Serial-Position Effects in Evaluative Judgments

Christian Unkelbach and Daniel Memmert

Current Directions in Psychological Science 2014 23: 195
DOI: 10.1177/0963721414533701

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
2014, Vol. 23(3) 195–200

DOI: 10.1177/0963721414533701
cdps.sagepub.com
SAGE

Serial-Position Effects in Evaluative Judgments

Christian Unkelbach[1] and Daniel Memmert[2]
[1]Department Psychologie, Universität zu Köln, and [2]Institut für Kognitions- und Sportspielforschung, Deutsche Sporthochschule Köln

Abstract
Serial evaluations are the basis of many judgment and decision processes (e.g., in sports, talent shows, or academic examinations). We address the advantages and disadvantages of being in the beginning or the end of such evaluation series. We propose that for serial evaluations, people must calibrate a transformation function that translates observable stimulus input (e.g., performances) into available judgment categories (e.g., "pass" or "fail"). Until this function is calibrated, people are motivated to avoid extreme categories. Therefore, being good in the beginning is disadvantageous because one is more likely to be categorized as "average" than "good," whereas being bad is advantageous because one is more likely to be categorized "average" than "bad." We present real-life and laboratory examples of the proposed calibration effects and compare the calibration explanation with other accounts of serial-position effects. Based on these theoretical considerations, we suggest possible ways to avoid these position effects in serial evaluations.

Keywords
serial-position effects, calibration, range frequency, consistency, evaluative judgments

Serial evaluations play a central role in many judgments and decisions; referees evaluate series of game situations ("Foul?"), professors evaluate series of students ("Fail?"), and consumers evaluate series of products ("Expensive?"). We address whether it makes a difference if performances or stimuli appear early or late in an evaluation series. We will show that evaluations of the same performance when it is at the beginning of a series differ systematically from evaluations when it is at the end. Specifically, judges evaluate good performances less positively in the beginning compared with the end, and, conversely, they evaluate poor performances less negatively in the beginning compared with the end. We will present examples of this effect, a theoretical explanation of why it occurs, empirical support for the theoretical assumptions, and possible ways to avoid the effect.

Being First Is Bad When You Are Good and Good When You Are Bad

Two real-world data sets illustrate our proposed serial-position effects in evaluations. First, in the TV show *Come Dine With Me*, five hosts prepare dinners, each on a different day from Monday to Friday, and audiences evaluate the hosts' performances. Figure 1a shows the frequency of winning performances as a function of day of the week, based on data from 4 years of the show's run in Germany. The pattern shows that people hosting the Monday dinner won the contest much less frequently, even though hosts were randomly assigned to weekdays (Unkelbach & Ostheimer, 2013). Second, in soccer, severe rule transgressions are punished with a "yellow card." Figure 1b shows how often referees issue this warning as a function of playing time across five seasons of the Bundesliga, Germany's highest soccer league (Memmert, Unkelbach, Rechner, & Ertmer, 2008). The pattern shows that within the first 15 minutes of a game, the likelihood of players' receiving a yellow card is reduced substantially.

In both examples, judges avoid extreme categories (i.e., extreme positive ratings, extreme punishments) in

Corresponding Author:
Christian Unkelbach, Department Psychologie, Universität zu Köln, Immermannstrasse 49-51, 50931 Köln, Germany
E-mail: christian.unkelbach@uni-köln.de

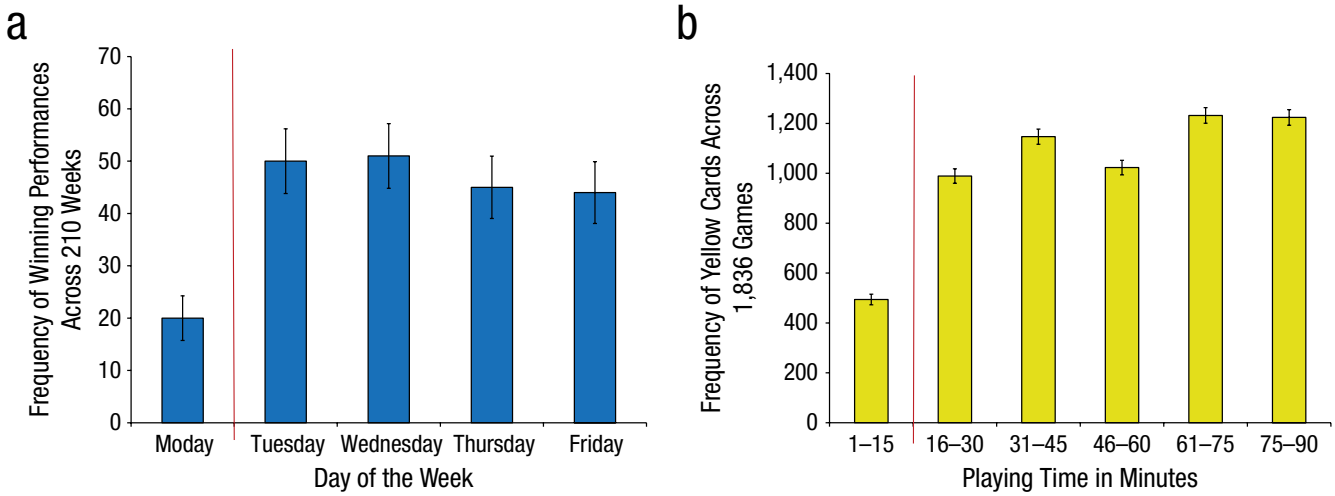

Fig. 1. Real-life examples of serial-position effects in evaluations. The graph in (a) shows the frequency of winning performances in the TV show *Come Dine With Me* as a function of the day of the week the show was aired in Germany across 210 weeks (Unkelbach & Ostheimer, 2013). The graph in (b) shows the frequency of referee warnings ("yellow cards") issued in games as a function of elapsed playing time across 1,836 Bundesliga soccer games (Memmert, Unkelbach, Rechner, & Ertmer, 2008). In both graphs, error bars represent standard errors of the means, estimated according to Lunney (1970).

the beginning. This avoidance of extreme judgments can be found in many other data sets—for example, in data from singing contests (e.g., the Eurovision Song Contest; de Bruin, 2005), sports competitions (e.g., Olympic figure skating; de Bruin, 2006), or experimental laboratory research (e.g., serial evaluations of stimuli; Förderer & Unkelbach, 2013). In addition, in professional sports, there is awareness of the effect. For example, gymnastics coaches exploit this effect and order performers in the team according to their talent, having the best performer always go last (Plessner, 1999).

Obviously, there are alternative explanations for these patterns. Soccer games might not be as intense in the beginning, and TV producers might save the best for the end. Yet the consistency of the data across domains begs the question of whether there is a more general explanation for why being at the beginning of a series is bad when you are good (e.g., hosting a great dinner) and good when you are bad (e.g., committing a nasty foul).

A Calibration Explanation of Serial-Position Effects in Evaluations

An analysis of the judgment situations provides a suggestion of why this effect occurs. Most evaluations are guided by categorical rating systems. School performances are graded from A to F, customer satisfaction is rated with 0 to 5 stars, and scientific manuscripts are placed in the categories "accept," "reject," or "revise." To use categorical rating systems, judges must develop a transformation function that translates observable input into the provided categories. Parducci's (1965) range-frequency theory, for example, provides such a transformation function. However, in the beginning of a series, the function's parameters (e.g., the range) are not yet fixed and need to be inferred or developed. We have called the development of this function "calibration" (Unkelbach, Ostheimer, Fasold, & Memmert, 2012). To explain the observed serial-position effects, we introduce a motivational component, derived from Haubensak's (1992) *consistency model*: Judges must have a need to preserve their judgmental degrees of freedom to avoid consistency violations. Thus, because the function is not calibrated in the beginning, people are motivated to avoid categories that make consistency violations more likely—that is, extreme categories.

To illustrate this need, imagine a professor administering a series of oral examinations. Using extreme categories—that is, awarding an A or an F—defines one end of the range. However, subsequent students might be much better or much worse. If the first student fails, then to not fail other students with similar performances would constitute a consistency violation. Likewise, using the same category (e.g., an A grade) for very different performances also constitutes a consistency violation. In comparison, the categories B, C, and D allow at least one further judgment that will for sure not violate judgmental consistency. Extreme categories thereby reduce judgmental degrees of freedom most strongly, leading to higher likelihoods of consistency violations. We have called the consequential avoidance of extreme categories in the beginning the *calibration effect* (Fasold, Memmert, & Unkelbach, 2013; Unkelbach et al., 2012).

The calibration explanation thereby locates the cause for serial-position effects in a motivational tendency, a need to avoid extreme categories in the beginning. This is

in contrast to cognitive models of serial judgments, such as social-comparison (Festinger, 1954; Suls, Martin, & Wheeler, 2002), range-frequency (Parducci, 1965; Parducci & Wedell, 1986), decision-by-sampling (Stewart, Chater, & Brown, 2006), or anchoring models (Petrov & Anderson, 2005). The main differentiating prediction is that motivational calibration effects occur as a result of the anticipation of subsequent stimuli and judgments, whereas cognitive models must rely on preceding stimuli and judgments.

Uncertainty, Length of the Series, and Consequences

Based on our theoretical considerations, three factors influence the strength of the expected calibration effect in serial evaluations: the uncertainty of the judgment situation, the length of the series, and the consequences of the judgments and decisions.

Uncertainty

Without uncertainty in the judgment situation, no calibration effects should occur. For example, when performances are measurable on physical scales (e.g., running time), the translation of observed input to evaluative categories is a matter of using mathematical rules consistently. Also, to return to the previous examples of decisions in soccer games and academic exams, a foul might be so severe that a yellow card is certain, or a student might answer all questions so fast and precisely that an A is certain; we would not expect calibration effects in such situations. While the need for some ambiguity on the stimulus side seems clear, uncertainty on the judges' side is less clear. One might infer that calibration effects should be stronger for novice than expert judges. Yet expertise might not only reduce uncertainty but also increase consistency-violation concerns; that is, expert judges might show stronger calibration effects just because they are particularly aware of the need to preserve judgmental degrees of freedom.

Series length

Longer series should lead to stronger calibration effects. If one assumes a normal distribution of performances and that only a fraction x should fall into extreme categories (e.g., the best and worst 5%), then the chance that the first performance will belong to the worst or the best is less than x/n, with n being the length of the series; in other words, to wrongly place a stimulus or performance into an extreme category in the beginning increases with n. Conversely, if judges evaluate only one performance, they should not avoid extreme categories because there is no possibility of violating consistency and, thus, no need to preserve judgmental degrees of freedom. An open question in this regard is what people perceive as the series to which they calibrate their judgment functions; for example, a soccer referee might see her whole career as a long series of judgments. At present, we believe that people use the salient units of the empirical world to define a series—that is, a given competition, a game, or one application round.

Consequences

Calibration effects should occur only for evaluations with consequences. However, most evaluations have consequences, even if it is only the fact of being right (Scherer, Windschitl, & Smith, 2013) or avoiding the unpleasant feeling of consistency violations (Gawronski & Strack, 2012; Heider, 1958). The quantitative prediction is nevertheless that calibration effects should be stronger if the stakes are high in a given evaluation, but they are most likely never completely absent, even if evaluations seem to have no consequences.

In sum, calibration effects should occur for consequential evaluations about ambiguous stimuli or performances in series of $n > 1$.

Empirical Evidence

In the following section, we present empirical evidence for our theoretical claims about serial-position effects in evaluative judgments.

The importance of the series

In one experiment (Unkelbach & Memmert, 2008; Experiment 1), referees of the German Football Association saw ten scenes in which a player committed a foul and, for each one, judged whether to award a yellow card or not—that is, whether to use an extreme category or not. Expert raters preselected the foul scenes to have about a 50% chance of justifying a yellow card. The central manipulation was that half of the referees judged the scenes as a series in the chronological order of a game. The other half judged each scene on its own, in a purely random order. The series condition led, on average, to significantly fewer yellow cards (36.4%) compared with the random condition (47.3%). Thus, the series condition led judges to make fewer extreme-category judgments, whereas referees in the random-order condition were close to the 50% expert benchmark.

The importance of anticipation

The experiment described above showed the importance of the series, but this effect could have been due to many

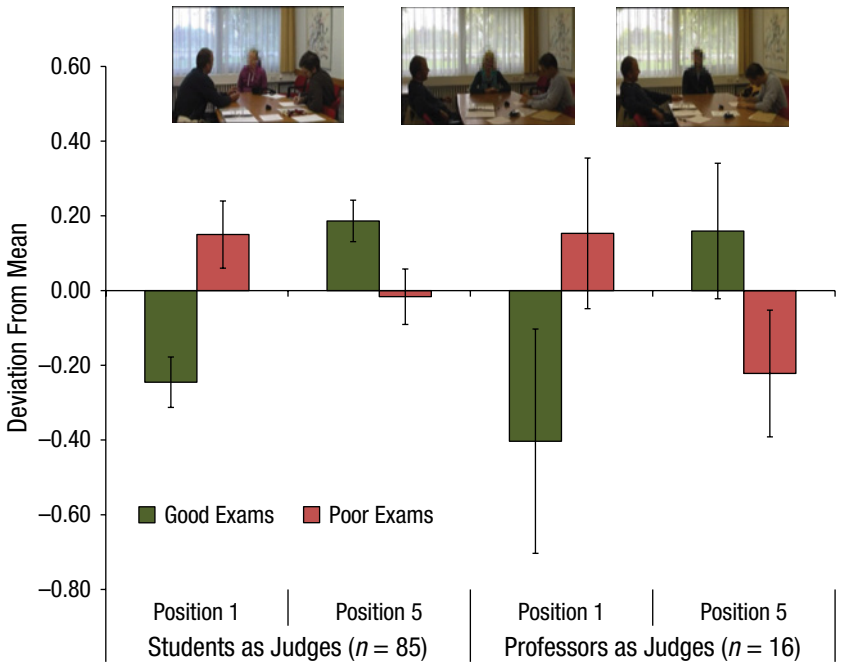

Fig. 2. Standardized changes in the evaluation of good and poor oral-exam performances as a function of serial position (Position 1 vs. Position 5) and judges' expertise (students vs. professors; Unkelbach, Ostheimer, Fasold, & Memmert, 2012). Positive values indicate better grades. The inserted pictures show screenshots from the videotaped oral exams.

other influences. In another experiment, we tested whether the mere anticipation of subsequent judgments causes calibration effects (Fasold, Memmert, & Unkelbach, 2012). Judges evaluated a single gymnastics performance, using the categories "bad," "average," and "good." Half of the judges expected to evaluate eight performances, and the other half expected to evaluate only one performance. In the latter condition, there was no need to preserve judgmental degrees of freedom, so judges could use extreme categories (in this case, "good" and "bad") without consistency-violation concerns. Conversely, we expected that judges in the former condition would avoid extreme categories for their first judgment, as this would increase the probability of consistency violations. In the one-performance condition, 42.9% of the judges used the "bad" category and 14.3% used the "good" category to evaluate the performance. In the eight-performances condition, only 20% of the judges used the "bad" category and only 5% used the "good" category. Thus, judges without the need to avoid consistency violations used the extreme categories more than twice as often (57.2% vs. 25%) for the same performances. Given that the design varied only the anticipation of subsequent performances and included only a single judgment, it precludes explanations that rely on preceding performances.

Expert and novice judges

In two further experiments, judges evaluated videotaped student performances in oral exams (Unkelbach et al., 2012; Experiments 1 and 3). We used video material from real university examinations and selected consensually good and poor student performances. Our variable of interest was the grading of the same performance at Position 1 in comparison with Position 5 in a series of six exams. We expected judges to evaluate the same good performances less positively when it was at Position 1 than when it was at Position 5 and the same poor performance less negatively when it was at Position 1 than when it was at Position 5. The central manipulation between conditions was therefore whether poor or good performances appeared at Positions 1 or 5.

Figure 2 presents the changes in grades due to position for novice judges (students) and expert judges (university professors who administered these exams in reality). Judges evaluated good performances significantly worse when they were at Position 1 compared with Position 5, whereas they evaluated poor performances significantly better when they were at Position 1 compared with Position 5. In addition, across the two experiments, good performances received an A in 11.5%

of the cases at Position 5, but at Position 1, no A was awarded at all. Conversely, poor performances received an F in 22.4% of the cases when they were at Position 1, but 30.6% received an F at Position 5. Replicating the real-life data, being first was bad for good students, but good for bad students.

Alternative Explanations

As stated above, other models predict serial-position effects as well. One prominent example is the decision-by-sampling model (Stewart et al., 2006). According to this account, judges' evaluations of items depend on their ordinal rank and are based on binary comparisons in an available sample. For example, applied to our yellow-card example, referees should evaluate the severity of rule transgressions in comparison to transgression severity in an available sample. In the beginning of a game (e.g., for the very first foul), there is no sample available from the given context, so the sample must come from long-term memory, and referees' long-term memory should cover the full range of transgressions, from very light to very severe. Later in the game, the sample is based on the game context, and the range of transgressions in a given game is by all likelihood more restricted than the range from long-term memory (i.e., it is unlikely that the context includes the worst foul observed within a referee's career). Thus, severe transgressions will appear less severe in the beginning of a game because of the comparison with the full long-term range; similarly, good performances will appear less good compared to excellent performances available from memory. Later in the series, the range will be narrower, leading to higher or lower ordinal ranks within the sample and the resulting evaluation effects.

The decision-by-sampling account predicts the same serial-position effect as the present calibration explanation (see also Stewart, 2009). The main difference lies in the nature of the explanation. The present calibration explanation locates the effect within the motivation to avoid inconsistencies with anticipated subsequent evaluations. Cognitive accounts, such as the decision-by-sampling account, must rely on preceding stimuli and judgments (e.g., retrieved from memory or the context). And empirically, we have suggestive evidence that these differences are due to anticipation (see above; Fasold et al., 2012), which should not occur in a cognitive decision-by-sampling explanation. Nevertheless, the example shows that other models and theories predict and explain serial-position effects. It will be a challenge for future research to establish the boundaries of the present account and to set it apart from existing theories and models.

Possible Solutions for Calibration Effects

If the calibration explanation is correct, then two strategies might reduce possible calibration effects.[1] First, judges should make end-of-sequence instead of step-by-step evaluations; that is, judgments should be made after the series has been observed. Preliminary evidence has suggested that end-of-sequence judgments eliminate calibration effects (see Unkelbach et al., 2012, Experiment 4). Yet there are memory and protocol restrictions that limit the applicability of this approach—for example, when oral examinations or Olympic competitions demand an immediate performance evaluation. In addition, end-of-sequence judgments do not handle initial evaluative anchors that judges might set during the series; that is, it might be impossible for judges not to evaluate individual stimuli or performances while observing the series, and these evaluations should also show calibration effects. And because adjustments of initial anchors are typically insufficient (e.g., Epley & Gilovich, 2004), calibration effects might occur in end-of-sequence judgments. In other words, a student's poor oral performance that does not receive a "fail" evaluation because of its early serial position should also have a lower likelihood of receiving a "fail" evaluation even in an end-of-sequence evaluation. These problems limit the appeal of end-of-sequence judgments as a solution.

A second, more parsimonious strategy is that judges should avoid construing a sequence of stimuli or performances as a series (see Unkelbach & Memmert, 2008, Experiment 1). If each performance is evaluated on its own, calibration effects should disappear. This strategy should be particularly effective for expert judges, because they can use their expertise without being restricted by strong consistency-violation concerns that might come with it. However, a test of pertinent methods to implement this strategy is lacking at the moment.

Summary

We have here presented a motivational explanation of serial-position effects in evaluative judgments. Judges avoid extreme categories in the beginning of a series because they need to preserve their judgmental degrees of freedom to avoid possible consistency violations across the series. This assumption explains and predicts serial-position effects in evaluations that do not follow, per se, from cognitive models of serial evaluations, such as social-comparison models (for two evaluations) or range-frequency theory (for larger numbers of evaluations). The present account does not contradict these explanations but, rather, complements them with a

motivational factor. This need for calibration is good news for poor performers in the beginning of a series, while good performers might be unlucky to be in the pole position.

Recommended Reading

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*, 472–489. A theoretical model of social-comparison processes for series of two stimuli.

Parducci, A. (1968). The relativism of absolute judgment. *Scientific American*, *219*, 84–90. An easily accessible introduction to range-frequency theory.

Unkelbach, C., & Memmert, D. (2008). (See References). An article that provides evidence for and examples of serial-position effects in soccer refereeing and addresses alternative explanation for serial-position effects in sports.

Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012). (See References). Experimental evidence for calibration effects in student's oral examinations and a discussion of possible solutions.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note

1. Both strategies would also work if a decision-by-sampling account were the underlying explanation for calibration effects.

References

de Bruin, W. B. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*, 245–260.

de Bruin, W. B. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*, 299–311.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, *30*, 447–460.

Fasold, F., Memmert, D., & Unkelbach, C. (2012). Extreme judgments depend on the expectation of following judgments: A calibration analysis. *Psychology of Sport and Exercise*, *13*, 197–200.

Fasold, F., Memmert, D., & Unkelbach, C. (2013). Calibration processes in a serial talent test. *Psychology of Sport and Exercise*, *4*, 488–492.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140.

Förderer, S., & Unkelbach, C. (2013). On the stability of evaluative conditioning effects: The role of identity memory, valence memory, and evaluative consolidation. *Social Psychology*, *44*, 380–389.

Gawronski, B., & Strack, F. (2012). *Cognitive consistency: A fundamental principle in social cognition*. New York, NY: Guilford Press.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 303–309.

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Erlbaum.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*, 263–269.

Memmert, D., Unkelbach, C., Rechner, M., & Ertmer, J. (2008). Gelb oder kein Gelb? Persönliche Verwarnungen im Fußball als Kalibrierungsproblem [Yellow card or no yellow card? Soccer cautioning as a calibration problem]. *Zeitschrift für Sportpsychologie*, *15*, 1–11.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*, 407–418.

Parducci, A., & Wedell, D. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 496–516.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review*, *112*, 383–416.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport & Exercise Psychology*, *21*, 131–144.

Scherer, A. M., Windschitl, P. D., & Smith, A. R. (2013). Hope to be right: Biased information seeking following arbitrary and informed predictions. *Journal of Experimental Social Psychology*, *49*, 106–112.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*, 1041–1062.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1–26.

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect. *Current Directions in Psychological Science*, *11*, 159–163.

Unkelbach, C., & Memmert, D. (2008). Game-management, context-effects, and calibration: The case of yellow cards in soccer. *Journal of Sport & Exercise Psychology*, *30*, 95–109.

Unkelbach, C., & Ostheimer, V. (2013). *I don't like Mondays: Why Monday performances are evaluated differently*. Unpublished manuscript, University of Cologne, Cologne, Germany.

Unkelbach, C., Ostheimer, V., Fasold, F., & Memmert, D. (2012). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes*, *119*, 103–113.