Manipulating Leaderboards to Induce Player Experience

Jason T. Bowey
University of Saskatchewan
Saskatoon, Canada
jason.bowey@usask.ca

Max V. Birk
University of Saskatchewan
Saskatoon, Canada
max.birk@usask.ca

Regan L. Mandryk
University of Saskatchewan
Saskatoon, Canada
regan.mandryk@usask.ca

ABSTRACT
Assessing and inducing player experience (pX) in games user research (GUR) is complicated because of the tradeoff between maintaining rigour through experimental control and having participants feel like they are engaged in play. To establish and evaluate an embedded method for inducing a sense of success or failure in participants during gameplay (e.g., to study how different players exhibit resilience to in-game failure), we manipulated leaderboard position in an experiment in which 155 participants played a Bejeweled clone. We show that manipulating success perception through leaderboards increases the player's perception of competence, autonomy, presence, enjoyment, and positive affect over manipulated failure. In addition, displaying the score enhances the effect on positive affect, autonomy and enjoyment, while not increasing detectability.

Author Keywords
Leaderboards; pX; GUR; success; failure

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI)

INTRODUCTION
Although there have been recent innovations in games user research (GUR) for empirically studying games and player experience (pX), researchers and developers need to establish, validate, and standardize methods of inducing and assessing experience to continue to advance the field. However, understanding and manipulating experience is particularly complicated in the context of games research because researchers and developers always need to balance the tradeoff between having enough control over the experimental environment (for rigour and validity) and having participants feel like they are actually engaged in the act of play under their own volition and not participating in the work of an experiment (for ecological validity). For example, standard experience assessment methods, such as the think-aloud protocol [30] and heuristic evaluation [24] have been adapted for application to games [1,6].

CHI PLAY 2015, October 03 - 07, 2015, London, United Kingdom
 ISBN 978-1-4503-3466-2/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2793107.2793138

Although there has been recent progress in methods for *assessing* player experience, we still need to make progress in establishing standardized methods for *inducing* different aspects of player experience. For example, we may want to control the game outcome – i.e., whether a player feels like they succeeded or failed in a game – in an experiment. This could be important to build models of how different types of players [7,14,23] differentially respond to winning, when studying resilience to in-game failure [1], or when studying how uncertainty in outcome affects play experience [1]. We need to establish and evaluate a method to reliably induce a sense of success/failure in the context of game experiments.

We argue that leaderboard manipulations are an excellent candidate for experimental manipulation of success and failure in games, for three main reasons: First, the concept of a leaderboard is extremely familiar to game players; from classic arcade games like PacMac (Namco, 1980) to recent games like League of Legends (Riot Games, 2009), leaderboards have been used to represent relative game performance. Second, as a method that is integrated into the game itself, it is not likely to be noticed as a manipulation. For example, in mood induction, it is common to show video clips of funny or sad scenes to induce affect prior to asking participants to complete an experimental task [10]; taking the same approach with a game might make the manipulation of success and failure obvious to participants. Third, there is a range of standard psychological induction methods that are built on the premise of social comparison [8] (e.g., the Trier-Social-Stress Test [5] uses a staged interview situation to induce social stress). Leaderboards are essentially a representation of relative performance and can leverage social comparison to induce an experience.

To investigate the use of manipulated leaderboards in game experiments, we developed a *Bejeweled* (PopCap, 2001) clone, in which players matched gems of the same colour, and presented manipulated leaderboard feedback after each round to give a sense of successful, failed, or neutral outcome. We encoded performance using spatial position – the most salient channel for visualizing information and investigated two other design factors (reinforcing through hue and the display of score) to determine whether the design affected its efficacy in manipulating experience.

Our results show that leaderboard manipulations affect pX. Specifically, inducing the perception of success increases the player's perception of competence, autonomy, presence, enjoyment, and positive affect over inducing the perception of failure. In terms of the leaderboard design, reinforcing

leaderboard position by displaying the score or using colour enhances the effect on positive affect, whereas displaying the score enhances the effect on autonomy and enjoyment.

We provide an experimental investigation of the efficacy of leaderboard manipulation for inducing a sense of success or failure in participants in game experiments. Researchers who wish to induce game outcome to ask questions in the growing area of games science can leverage our work.

RELATED WORK

We discuss success and failure in games, and present pX research and the use of leaderboards in games and research.

Success and Failure in Games

The ability of successfully overcoming challenge is at the heart of player experience [26]. Research in game outcome is usually focused on showing that player performance predicts game enjoyment [28]; however, some findings suggest that the effects of success and failure on players may be more complex; for example, if games are played in a social setting [3], or if competition appears to be unbalanced and players have different abilities [11]. This suggests that pX is not only affected by how players perceive their own performance, but also by the attribution of success and failure. Along these lines, Klimmt et al. [17] show that expertise influences pX, suggesting that expected competence affects the evaluation of success or failure.

Assessing Player Experience

pX research investigates how people experience games and techniques to evaluate concepts relevant for pX, e.g. motivation [26], flow [27], or fun [19] have been created. Methods to assess these constructs include instrumenting [20], observing [19], and surveying [14] players.

The Player Experience of Needs Satisfaction (PENS) [26] scale, for example, is an instrument that measures need satisfaction in games through perceived competence (experiencing mastery), autonomy (making decision under one's own volition), relatedness (feeling related to others), presence (the feeling to be transported to another world), and intuitive control (the naturalness of the input controls). Need satisfaction is a prerequisite for intrinsic motivation, which can be measured using the intrinsic motivation inventory (IMI) [22] that measures the three constructs: enjoyment (feeling joy during play), effort (being cognitively invested and trying hard), and tension (feeling tense while playing). Intrinsic motivation has been shown to increase positive affect [26], which can be measured using the positive affect/negative affect scale (PANAS) [31]. These scales have been used in pX research [26].

Manipulating Player Experience

GUR research is mostly concerned about measuring experience, because experience is manipulated in the context of the game itself, e.g., though different controller layouts [2]. However, to fully understand player experience, we are also interested in manipulating player experience under well-known constraints. Psychological research has developed a variety of induction methods and protocols to research human experience and behaviour under laboratory conditions. Examples for such manipulations are stress induction methods [18], aggression paradigms [29], social exclusion [34], and mood induction [33].

In games like PacMan (Namco, 1980), before synchronous multiplayer competition (e.g., Quake, id Software, 1996) was mainstream, leaderboards were a way of comparing players asynchronously. Leaderboards are still used to create competition and social comparison between players; for example, in World of Warcraft (Blizzard, 2004), players who engage in Player vs. Player competition can see their ranking in comparison to others. Social comparison is also a well-investigated concept in gamification research [4,12].

Social comparison theory argues that people learn about themselves through comparison with others [9]. Prior research has shown that performance in a competitive setting has direct effects on our experience [13]. In the context of leaderboards, players can learn about their performance relative to a group of competitors, which we hypothesize will affect the resulting experience.

METHODS

In this section we describe the game, the leaderboard designs, the experiment procedure, and the instruments.

Game

We created a clone of *Bejeweled* (PopCap Games, 2001) using the *Unity 4.6 3D Game Engine* (Unity Technologies, 2014). We present players with an 8 by 8 grid of gems in one of five colours. Players click or drag adjacent gems to swap their locations and make matches. Players gain 5 points for each gem when 3 or more gems of the same colour are aligned. While three-of-a-kind matches only give points, four-of-a-kind matches create special pieces that destroy all adjacent pieces; five-of-a-kind matches in "L", "T", or "+"-shapes create special pieces that destroy gems in a horizontal and vertical line; five-of-a-kind matches in one line destroy all gems in the matched colour.

Manipulating Success/Failure through Leaderboards

Leaderboards were presented after each round to manipulate outcome for each player. There were three leaderboard conditions: success, neutral, and failure. For each condition, the player was randomly assigned a position on the leaderboard (between 1-3, 8-12, 14-20, respectively). We chose neutral to be above average as research suggests that people consider median performance as failure [25].

We used the most salient visual channel of spatial position [32] to represent leaderboard position (see Figure 1). We drew attention to the player's position on the leaderboard by flashing their name and score, leveraging the pre-attentive processing of flicker [32]. We also investigated two additional factors in our leaderboard design. We reinforced the feedback of success or failure through hue – the first three positions were green, the last third red, and the rest white, and by toggling the display of the scores.

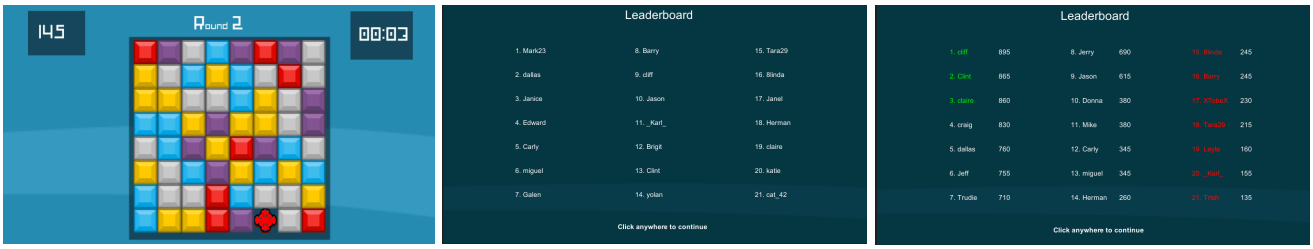

Figure 1. Left: A screenshot of our game with score in the top left, time in the top right, and the game-board in the middle; Middle: Leaderboard with no colour feedback or displayed score; Right: Leaderboard with both colour feedback and displayed score.

We manipulated the scores of the remaining positions using the following algorithm: Positions above the player's score (PS) were assigned a random number between 0.22*PS and 10*PS, multiplied by 5; positions below were assigned scores between PS/50 and PS/6, multiplied by 5. Colour (C) and score (S) were manipulated between participants resulting in four conditions: S/C, S/-C, -S/C, -C/-S.

Procedure

The experiment was conducted using Amazon Mechanical Turk (mTurk), which has been shown to be robust for user studies [21] when precautions are put into place [16]. We collected 155 participants (34.8% female, age mean=31.79, SD=8.69), 14 participants were excluded from analysis for incomplete data or low compliance, defined as responding incorrectly to attention-testing questions, -1SD completion time, or results that differed +3SD from the mean.

Immediately after giving consent, participants provided their baseline affect ratings, the leaderboard condition was assigned, and each participant provided a nickname and completed a game tutorial. Seven rounds of the game were played with all rounds ending in a neutral outcome, except for rounds 4 and 6, which had a manipulated outcome of either success or failure. Participants filled out experience questionnaires following rounds 2, 4, and 6. We presented a final neutral round to protect the post-experiment surveys from effects of the leaderboard manipulation. The condition of success or failure after rounds 4 and 6 were balanced using a Latin-square to avoid sequence effects. After the experiment, participants completed a series of demographic questionnaires. We finished with a free-form text response about the purpose of the experiment to determine whether participants were suspicious about the leaderboard manipulation; 30/141 players were labeled as suspicious, 19 for mentioning that we were measuring response to performance and 11 for mentioning manipulation or deception. Finally, we debriefed participants about the deception in the leaderboard manipulation.

Instruments

To assess player experience, we used validated instruments that have been used to measure player experience before [26], including PANAS [31], PENS [26], and IMI [22].

Data analyses

We performed a repeated-measures Multivariate Analysis of Covariance with condition (success, failure) as a within-subjects factor and color feedback (on, off) and score feedback (on, off) as two between-subject factors, while controlling for suspicious players. We did not include the neutral condition because it was always presented first, thus was confounded with order. We also include baseline positive/negative affect as a covariate when assessing positive and negative affect to control for preexisting differences between players. We controlled for Type 1 error with Bonferroni-corrected (α=.05) pairwise comparisons.

RESULTS AND DISCUSSION

Does leaderboard position affect player experience?

There are main effects on perceived competence ($F_{1,135}$=23.8, p<.001, η^2=.15), autonomy ($F_{1,135}$=4.5, p=.036, η^2=.03), and presence ($F_{1,135}$=9.0, p=.003, η^2=.06) with players feeling more competent, more autonomous, and more immersed after success than failure. We also show a main effect on enjoyment ($F_{1,135}$=11.9, p=001, η^2=.09), with players rating the game as more enjoyable after success than failure. There were no differences in relatedness, intuitive control, invested effort, or tension.

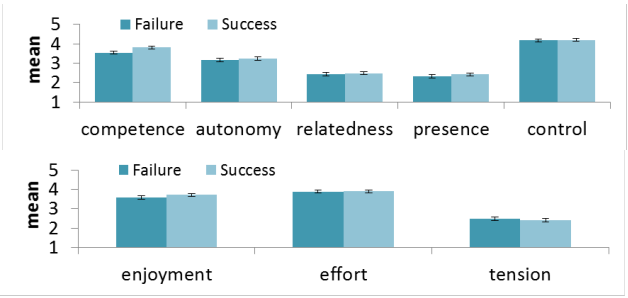

Figure 2. Means (±SE) for success and failure for PENS competence, autonomy, relatedness, presence, intuitive control, and IMI enjoyment, effort, and tension.

Does emphasizing position through score or colour feedback affect the manipulation?

Significant interactions between leaderboard manipulation and score on autonomy ($F_{1,135}$=7.1, p=.009, η^2=.05) and enjoyment ($F_{1,135}$=16.1, p<.001, η^2=.11) suggest that displaying the score differentially affects the manipulation of success and failure. Pairwise comparisons reveal that the effects of leaderboard manipulation on autonomy and enjoyment (see previous section) are only present when score is shown (autonomy:p=.001, enjoyment:p<.000); when score is not shown, there are no differences between success and failure (autonomy:p=.698, enjoyment:p=.688).

There were no other interactions of leaderboard manipulation with score, no interactions of leaderboard with colour, and no main effects of either colour or score.

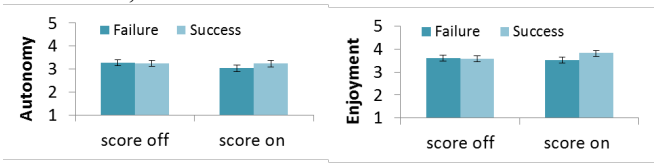

Figure 3. Autonomy (left) and enjoyment (right) mean ratings (±SE) after failure and success with or without score.

Does leaderboard position result in affect changes?

There was a significant main effect of leaderboard manipulation on positive affect (controlling for baseline positive affect), which showed that players felt more positive after success than failure ($F_{1,135}$=6.1, p=.015, η^2=.04). There was no main effect on negative affect.

Does colour or score reinforce changes in affect?

Significant interactions of score and colour show that the difference between success and failure appears when either score ($F_{1,135}$=9.8, p=.002, η^2=.07) or feedback ($F_{1,135}$=4.9, p=.029, η^2=.03) is provided. The significant three-way interaction ($F_{1,135}$=7.6, p=.007, η^2=.05) makes this clear. As Figure 4 shows, there is no affective difference between success and failure when neither score nor colour are included (p=.910); but that including score (p<.001), colour (p<.001), or both (p<.001) will change positive affect. There were no interactions with negative affect.

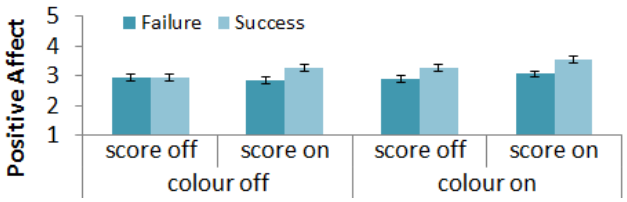

Figure 4. Mea n(±SE) for outcome [success, failure] by score[off, on] by colour [off, on] for positive affect.

Can the effects be explained by performance?

A paired-sample t-test comparing score after success (mean=841.32 SD=392.70) and failure (mean=832.20, SD=361.83) shows a non-significant result (t_{151}=.349, p=.730), suggesting that the effects on pX can be attributed to the leaderboard manipulation and not performance.

Does reinforcing position make players suspicious?

Because reinforcing leaderboard position (particularly through score) strengthens the manipulation, we checked to see whether it also made the manipulation more noticeable. Of the 11/141 players who were labeled as suspicious of deception or manipulation, 4 were in the score/colour condition, 5 in the colour condition, and 2 in the score condition. Although providing reinforcing cues may make more players suspicious, the benefits on experienced autonomy, enjoyment, and positive affect to those who are not suspicious outweighs the need to potentially filter the few suspicious participants from further analyses.

Is the manipulation different for different people?

We divided players into groups based on demographic factors, including sex, age, and a median split on hours spent playing and included these factors in the model as between-subjects factors. There were no interactions of any demographic factor and leaderboard manipulation, suggesting that the manipulation is no more or less effective for various demographic groups.

Limitations and Future Work

We show that leaderboard manipulations can change pX. Specifically, manipulating the perception of success can increase the player's perception of competence, autonomy, presence, enjoyment, and positive affect over manipulating the perception of failure. Also, reinforcing position by displaying the score or using colour enhances the effect on positive affect, whereas displaying the score enhances the effect on autonomy and enjoyment. However, as Figure 4 shows, differences are small – η^2 shows that between 3% and 15% or the variance is explained by the leaderboard manipulation. Previous work has shown that people like to fail in games [5] and overcome challenges [5], so it is not surprising that the effects of failure are small. The largest effect was on perceived competence, which is the construct that should be most affected by a manipulation of success.

As with any protocol that involves deception, it is important that researchers using manipulated leaderboards act ethically. We debriefed players and confirmed that they understood that performance was manipulated.

The manipulation worked for the type of game we used, which we feel can be explained by two factors: our game has no score ceiling and performance depends on chance, making it easier for players to accept performance changes between rounds. Games more dependent on skill may have tighter coupling between player action and performance, and leaderboard manipulations may not work as well. We will study leaderboard manipulations in other game genres.

Our leaderboards were displayed after each round; in future work, we will study whether it is more effective to display a leaderboard in real-time during play, showing position changes during the game rather than just being shown the final outcome (e.g., fighting up from the bottom to the top).

CONCLUSION

Conducting experiments in game science is complicated by the need to ensure experimental rigour, while maintaining a sense of playfulness. To induce a sense of success or failure in experiment participants, we manipulated leaderboard position and showed that success led to greater perceived competence, autonomy, presence, enjoyment, and positive affect than failure. Also, displaying the score enhanced the effects on positive affect, autonomy and enjoyment, while not raising players' suspicions that they were being manipulated. Leaderboard manipulations are now part of a growing body of knowledge on developing and validating methods for assessing and inducing pX in GUR.

ACKNOWLEDGEMENTS

We thank NSERC and the GRAND NCE for funding, members of the Interaction Lab, and our participants. Jason thanks Halestorm and Max and Regan thank The National for contributing to this project's playlist.

REFERENCES

1. Abuhamdeh, S., Csikszentmihalyi, M., & Jalal, B. (2015). Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation. *Motivation and Emotion*, 39(1), 1-10.
2. Birk, M., & Mandryk, R. L. (2013). Control your game-self: effects of controller type on enjoyment, motivation, and personality in game. In *Proc. of CHI'13*, 685-694.
3. Bessière, K., Seay, A. F., & Kiesler, S. (2007). The ideal elf: Identity exploration in World of Warcraft. *CyberPsychology & Behavior,* 10, 4, 530-535.
4. Costa, J. P., Wehbe, R. R., Robb, J., & Nacke, L. E. (2013). Time's up: studying leaderboards for engaging punctual behaviour. In *Proc. of Gamification'13*, 26-33.
5. Davison, G. C., Vogel, R. S., & Coffman, S. G. (1997). Think-aloud approaches to cognitive assessment and the articulated thoughts in simulated situations paradigm. *Journal of Consulting and Clinical Psychology*, 65(6), 950.
6. Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *Proc. of CHI EA'04*, 1509-1512.
7. De Grove, F., Cauberghe, V., & Van Looy, J. (2014). Development and validation of an instrument for measuring individual motives for playing digital games. *Media Psychology*, 1-25.
8. Epstude, K., & Mussweiler, T. (2009). What you feel is how you compare: how comparisons influence the social induction of affect. *Emotion*, 9(1), 1.
9. Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117-140.
10. Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, 44(5), 1362-1367.
11. Gerling, K.M., Miller, M., Mandryk, R.L., Birk, M., and Smeddinck, J. (2014) Effects of Balancing for Physical Abilities on Player Performance, Experience and Self-Esteem in Exergames. In *Proc. of CHI 2014*, 2201-2210.
12. Hamari, J., Koivisto, J., Sarsa, H. (2014). Does gamification work?--a literature review of empirical studies on gamification. In *Proc. of HICSS'14*, 3025-3034.
13. Jagacinski, C. M., & Nicholls, J. G. (1987). Competence and affect in task involvement and ego involvement: The impact of social comparison information. *Journal of Educational psychology*, 79(2), 107.
14. John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2, 102-138.
15. Johnson, D., & Gardner, J. (2010). Personality, Motivation and Video Games. In *Proc. of OZCHI'10*, 276-279.
16. Kittur, A., Chi, E. H., & Suh, B. (2008) Crowdsourcing user studies with Mechanical Turk. In *Proc. of CHI'08*, 453-456.
17. Klimmt, C., Blake, C., Hefner, D., Vorderer, P., and Roth, C. (2009) Player Performance, Satisfaction, and Video Game Enjoyment. In *Proc. of ICEC'09*, 1-12.
18. Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2), 76-81.
19. Lazzaro, N. (2004). Why we play games: Four keys to more emotion without story.
20. Mandryk, R.L., Atkins, M., Inkpen, K. (2006). A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In *Proc. of CHI'06*, 1027-1036.
21. Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1-23.
22. McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60, 48-58.
23. Nacke, L. E., Bateman, C., & Mandryk, R. L. (2011). BrainHex: preliminary results from a neurobiological gamer typology survey. In *Proc. of ICEC'11,* 288-293.
24. Nielsen, J. (1994). Heuristic evaluation. *Usability Inspection Methods*, 17(1), 25-62.
25. Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological review*, 91(3), 328.
26. 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.

27. Sweetser, P., and Wyeth, P. (2005). GameFlow: A Model for Evaluating Player Enjoyment in Games. *Computers in Entertainment,* 3, 3.

28. Trepte, S., and Reinecke, L. (2011). The Pleasures of Success: Game-Related Efficacy Experiences as a Mediator Between Player Performance and Game Enjoyment. *Cyberpsychology, Behavior, and Social Networking,* 14, 9, 555-557.

29. Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, 35(2), 297-310.

30. Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical guide to modelling cognitive processes* (Vol. 2). London: Academic Press.

31. Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*.

32. Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

33. Westermann, R., Spies, K., Stahl, GK, & Hesse, FW (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, (26), 557-580.

34. Williams, K. D., & Jarvis, B. (2006). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods*, 38(1), 174-180.