Value Orientation in Players: Development and Validation of a Tradeoff-Based Scale of In-Game Priorities

Ziao Tang and Ben Kirman

School of Arts and Creative Technologies, University of York, York, UK

Abstract

Players differ not only in what they enjoy but in what they prioritize when in-game goals conflict. Existing player-typology and motivation measures are additive and cannot capture which facet a player sacrifices under conflict. We develop and validate the Value Orientation (VO) scale, a tradeoff-based measure of in-game priorities across four dimensions: Efficiency, Exploration, Story, and Challenge. Each item embeds an explicit opportunity cost. Across three studies with Chinese players (total N=813), exploratory factor analysis yielded a 16-item, four-factor scale, and confirmatory factor analysis in an independent sample replicated this structure (CFI=0.99) and outperformed collapsed alternatives. The dimensions showed acceptable reliability, clear discriminant validity, theory-consistent convergent relations with needs and player-type measures, incremental validity beyond basic-needs satisfaction, prediction of choices on a differently worded situational task, full scalar invariance, and good retest stability. Known-groups evidence was largely supportive.

Keywords: Player modeling; value orientation; tradeoffs; scale validation; games user research

1. Introduction

Digital games rarely let players pursue every goal at once. A timed objective competes with the urge to explore a side area; a guided main quest competes with the impulse to wander; a tense narrative beat competes with the desire to grind out a harder challenge. Players who otherwise enjoy all of these facets must nonetheless, moment to moment, decide which to protect and which to give up. The question this paper addresses is whether players hold stable priorities over such conflicts—and whether those priorities can be measured.

Research on individual differences among players has a long history (for a synthesis, see Hamari & Tuunanen, 2014), and such differences continue to matter for engagement; in massively multiplayer online role-playing games (MMORPGs), for example, avatar attributes and the Proteus effect shape players' intention to keep playing (Nan et al., 2024; Sun et al., 2024). The dominant measurement logic of this research, however, has been additive rather than tradeoff-based. Bartle's (1996) early taxonomy of player types proposed that orientations such as achievement and socializing were, to some degree, mutually suppressing: being more of one made one less of another. When this idea was put to empirical test, however, it did not hold. Yee's (2006) factor-analytic model of online-game motivations recovered three broad components—achievement, social, and immersion—that were only weakly intercorrelated ($rs < 0.10$), leading Yee to conclude explicitly that motivations do not suppress one another: a player high on achievement may be equally high on the social or immersion components. This finding established what remains the prevailing paradigm: player motivations and preferences are measured as separable continua, each scored on its own, so that a player can in principle be high (or low) on all of them simultaneously. The same logic organizes recent work, including trait models of game preference (Tondello et al., 2018; Tondello & Nacke, 2019) and multi-attribute player-type profiles (Xiao et al., 2025).

Subsequent instruments inherited this additive structure. The Gamification User Types Hexad scale (Marczewski, 2015; Tondello et al., 2016, 2019) profiles users across six motivational types (philanthropist, socializer, free spirit, achiever, player, and disruptor), each rated independently. Motivation models such as the Online Gaming Motivations Scale (Yee et al., 2012) and the industry Gamer Motivation Model (Yee, 2016) likewise array players across many independently scored motivations, as do neurobiologically framed typologies such as BrainHex (Nacke et al., 2014). A parallel line of research has begun to differentiate the specific motivational facets that drive engagement, such as the multiple forms of player curiosity (Tang &

Kirman, 2025b). And within the self-determination-theory tradition (Deci & Ryan, 1985; Ryan & Deci, 2000; for its uses in games research, see Tyack & Mekler, 2020), measures of in-game need satisfaction—from the Player Experience of Need Satisfaction model (Przybylski et al., 2010; Ryan et al., 2006) to the recent Basic Needs in Games Scale (BANGS; Ballou et al., 2024), and broader player-experience instruments such as the Player Experience Inventory (Vanden Abeele et al., 2020)—quantify how much each of autonomy, competence, and relatedness is satisfied, again as separable dimensions that can all be high at once. These instruments are valuable and well validated, and they have advanced our understanding of why people play. But by design they answer the question "how much does this player value each facet?" rather than the question "when facets cannot be jointly satisfied, which does this player give up?"

That second question is not reducible to the first. Two players may value exploration and efficient progression equally in the abstract, yet differ systematically in which they sacrifice when a timer forces a choice. Additive profiles, precisely because they let all facets be high, are silent about this ordering. Yet much of what designers manipulate, and much of what players experience as the texture of play, lives in exactly these conflicts: pacing versus thoroughness, guidance versus discovery, narrative immersion versus mechanical difficulty. A measure of how players resolve such conflicts—of their tradeoff priorities—would therefore capture something that additive measures, by construction, cannot.

We propose Value Orientation (VO) as such a measure. VO characterizes the stable priorities a player expresses when in-game goals conflict, across four dimensions that recur across game genres and design tensions: Efficiency (prioritising minimal cost or time to the objective), Exploration (prioritising curiosity and departure from the prescribed path), Story (prioritising narrative, world, and character experience), and Challenge (prioritising difficulty itself and the active testing of skill). The defining feature of VO is not the four labels—which overlap conceptually with facets in existing typologies—but the tradeoff logic by which they are measured: rather than asking how much a player likes each, VO asks which a player relinquishes first. This logic is built into the instrument in two ways. Every primary-scale item presents the target dimension in explicit conflict with a named "opponent" dimension, so that endorsing the target necessarily implies giving up the opponent (an embedded opportunity cost). And a separate situational task forces a single, concrete choice among the four priorities, providing a criterion that does not share the scale's wording.

Operationalizing priorities as tradeoffs carries a known psychometric cost—forcing relative judgements removes the shared response variance that ordinarily inflates internal-consistency and variance-extracted indices—and we address this directly rather than treating it as a defect. The central empirical questions, however, are whether the four dimensions form a replicable structure, whether they are genuinely distinct (rather than collapsing into a coarser "instrumental versus experiential" split), whether they relate sensibly to established needs and player-type measures while adding predictive value beyond them, whether they predict players' choices on a differently worded situational task, and whether they are stable and comparable across player subgroups.

We report three studies and a retest addressing these questions in a Chinese player sample— recruited via a general online panel for Studies 1 and 2 and the retest, and via game-community forums for the known-groups study. Study 1 develops the item pool and establishes its factor structure. Study 2—the primary validation—tests the confirmatory structure, reliability, discriminant and convergent validity, criterion validity (including the situational task), incremental validity over BANGS, and measurement invariance, with a retest sub-sample for temporal stability. Study 3 provides supplementary known-groups evidence. Together they ask whether tradeoff priorities in play can be measured with a scale that is structurally sound, distinct from additive measures, and predictively useful.

2. The Value Orientation construct

We define Value Orientation as a player's stable, relative prioritization among competing in-game goals: which goal the player protects, and which the player sacrifices, when the goals cannot be jointly satisfied. This definition is deliberately ipsative in spirit—it concerns the ordering of priorities under conflict, echoing how theories of human values treat a value system as a set of priorities revealed when values compete (Schwartz, 1992)—and it is what distinguishes VO from the additive constructs reviewed above. An additive

preference profile can place a player high on several facets at once; a value orientation, by contrast, is expressed only when facets compete, and its content is the resolution of that competition.

2.1. Priorities under conflict

The claim that a person's values are best described by their order, rather than by their separate strengths, has a long history in value theory. Rokeach (1973) defined a value system as an enduring organization of values along a continuum of relative importance, and measured it by asking people to rank values against one another. Schwartz's (1992, 2012) theory of basic values made the ordering structural: values form a circular system of compatibilities and conflicts, and attitudes and behavior are guided less by the importance of any single value than by the tradeoff between the competing values that a situation engages. Two consequences carry over to play. First, a value becomes informative about behavior where it competes with another: a player who endorses both exploration and efficient progression reveals little until the two cannot both be served. Second, what is to be measured is the relative standing of competing values, which is why VO is defined as a prioritization rather than as a set of independent strengths.

Value theory explains why the order matters; research on goal systems explains how an order is enacted in the moment. Goals that draw on the same limited resources—time, attention, effort— compete for them, and activating one goal tends to inhibit its alternatives, a process termed goal shielding (Kruglanski et al., 2002; Shah et al., 2002). Games stage this competition by design. A side path placed beside a marked objective, a timed mission that leaves no room for a detour, or a harder route offered just as a story beat begins each asks the player to shield one goal at the cost of another. Whether level design elicits exploration at all depends on whether an explicit goal is competing for the player's attention (Gómez-Maureira et al., 2021), and players describe mechanics that force them to give one thing up for another as a distinct and consequential part of play (Wright & Denisova, 2024). VO can be understood as the stable, cross-situational component of this shielding: the goal a player tends to protect when play forces the choice. On this account, valuing a facet and protecting it come apart for a principled reason. How much a facet is valued sets the attraction of the corresponding goal, but which goal is shielded depends on the relative attraction of the goals in competition. Additive measures capture the first and are silent about the second, which is the gap identified in Section 1.

2.2. Four dimensions

VO comprises four dimensions. Efficiency is the priority placed on reaching the objective at minimal cost or time, even at the expense of thoroughness, novelty, or narrative. Exploration is the priority placed on curiosity and on departing from the prescribed path, even at the expense of speed or optimal play. Story is the priority placed on narrative, world, and character experience, even at the expense of progression or difficulty. Challenge is the priority placed on difficulty itself and on actively testing one's skill, even at the expense of efficiency or narrative flow. The four are not proposed as exhaustive or as mutually exclusive in the additive sense; players may care about all four. They are also deliberately bounded to the single-player experience of progression, discovery, narrative, and difficulty; the present scale does not represent social priorities—cooperation, competition, or sociality—that a fuller account of multiplayer play would require. We treat this as a scope boundary rather than a claim that priorities reduce to these four, and we return to it in the Limitations (Section 7.5). Each dimension nonetheless draws on a well-established account of what engages players: the appeal of optimal challenge and skill-testing (Csikszentmihalyi, 1990; Denisova et al., 2020; Malone, 1981; Sweetser & Wyeth, 2005), of curiosity and discovery (Gómez-Maureira et al., 2021; Malone, 1981; Tang & Kirman, 2025a, 2025b), of narrative transportation (Green & Brock, 2000) and of agency in narrative-focused play (Carstensdottir et al., 2021), and of efficient, goal-directed mastery (Ryan et al., 2006). The claim is rather that, under conflict, players differ stably in which they rank first—and that this ranking is measurable.

The choice of the four dimensions rests on two criteria. The first is recurrence: each corresponds to a facet that appears, under different names, across independently derived accounts of player motivation and preference, from Yee's (2006) components and BrainHex (Nacke et al., 2014) to the Gamer Motivation Model (Yee, 2016), the Trojan typology (Kahn et al., 2015), and the player-traits model (Tondello & Nacke, 2019). The second is conflict: these are the facets that single-player design most often sets against one another, as in pacing against thoroughness, guidance against discovery, expedience against narrative, and

narrative immersion against mechanical difficulty. Recurrence indicates that each facet is an established source of engagement; conflict indicates that the priority among them is a consequential individual difference.

2.3. Measuring priorities as tradeoffs

One might ask why not measure this ranking with a simpler instrument—for example, a forced-ranking task that asks players to order the four priorities from most to least important. Such a task would be ipsative in the strict sense, but it would yield only four ordinal data points per respondent, too few to support latent-variable modeling, and it would discard information about the strength of each priority: two players may rank the dimensions in the same order yet differ markedly in how close their top two priorities are. Scores derived in this way from rankings or forced choices sum to the same constant for every respondent, which distorts factor-analytic and reliability estimates (Hicks, 1970). The operationalization adopted here preserves the continuous, per-item response variance needed for factor analysis and reliability estimation while embedding the tradeoff logic within each item, rather than across items.

Two design commitments make the construct measurable as a tradeoff rather than as another additive profile. First, each primary-scale item embeds an explicit opportunity cost: it names the target dimension and an opponent dimension that must be sacrificed if the target is preserved, so that the response expresses a relative priority rather than an absolute endorsement. To keep this from degenerating into a single "approach style," opponent dimensions are rotated across items, so that each dimension is set against several different opponents. Second, a situational choice task presents concrete dilemmas with four mutually exclusive options, one per dimension, requiring a single choice; because its wording is concrete and game-specific rather than abstract, it provides a criterion that does not merely re-express the scale.

2.4. Value orientation among adjacent constructs

Figure 1 places VO among the constructs that games research already measures, each of which answers a different question. Need-satisfaction measures such as BANGS ask what makes play satisfying or frustrating (Ballou et al., 2024; see also Ballou & Deterding, 2023); motive and player-type measures such as Hexad ask how strongly each facet of play attracts a player; behavioral-tendency measures ask what a player habitually does; and gameplay logs record what a player did in a particular game. VO asks what a player protects when facets compete. Read from left to right, the figure runs from what a player finds satisfying to what a player does. A priority is expressed only when a situation poses a conflict, so it should predict choices in dilemmas; such choices should in turn shape habits; and whether a stated choice is enacted in a game depends also on what the game offers (Section 7.4). The measures administered alongside VO in Study 2 correspond to these links: BANGS to the link with needs, the situational task to the link with choice, the criterion composites to the link with habitual behavior, and Hexad to the adjacent additive typologies.

VO is expected to relate to, but remain distinct from, self-determination-theory needs: a player's competence need may dispose them toward efficiency or challenge, and their autonomy need toward exploration or story, yet the need (what is satisfying) is conceptually upstream of the priority (what is protected under conflict). The construct's core, testable claim is that these four tradeoff priorities are empirically separable and consequential for players' stated choices—a claim we evaluate, in the construct-validation tradition (Cronbach & Meehl, 1955), through the replicability of the four-factor structure, the rejection of collapsed models, the situational-choice criterion, and incremental prediction over an established needs measure.

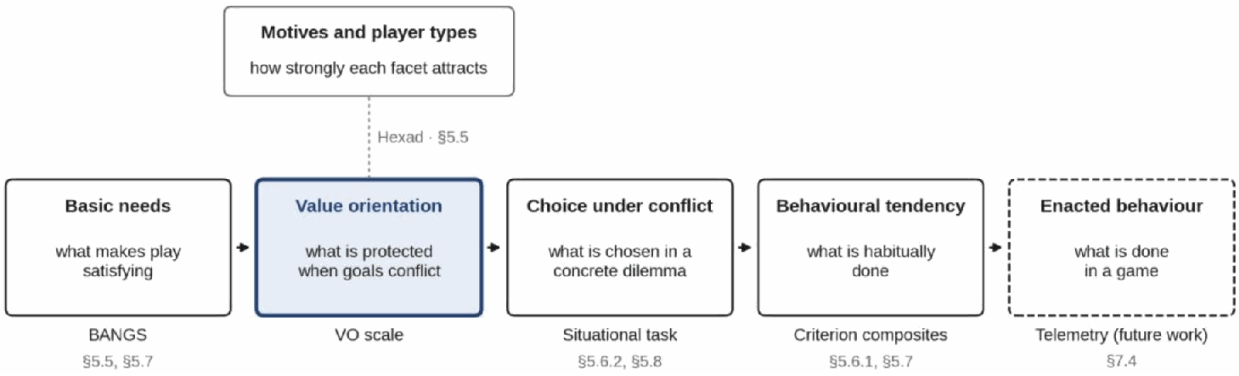


Figure 1. Value orientation in relation to adjacent constructs. Solid boxes were measured in Study 2 and are labeled with the measure used and the sections that report the corresponding tests; the dashed box marks the next step for future work (Section 7.4). Arrows indicate the expected direction of influence, not paths estimated in this study.

3. Method

3.1. Overview and design

We report three studies and a retest that develop and validate the Value Orientation (VO) scale, a measure of the stable priorities players express when in-game goals come into conflict. Across studies the scale is treated as comprising four dimensions: Efficiency (prioritising minimal cost or time to the objective), Exploration (prioritising curiosity and departure from the prescribed path), Story (prioritising narrative, world, and character experience), and Challenge (prioritising difficulty itself and the active testing of skill). Study 1 (N=307 after screening) developed the item pool and established its latent structure through exploratory factor analysis (EFA). Study 2 (N=381) was the primary validation study, conducted in an independent sample, and tested the confirmatory structure, reliability, discriminant and convergent validity, criterion validity (including a within-questionnaire situational choice task), incremental validity over an established needs measure, and measurement invariance. A retest sub-sample (N=97 matched pairs, 2–4weeks apart) assessed temporal stability. Study 3 (N=125) provided supplementary known-groups evidence using players recruited from four distinct game communities.

Studies 1 and 2 and the retest sub-sample were recruited through a general online research panel (Credamo). Study 3, whose known-groups design requires members of genuine game communities that a general panel cannot supply, was instead recruited through targeted posts in four real game-community forums on a Chinese gaming platform (Heybox). Both recruitment sources draw from a mainland-Chinese player population, so the deliberate decision to hold culture constant, and thereby avoid confounding culture with platform, is preserved even though Study 3 uses a different, community-targeted channel; the sample is in all cases a Chinese convenience sample, and we make no claim of cross-cultural or global generalizability. The design is cross-sectional. We did not lodge an advance public registration of this work; for the measurement models, fit indices, decision thresholds, and multiple-comparison control we adopted conventional, literature-based criteria, applied them uniformly across studies, and report every analysis, including those that did not support the scale, with conclusions downgraded where the evidence is weak (the criteria and our reporting commitments are set out in Section 3.6). Ethical approval for this study was granted by the Ethics Committee of the School of Arts and Creative Technologies at the University of York; the committee does not issue formal approval reference numbers, and approval was confirmed to the authors by email. All participants provided written informed consent, given electronically before taking part.

A central design commitment, maintained throughout, distinguishes VO from additive player-preference instruments such as the Hexad player-types model or the Basic Needs in Games Scale (BANGS): rather than asking how much a player values each facet (so that all facets may be high at once), VO measures tradeoff priority—which facet a player relinquishes first when facets cannot be jointly satisfied. This commitment is operationalized in two ways that we describe below: every primary-scale item embeds an explicit opportunity cost, and a separate situational choice task forces a single concrete decision among the four priorities.

3.2. Item development and the tradeoff operationalization

An initial pool of 24 items (six per dimension) was written, following established scale-construction guidelines (Boateng et al., 2018; Clark & Watson, 1995; DeVellis, 2017; Worthington & Whittaker, 2006), to express each dimension as a relative priority under conflict rather than as an absolute endorsement. Each item is phrased as an agreement statement in which the opportunity cost is embedded in the statement itself (that is, the item asserts a priority for the target dimension explicitly at the expense of a competing consideration), and respondents rate their agreement on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree). For most items the competing consideration is another VO dimension (the item's "opponent"); for three items it is a non-dimensional cost of play (time, steady progression, or completeness), so that these items still carry an opportunity cost but are not pitted against a second VO dimension. Because a single item could otherwise pair one dimension repeatedly against the same opponent (inducing a method factor), opponent dimensions were rotated across the pool so that each dimension was set against several different opponents (Discipline 1, opponent rotation). To prevent lexical overlap between the abstract scale and the concrete situational task from producing spurious correlations, the primary scale used deliberately abstract tradeoff language, while the situational task (Section 3.3.2) used concrete game imagery (Discipline 2, abstract–concrete separation).

On measurement-theoretic grounds we expected tradeoff wording to systematically depress internal-consistency coefficients and average variance extracted (AVE) relative to conventional Likert items, because forcing a relative judgment removes the shared "acquiescent" variance that inflates those indices; we therefore treat ω and standardized CFA loadings as the primary reliability evidence, with somewhat relaxed thresholds reported transparently rather than treated as failures (Discipline 3, reliability concession).

3.3. Measures

3.3.1. The Value Orientation scale

The 24 development items (Study 1) and the resulting 16-item scale (four items per dimension; Studies 2, 3, and retest) used the embedded-opportunity-cost format described above, each rated on a 7-point agreement scale (1=strongly disagree, 7=strongly agree) under an instruction emphasizing that the items concern how the respondent trades off competing play goals and that there are no right answers. The items were authored in English (the submission language) and translated into Chinese for administration; an independent back-translation by a bilingual translator blind to the original Chinese confirmed semantic equivalence, with the few wording-level discrepancies reconciled by a third party. The bilingual 16-item scale, the English scenario and criterion materials, and the eight items dropped at the EFA stage (with their pre-revision wordings) are provided in the Supplementary Materials, following recommendations for transparent reporting of player-motivation questionnaires (Hughes et al., 2023). One representative item per dimension is shown in Table 1.

3.3.2. Situational choice task

To obtain a criterion that did not share the primary scale's wording, Study 2 included a forced-choice situational task. Each of 10 scenarios described a concrete in-game dilemma and offered four mutually exclusive courses of action, one expressing each VO dimension; the options described only concrete actions and carried no value labels, and respondents were asked to select the single action they would be "most likely to do first." The on-screen order of the four options was randomized across participants (yielding up to 24 distinct layouts per scenario), and responses were scored by the dimension of the chosen option (chosen_dimension), reconstructed from the displayed option order and the chosen position rather than from the on-screen letter. Before any inferential use, this scoring key was verified exhaustively: across all 4,610 scenario responses, every option set was a valid permutation of the four dimensions and every chosen dimension resolved correctly from position and order (Section 5.1). One scenario (a resource-allocation dilemma) was revised during content validation, on the basis of cognitive-interview feedback, from a build-specific framing—which mobile, casual, and narrative-oriented players found unfamiliar—to a generic resource/allocation framing. The task yields, per participant and scenario, a categorical choice among the

four dimensions; it remains a within-questionnaire behavioral-intention measure and is not a log of in-game behavior.

Table 1. Representative Value Orientation items (English source version; sacrificed opponent dimension in brackets).

Dimension	Example item
Efficiency	"When exploring side content would clearly slow my progress, I usually prioritise completing the main objective." [vs. Exploration]
Exploration	"More than proving I can win a fight, I want to know where an untried path leads." [vs. Challenge]
Story	"More than taking on hard challenges, I care about whether I'm truly drawn into the character and story." [vs. Challenge]
Challenge	"When easy completion conflicts with a high-difficulty goal, I lean toward taking on the high-difficulty goal, even if it costs more time." [vs. Efficiency]

3.3.3. Convergent and incremental-baseline measure (BANGS)

The Basic Needs in Games Scale (BANGS; Ballou et al., 2024) provided a conservative, theory-based motivational baseline. We administered its autonomy, competence, and relatedness satisfaction subscales and scored them as three separate subscale means, using neither a total score nor item parcels so that the three needs remained separable. BANGS is openly available for research use and requires no licensing permission (Ballou et al., 2024); we administered a Chinese translation—produced by the same forward-translation and back-translation procedure used for the main scale—of the autonomy, competence, and relatedness satisfaction subscales (three items per subscale, following the published scale), each rated on the scale's standard seven-point response format and scored as a subscale mean. In the current sample, internal consistency was acceptable for competence ($\alpha = 0.73$) but low for autonomy ($\alpha = 0.47$) and relatedness ($\alpha = 0.42$); the low autonomy and relatedness reliabilities are unsurprising in a predominantly single-player sample and should be borne in mind when interpreting the convergent correlations involving these needs, which they will attenuate.

3.3.4. Convergent–discriminant measure (Hexad)

The Gamification User Types Hexad scale (Marczewski, 2015; Tondello et al., 2016, 2019)—comprising the achiever, philanthropist, socializer, free spirit, player, and disruptor types—was included to map the VO dimensions onto an established additive typology, yielding a pattern of theoretically expected convergent and discriminant correlations. We treated Hexad as a secondary measure and retained it in Study 2. The Hexad scale is a published academic instrument, free to use in research with citation; we administered the 24-item version (Hexad-24; Tondello et al., 2016, 2019)—six player types, four items each—in a Chinese translation produced by the same procedure and scored as six subscale means. Current-sample subscale reliabilities were modest ($\alpha = 0.50$–0.70), consistent with prior reports of limited reliability for the four-item Hexad-24 subscales.

3.3.5. Social desirability

A 13-item social-desirability response-tendency block was included as a discriminant and method-bias marker, on the expectation that VO, as a value-priority construct, should be largely uncorrelated with socially desirable responding. The items were rated on a 1–7 Likert scale and scored as the reverse-keyed item mean (reverse-keyed items: sds_03, sds_06, sds_07, sds_10, sds_12); in the current sample this block had $\alpha = 0.72$, and the reverse-keyed mean is the value analyzed. We emphasize that this block is not the standard short-form Marlowe–Crowne scale (MC-C Form C; Reynolds, 1982), which comprises 13 true/false items scored 0–13; we therefore report it as a continuous response-tendency covariate rather than as an MC-C total. The 13 items were adapted from the MC-C Form C item content, with the original true/false response format replaced by a 7-point Likert scale to increase response variance; because this adaptation changes the scale's psychometric properties, we treat the block as a Likert-adapted social-desirability marker rather than as the MC-C itself.

3.3.6. Criterion measures

Four self-report criterion composites (four items each) indexed the player's actual gaming habits with respect to each priority (efficiency, exploration, story, and challenge), rated on a 1–7 scale (1=never/not at all true, 7=always/completely true). These items described concrete behavioral frequencies ("what I actually do")

rather than priorities under conflict ("what I would trade off"), and were deliberately worded to differ from the VO items, so that they served as continuous criteria distinct in wording from the scale and supported a multitrait-style convergent–discriminant comparison.

3.3.7. Known-groups membership (Study 3)

Study 3 participants were recruited, through hypothesis-blind posts in four real game-community forums, as members of one of four communities—optimisation/strategy, narrative/visual-novel, open-world/sandbox, and high-difficulty/competitive—corresponding respectively to the Efficiency, Story, Exploration, and Challenge dimensions. Several features were built in to make group membership genuine rather than an artifact of the design. Recruitment posts and the consent form never mentioned the predicted community-to-dimension mapping or any expectation that a community would score high on a particular dimension, and neither participants nor front-line recruiters were aware of that mapping. Group assignment was made by the researchers from each participant's self-reported most-invested community type and named representative game (the known_group and representative_game fields), rather than by letting participants select a value label; the optimisation/strategy community, being the most internally heterogeneous, was recruited with more specific sub-anchors (e.g., speedrun, build-optimisation, and 4X strategy forums). Inclusion required regular play over the past 12months, the ability to name a single most-invested genre and representative game, and age ≥ 18; participants with ambiguous membership ("I play everything, no clear preference") were recorded but excluded from the main analysis rather than forced into a group. To avoid demand characteristics, the community-belonging question was administered after the VO scale rather than among the background items, and the strength of belonging (1–7) was measured separately from group assignment, enabling a high-purity sensitivity analysis (belonging ≥ 5).

3.3.8. Demographics and quality controls

All studies collected age band, gender, primary platform, weekly play hours, years of experience, and favored game genres (multiple choice) together with a single forced choice of the respondent's primary genre, which served as the grouping variable for the game-type choice-effect analysis and the game-type invariance test; region was collected in Studies 1 and 2 but not Study 3. Each questionnaire embedded three attention-check items with fixed correct responses (not counted toward any scale) and passively recorded completion time; participants failing two or more attention checks were removed.

3.4. Participants, screening, and procedure

Eligible participants were adults who play digital games. After providing consent, participants completed the relevant questionnaire battery; the retest sub-sample was re-contacted 2–4weeks after Study 2 using an anonymous pairing code. Data quality was screened with a uniform pipeline applied identically across studies: responses were removed for failing attention checks, for implausibly fast completion, for straight-lining, or for periodic response patterns. The screening flags were fixed in the data package and applied without re-derivation, so exclusions were not adjusted to suit results. Drop rates were below the conventional 25% benchmark in every study.

Final analyzed samples were: Study 1, 372 collected and 307 retained (17.5% removed); Study 2, 461 collected and 381 retained (17.4% removed); Study 3, 158 collected and 125 retained (20.9% removed); and the retest, 127 collected, 109 retained after screening, and 97 successfully matched to their Study 2 records.

The three samples were demographically similar. A majority of participants identified as male (Study 1, 66.1%; Study 2, 62.2%; Study 3, 66.4%), with 29.0%, 34.6%, and 28.8% identifying as female and the remainder selecting "other" or declining to disclose (4.9%, 3.2%, and 4.8%). Participants were predominantly under 35 (the 18–24 and 25–34 bands together accounting for 88.6%, 86.4%, and 86.4%). PC was the most common primary platform (45.6%, 42.3%, and 44.8%), followed by mobile (21.2%, 24.7%, and 17.6%). Most participants reported at least three years of gaming experience (61.2%, 65.1%, and 66.4%) and five or more hours of play per week (49.2%, 50.7%, and 56.0%). A full breakdown of all demographic variables by study is provided in the supplementary materials. In Study 3 the four community groups were approximately balanced (optimisation/strategy n=32, narrative/visual-novel n=31, open-world/sandbox n=32, high-difficulty/competitive n=30; all ≥ 30). Study 3 was treated as supplementary evidence, for which

smaller and not-fully-balanced groups are acceptable. The development and validation samples were adequate for their analyses: Study 1 (N=307) and Study 2 (N=381) both exceeded the N>200 guideline for stable maximum-likelihood factor estimation, and the Study-2 sample provided a participant-to-item ratio of approximately 24:1 for the 16-item confirmatory model, well above the commonly cited 10:1–20:1 heuristics; the one exception, the below-target female subgroup in the gender-invariance test, is noted where it arises.

3.5. Content validity

Before the quantitative studies, the 24-item pool and the 10 scenarios underwent a two-stage expert review followed by cognitive interviewing. The two expert stages were deliberately sequenced—the blind stage was completed and returned before the labeled stage was distributed—so that seeing the intended labels could not bias the relevance ratings, which would otherwise render the attribution check circular.

In Stage A, five expert raters sorted each item to one of the four dimensions (or an explicit "cannot decide" option, included so that ambiguous items would not be forced into a category) without seeing its intended label, and indicated a second-choice dimension to surface potential double-loadings. Across the 120 expert–item judgements, the first choice matched the intended dimension in 70.0% of cases, with a 12.5% "cannot decide" rate; per-dimension agreement ranged from 0.63 (Story) to 0.73 (Challenge). In Stage B, raters judged each (now labeled) item's relevance to its dimension on a 1–4 scale, its tradeoff clarity, and any double-loading or wording problems. Using the conventional relevance cutoff (≥ 3) and the conventional thresholds for the item-level and scale-level content validity indices (I-CVI ≥ 0.78; S-CVI ≥ 0.90; Lynn, 1986; Polit & Beck, 2006), expert review served to screen and revise the 24-item development pool. Across the full pool the S-CVI/Ave was 0.83, because several of the weaker candidate items—predominantly the fifth and sixth items written for each dimension, which the pool deliberately included as alternates—did not meet the content-validity standard and were revised or earmarked for replacement under these criteria. Each of the 16 items ultimately retained for the validated scale met the I-CVI criterion (all I-CVI ≥ 0.78), and the 16-item scale that entered formal validation (Studies 2 and 3 and the retest) had an S-CVI/Ave of 0.925, satisfying the 0.90 threshold. A few retained items nonetheless had weaker blind first-choice attribution at Stage A (for example, an Exploration and a Story item at 0.40; Supplementary Table S2); these were retained on the strength of their Study 1 factor loadings, and we flag the item-level unevenness rather than leaving it implicit. The full pool falls below that threshold only because it contains the development alternates that screening removed; we therefore report content validity at the level of the final scale and do not claim that the 24-item pool as a whole reached 0.90.

Eight players spanning hardcore and moderate segments then completed think-aloud cognitive interviews (Willis, 2005) with structured probes (paraphrasing each item, naming the tradeoff it posed, justifying the chosen response, checking option comprehension, and rating scenario familiarity). Items were correctly understood in 93.8% of presentations and the intended tradeoff was correctly perceived in 92.2%; the few comprehension or wording problems led to minor rewording, and one scenario was revised to a more generic framing (Section 3.3.2). We distinguished genuine low endorsement from comprehension failure when deciding what to revise—for example, a Challenge item rated low but understood was retained as a true individual difference, whereas an item that players could not understand was revised. Particular attention was paid to the Efficiency–Challenge and Exploration–Story boundaries, which the blind classification and interviews confirmed were not systematically confused. The items judged weakest at review overlapped substantially with those subsequently dropped by the Study 1 EFA.

3.6. Analytic strategy

All primary analyses were conducted in R (R Core Team, 2023), using lavaan (Rosseel, 2012), semTools, psych, geepack (Halekoh et al., 2006), nnet, and nlme; Python was used only for cross-checking, and any discrepancy was resolved in favor of R.

3.6.1. Study 1

We assessed factorability (Kaiser–Meyer–Olkin index, Bartlett's test, per-item sampling adequacy), determined the number of factors by parallel analysis (Horn, 1965; 1,000 iterations, maximum-likelihood

extraction) corroborated by Velicer's minimum average partial (MAP) test (Velicer, 1976), and extracted factors by maximum likelihood with oblimin rotation, following recommended practice for exploratory factor analysis (Costello & Osborne, 2005; Fabrigar et al., 1999; Watkins, 2018). The 24-item pool was reduced to a 16-item scale (four items per dimension) using retention rules (primary loading, cross-loading, communality, and primary-minus-cross gap), and the reduced scale's reliability was indexed by Cronbach's α (Cronbach, 1951) and McDonald's ω (Dunn et al., 2014; McDonald, 1999).

3.6.2. Study 2

Confirmatory factor analysis (Brown, 2015; Kline, 2016) used the robust maximum-likelihood (MLR) estimator as the primary estimator, with the weighted-least-squares-mean-and-variance (WLSMV) estimator as an ordinal-treatment robustness check (Li, 2016). Model fit was evaluated against conventional criteria (CFI and TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08; Hu & Bentler, 1999). The four-factor model was compared against three alternatives: a two-factor model collapsing Efficiency with Challenge ("instrumental") and Exploration with Story ("experiential"), a single-factor model, and a second-order model positing one general factor above the four first-order factors (with a bifactor model attempted as a further check). A three-factor model merging Efficiency and Exploration, the most strongly correlated pair, was tested post hoc (Section 5.2). Reliability was indexed by α, ω (composite reliability), and AVE. Discriminant validity was assessed by inter-factor correlations, the heterotrait–monotrait ratio (HTMT, criterion < 0.85; Henseler et al., 2015), and the Fornell–Larcker criterion (Fornell & Larcker, 1981). Convergent and discriminant relations were examined against BANGS, Hexad, and the social-desirability marker. Criterion validity was assessed in three ways: a multitrait-style 4×4 matrix (Campbell & Fiske, 1959) relating the VO dimensions to the four continuous criteria; the situational choice task, analyzed with one binary generalized estimating equation per dimension (Liang & Zeger, 1986; geepack; predictor=the matched VO score, covariates=the remaining three VO scores, scenario identity as fixed effects, exchangeable working correlation, clustered by participant), with a trial-level multinomial model as a supplement; and a choice-effect analysis relating VO scores to players' primary game genre by multinomial regression. Incremental validity (Hunsley & Meyer, 2003) was tested with hierarchical regressions in which the three BANGS subscales were entered first and the four VO scores second. A more direct format comparison re-estimated the situational-task GEE models with the criterion composites (an additive measure of the same four domains) entered alongside the VO scores, testing whether the tradeoff format predicts choices beyond an additive measure of the same content. Measurement invariance (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000) was tested across gender (primary), gaming experience (secondary), and game-type orientation (conditional, requiring each group ≥ 150), progressing from configural to metric to scalar models and judged by ΔCFI < 0.01 and ΔRMSEA < 0.015 (Chen, 2007; Cheung & Rensvold, 2002). Scaled χ^2 difference tests and ΔSRMR are reported for each step, and, because the female group was smaller than recommended, a post hoc Monte Carlo sensitivity analysis estimated the size of violation the gender comparison could detect (Section 5.9).

3.6.3. Retest

Temporal stability was indexed by the intraclass correlation coefficient (ICC(2,1); two-way random effects, absolute agreement, single rater; Shrout & Fleiss, 1979; Koo & Li, 2016) for each dimension, with 0.70 as the target and 0.50 as the floor; values below the floor were to be described as "modest stability" rather than treated as a hard failure.

3.6.4. Study 3

Known-groups evidence comprised within-group profile comparisons (each group's matched dimension versus its other three, by paired t tests), between-group comparisons on each dimension (Welch's ANOVA with Games–Howell post-hoc tests), a mixed-effects model testing the group×dimension interaction, and a sensitivity analysis restricted to high-belonging members.

3.6.5. Multiple comparisons and the validation criterion

Within each family of related follow-up tests, p values were controlled by the Benjamini–Hochberg false-discovery-rate procedure at q=0.05 (Benjamini & Hochberg, 1995). The Study 2 CFA is a single, theory-

derived test of the structure established in Study 1—a cross-validation in an independent sample—rather than one of a family of exploratory comparisons, and so is reported without FDR adjustment; all p values, including non-significant ones, are reported throughout. Our central criterion for regarding the scale as validated was replication of the four-factor structure under CFA in the independent Study 2 sample. We report every validity analysis regardless of its outcome, and where the external-validity evidence is weak we downgrade the corresponding conclusion rather than omitting the result. We did not register these criteria in advance, and we make no confirmatory claim beyond what the independent-sample design itself supports.

4. Study 1: Scale development

Study 1 examined the latent structure of the 24-item pool in the development sample (N=307) and reduced it to a 16-item scale.

4.1. Item distributions and factorability

All 24 items were approximately symmetric ($|\text{skew}| \leq 0.54$, $|\text{kurtosis}| \leq 1.24$), so a Pearson-based exploratory factor analysis was appropriate and polychoric methods were not required. The correlation matrix was clearly factorable: the Kaiser–Meyer–Olkin index was 0.81, Bartlett's test of sphericity was significant, $\chi^2(276) = 1{,}883.1$, $p<0.001$, and every item's measure of sampling adequacy exceeded 0.50.

4.2. Number of factors

Parallel analysis (1,000 iterations, maximum-likelihood extraction) and Velicer's MAP test both indicated four factors. The first four observed eigenvalues (3.44, 1.94, 1.64, 0.91) each exceeded their 95th-percentile random reference (0.71, 0.48, 0.41, 0.35), whereas the fifth (0.23) fell below its reference (0.30). The fourth factor's eigenvalue was below the Kaiser "greater-than-one" rule but well above the parallel-analysis reference, which is the more accurate retention criterion; the two recommended criteria thus converged on four. A four-factor solution also fit substantially better than a three-factor solution (TLI = 0.976 vs. 0.825; RMSEA = 0.021 vs. 0.057), while a five-factor solution was over-extracted (TLI > 1, an inadmissible value). We therefore retained four factors.

4.3. The four-factor structure

Under maximum-likelihood extraction with oblimin rotation, the four factors corresponded cleanly to the intended dimensions. The four items written for each dimension loaded on their target factor in the 0.54–0.77 range, with all non-target (cross-) loadings ≤ 0.13 (Table 2). The factors were obliquely related but distinguishable: the largest inter-factor correlation was between Efficiency and Exploration (r=0.34), followed by Exploration–Story (r=0.24); the remaining correlations were ≤ 0.11 (Table 3). The oblique rotation was therefore justified, and the modest correlations foreshadowed the discriminant-validity tests of Study 2. The four-factor solution accounted for 37.6% of the variance among the 24 items.

Notably, although Efficiency and Exploration were the most strongly related pair, neither the Efficiency–Challenge correlation (r=0.11) nor the Exploration–Story correlation (r=0.24) was large, providing an early indication that the "instrumental" (Efficiency+Challenge) and "experiential" (Exploration+Story) groupings—tested formally as a competing model in Study 2—lacked strong empirical support.

4.4. Item reduction (24 → 16)

We applied the retention rules (primary loading ≥ 0.50, maximum cross-loading < 0.30, primaryminus-cross gap ≥ 0.20, communality ≥ 0.30) and retained, for each dimension, the four best-performing items. This data-driven selection reproduced the frozen 16-item set exactly: the eight items removed were the two weakest items in each dimension, which were precisely the items with sub-threshold primary loadings (0.28–0.49), low communalities, or the largest cross-loadings—including a candidate Challenge item that had been flagged during item development as needing EFA adjudication. A re-run of the EFA on the 16 retained items reproduced the clean four-factor structure with improved fit (TLI = 0.986, RMSEA = 0.021) and accounted for 46.6% of the variance.

4.5. Reliability of the 16-item scale

Internal consistency was acceptable for all four subscales: Efficiency α = 0.785, ω = 0.791; Exploration α = 0.738, ω = 0.744; Story α = 0.774, ω = 0.779; Challenge α = 0.769, ω = 0.770. McDonald's ω equaled or exceeded Cronbach's α for every dimension. Although we had anticipated that the tradeoff wording would depress these coefficients, the values were comfortably above the conventional 0.70 threshold, indicating that the anticipated suppression, while a real property of the format, did not materially compromise reliability in this pool.

Table 2. Retained 16-item four-factor pattern (Study 1; oblimin rotation, maximum likelihood).

Dimension	Item	Primary loading	Max cross-loading	h^2
Efficiency	EFF1	0.77	0.07	0.64
	EFF3	0.67	0.05	0.45
	EFF4	0.67	0.04	0.45
	EFF2	0.59	0.06	0.38
Exploration	EXP1	0.70	0.08	0.47
	EXP3	0.67	0.13	0.51
	EXP2	0.63	0.06	0.41
	EXP4	0.54	0.08	0.35
Story	STO2	0.73	0.11	0.59
	STO1	0.72	0.02	0.51
	STO3	0.65	0.05	0.43
	STO4	0.65	0.09	0.39
Challenge	CHA1	0.77	0.05	0.58
	CHA3	0.65	0.03	0.42
	CHA4	0.64	0.03	0.42
	CHA2	0.63	0.02	0.40

Note: The full 24-item pattern, including the eight dropped items, appears in Supplementary Table S1.

Table 3. Inter-factor correlations (Study 1).

	EFF	EXP	STO	CHA
EFF	—			
EXP	0.34	—		
STO	−0.04	0.24	—	
CHA	0.11	0.09	0.11	—

4.6. Preliminary method-factor check

Because each item names an "opponent" dimension that is sacrificed, an item could in principle load on the dimension it sacrifices, creating a method factor. Among the 16 retained items, however, loadings on the sacrificed dimension were uniformly negligible (−0.08 to 0.07); the only items showing a non-trivial loading on their opponent dimension (around 0.25–0.33) were among the eight dropped at the reduction stage. One asymmetry was carried forward as a check for Study 2: across the retained items, Efficiency was the sacrificed opponent five times, more often than Story (twice), Exploration, or Challenge (three times each). We therefore ran a targeted test in Study 2 of whether the five items that sacrifice Efficiency share systematic residual covariance.

5. Study 2: Validation

Study 2 was the primary validation study, conducted in an independent sample (N=381). It tested the confirmatory structure, reliability, discriminant and convergent validity, criterion validity, incremental validity over an established needs measure, and measurement invariance; a retest sub-sample provided evidence of temporal stability.

5.1. Verification of the situational-task scoring key

Because the situational choice task is scored by the dimension of the chosen option rather than by its on-screen position, we first verified the scoring key exhaustively. Across all 4,610 scenario responses (the full as-collected long file of 461 collected participants × 10 scenarios, prior to participant screening), every option set was a valid permutation of the four dimensions, and the recorded chosen_dimension matched the dimension implied by the displayed option order and the chosen position in 4,610 of 4,610 cases. After

participant screening, the 3,810 responses contributed by the 381 retained participants (381×10 scenarios) entered the choice models. The four dimensions were chosen at broadly comparable rates (Exploration 27.7%, Efficiency 26.2%, Challenge 26.0%, Story 20.1%). The scoring key was therefore sound, and the situational analyses below proceed on the dimension-level codes.

5.2. Confirmatory factor structure

The four-factor model, estimated with the robust maximum-likelihood (MLR) estimator, fit the data well: $\chi^2(98) = 108.8$, p=0.215, CFI = 0.992, TLI = 0.990, RMSEA = 0.017 (90% CI [0.000, 0.034]), SRMR=0.032. All standardized loadings were substantial and significant, ranging from 0.557 to 0.771 (all p<0.001; Table 4). Treating the items as ordinal and re-estimating with the WLSMV estimator yielded virtually identical fit (CFI = 0.994, TLI = 0.992, RMSEA = 0.019, SRMR = 0.033), indicating that the solution did not depend on the continuous treatment of the response scale.

To test whether the four dimensions could be reduced, we compared the four-factor model against two reduced alternatives: a two-factor model collapsing Efficiency with Challenge ("instrumental") and Exploration with Story ("experiential"), and a single-factor model. Both alternatives fit poorly and far worse than the four-factor model (Table 5). The two-factor model produced CFI = 0.598 and RMSEA = 0.122 (ΔCFI relative to the four-factor model = 0.394), and the single-factor model CFI = 0.469 and RMSEA = 0.139 (ΔCFI = 0.523); the Akaike and Bayesian information criteria likewise favored the four-factor model by a wide margin, and a likelihood-ratio test rejected the single-factor restriction, $\Delta\chi^2 = 2{,}693.8$, Δdf=6, p < 0.001. The instrumental/experiential collapse, in particular, was decisively rejected, indicating that Efficiency and Challenge do not merge, nor do Exploration and Story.

Table 4. Standardized CFA loadings (Study 2, MLR).

Dimension	Item	λ (std.)	Dimension	Item	λ (std.)
Efficiency	EFF1	0.72	Story	STO1	0.65
	EFF2	0.64		STO2	0.70
	EFF3	0.70		STO3	0.56
	EFF4	0.62		STO4	0.68
Exploration	EXP1	0.63	Challenge	CHA1	0.77
	EXP2	0.61		CHA2	0.65
	EXP3	0.63		CHA3	0.61
	EXP4	0.66		CHA4	0.67

Table 5. Confirmatory model comparison (Study 2, MLR).

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	AIC	BIC	ΔCFI
Four-factor	108.8	98	0.992	0.990	0.017	0.032	23,234	23,384	—
Second-order (one general factor)	131.1	100	0.978	0.973	0.029	0.050	23,253	23,395	0.014
Three-factor (Efficiency + Exploration)	293.0	101	0.865	0.839	0.071	0.067	23,413	23,551	0.127
Two-factor (instrumental/experiential)	706.9	103	0.598	0.531	0.122	0.112	23,791	23,922	0.394
One-factor	870.2	104	0.469	0.388	0.139	0.125	23,980	24,106	0.523

Note: The three-factor model was tested post hoc; χ^2 values are MLR-scaled.

We also tested a second-order model, in which a single general Value-Orientation factor accounts for the four first-order factors. It fit significantly worse than the correlated four-factor model (Table 5; CFI = 0.978 versus 0.992, ΔCFI = 0.014; scaled $\Delta\chi^2(2) = 24.4$, p<0.001; AIC and BIC both higher), and a bifactor model failed to reach an admissible solution (it did not converge and produced negative residual variances). The second-order misfit is diagnostic rather than incidental: the inter-factor correlations are too heterogeneous for a single common factor to reproduce. Efficiency and Story correlate only 0.10, whereas Efficiency and Exploration correlate 0.49, so no general factor can account for both at once; consistent with this, the implied second-order loadings ranged widely (0.38 to 0.83). This pattern matches the construct's logic. Value Orientation is a set of competing priorities, not a set of facets of one "more-is-better" engagement dimension, so a reflective higher-order factor is neither theoretically apt nor empirically supported, and the correlated four-factor model is retained.

Because Efficiency and Exploration were the most strongly correlated pair, we also tested post hoc whether they form a single dimension. A three-factor model merging them fit markedly worse than the four-factor model (CFI = 0.865, RMSEA = 0.071, SRMR = 0.067; scaled $\Delta\chi^2(3) = 1{,}088.2$, p<0.001; AIC and BIC both

higher; Table 5). The two dimensions share variance, which we examine in Section 7.3, but they do not collapse into one.

Because Study 1 had flagged Efficiency as the most frequently sacrificed opponent dimension, we tested directly whether the five items that sacrifice Efficiency (EXP1, EXP4, STO1, STO4, CHA1) shared systematic residual covariance. They did not: among these items the largest absolute residual correlation was 0.037 (mean = 0.021), and no modification index for a residual covariance among them reached the conventional threshold of 10. The opponent-rotation concern raised in Study 1 was thus monitored and found not to manifest as a method factor.

This replication of the four-factor structure in an independent sample met our central criterion for a successful validation (Section 3.6). We therefore proceeded to the remaining validity tests.

5.3. Reliability and average variance extracted

Internal consistency was acceptable for all four subscales: Efficiency $\alpha = 0.761$, $\omega = 0.765$; Exploration $\alpha = 0.723$, $\omega = 0.725$; Story $\alpha = 0.730$, $\omega = 0.734$; Challenge $\alpha = 0.770$, $\omega = 0.774$ (Table 6). Average variance extracted (AVE), by contrast, fell below the conventional 0.50 benchmark for every dimension (0.40–0.47). This pattern was anticipated on theoretical grounds: forcing a relative, opportunity-cost judgment removes the shared response variance that ordinarily inflates AVE, so AVE is expected to be conservative for tradeoff items. We therefore treat ω and the CFA loadings—not AVE—as the primary reliability evidence (Hayes & Coutts, 2020; McDonald, 1999), and we note below that the sub-threshold AVE did not undermine discriminant validity.

5.4. Discriminant validity

The four dimensions were empirically distinguishable. Inter-factor correlations were modest, the largest being Efficiency–Exploration (r=0.49); all heterotrait–monotrait (HTMT) ratios were well below the 0.85 criterion, the largest again being Efficiency–Exploration (0.48; Table 7). The Fornell–Larcker criterion was satisfied for every dimension: each dimension's √AVE (0.63–0.68) exceeded all of its correlations with the other factors. Thus, although AVE was below 0.50, discriminant validity held, because the inter-factor correlations were themselves low—a reminder that the AVE benchmark and the Fornell–Larcker comparison can diverge when correlations are modest.

Table 6. Reliability and average variance extracted (Study 2).

Dimension	α	ω (CR)	AVE	√AVE
Efficiency	0.761	0.765	0.456	0.675
Exploration	0.723	0.725	0.402	0.634
Story	0.730	0.734	0.407	0.638
Challenge	0.770	0.774	0.467	0.683

Table 7. Discriminant validity (Study 2): factor correlations (above diagonal), HTMT (below diagonal), √AVE (diagonal, bold).

	EFF	EXP	STO	CHA
EFF	**0.68**	0.49	0.10	0.26
EXP	0.48	**0.63**	0.39	0.23
STO	0.08	0.39	**0.64**	0.35
CHA	0.24	0.21	0.35	**0.68**

Table 8. Convergent and discriminant correlations of VO with external measures (Study 2).

	BANGS-Aut	BANGS-Com	BANGS-Rel	Hexad (peak)	Social desirability
Efficiency	0.10	**0.31**	0.05	achiever 0.41	0.04
Exploration	**0.26**	0.05	0.12	free spirit 0.30	0.06
Story	**0.25**	0.02	0.26	philanthropist 0.44	0.04
Challenge	0.10	**0.35**	0.04	achiever 0.27	−0.12

Bold values mark the BANGS need theoretically expected to relate most strongly to each VO dimension (competence for Efficiency and Challenge; autonomy for Exploration and Story).

Table 9. VO×continuous criterion correlations (Study 2). Matched (diagonal) cells in bold.

	crit-EFF	crit-EXP	crit-STO	crit-CHA
VO Efficiency	**0.41**	0.23	0.07	0.10
VO Exploration	0.37	**0.36**	0.30	0.07
VO Story	0.09	0.22	**0.35**	0.12
VO Challenge	0.22	0.07	0.18	**0.37**

5.5. Convergent and discriminant validity with external measures

Relations with established measures followed the expected pattern (Table 8). Against BANGS, the two skill-oriented dimensions correlated most strongly with the competence need (Efficiency r=0.31; Challenge r=0.35), whereas the two experience-oriented dimensions correlated with autonomy (Exploration r=0.26; Story r=0.25, and Story about equally with relatedness, r=0.26). Against Hexad, Efficiency related most to the achiever type (r=0.41), Exploration to free spirit (r=0.30), Story to philanthropist and socializer (r=0.44 and 0.36), and Challenge to achiever and disruptor (r=0.27 and 0.24). These moderate, theory-consistent associations indicate that VO is related to, but distinct from, existing needs and player-type constructs. Several of the external subscales had modest internal consistency in this sample (notably BANGS autonomy and relatedness, α = 0.47 and 0.42, and several Hexad subscales, α = 0.50–0.63); because low reliability attenuates correlations, the convergent associations reported here should be read as lower bounds.

As a discriminant check against impression management, correlations between the VO dimensions and the social-desirability marker were uniformly small ($|r| \leq 0.12$) and none was significant after false-discovery-rate correction, indicating that VO scores are not appreciably contaminated by socially desirable responding.

5.6. Criterion validity

5.6.1. Continuous criteria

We related the four VO dimensions to four continuous criterion composites indexing typical behavioral tendency toward each priority (Table 9). For three of the four dimensions, the matched (diagonal) correlation was the largest in its row: Efficiency (matched r=0.41 vs. $\leq$ 0.23 off-diagonal), Story (0.35 vs. $\leq$ 0.22), and Challenge (0.37 vs. $\leq$ 0.22). For Exploration, the matched correlation (0.36) was essentially tied with—and marginally lower than—its correlation with the efficiency criterion (0.37); this single exception is consistent with the elevated Efficiency–Exploration relationship documented throughout and does not generalize to the other dimensions. Read by column, the convergent– discriminant pattern was clean for all four criteria: each criterion's strongest VO predictor was its matched dimension. All matched correlations were significant after false-discovery-rate correction.

5.6.2. Situational choice task

The situational task provided a criterion that did not share the scale's wording. For each dimension we fit a binary generalized estimating equation predicting whether the dimension's option was chosen, with the matched VO score as the focal predictor, the remaining three VO scores as covariates, scenario identity as fixed effects, an exchangeable working correlation, and clustering by participant. In every model the matched VO score significantly and positively predicted choosing its own option, and was the strongest positive predictor among the four scores (Table 10): Efficiency, odds ratio (OR) = 1.21 (95% CI [1.12, 1.29]); Exploration, OR = 1.31 [1.22, 1.42]; Story, OR = 1.19 [1.11, 1.29]; Challenge, OR = 1.25 [1.18, 1.34]; all $q<0.001$ after false-discovery-rate correction. A trial-level multinomial model yielded the same pattern. The tradeoff scale thus predicted tradeoff choices on a task with concrete, differently worded content—though, as noted in the Method, this remains a within-questionnaire choice rather than a record of in-game behavior.

5.6.3. Choice effect: Alignment with preferred game genre

We then asked whether players' VO scores aligned with the genre they reported playing most, using multinomial regression with the exploration-leaning genre as the reference category (Table 11). Each genre group was elevated, relative to the reference, on its theoretically matched dimension: the challenge-leaning group on Challenge (OR per SD=1.93, p=0.002), the story-leaning group on Story (OR = 2.11, $p<0.001$), and

the efficiency-leaning group on Efficiency (OR = 1.47, p=0.072); reciprocally, all three non-reference groups scored lower than the exploration-leaning reference on Exploration. The efficiency-group effect was in the predicted direction but only marginally significant, plausibly reflecting the smaller and more heterogeneous efficiency group (n=75). This pattern is consistent with players gravitating to genres that match their value orientation, but because the design is cross-sectional we interpret it as a selection-consistent association rather than evidence of a causal effect.

5.7. Incremental validity over BANGS

We tested whether VO predicted the criteria beyond the conservative motivational baseline. In hierarchical regressions predicting each matched criterion, the three BANGS subscales were entered first and the four VO scores second (Table 12). The VO scores added significant incremental variance for every criterion (ΔR^2 = 0.10–0.19, all q<0.001), and in each model the matched VO score was the strongest predictor among the four. VO therefore captures dimension-specific variance in players' priorities that an established needs measure does not.

Table 10. Situational-choice GEE: effect of the matched VO score on choosing its own option (Study 2).

Choice	Focal VO	*B*	*SE*	OR [95% CI]	*q* (FDR)
Efficiency option	VO Efficiency	0.187	0.036	1.21 [1.12, 1.29]	< 0.001
Exploration option	VO Exploration	0.272	0.039	1.31 [1.22, 1.42]	< 0.001
Story option	VO Story	0.177	0.038	1.19 [1.11, 1.29]	< 0.001
Challenge option	VO Challenge	0.227	0.033	1.25 [1.18, 1.34]	< 0.001

Table 11. Choice effect: multinomial OR per SD of VO, relative to the exploration-leaning reference group (Study 2).

Genre group (vs. Exploration)	VO-EFF	VO-EXP	VO-STO	VO-CHA
Challenge-leaning (*n* = 144)	1.05	0.57	0.95	**1.93**
Efficiency-leaning (*n* = 75)	**1.47**	0.48	1.28	0.54
Story-leaning (*n* = 99)	0.84	0.62	**2.11**	0.64

The exploration-leaning reference group comprised 51 participants; a residual mixed-genre group (n=12) was excluded from this analysis. Bold values mark each genre group's theoretically matched VO dimension.

Table 12. Incremental validity of VO over BANGS subscales (Study 2).

Criterion	R^2 (BANGS)	R^2 (+ VO)	ΔR^2	ΔF	Matched VO β	*q* (FDR)
Efficiency	0.057	0.249	0.192	23.9	0.282	< 0.001
Exploration	0.055	0.174	0.119	13.4	0.267	< 0.001
Story	0.059	0.184	0.125	14.3	0.244	< 0.001
Challenge	0.043	0.147	0.104	11.4	0.335	< 0.001

5.8. Format comparison: Tradeoff versus additive prediction of situational choices

The incremental-validity analysis above used BANGS—a measure of basic psychological needs—as the baseline. Because BANGS does not itself measure the four VO facets, it cannot distinguish whether the tradeoff format or merely the four-domain content drives the predictive gain. We therefore conducted a more direct test using the criterion composites (Section 3.3.6) as an additive competitor: these composites index the same four domains (efficiency, exploration, story, and challenge) in an additive, behavioral-tendency format that carries no opportunity cost and permits all four scores to be high simultaneously—exactly the contrast the tradeoff argument requires.

For each dimension we re-estimated the binary GEE from the situational-task analysis (same specification: scenario fixed effects, exchangeable working correlation, clustered by participant) under three conditions: (A) the four criterion-composite scores as the sole predictors, (B) the four VO scores as the sole predictors, and (C) both sets entered simultaneously (Table 12b). In every model the matched predictor was the focal coefficient; the remaining seven scores served as covariates.

The critical result is in Model C. For all four dimensions, the matched VO score remained a significant positive predictor of choosing its own option after the additive composites were controlled: Efficiency OR = 1.14, Exploration OR = 1.30, Story OR = 1.14, Challenge OR = 1.22 (all q≤0.001 after false-discovery-rate correction across the eight Model-C tests). The tradeoff scale thus predicts tradeoff choices over and above an additive measure of the same four content domains. Conversely, the matched criterion composite also

survived for three of four dimensions (Efficiency, Story, Challenge, all q<0.01), whereas for Exploration the additive composite dropped to non-significance once VO was controlled (q=0.086). This pattern indicates that the two formats are largely complementary—each captures variance the other misses—with the tradeoff format showing a consistent independent contribution and, for Exploration, fully absorbing the additive measure's predictive power. We note that the criterion composites measure behavioral tendency ("what I actually do") rather than stated preference ("how much I value X"), so this comparison is between tradeoff priority and additive behavioral tendency in the same domains rather than between two pure preference formats; a head-to-head with an additive preference measure of the same content would provide an even stricter test.

5.9. Measurement invariance

We tested invariance across gender (primary), gaming experience (secondary), and game-type orientation (conditional). Full scalar invariance held in all three cases, with every step satisfying the ΔCFI < 0.01 and ΔRMSEA < 0.015 criteria (Table 13); no scaled χ^2 difference test was significant (all p≥0.17), and no ΔSRMR exceeded 0.006. For game-type orientation, both groups met the conditional ≥ 150 requirement (experiential n=150; instrumental n=219), so the full three-level test is reported as planned rather than as exploratory. These results support the comparability of the VO measurement structure across gender, experience, and the instrumental/experiential divide.

The female group (n=132) was smaller than the 150 per group often recommended for invariance testing. For gender, the observed changes (ΔCFI=−0.001 at the metric step and +0.001 at the scalar step) also met the stricter criterion of 0.002 proposed by Meade et al. (2008). A post hoc Monte Carlo sensitivity analysis, based on the standardized four-factor solution and the observed group sizes (1,000 replications per condition), showed that the scaled χ^2 difference test held its nominal error rate (4–6%) and would have detected an intercept difference of 0.35 SD on two items in 95% of replications and a loading difference of 0.30 on two items in 67% (full results in the Supplementary Materials).

Table 12b. Format comparison: VO (tradeoff) versus criterion composites (additive) predicting situational choices (Study 2). Each row reports the matched predictor for that dimension.

Dimension	A: Additive only OR	B: VO only OR	C: VO OR [95% CI]	C: VO q	C: Additive OR [95% CI]	C: Additive q
Efficiency	1.25	1.21	1.14 [1.06, 1.23]	0.001	1.21 [1.09, 1.34]	0.001
Exploration	1.17	1.31	1.30 [1.20, 1.41]	<0.001	1.08 [0.99, 1.19]	0.086
Story	1.33	1.19	1.14 [1.05, 1.22]	0.001	1.27 [1.14, 1.42]	<0.001
Challenge	1.27	1.25	1.22 [1.14, 1.31]	<0.001	1.16 [1.06, 1.26]	0.001

Models: (A) additive composites only, (B) VO only, (C) both entered simultaneously. ORs are per 1-point increase; q-values are FDR-corrected across the eight model-C tests.

Table 13. Measurement invariance (Study 2, MLR; robust CFI and RMSEA).

Grouping	Step	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR	$\Delta\chi^2$ (12)	*p*
Gender	Configural	0.977	0.029	0.047	—	—	—	—	—
female 132, male 237	Metric	0.976	0.029	0.053	−0.001	0.000	+0.006	13.55	0.331
	Scalar	0.977	0.028	0.054	+0.001	−0.001	+0.001	10.77	0.549
Experience	Configural	0.991	0.019	0.042	—	—	—	—	—
newer 244, veteran 137	Metric	0.992	0.017	0.046	+0.001	−0.002	+0.004	10.45	0.577
	Scalar	0.995	0.014	0.046	+0.003	−0.004	+0.001	7.72	0.807
Game type	Configural	0.967	0.036	0.050	—	—	—	—	—
experiential 150, instrumental 219	Metric	0.964	0.037	0.057	−0.003	0.000	+0.006	16.49	0.170
	Scalar	0.962	0.037	0.058	−0.002	0.000	+0.001	14.94	0.245

Note: Group sizes sum to 369 rather than 381 for gender and for game type because each test excludes the participants it cannot assign to one of its two groups: for gender, the 12 participants who selected "other" or declined to disclose; for game type, the 12 participants whose single most-played genre was "other," which the experiential/instrumental mapping does not classify. These are two largely distinct subsets (one participant falls in both). The experience grouping requires no such exclusion (244+137=381). Δ values are differences from the preceding step; $\Delta\chi^2$ is the Satorra–Bentler scaled difference test.

5.10. Temporal stability

A retest sub-sample of 97 participants completed the scale again 2–4weeks later. Single-rater, absolute-agreement intraclass correlations (ICC(2,1)) were above the 0.70 target for all four dimensions: Efficiency 0.75 (95% CI [0.65, 0.83]), Exploration 0.75 [0.65, 0.82], Story 0.79 [0.70, 0.86], and Challenge 0.81 [0.73,

0.87]. The VO dimensions thus show good temporal stability, supporting their interpretation as trait-like positions rather than transient states.

6. Study 3: Known-groups validation

Study 3 provided supplementary known-groups evidence using players (N=125) recruited as members of four game communities: optimisation/strategy (matched to Efficiency, n=32), narrative/visual-novel (Story, n=31), open-world/sandbox (Exploration, n=32), and high-difficulty/competitive (Challenge, n=30). Group membership was assigned by the researchers from each participant's reported community and representative game, under recruitment and consent materials that never revealed the predicted community-to-dimension mapping (Section 3.3.7); the between-group differences below are therefore unlikely to reflect demand characteristics.

The between-group evidence was clean. On each VO dimension, the theoretically matched group scored highest, and all four one-way Welch analyses were significant after false-discovery-rate correction: Efficiency, $F(3, 66.8) = 3.58$, $\eta^2 = 0.075$; Exploration, $F(3, 66.7) = 4.05$, $\eta^2 = 0.082$; Story, $F(3, 66.8) = 6.56$, $\eta^2 = 0.164$; Challenge, $F(3, 67.1) = 4.44$, $\eta^2 = 0.090$ (all $q<0.02$; Table 14). Games– Howell post-hoc tests confirmed that each matched group exceeded the others, with one exception: the open-world/sandbox group did not significantly exceed the narrative/visual-novel group on Exploration ($p=0.35$). A mixed-effects model corroborated the overall pattern: the group×dimension interaction was significant, $F(9, 363) = 7.07$, $p<0.0001$, while the group main effect was not ($p=0.35$), indicating that the groups differed in their profile across dimensions rather than in overall response level. The pattern was robust in a sensitivity analysis restricted to high-belonging members (belonging ≥ 5; n=93), in which the matched group remained highest and significant on all four dimensions.

The within-group profile evidence—whether each group's members rated their matched dimension above their own other three dimensions—was largely but not entirely supportive: 9 of the 12 paired contrasts were significant after correction. The optimisation/strategy and high-difficulty/competitive groups showed the predicted peak on all three contrasts (Cohen's d=0.45–0.98), and the narrative/visual-novel group on two of three (the contrast against Challenge was marginal, q=0.055). The open-world/sandbox group was the exception: although it scored highest on Exploration between groups, its members did not rank Exploration clearly above Challenge within their own profile (Exploration was not significantly above Efficiency or Challenge, and the Exploration–Challenge difference was negligible). This is consistent with the heterogeneity of the sandbox community, in which open-ended exploration and survival-style difficulty often co-occur. By design, Study 3 is supplementary; its smaller groups and the single weak profile do not bear on the four-factor structure or the criterion evidence established in Study 2, but they are reported transparently.

Table 14. Study 3 group means by VO dimension (matched-group cells in bold).

Community group	EFF	EXP	STO	CHA
Optimisation/strategy	**5.06**	3.51	3.91	4.26
Narrative/visual-novel	4.28	4.07	**5.16**	4.44
Open-world/sandbox	4.29	**4.45**	3.89	4.52
High-difficulty/competitive	4.28	3.59	3.53	**5.27**

7. General discussion

Across three studies and a retest we developed and validated the Value Orientation (VO) scale, a 16-item measure of the priorities players express when in-game goals conflict, organized around four dimensions: Efficiency, Exploration, Story, and Challenge. Study 1 established a clean four-factor structure in a development sample and reduced a 24-item pool to 16 items by standard item-retention rules. Study 2, in an independent sample, replicated the four-factor structure with excellent fit, decisively rejected models that collapsed the dimensions, and produced supporting evidence of reliability, discriminant and convergent validity, criterion validity, and incremental validity over an established needs measure, together with full scalar measurement invariance across gender, experience, and game-type orientation. A retest sub-sample showed good temporal stability, and Study 3 added largely supportive known-groups evidence. Below we

situate these findings, discuss what the tradeoff operationalization buys and costs, consider how stated priorities relate to enacted behavior, and state the boundaries of what can be concluded.

7.1. A tradeoff construct, not an additive one

The central contribution of VO is conceptual as much as psychometric. Existing player-measurement instruments—additive player-type typologies and needs scales—quantify how much a player values each facet of play, allowing all facets to be endorsed simultaneously. VO instead measures which facet a player relinquishes first when facets cannot be jointly satisfied. Two features carry this distinction into the measurement: every primary-scale item embeds an explicit opportunity cost against a named opponent dimension, and a separate situational task forces a single concrete choice among the four priorities. The validation evidence indicates that this operationalization succeeds. The four dimensions formed a replicable structure (Study 2 CFI = 0.992 in an independent sample), and—importantly for a tradeoff account—the dimensions did not collapse into a coarser "instrumental versus experiential" distinction: the two-factor model pairing Efficiency with Challenge and Exploration with Story fit very poorly (ΔCFI = 0.39 relative to the four-factor model). Players' positions on these four priorities are thus empirically separable, not reducible to a single approach style or to two broad orientations.

The most distinctive validity evidence comes from the situational choice task. Each VO dimension predicted, over and above the other three, the likelihood of choosing its own option in concrete, differently worded dilemmas. Because the task shares no wording with the scale, this association cannot be attributed to item overlap, and it provides a behavioral-intention analog to the scale's tradeoff claim: players who position themselves toward a priority in the abstract also choose it when forced to decide in the concrete. The convergent pattern with established constructs further locates VO in the nomological network without redundancy—the skill-oriented dimensions (Efficiency, Challenge) related most to the competence need and the achiever type, and the experience-oriented dimensions (Exploration, Story) to autonomy and to the free-spirit, socializer, and philanthropist types—while the incremental analyses showed that VO predicts players' self-reported behavioral tendencies beyond what these needs explain. These convergent correlations are modest in absolute terms, partly because two of the BANGS subscales were themselves measured with low reliability in this predominantly single-player sample; when the matched VO–BANGS correlations are disattenuated for the reliabilities of both measures, they rise to roughly 0.42 to 0.47, so the modest raw coefficients understate the underlying associations rather than signaling weak convergence.

The incremental and format-comparison analyses answer different questions. The incremental analyses (Section 5.7) predicted the behavioral-tendency composites, that is, what players report habitually doing, and their baseline was a needs measure; they show that VO carries information about habitual play that need satisfaction does not, but they do not set VO against an additive measure of the same four facets. The format comparison (Section 5.8) makes that comparison as far as these data allow. Its criterion is the situational choice task, and its competitor is an additive measure of the same four domains, worded as behavioral tendency ("what I actually do") rather than as absolute preference ("how much I value X"). When the VO scores and the additive criterion composites—which carry no opportunity cost—were entered as competing predictors of the situational choices, every matched VO score remained significant, indicating that the tradeoff format captures variance in players' choices that an additive measure of the same content does not. For Exploration the additive composite was fully absorbed by VO, and for the remaining three dimensions the two formats contributed independently, suggesting complementarity rather than redundancy. This pattern is consistent with the conceptual argument that tradeoff priority and additive behavioral tendency, although related, tap different aspects of the player: the former indexes what is protected under conflict, the latter indexes what is practised in the absence of conflict.

7.2. What the tradeoff format costs, and why that is acceptable

We anticipated, and observed, that the tradeoff format depresses average variance extracted: AVE fell below the conventional 0.50 benchmark for all four dimensions even though ω was acceptable (0.72– 0.77) and the standardized loadings were substantial (0.56–0.77). Maximal reliability (coefficient H; Hancock & Mueller, 2001)—computed from the loadings rather than from extracted variance—was 0.73 to 0.78 across the four dimensions, essentially equal to ω, confirming that reliability indices independent of the AVE computation

are uniformly acceptable and that the sub-threshold AVE reflects the variance-extraction metric specifically rather than unreliable measurement. This is an expected property of forcing a relative judgment, which removes the shared response variance that ordinarily inflates AVE, rather than a sign of weak measurement. Two observations support treating it this way. First, discriminant validity held despite the sub-threshold AVE: every dimension's √AVE exceeded its correlations with the other factors (Fornell–Larcker), and all HTMT ratios were well below 0.85, because the inter-factor correlations were themselves modest. Second, the opponent-rotation discipline appears to have worked: although Efficiency was the most frequently sacrificed opponent across the retained items, the items that sacrifice Efficiency showed no systematic residual covariance in the CFA, so the format did not introduce a method factor. We therefore report ω and the loadings as the primary reliability evidence, with AVE interpreted in light of the format; we do not treat the sub-threshold AVE as evidence for or against convergent validity, which we assess instead from the pattern of theory-consistent external correlations. Researchers adopting tradeoff operationalisations should expect, and plan in advance for, similarly conservative AVE.

7.3. The Efficiency–Exploration relationship

The one relationship that recurred across analyses was between Efficiency and Exploration, which were the most strongly associated pair at every stage (factor r=0.49; HTMT = 0.48), and which produced the single off-diagonal exception in the continuous-criterion matrix, where VO-Exploration correlated marginally more with the efficiency criterion (0.37) than with its own (0.36). We read this relationship as a shared orientation toward progress. Efficiency and Exploration are the only VO dimensions defined in relation to the same feature of a game, the path to the next objective: Efficiency is the priority on following it at least cost, Exploration the priority on leaving it. Players to whom routing decisions are salient therefore hold views on both, whereas players who care more about a game's story or difficulty than about how they move through it have less reason to endorse either. This reading is consistent with the Exploration criterion exception and with evidence that exploration in games depends on whether an explicit goal competes for the player's attention (Gómez-Maureira et al., 2021). The shared orientation does not make the two a single dimension: a model merging them fit markedly worse than the four-factor model (CFI = 0.865 versus 0.992; Section 5.2), their HTMT was far below the 0.85 threshold, and, read by column, each criterion's strongest predictor was its matched dimension. Efficiency and Exploration are thus two directions within a common concern with progression, differing in how a player resolves route conflicts.

7.4. From stated priority to enacted behavior

VO is a stated priority, and the present evidence covers three steps from it (Figure 1): the abstract priority reported on the scale, the choice made in a concrete but hypothetical dilemma, and the habit players report. The step to what players do in a game is the natural next test. In social psychology, intentions predict behavior reliably but imperfectly; the shortfall, known as the intention–behavior gap, arises when intentions are not acted on or are overridden by circumstances (Sheeran, 2002; Sheeran & Webb, 2016). Intention to play is also the outcome in recent models of continued play (Nan et al., 2024; Sun et al., 2024), so the link from stated intention to behavior is a question for the field rather than for VO alone. Two features of play bear on it. A priority can be enacted only where a game offers the corresponding tradeoff, so behavior in any single game reflects its design as well as the player; and because abstract considerations weigh more in hypothetical and distant decisions than in immediate ones (Trope & Liberman, 2010), stated priorities should govern deliberate choices, such as which game to play or which route to plan, more closely than choices made under time pressure. VO should therefore predict logged behavior best in games that leave tradeoffs open and for deliberate decisions, a test made feasible by earlier work linking questionnaire measures to gameplay data (Canossa et al., 2013; Kahn et al., 2015; Melhart et al., 2019).

7.5. Limitations and boundaries of interpretation

Several boundaries constrain what these results license. First, the design is cross-sectional, so the alignment between players' VO scores and their preferred genres (Study 2) is consistent with players selecting genres that match their values, but cannot establish a causal direction; the efficiency-group effect in that analysis was, moreover, only marginal (p=0.07), plausibly reflecting the smaller and more heterogeneous efficiency group; at the observed effect size (OR = 1.47 per SD), the efficiency group's sample of 75 afforded only

modest statistical power, and roughly twice that number would be needed to detect this effect reliably at $\alpha = 0.05$. Second, the situational task, although differently worded from the scale, remains a forced choice within a questionnaire; it is a behavioral-intention measure, not a record of in-game behavior (Section 7.4). Because the scale and the task were both self-reported within the same session, common-method variance may also have inflated their association to some degree; the near-zero correlations between the VO dimensions and the social-desirability marker (all $|r| \leq 0.12$) argue against a strong general response bias, but they cannot rule out common-method influence entirely, and a stronger test would separate the scale and the criterion in time or source. Third, the sample is a Chinese convenience sample—drawn from a general online panel for the main studies and from game-community forums for the known-groups study, but held to a single cultural context by design to avoid confounding culture with platform—which precludes any claim of cross-cultural or global generalizability; the demonstrated invariance is across subgroups within this population, not across cultures. Fourth, the dimensions are correlated, not orthogonal; our claim is that they are distinguishable, not that they are independent, and the Efficiency–Exploration relationship in particular should be kept in view when scores are used. Fifth, the format comparison (Section 5.8) set VO against additive behavioral-tendency items, not additive preference items; a head-to-head with an additive preference measure of the same four facets, worded without opportunity costs, would be the stricter test of the format itself. Sixth, the construct is deliberately bounded to the single-player experience of progression, discovery, narrative, and difficulty, and does not represent the social priorities—cooperation, competition, and sociality— that organize multiplayer play and that social-leaning typologies foreground (Bartle, 1996; Tondello et al., 2016), and in multiplayer role-playing games, players' relationship to their avatars shapes their intention to keep playing (Nan et al., 2024; Sun et al., 2024); a value-orientation account of multiplayer settings would need to model how such social goals trade off against the four dimensions studied here, and we regard extending the construct into the social domain as the most substantial boundary of the present scale rather than a facet it subsumes.

The supplementary known-groups study (Study 3) was largely but not uniformly supportive. The between-group evidence was clean—each community scored highest on its matched dimension on every dimension tested—but the within-group profile evidence held for 9 of 12 contrasts, with the open-world/sandbox community the exception: its members scored highest on Exploration relative to other communities, yet did not rank Exploration clearly above Challenge within their own profiles. This is consistent with the heterogeneity of sandbox play, where open-ended exploration and survival-style difficulty frequently co-occur, and it suggests that "sandbox" is a less pure marker of a single value orientation than the other three communities. Finally, the female subgroup in the gender-invariance test was smaller than recommended; the sensitivity analysis in Section 5.9 indicates that the comparison was well powered to detect moderate intercept differences and less so for loading differences confined to a few items. None of these qualifications bears on the core four-factor structure or the criterion evidence, but each is reported so that the strength of the conclusions is not overstated.

7.6. Future directions and practical implications

The present work was deliberately scoped to the measurement instrument, and several extensions follow naturally. The most direct extension is the stricter format test noted in Section 7.5, which would isolate the contribution of the tradeoff wording itself. Beyond it, linking VO to behavioral telemetry would test the predictions set out in Section 7.4, and a direct measure of attention to routing decisions would test the progress-orientation account in Section 7.3. Because value orientations may shift with context, game, or session, a state-sensitive or dynamic treatment, together with a broader framework integrating these stable orientations with situational variation, is a worthwhile next step that we leave to future work, keeping the present contribution focused on the validated tool. Finally, replication in other cultural contexts and on other platforms is needed before any cross-cultural claim can be made.

Practically, a validated tradeoff measure of player priorities has uses in both research and design. As a research instrument, VO offers a parsimonious way to characterize players by what they protect under conflict, complementing additive typologies. For design, the tradeoff framing maps onto real decisions that games impose—time pressure versus thoroughness, guidance versus discovery, narrative pacing versus difficulty—and a measure of how players resolve such conflicts could inform adaptive systems that tailor those very tradeoffs. We caution, however, that such applications should respect the boundaries above: VO

indexes stated priorities in a Chinese sample, the dimensions are correlated, and the link to in-game behavior remains to be demonstrated.

7.7. Conclusion

Players differ not only in what they enjoy but in what they are willing to give up. The Value Orientation scale operationalizes this tradeoff view across four dimensions and, across three studies, shows a replicable four-factor structure that does not collapse into coarser groupings, acceptable reliability and clear discriminant validity under a format that conservatively depresses variance-based indices, convergent and incremental validity against established measures, criterion validity extending to a differently worded situational choice task, and measurement invariance and temporal stability. With its limitations made explicit, the scale provides a usable tool for studying how players prioritize—and relinquish—competing goals.

Author contributions

CRediT: Ziao Tang: Conceptualization, Data curation, Formal analysis, Project administration, Resources, Software, Visualization, Writing – original draft; Ben Kirman: Methodology, Supervision, Validation, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the China Scholarship Council under Grant 202208060093. The funder had no role in the design of the study, the collection, analysis, or interpretation of the data, or the preparation of this article.

Data availability statement

The data, analysis code, and materials that support the findings of this study are openly available in the Open Science Framework at https://osf.io/rzxwg/.

References

Ballou, N., Denisova, A., Ryan, R. M., Rigby, C. S., & Deterding, S. (2024). The Basic Needs in Games scale (BANGS): A new tool for investigating positive and negative video game experiences. International Journal of Human-Computer Studies, 188, 103289. https://doi.org/10.1016/j.ijhcs.2024.103289

Ballou, N., & Deterding, S. (2023). 'I just wanted to get it over and done with': A grounded theory of psychological need frustration in video games. Proceedings of the ACM on Human-Computer Interaction, 7(CHI PLAY), 217–236. https://doi.org/10.1145/3611028

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. Journal of MUD Research, 1(1), 19. https://mud.co.uk/richard/hcds.htm

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6, 149. https://doi.org/10.3389/fpubh.2018.00149

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105. https://doi.org/10.1037/h0046016

Canossa, A., Martinez, J. B., & Togelius, J. (2013). Give Me a Reason to Dig: Minecraft and Psychology of Motivation [Paper Presentation]. 2013 IEEE Conference on Computational Intelligence in Games (CIG) (pp. 1– 8). IEEE. https://doi.org/10.1109/Cig.2013.6633612

Carstensdottir, E., Kleinman, E., Williams, R., & Seif El-Nasr, M. (2021). “Naked and on Fire”: Examining Player Agency Experiences in Narrative-Focused Gameplay. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21) (pp. 1–13). ACM. https://doi.org/10.1145/3411764.3445540

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research, and Evaluation, 10(7), 1–9. https://doi.org/10.7275/jyj1-4868

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281– 302. https://doi.org/10.1037/h0040957

Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. Harper & Row.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. Plenum Press. https://doi.org/10.1007/978-1-4899-2271-7

Denisova, A., Cairns, P., Guckelsberger, C., & Zendle, D. (2020). Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). International Journal of Human-Computer Studies, 137, 102383. https://doi.org/10.1016/j.ijhcs.2019.102383

DeVellis, R. F. (2017). Scale development: Theory and applications (4th ed.). Sage.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399–412. https://doi.org/10.1111/bjop.12046

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104

Gómez-Maureira, M. A., Kniestedt, I., van Duijn, M., Rieffe, C., & Plaat, A. (2021). Level design patterns that invoke curiosity-driven exploration: An empirical study across multiple conditions. Proceedings of the ACM on Human-Computer Interaction, 5(CHI PLAY), 1–32. https://doi.org/10.1145/3474698

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. Journal of Personality and Social Psychology, 79(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. Journal of Statistical Software, 15(2), 1–11. https://doi.org/10.18637/jss.v015.i02

Hamari, J., & Tuunanen, J. (2014). Player types: A meta-synthesis. Transactions of the Digital Games Research Association, 1(2), 29–53. https://doi.org/10.26503/todigra.v1i2.13

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), Structural equation modeling: Present and future (pp. 195–216). Scientific Software International.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach’s alpha for estimating reliability. But…. Communication Methods and Measures, 14(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. Psychological Bulletin, 74(3), 167–184. https://doi.org/10.1037/h0029780

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, N. G. J., Flockton, J. R., & Cairns, P. (2023). Growing together: An analysis of measurement transparency across 15 years of player motivation questionnaires. International Journal of Human-Computer Studies, 169, 102940. https://doi.org/10.1016/j.ijhcs.2022.102940

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. Psychological Assessment, 15(4), 446–455. https://doi.org/10.1037/1040-3590.15.4.446

Kahn, A. S., Shen, C., Lu, L., Ratan, R. A., Coary, S., Hou, J., Meng, J., Osborn, J., & Williams, D. (2015). The Trojan Player Typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations. Computers in Human Behavior, 49, 354–361. https://doi.org/10.1016/j.chb.2015.03.018

Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. In M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 34, pp. 331–378). Academic Press. https://doi.org/10.1016/S0065-2601(02)80008-9

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior Research Methods, 48(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. Cognitive Science, 5(4), 333–369. https://doi.org/10.1207/s15516709cog0504_2

Marczewski, A. (2015). User types. In Even ninja monkeys like to play: Gamification, game thinking and motivational design (1st ed., pp. 65–80). CreateSpace Independent Publishing Platform.

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology, 93(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Melhart, D., Azadvar, A., Canossa, A., Liapis, A., & Yannakakis, G. N. (2019). Your Gameplay Says It All: Modelling Motivation in Tom Clancy's The Division [Paper Presentation]. In 2019 IEEE Conference on Games (CoG) (pp. 1–8). IEEE. https://doi.org/10.1109/Cig.2019.8848123

Nacke, L. E., Bateman, C., & Mandryk, R. L. (2014). BrainHex: A neurobiological gamer typology survey. Entertainment Computing, 5(1), 55–62. https://doi.org/10.1016/j.entcom.2013.06.002

Nan, D., Sun, S., Jansen, B. J., & Kim, J. H. (2024). Beyond avatar coolness: Exploring the effects of avatar attributes on continuance intention to play massively multiplayer online role-playing games. International Journal of Human–Computer Interaction, 40(23), 8219–8228. https://doi.org/10.1080/10447318.2023.2278941

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497. https://doi.org/10.1002/nur.20147

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. Review of General Psychology, 14(2), 154–166. https://doi.org/10.1037/a0019440

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review, 41, 71–90. https://doi.org/10.1016/j. dr.2016.06.004

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe–Crowne Social Desirability Scale. Journal of Clinical Psychology, 38(1), 119–125. https://doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Rokeach, M. (1973). The nature of human values. Free Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1– 36. https://doi.org/10.18637/jss.v048.i02

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. Motivation and Emotion, 30(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 25, pp. 1–65). Academic Press. https://doi.org/10.1016/S0065-2601(08)60281-6

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture, 2(1), 11. https://doi.org/10.9707/2307-0919.1116

Shah, J. Y., Friedman, R., & Kruglanski, A. W. (2002). Forgetting all else: On the antecedents and consequences of goal shielding. Journal of Personality and Social Psychology, 83(6), 1261–1280. https://doi.org/10.1037/0022-3514. 83.6.1261

Sheeran, P. (2002). Intention–behavior relations: A conceptual and empirical review. European Review of Social Psychology, 12(1), 1–36. https://doi.org/10.1080/14792772143000003

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. Social and Personality Psychology Compass, 10(9), 503–518. https://doi.org/10.1111/spc3.12265

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sun, S., Kim, J. H., Lee, K. M., & Nan, D. (2024). Exploring the association between the Proteus effect and intention to play massive multiplayer online role-playing games (MMORPGs). Internet Research, 34(1), 58–78. https://doi.org/10.1108/INTR-07-2022-0487

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. Computers in Entertainment, 3(3), 3–3. https://doi.org/10.1145/1077246.1077253

Tang, Z., & Kirman, B. (2025a). Designing for curiosity: Understanding player engagement with treasure chest mechanics in open-world games. International Journal of Human-Computer Studies, 203, 103575. https://doi.org/10.1016/j.ijhcs.2025.103575

Tang, Z., & Kirman, B. (2025b). Exploring curiosity in games: A framework and questionnaire study of player perspectives. International Journal of Human–Computer Interaction, 41(4), 2475–2490. https://doi.org/10.1080/10447318.2024.2325171

Tondello, G. F., Mora, A., Marczewski, A., & Nacke, L. E. (2019). Empirical validation of the Gamification User Types Hexad scale in English and Spanish. International Journal of Human-Computer Studies, 127, 95–111. https://doi.org/10.1016/j.ijhcs.2018.10.002

Tondello, G. F., & Nacke, L. E. (2019). Player characteristics and video game preferences [Paper presentation]. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '19) (pp. 365– 378). ACM. https://doi.org/10.1145/3311350.3347185

Tondello, G. F., Valtchanov, D., Reetz, A., Wehbe, R. R., Orji, R., & Nacke, L. E. (2018). Towards a trait model of video game preferences. International Journal of Human–Computer Interaction, 34(8), 732–748. https://doi.org/10.1080/10447318.2018.1461765

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016). The Gamification User Types Hexad scale [Paper presentation]. In Proceedings of the 2016 Annual Symposium on Computer– Human Interaction in Play (CHI PLAY '16) (pp. 229–243). ACM. https://doi.org/10.1145/2967934.2968082

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. Psychological Review, 117(2), 440–463. https://doi.org/10.1037/a0018963

Tyack, A., & Mekler, E. D. (2020). Self-determination theory in HCI games research: Current uses and open questions [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20) (pp. 1–22). ACM. https://doi.org/10.1145/3313831.3376723

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1), 4–70. https://doi.org/10.1177/109442810031002

Vanden Abeele, V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. International Journal of Human-Computer Studies, 135, 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 41(3), 321–327. https://doi.org/10.1007/BF02293557

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. Journal of Black Psychology, 44(3), 219–246. https://doi.org/10.1177/0095798418771807

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Sage.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806–838. https://doi.org/10.1177/0011000006288127

Wright, S., & Denisova, A. (2024). “It’s a terrible choice to make but also a necessary one”: Exploring player experiences with moral decision making mechanics in video games. Computers in Human Behavior, 161, 108424. https://doi.org/10.1016/j.chb.2024.108424

Xiao, Y., Chen, X., Jin, Y., & Hew, K. F. (2025). Player types in gamified learning: Conceptualization, validation, and profiling. International Journal of Human–Computer Interaction, 41(5), 3557–3575. https://doi.org/10.1080/10447318.2024.2338664

Yee, N. (2006). Motivations for play in online games. CyberPsychology & Behavior, 9(6), 772–775. https://doi.org/10.1089/cpb.2006.9.772

Yee, N. (2016). The Gamer Motivation Profile: What we learned from 250,000 gamers [Paper presentation]. Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY ’16) (p. 2). ACM. https://doi.org/10.1145/2967934.2967937

Yee, N., Ducheneaut, N., & Nelson, L. (2012). Online gaming motivations scale: Development and validation [Paper presentation]. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’12) (pp. 2803–2806). ACM. https://doi.org/10.1145/2207676.2208681

About the authors

Ziao Tang is a researcher in the School of Arts and Creative Technologies at the University of York. His work focuses on player motivation, curiosity, and games user research, including the measurement of player experience and in-game decision-making. Ben Kirman is a Reader in Interactive Media in the School of Arts and Creative Technologies at the University of York. His research uses game design and playful design to explore the social and cultural effects of emerging technologies, spanning games user research, human–computer interaction, and game studies.