エンタテインメント系システムの主観評価実験におけるユーザ統制及び実験手法の検討

藤井 叙人[1,a)]　福嶋 良平[1,b)]　片寄 晴弘[1,c)]

概要：エンタテインメント系システムの主観評価では，ユーザのシステムに対する経験や知識，実験手順が結果に大きく影響するため，実験の信頼性を確保することが非常に難しい．本研究では，ユーザ統制や実験手法による主観評価実験の信頼性確保を最終目的とし，本論文では，その足掛かりとして，エンタテインメント性を公正に評価するために考慮すべき統制視点について議論する．具体的な実験計画として，アクションゲームのプレイ動画視聴における主観評価実験を実施し，その評定結果と発話プロトコル分析から，評定の結果にぶれが生じる例を紹介する．

Study of user control and experimental techniques in subjective evaluation for entertainment systems

Nobuto Fujii[1,a)]　Ryohei Fukushima[1,b)]　Haruhiro Katayose[1,c)]

Abstract: In a subjective evaluation for entertainment systems, it is very difficult to ensure the reliability of an experiment because users' experiences of the system and the experimental procedure are significantly affect the results. In this study, our main goal is ensuring the reliability of a subjective evaluation by proposing user controls and experimental techniques. As the first step to this goal, in this paper, we discuss some important viewpoints to fairly evaluate the user's satisfaction in an experiment. We conducted a subjective evaluation, in which participants assessed players' skill and characters' human-likeness in play videos of the action game, as a specific example of experimental procedures. From the result and the content of utterance, we examined differences of assessments among participants.

1. はじめに

工学系の研究においてシステムを開発した際，その有用性や信頼性を示す評価実験は必要不可欠である．HCI 系の研究領域においては，システムのユーザが人間であることから，人間を対象とした主観評価実験が実施されてきた [1], [2]．主観評価実験の中でも，「使いやすさ」や「見やすさ」といった知覚に関連した事項では，実験参加者による評価の差異（ぶれ）は少なく，ある程度妥当な実験結果が得られると考えられる．これに対し，エンタテインメント系システムにおいて定量化したい評価項目である「楽しさ」や「没入感」といった認知に関連した事項は，かなり個人性が強く，そもそも，その概念に対する個人の捉え方自体の統制が難しい．

このような問題意識から，これまでにも，皮膚電気活動や脳血流といった生理指標に着目した研究 [3], [4], [5] や，センサやカメラを用いてユーザのシステム使用状況を物理的に評価する研究 [6], [7] が行われてきた．これらの研究では，システムのエンタテインメント性を，客観的，定量的に評価することができているといえる．しかし，実験計画に帰する要因で結果が大きくぶれてしまうことが，実際の EC 研究の中ではたびたび起こる．ユーザのシステムに対する経験や知識，実験手順及び実験教示といった，実験計画段階での議論もなされるべきである．

本研究では，そのような状況を踏まえ，実験計画に帰す

[1] 関西学院大学大学院理工学研究科
Graduate School of Science and Technology, Kwansei Gakuin University
[a)] nobuto@kwansei.ac.jp
[b)] erv95460@kwansei.ac.jp
[c)] katayose@kwansei.ac.jp

る要因を統制していくための考慮事項について議論する．まず，筆者らのこれまでの研究事例の中でこのような案件が観測された事例について紹介し，その解釈を試みる．続いて，安定した評価実験を行なうにあたって考慮すべき統制視点について議論し，その一部について，具体的な実験計画として落とし込んだ主観評価実験として，アクションゲームのプレイ動画視聴における「上手さ」と「人間らしさ」の評定実験を実施する．最後に，その評定結果と発話プロトコル分析から，実験参加者間の評定結果のぶれ，及び，実験参加者内での評価基準のぶれが生じる具体例を紹介する．

2. 従来研究における具体例

著者らの従来研究における主観評価実験において，実験計画に帰する要因で結果がぶれてしまう事例を二つ紹介する．本章では，従来研究の概略と，主観評価実験の事例を記述するにとどめるため，詳細な内容は，参考文献 [2], [4], [5] を参照されたい．

2.1 キャラクタの振る舞いの人間らしさに関する事例

一つ目は，アクションゲームの操作キャラクタの振る舞いにおける「人間らしさ（自然さ）」の評定実験である．この研究では，ゲームの仕様やゲーム環境パラメータが公開されている “Infinite Mario Bros.”[8]（世界的に有名なゲームである “Super Mario Bros.” を模したアクションゲーム）を機械学習の対象とし，強化学習手法の一つである Q 学習を用いて，キャラクタの振る舞いを自動獲得することが目標であった．ただし，一般的な機械学習手法で得られる「強いゲーム AI」ではなく，プレイヤのプレイフィール（プレイ時の感覚や印象）やユーザエクスペリエンスを良くするための「人間らしいゲーム AI」の実現を目指したものである．『生物学的制約：“ゆらぎ” “遅れ” “疲れ”』を強化学習の枠組みに導入し，人間プレイヤがゲームをするときに必ず生じる制約を，ゲーム AI に再現させることで，自動獲得されたキャラクタの振る舞いを『人間らしい』ものとすることに成功している．

主観評価実験では，自動獲得されたキャラクタの振る舞いの『人間らしさ』の検証が必要不可欠であった．そこで，20〜24 歳の男女 20 名（男性 13 名，女性 7 名）を対象に，2 つのプレイ動画を比較し「どちらのマリオが人間らしいプレイか」を 7 段階で評定させる主観評価実験を実施した．実験に使用したプレイ動画は，ゲーム AI（強化学習）によるものを 3 つ，人間プレイヤの操作によるものを熟練度を考慮して 3 つ（初級者，中級者，上級者）の計 6 つである．実験の結果，ゲーム AI によるキャラクタの振る舞いは，人間プレイヤより人間らしいという結果を得ることができた．しかし，初級者のプレイ動画は，人間プレイヤの操作であるにも関わらず，人間らしくないと評価された．初級者のプレイ動画を録画した際の初級者プレイヤについて検証したところ，当該プレイヤは，極力キー操作を減らして安全に進むために「十字ボタンの右キーは押しっぱなしで，タイミング良くジャンプをすることに注力する」というプレイスタイルであったこと，また，ダッシュボタンとジャンプボタンを同時に押すことができないコントローラの持ち方（中級者以上は右手親指を両ボタンの上に被せる持ち方が多い）であったことが分かった．そのため，「ジャンプの高さや進むスピードが一定で人間らしくない」という評価が多かったと考えられる．一方，「初級者はダッシュを使えないから自然にみえる」という評価も一部あった．この事例は，評定者における，プレイスタイルやコントローラの持ち方に関する知識の有無によって，主観評価実験の結果がぶれてしまったと考えられる．

2.2 テレビゲーム実施時の脳活動に関する事例

二つ目は，テレビゲーム実施時のプレイヤの脳活動を fNIRS(機能的近赤外分光法) によって計測し，比較，検討した実験である．この研究では，テレビゲームにおける熟達度に焦点を当て，シューティングゲームを実施している際のプレイヤの脳活動を，熟達者，中級者，初心者の 3 種類の条件で計測した．その結果，「熟達者が熟達しているゲームタイトル」を実施時に，中級者と初心者は前頭前野の脳活動が低下する一方で，熟達者は上昇するという状況が観測された．また，熟達者に「熟達したジャンルの初めて実施するゲーム」「経験の浅いジャンルのゲーム」もプレイさせたところ，前頭前野の脳活動が低下することが確認された．ただし，「熟達したジャンルの初めて実施するゲーム」であっても，長期間 (10 日で計 10 時間) にわたり訓練することで，訓練後期には前頭前野の活動が上昇するという結果を得ている．この事例は，実験参加者の熟達度が実験結果に大きく影響することを示しており，さらには熟達したジャンルかどうかも検討する必要があるといえる．

3. 実験計画段階で考慮すべき統制視点

本研究の目的は，「結果がぶれる原因となる，実験計画に帰する要因」を洗い出し，それぞれのエンタテインメント系システムに応じて要因の統制を実施し，主観評価実験の信頼性を確保することである．しかしながら，最初に述べたとおり，エンタテインメント系システムにおいて定量化したい評価項目である「楽しさ」や「没入感」は，かなり個人性が強く，そもそも，その概念に対する個人の捉え方自体の統制が難しい．前章で述べた評価実験の事例もあくまで一例であり，実際の評価実験では他にも様々な「結果がぶれる原因となる，実験計画に帰する要因」があると考えられる．本論文では，まずは，実験計画に帰する要因の洗い出しを主目的とし，統制視点の検討，具体的な実験計画として落とし込んだ主観評価実験の実施，及び，統制視

点の差異により結果にぶれが生じた例を紹介する．

本章では，以下の 3 つの統制視点を検討する．

(1) **実験参加者の熟練度と知識量**

システムが対象とするエンタテインメントに対する，実験参加者の熟練度は，そのシステムを評価する際に大きな影響を与えると考えられる．エンタテインメント（ゲーム，音楽，スポーツなど多岐にわたる）の経験の有無，経験の量だけでなく，類似ジャンルや類似システムの経験に関しても言及する必要があるだろう．また，そのエンタテインメントにおける知識や技術の量も評価結果を左右する可能性がある．特に，その知識や技術を言語化できるかどうか，知識や技術を得るためのメタな知識の有無は，評価する際の評価基準に影響すると思われる．

(2) **実験刺激の提示順序**

実験刺激の提示順序は，完全なランダムとすることで順序効果を消すのが一般的な評価実験である．「使いやすさ」や「見やすさ」といった知覚に関連した事項では，実験参加者は過去の経験からある程度の絶対的評価基準を持っていると考えられるため，順序効果を消すことで公正な評価結果が得られると思われる．一方，「楽しさ」や「没入感」といった認知に関連した事項では，その評価基準が実験の過程で大きく変化する可能性がある．また，エンタテインメント系システムの評価においては，よりシステムを楽しく感じさせる提示順序の検討もなされるべきである．

(3) **実験参加者の評価基準と評価軸**

「楽しさ」や「没入感」といった認知に関連した事項の評価では，評価の基準が曖昧である場合が多い．また，評価軸自体が実験参加者によって異なる，あるいは，複数の評価軸を使い分けている可能性も高い．実験過程での実験参加者の評価基準の変化，及び，評価軸の特定，評価軸の変化について精緻に分析を試みる必要がある．同じ評価軸を持った実験参加者向けにシステムをカスタマイズすることで，ユーザ満足度の向上を図ることも，エンタテインメント系システムでは重要と考える．

4. 主観評価実験の実施

第 3 章で述べた実験計画段階で考慮すべき統制視点を，実際の主観評価実験に落とし込んだ実験例について述べる．

本論文では，1995 年に任天堂からスーパーファミコンで発売されたアクションゲーム “スーパーマリオ ヨッシーアイランド” （国内売上本数は 2013 年時点で 177 万本の有名タイトル）を主観評価実験の対象とした．ゲームの詳細はここでは割愛するが，2.1 節で述べた “Infinite Mario Bros.” と同様に，キャラクタを操作して横スクロール式のステージを攻略するアクションゲームである．ただし，操作は “Infinite Mario Bros.” よりも複雑で難しく，実験参加者の熟練度の差が出やすいタイトルとなっている．

表 1 チェック項目の一部

レベル	キャラクタの振る舞い
1	ジャンプができる
2	走りながらのジャンプで飛距離調整ができる
2	ジャンプボタン押しの長さでジャンプの高さを調整できる
2	ジャンプ中に再度ジャンプボタンを押すことで踏ん張りジャンプができる
3	踏ん張りジャンプで飛距離調整ができる
3	連続で踏ん張りジャンプができる
4	連続踏ん張りジャンプで飛距離調整ができる
・・・	など，全 45 項目

4.1 実験手続き

20〜24 歳の男女 8 名（男性 7 名, 女性 1 名）を対象に実験を執り行った．まずは，実験参加者の熟練度を知るために，家庭用ゲームの経験に関するアンケートを実施した．質問項目は以下のとおりである．

- 直近 10 年間での家庭用ゲームのプレイ時間の月平均
- 直近 10 年間でのアクションゲームのプレイ時間の月平均
- ヨッシーアイランドの経験の有無，及び，プレイ時間

次に，実験参加者にヨッシーアイランドの操作方法を説明した後に，実際にゲームをプレイさせた．プレイ時間は練習約 10 分＋本番 30 分とし，練習ではチュートリアルステージと最初のステージを，本番では序盤でやや難しいステージを繰り返し訓練させた．その際，実験参加者にはプレイ中に考えていること，うまくいったこと，失敗したこと，ゲームへの気づき，感想などを自由に発話させ，ゲーム画面の録画録音，及び，実験参加者の発話の録音を実施した．

実験参加者のヨッシーアイランドに対する知識や技術を知るために，録画録音したプレイ動画の内容は，実験後に実験者がチェックした．チェックの際には，コントローラ操作の難しさにより 4 段階にレベル分けされた全 45 項目のチェックシートを用いた．コントローラ操作の難しさとは，連続した一連のボタン操作，及び，同時に複数のボタン操作を難しいと定義したものであり，ボタン一つで可能なアクションはレベル 1，ボタン二つ同時押し，あるいは，ボタン二つを連続して押すアクションはレベル 2 という規則で定めた．チェック項目の一部を以下の表 1 に示す．実験参加者の知識や技術の指標として，レベルの合計値（レベル 1 を一つ，レベル 2 を二つ実行できる実験参加者は $1 \times 1 + 2 \times 2 = 5$ 点）を算出した．

最後に，複数のプレイ動画を実験参加者に見てもらい，うまいプレイと感じる順，及び，人間らしいプレイと感じる順に順位付けをさせた．評定に用いたプレイ動画は，各

表 2　評定に用いたプレイ動画

	ヨッシーアイランドのプレイ時間	直近 10 年でのアクションゲームのプレイ時間の月平均	動画時間	知識と技術のレベル合計値
1	0 時間	0 時間	157 秒	12
2	0 時間	0 時間	181 秒	12
3	0 時間	35 時間	229 秒	18
4	0 時間	3 時間	139 秒	16
5	2 時間	16 時間	201 秒	22
6	7 時間	18 時間	201 秒	21
7	15 時間	25 時間	132 秒	22
8	60 時間	28 時間	85 秒	47
9	RTA 日本レコードの動画		55 秒	71

実験参加者が，30 分の訓練の内で最もステージを攻略できた際のプレイを編集したもの（計 8 つ）に加え，RTA（ステージを最速でクリアすることを目指したリアルタイムアタック）の日本レコードであるプレイ動画を用意した．動画内のステージの区間は全動画で統一し，ステージの途中で死んでしまったシーンは全てカットした．各プレイ動画の詳細を表 2 に示す．動画 1 と 2 の操作者は，ヨッシーアイランドのプレイ経験がなく，かつ，家庭用ゲーム機自体のプレイ経験がほぼない人である．動画 3 と 4 の操作者は，ヨッシーアイランドのプレイ経験はないが，家庭用ゲーム機のアクションゲームはそこそこプレイしたことがある人，また，動画 5 以降は，ヨッシーアイランドのプレイ経験，及び，アクションゲームのプレイ経験順に昇順に並べた．動画 9 は RTA 日本レコードの動画とした．プレイ動画の提示順は，まず，実験参加者本人が操作したプレイ動画を見せ，それ以降は表 2 の番号通りとした．この提示順序は，あえて全実験参加者で共通としている．ただし，「各動画は過去に遡って何回見ても良い」と実験参加者に教示している．

評定の際には，「コンピュータ（機械）が操作しているプレイ動画も含まれている可能性がある」こと，また，「評定の結果は実験の途中で何度変更しても良い」と実験参加者に教示した上で，その動画の操作者のヨッシーアイランドの腕前（上手さ）がどの程度かを，かなり下手（0） ～ 下手（25） ～ 普通（50） ～ 上手（75） ～ かなり上手（100）として 5 点刻みの 100 点満点で，また，その動画の操作者が人間プレイヤであるという確信度（%）を，機械（0）～ やや機械（25）～ どちらとも言えない（50）～ やや人間（75）～ 人間（100）として 5 %刻みの最大 100 %で評定させた．腕前（上手さ）に関しては，実験参加者本人が操作したプレイ動画も含めて 9 つの動画の評定をさせたが，人間プレイヤの確信度に関しては，実験参加者本人「以外」が操作したプレイ動画 8 つにおける評定とした．

プレイ動画を見せている際，及び，評定中は，実験参加者には以下の 4 点に関して発話するように促し，その発話内容を録音した．

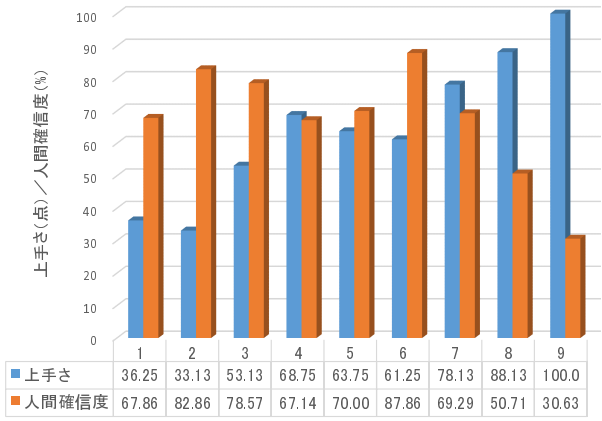

図 1　実験参加者 8 名の評定の平均

- 上手いプレイだと思う箇所，及び，そう思う理由
- 下手なプレイだと思う箇所，及び，そう思う理由と，上手くプレイするためのアドバイス
- 人間らしく（人間プレイヤが操作しているように）見える箇所，及び，そう思う理由
- 人間らしくなく（コンピュータが操作しているように）見える箇所，及び，そう思う理由

4.2　実験結果

まずは，動画の操作者のヨッシーアイランドの腕前（上手さ）と，動画の操作者が人間プレイヤであるという確信度（人間らしさ）について，実験参加者全 8 名の評定の平均を図 1 に示す．上手さに関しては，表 1 におけるヨッシーアイランドやアクションゲームのプレイ時間と大体比例して，右肩上がりのグラフとなった．人間らしさに関しては，動画 8,9 のような「うますぎるプレイ動画」は機械寄りの評定をされることが分かった．

次節からは，第 3 章で述べた実験計画段階で考慮すべき統制視点から，実験結果を個人毎に細かく検証していく．

4.3　熟練度と知識量による影響

実験参加者 8 名の個人毎の上手さの評定結果は表 3 のとおりであった．ここで，表 3 の動画 8 と 9 の上手さの評定結果に注目すると，参加者 1,2 は動画 8 と 9 の点差が 5.00 点であるのに対し，参加者 3～8 は平均 14.17 点の点差を付けている．同様に動画 7 と 9 に注目すると，参加者 1～4 は平均 13.75 点の点差であるのに対し，参加者 5～8 は平均 30.00 点の点差を付けている．この結果には，実験参加者の熟練度と知識量の違いによる「順位付けはできるが，点差を正確につけることができない」という状況が生じている可能性がある．さらに，実験参加者 1～4 における，上手さの評定結果と表 2 の「知識と技術のレベル合計値」との相関係数 R の平均を求めると $\bar{R} = 0.6822$ であるのに対し，実験参加者 5～8 における相関係数 R の平均は $\bar{R} = 0.8320$ であった．実験参加者 1～4 は，動画内の操作者の熟練度

表 3　個人毎の評定結果（上手さ）

参加者	1	2	3	4	5	6	7	8
動画 1	20	50	35	70	20	25	30	40
動画 2	25	30	75	25	40	15	25	30
動画 3	75	70	60	45	50	55	20	50
動画 4	85	75	60	70	75	65	60	60
動画 5	90	65	50	75	60	55	50	65
動画 6	50	40	65	80	65	65	65	60
動画 7	80	90	90	85	65	70	70	75
動画 8	95	95	80	90	90	80	90	85
動画 9	100	100	100	100	100	100	100	100

表 4　評定結果の変更回数

参加者	1	2	3	4	5	6	7	8
上手さ	3	0	10	5	13	10	1	2
人間らしさ	2	0	2	1	0	1	1	0

と知識量を正確に把握できていないため，点差を正確につけることができていない可能性が示唆された．

4.4　実験刺激の提示順序の影響

実験参加者には「評定の結果は実験の途中で何度変更しても良い」と教示しているため，各参加者は自分の評定に疑問を感じた際に，実際に何回か評定結果を変更している．評定結果を変更した回数を表 4 に示す．上段は上手さの評定を変更した回数，下段は人間らしさの評定を変更した回数であり，1 つの動画の評定結果を変更する度に 1 回とカウントしている．参加者 3〜6 の上手さの評定を変更した回数に注目すると，参加者 1,2,7,8 よりかなり大きい数値となっているのが分かる．参加者 3〜6 が上手さの評定を変更したタイミングを確認すると，合計 38 回中 26 回は最後の動画 9 を見た後であった．評定を変更する際の発話内容を確認すると，

参加者 3「これは，逆転ですね．（動画 9 は）100．他を下げざるをえない．」

参加者 4「ここ（動画 9 と他）をもうちょっと差を開けときます．動画 7 を 85 にします．動画 8 は 90 で．」

参加者 5「これ（動画 9）100 点しか無理ですね．他を修正します．（他を）10 点は下げます．」

参加者 6「これ（動画 9）上手すぎですよ．100 にします．（他を）全部 15 ぐらい落とします．」

となっており，自分の想定以上の上手さの動画であったために他を下げざるを得ない状況であることが分かる．実験刺激の提示順序の影響で，評定するための評価基準が大きく変更してしまった例と言えるだろう．一方，参加者 1 と 2 が変更回数が少ない理由としては，上で述べたとおり，「順位付けはできるが，点差を正確につけることができない」状況であるため，参加者 7 と 8 が変更回数が少ない理由としては，動画 9 の上手さが想定内であり，正確に点数付けができていたためと考えられる．

4.5　評価基準，評価軸の変更による影響

評定結果を変更した回数（表 4）について，上手さと人間らしさの評定を変更した際の状況を確認すると，合計 51 回中 44 回は，「下手」「上手」「やや人間」「やや機械」などの日本語でラベル付けをした点数絡みであった．今見た動画の点数を上手（75 点）にしたいので，前見た動画の点数を 5 点上げて 80 点にする，という状況や，今見た動画の点数を上手（75 点）とすると，前見た動画の点数は 60 点から 70 点に上げてもいい，という状況が多いようである．これは，実験参加者は評定の際に，まずは日本語のラベルを一つの評価基準として捉えていること，また，実験の過程で評価の閾値が少しずつ変化していることを示していると言える．参加者 7 や 8 が変更回数が少ない理由としては，はじめから絶対的な評価の閾値を持っているためと考えられる．

次に，実験参加者 8 名の個人毎の人間らしさの評定結果を表 5 に示す．動画 8 と 9 の評定結果に注目すると，参加者によって 0〜80 とまちまちの評定となっているのが分かる．参加者 2,4,6,7 は動画 8 は人間寄り，動画 9 はほぼ機械と評定しているが，参加者 1,3,5 は動画 8 は機械寄り，動画 9 は人間寄りに評定をしている．実験参加者 8 名の平均である図 1 では，動画 8,9 は「うますぎるプレイ動画」であるため機械寄りの評定であるように見えたが，個人毎に精緻に見ると，そう結論づけるのは憚られる．

そこで，人間らしさの評定結果と表 2 の「動画時間」との相関係数 R を，参加者毎に，動画 7 の評定後，動画 8 の評定後，動画 9 の評定後で算出した結果を表 6 に示す．参加者 2,4,6,7 は相関係数がほとんど正相関であり，実験中は一貫して，攻略スピードが速い（＝動画時間が短い）動画を機械らしいと評定する傾向にあるようだ．参加者 1 と 3 は，人間らしさの評定と動画時間とはほとんど相関が無いようである．動画 9 評定時の発話内容を確認すると，

参加者 1「逆に上手すぎて機械じゃこんな動作できないんじゃないですかね．これは結構人間なんじゃないかなって．人間のめちゃくちゃ上手い人．」

参加者 3「タイムアタックを極めた人かなぁ．さっきの（動画 8）は上手くて，マップを熟知してるんだけど，その割には動きがもっさりしているというか．」

となっており，「上手すぎる＝機械では無理」と判断しているようであった．参加者 5 は動画 8 までは攻略スピードが速い動画を機械らしいと評定する傾向にあったが，最も攻略スピードが速い動画 9 は 50 %（人間か機械かどちらとも言えない）と評定している．参加者 8 は，逆に，動画 8 までは攻略スピードが速い動画を人間らしいと評定する傾向にあったが，動画 9 は 25 %（機械らしい）と評定している．動画 8 までの評定と，動画 9 の評定は，明らかに別の評価軸を用いていると思われる．動画 9 評定時の発話内容を確認すると，

表 5　個人毎の評定結果（人間らしさ）

参加者	1	2	3	4	5	6	7	8
動画 1	−	80	50	80	25	80	80	80
動画 2	70	−	90	65	80	90	100	85
動画 3	45	90	−	95	80	80	100	60
動画 4	40	55	75	−	40	70	100	90
動画 5	60	90	30	45	−	75	100	90
動画 6	50	95	90	95	90	−	100	95
動画 7	55	55	100	90	25	60	−	100
動画 8	30	70	40	80	10	55	70	−
動画 9	75	30	65	0	50	0	0	25

表 6　人間らしさの評定と動画時間との相関係数 R の遷移

	動画 7 評定後	動画 8 評定後	動画 9 評定後
参加者 1	-0.2139	0.5164	-0.0608
参加者 2	0.8963	0.7534	0.8625
参加者 3	-0.3428	0.2577	0.1229
参加者 4	-0.0419	0.6343	0.5490
参加者 5	0.8519	0.8913	0.6816
参加者 6	0.5751	0.8077	0.8249
参加者 7	0.4138	0.8216	0.8294
参加者 8	-0.6066	-0.6066	0.5090

参加者 5「ここまで来ると，やりこんでるプレイヤならできなくはないのかなと思いますね．単純に考えればもちろん機械っぽいんですけど．世の中にはそういう人もいるかな，とは思います.」

参加者 8「RTA とかやってるのは機械的に動いているので，どちらかというと機械っぽいなとは思います.」

となっており，両者とも「上手すぎる＝機械的な動きを追求しているもの」と判断しているようであった．

4.6　考察

前章での実験結果から，エンタテインメント系システムの主観評価の実験計画で注意すべき点をまとめる．まず，システムが対象とするエンタテインメントの未経験者を実験参加者とした場合，順位付け（相対評価）はできても採点（絶対評価）するのは難しいと思われる．絶対評価をさせたいのであれば，少なくとも何度かは経験したことのある参加者を用意するべきである．ただし，経験が浅い参加者は，実験途中で自身の想定以上（あるいは想定以下）の刺激に直面したり，日本語のラベルの影響を受けたりと，評価の基準（閾値）が実験過程で随時変更される危険性がある．あらかじめ実験参加者に最大値と最小値の刺激を参考として与える，評価の理由を詳細に聴取する，実験中の発話を録音しておく，などの対策が必要となるだろう．上級者，熟達者ほどの経験がある参加者であれば，評定結果がそれほどぶれること無く，安心して採点等の絶対評価を任せられそうである．しかし，熟達者になればなるほど，評価軸が複数ある可能性が高いという問題もある．あるコンテキストで考えると良い評価だが，そうでない場合は悪い評価であり，どちらの評価をすべきか悩むケースも起こりうる．評価軸の特定も，評価理由の聴取と，発話の録音に頼るしかないが，あらかじめある程度の評価軸を参加者に教示しておくのも手かもしれない．

5.　おわりに

本論文ではエンタテインメント系システムの主観評価において，「結果がぶれる原因となる，実験計画に帰する要因」を統制していくための考慮事項について議論した．安定した評価実験を行うために考慮すべき統制視点として，実験参加者の熟練度と知識量，実験刺激の提示順序，実験参加者の評価基準と評価軸を挙げ，それらの視点を具体的な実験計画として落とし込んだ主観評価実験を実施した．主観評価実験では，アクションゲームのプレイ動画視聴における「上手さ」と「人間らしさ」の評定というタスクを実験参加者に与え，その評定結果と発話プロトコル分析から，実験参加者間の評定結果のぶれ，及び，実験参加者内での評価基準のぶれが生じる具体例を述べた．最後に，エンタテインメント系システムの主観評価の実験計画で注意すべき点として，少なくともシステムが対象とするエンタテインメントの経験者を実験参加者とすること，また，評価理由の聴取，及び，実験中の発話から，実験参加者の評価基準や評価軸を押さえるべきであることを述べた．

今後は，「結果がぶれる原因となる，実験計画に帰する要因」の更なる洗い出しを行い，エンタテインメント系システムの主観評価実験の信頼性を確保するための，ユーザ統制や実験手法を確立していく．

参考文献

[1] 星野准一，田中彰人，濱名克季：模倣学習により成長する格闘ゲームキャラクタ，情報処理学会論文誌，Vol. 49, No. 7, pp. 2539–2548 (2008).

[2] 藤井叙人，佐藤祐一，若間弘典，風井浩志，片寄晴弘：生物学的制約の導入によるビデオゲームエージェントの「人間らしい」振舞いの自動獲得，情報処理学会論文誌，Vol. 55, No. 7, pp. 10–19 (2014).

[3] 代蔵巧，棟方渚，小野哲雄：E3-Player: 鑑賞者の興奮を促進させる動画鑑賞システム，pp. 272–277 (2013).

[4] 八田原慎悟，藤井叙人，長江新平，風井浩志，片寄晴弘：熟達度を視点としたテレビゲーム実施辞の脳活動の分析，情報処理学会論文誌，Vol. 49, No. 12, pp. 3859–3866 (2008).

[5] 八田原慎悟，藤井叙人，古屋晋一，風井浩志，片寄晴弘：テレビゲーム熟達者の脳活動に関するケーススタディ，情報処理学会論文誌，Vol. 50, No. 12, pp. 2782–2795 (2009).

[6] 岩楯翔仁，藤村航，三角甫，小坂崇之，白井暁彦：奥行き画像センサを用いた展示空間の物理評価 (2011).

[7] 田所康隆，藤村航，北田大樹，白井暁彦：エンタテイメントシステム展示を対象とした質的評価ツールの提案，pp. 107–110 (2013).

[8] Karakovskiy, S. and Togelius, J.: The Mario AI Benchmark and Competitions, *Computational Intelligence and AI in Games, IEEE Transactions on*, Vol. 4, No. 1, pp. 55–67 (2012).