (e.g., Fruit Ninja), or by the anticipation of “safe” scares offered by the fear-inducing games (e.g., Richie’s Plank Experience), as would also be sought by people who enjoy horror movies. Even as VR’s novelty is expected to wear off with time, the latter two elements might continue to complicate emotion measurement, and should be controlled for in standardized VR paradigms for emotion induction.

A final important limitation is that we did not explicitly compare VR to other paradigms of emotion induction. We chose not to do this directly for a number of reasons, (a) because such a comparison is inherently complicated by the fact that, as argued in the introduction, classic paradigms for emotion induction are largely incongruous with multi-componential definitions of emotion and appraisal theory, (b) the fact that there exists already much literature and data on classic paradigms that can be consulted for comparison, and (c) the fact that any systematic comparison between paradigms should—in our opinion—delve into exactly *what* makes these paradigms different (e.g., medium, presentation format, sensory complexity, narrative complexity). This suggests an ambitious research project that was beyond the scope of the current article but that should inform highly relevant future research

Acknowledgment

The authors would like to thank the master students Esraa Chaddad, Alexandra Denis, Claudia Teran, Vladimira Ivanova, Hélène Maurer, and Fabien Pavy for help in assisting with the data collection for this study.

References

[1] Fontaine, J.R., K.R. Scherer, and C. Soriano, Components of emotional meaning: A sourcebook. 2013, Oxford: Oxford University Press.
[2] Frijda, N.H., The psychologists' point of view, in Handbook of Emotions, M. Lewis, J.M. Haviland-Jones, and L. Feldman-Barret, Editors. 2008, Guilford: New York. p. 68-87.
[3] Moors, A., Theories of emotion causation: A review. Cognition & Emotion, 2009. 23(4): p. 625-662.
[4] Scherer, K.R., What are emotions? And how can they be measured? Social Science Information Sur Les Sciences Sociales, 2005. 44(4): p. 695-729.
[5] Moors, A., et al., Appraisal Theories of Emotion: State of the Art and Future Development. Emotion Review, 2013. 5(2): p. 119-124.
[6] Scherer, K.R., A. Schorr, and T. Johnstone, Appraisal processes in emotion: Theory methods, research. 2001, Oxford: Oxford University Press.
[7] Bradley, M.M., et al., Emotion and Motivation I: Defensive and Appetitive Reactions in Picture Processing. Emotion, 2001. 1(3): p. 276-298.
[8] Frijda, N.H., The Laws of Emotion. American Psychologist, 1988. 43(5): p. 349-358.
[9] Roseman, I.J., A model of appraisal in the emotion system: Integrating theory, research, and applications, in Appraisal processes in emotion: Theory, methods, research, K.R. Scherer, A. Schorr, and T. Johnstone, Editors. 2001, Oxford University Press: Oxford. p. 68-91.
[10] Russell, J.A. and L.F. Barrett, Core affect, prototypical emotional episodes, and other things called Emotion: Dissecting the elephant. Journal of Personality and Social Psychology, 1999. 76(5): p. 805-819.
[11] Ellsworth, P.C. and E.M.W. Tong, What does it mean to be angry at yourself? Categories, appraisals, and the problem of language. Emotion, 2006. 6(4): p. 572-586.
[12] Frijda, N.H., Emotion, cognitive structure, and action tendency. Cognition & Emotion, 1987. 1(2): p. 115-143.
[13] Frijda, N.H., P. Kuipers, and E. Terschure, Relations among Emotion, Appraisal, and Emotional Action Readiness. Journal of Personality and Social Psychology, 1989. 57(2): p. 212-228.
[14] Lazarus, R.S., Emotion and adaptation. 1991, New York: Oxford University Press.
[15] Roseman, I.J., Emotional Behaviors, Emotivational Goals, Emotion Strategies: Multiple Levels of Organization Integrate Variable and Consistent Responses. Emotion Review, 2011. 3(4): p. 434-443.
[16] Aue, T. and K.R. Scherer, Appraisal-driven somatovisceral response patterning: Effects of intrinsic pleasantness and goal conduciveness. Biological Psychology, 2008. 79(2): p. 158-164.
[17] Scherer, K.R., Studying the Emotion-Antecedent Appraisal Process - an Expert-System Approach. Cognition & Emotion, 1993. 7(3-4): p. 325-355.
[18] Scherer, K.R., Appraisal considered as a process of multilevel sequential checking, in Appraisal processes in emotion: Theory, methods, research, K.R. Scherer, A. Schorr, and T. Johnstone, Editors. 2001, Oxford University Press: Oxford. p. 92-120.
[19] Scherer, K.R., The dynamic architecture of emotion: Evidence for the component process model. Cognition & Emotion, 2009. 23(7): p. 1307-1351.
[20] Kaiser, S. and T. Wehrle, Facial expressions as indicators of appraisal processes, in Appraisal processes in emotion: Theory, methods, and research, K.R. Scherer, A. Schorr, and T. Johnstone, Editors. 2001, Oxford University Press: Oxford. p. 285-300.
[21] Laird, J.D. and C. Bresler, The process of emotional experience: A self-perception theory, in Review of personality and social psychology: Emotion, S. Clark, Editor. 1992, SAGE: Newbury Park, CA. p. 213-234.
[22] Scherer, K.R. and H. Ellgring, Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? Emotion, 2007. 7(1): p. 113-130.
[23] Lazarus, R.S., Relational meaning and discrete emotions, in Appraisal processes in emotion: Theory, methods, research, K.R. Scherer, A. Schorr, and T. Johnstone, Editors. 2001, Oxford University Press: Oxford. p. 37-67.
[24] Ortony, A., G. Clore, and A. Collins, The cognitive structure of emotions. 1988, Cambridge: Cambridge University Press.
[25] Lang, P.J., M.M. Bradley, and B.N. Cuthbert, International affective picture system (IAPS): Technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention, 1997: p. 39-58.
[26] Coan, J.A. and J.J. Allen, Handbook of emotion elicitation and assessment. 2007, Oxford: Oxford University Press.
[27] Kreibig, S.D., Autonomic nervous system activity in emotion: A review. Biological Psychology, 2010. 84(3): p. 394-421.
[28] Moors, A. and K.R. Scherer, The role of appraisal in emotion, in Handbook of cognition and emotion, M. Robinson, E. Watkins, and E. Harmon-Jones, Editors. 2013, Guilford Press: New York.
[29] Freeman, W.J., Emotion is from preparatory brain chaos; irrational action is from premature closure. Behavioral and Brain Sciences, 2005. 28(2): p. 204-+.
[30] Gross, J.J., The Future's So Bright, I Gotta Wear Shades. Emotion Review, 2010. 2(3): p. 212-216.
[31] Scherer, K.R., Emotions are emergent processes: they require a dynamic computational architecture. Philosophical Transactions of the Royal Society B-Biological Sciences, 2009. 364(1535): p. 3459-3474.
[32] Grandjean, D., D. Sander, and K.R. Scherer, Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Consciousness and Cognition, 2008. 17(2): p. 484-495.
[33] Scherer, K.R., Feelings integrate the central representation of appraisal-driven response organization in emotion, in Feelings and emotions: The Amsterdam symposium, A.S.R. Manstead, N.H. Frijda, and A.H. Fischer, Editors. 2004, Cambridge University Press: Cambridge. p. 136-157.
[34] Parkinson, B., Untangling the appraisal-emotion connection. Personality and social psychology review, 1997. 1(1): p. 62-79.
[35] Parkinson, B. and A.S.R. Manstead, Making Sense of Emotion in Stories and Social-Life. Cognition & Emotion, 1993. 7(3-4): p. 295-323.
[36] Ceschi, G. and K.R. Scherer, Children's ability to control the facial expression of laughter and smiling: Knowledge and behaviour. Cognition & Emotion, 2003. 17(3): p. 385-411.
[37] Folkman, S. and R.S. Lazarus, If It Changes It Must Be a Process - Study of Emotion and Coping during 3 Stages of a College-

Examination. Journal of Personality and Social Psychology, 1985. 48(1): p. 150-170.
[38] Smith, C.A. and P.C. Ellsworth, Patterns of Appraisal and Emotion Related to Taking an Exam. Journal of Personality and Social Psychology, 1987. 52(3): p. 475-488.
[39] Tong, E.M.W., et al., The use of ecological momentary assessment to test appraisal theories of emotion. Emotion, 2005. 5(4): p. 508-512.
[40] Slater, M. and M.V. Sanchez-Vives, Enhancing Our Lives with Immersive Virtual Reality. Frontiers in Robotics and Ai, 2016. 3.
[41] Krijn, M., et al., Treatment of acrophobia in virtual reality: The role of immersion and presence. Behaviour Research and Therapy, 2004. 42(2): p. 229-239.
[42] Riva, G., Virtual reality in psychotherapy: Review. Cyberpsychology & Behavior, 2005. 8(3): p. 220-230.
[43] Choi, Y.H., et al., Short-term treatment of acrophobia with virtual reality therapy (VRT): A case report. Cyberpsychology & Behavior, 2001. 4(3): p. 349-354.
[44] Coelho, C.M., et al., Virtual reality and acrophobia: One-year follow-up and case study. Cyberpsychology & Behavior, 2006. 9(3): p. 336-341.
[45] Coelho, C.M., et al., The use of virtual reality in acrophobia research and treatment. Journal of Anxiety Disorders, 2009. 23(5): p. 563-574.
[46] Emmelkamp, P.M.G., et al., Virtual reality treatment in acrophobia: A comparison with exposure in vivo. Cyberpsychology & Behavior, 2001. 4(3): p. 335-339.
[47] Hodges, L.F., et al., Virtual Environments for Treating the Fear of Heights. Computer, 1995. 28(7): p. 27-34.
[48] Rothbaum, B.O., et al., Virtual-Reality Graded Exposure in the Treatment of Acrophobia - a Case-Report. Behavior Therapy, 1995. 26(3): p. 547-554.
[49] Rothbaum, B.O., et al., Effectiveness of Computer-Generated (Virtual-Reality) Graded Exposure in the Treatment of Acrophobia. American Journal of Psychiatry, 1995. 152(4): p. 626-628.
[50] Kozlov, M.D. and M.K. Johansen, Real Behavior in Virtual Environments: Psychology Experiments in a Simple Virtual-Reality Paradigm Using Video Games. Cyberpsychology Behavior and Social Networking, 2010. 13(6): p. 711-714.
[51] Rovira, A., et al., The use of virtual reality in the study of people's responses to violent incidents. Frontiers in Behavioral Neuroscience, 2009. 3.
[52] Slater, M., et al., Bystander Responses to a Violent Incident in an Immersive Virtual Environment. Plos One, 2013. 8(1).
[53] Gillath, O., et al., What can virtual reality teach us about prosocial tendencies in real and virtual environments? Media Psychology, 2008. 11(2): p. 259-282.
[54] Navarrete, C.D., et al., Virtual Morality: Emotion and Action in a Simulated Three-Dimensional "Trolley Problem". Emotion, 2012. 12(2): p. 364-370.
[55] Patil, I., et al., Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. Social Neuroscience, 2014. 9(1): p. 94-107.
[56] Skulmowski, A., et al., Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. Frontiers in Behavioral Neuroscience, 2014. 8.
[57] Slater, M., et al., A Virtual Reprise of the Stanley Milgram Obedience Experiments. Plos One, 2006. 1(1).
[58] Scherer, K.R. and J.R. Fontaine, Driving the emotion process: The appraisal component, in Components of emotional meaning: A sourcebook, J.R. Fontaine, K.R. Scherer, and C. Soriano, Editors. 2013, Oxford University Press: Oxford. p. 186-209.
[59] Lin, J.H.T., Fear in virtual reality (VR): Fear elements, coping reactions, immediate and next-day fright responses toward a survival horror zombie virtual reality game. Computers in Human Behavior, 2017. 72: p. 350-361.
[60] Sacharin, V., K. Schlegel, and K.R. Scherer, Geneva emotion wheel rating study. 2012, Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences: Aalborg.
[61] Khalsa, S.S., et al., The pathways of interoceptive awareness. Nature Neuroscience, 2009. 12(12): p. 1494-1496.
[62] Nummenmaa, L., et al., Bodily maps of emotions. Proceedings of the National Academy of Sciences of the United States of America, 2014. 111(2): p. 646-651.
[63] Fitzmaurice, G.M., N.M. Laird, and J.H. Ware, Applied longitudinal analysis. 2004, New York: Wiley.
[64] Nakagawa, S. and H. Schielzeth, A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 2013. 4(2): p. 133-142.
[65] Xu, R.H., Measuring explained variation in linear mixed effects models. Statistics in Medicine, 2003. 22(22): p. 3527-3541.
[66] Burnham, K.P. and D.R. Anderson, Multimodel inference - understanding AIC and BIC in model selection. Sociological Methods & Research, 2004. 33(2): p. 261-304.
[67] Raftery, A.E., Bayesian model selection in social research. Sociological Methodology 1995, Vol 25, 1995. 25: p. 111-163.
[68] R Core Team, R: A language and evironment for statistical computing. 2017, R Foundation for Statistical Computing: Vienna, Austria.
[69] Maechler, M., et al., cluster: Cluster analysis basics and extensions, in R package. 2016.
[70] Bates, D., et al., Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 2015. 67(1): p. 1-48.
[71] Kuznetsova, A., P.B. Brockhoff, and R.H.B. Christensen, lmerTest: Tests in linear mixed effects models, in R package. 2016.

Ben Meuleman. The author obtained master degrees in Experimental Psychology (2008) and Statistical Data Analysis (2010) at Ghent University, and completed a doctoral thesis in Psychology on computational modelling of appraisal theory (2015), supervised by Klaus Scherer and Agnes Moors. He has worked as a doctoral researcher and as a post-doctoral researcher at the Swiss Center for Affective Sciences (University of Geneva; 2011–present). He has published to date five papers in academic journals and contributed one book chapter. Current research focuses on the modelling of multimodal emotion processes in virtual reality environments.

David Rudrauf. The author obtained a master degree (2000) and a Ph.D. (2005) in Cognitive Sciences from Pierre and Marie Curie University, a Ph.D. in Neurosciences from the University of Iowa (2005), and a Habilitation from the University Joseph Fourier in Grenoble (2015). He has worked as: an Assistant Professor of Neurology & Radiology, in the Department of Neurology (University of Iowa; 2008–2011); a non-permanent research scientist at the Laboratoire d'Imagerie Fonctionelle (INSERM, UMR S 678; 2012–2014); a non-permanent research scientist at the Grenoble Institute of Neuroscience GIN (INSERM; 2014–2016); and an Associate Professor of Psychology at the Swiss Center for Affective Sciences (University of Geneva; 2016-present). He was the Director of the Laboratory of Brain Imaging and Cognitive Neuroscience, Division of Behavioral Neurology and Cognitive Neuroscience, Department of Neurology, (UIHC, University of Iowa; 2008–2011), and is currently the Director of the Laboratory of Multimodal Modelling of Emotion & Feeling, Swiss Center for Affective Sciences, FAPSE, University of Geneva (since 2016). He has published to date 53 journal articles. Current research focuses on the development of computational models of normal and pathological psychology combined with virtual reality.