Journal of Experimental Psychology: General
2000, Vol. 129, No. 4, 508–523

0096-3445/00/$5.00 DOI: 10.1037//0096-3445.129.4.508

When Does Duration Matter in Judgment and Decision Making?

Dan Ariely
Massachusetts Institute of Technology

George Loewenstein
Carnegie Mellon University and
Center for Advanced Study in Behavioral Sciences

Research on sequences of outcomes shows that people care about features of an experience, such as improvement or deterioration over time, and peak and end levels, which the discounted utility model (DU) assumes they do not care about. In contrast to the finding that some attributes are weighted more than DU predicts, Kahneman and coauthors have proposed that there is one feature of sequences that DU predicts people should care about but that people in fact ignore or underweight: duration. In this article, the authors extend this line of research by investigating the role of conversational norms (H. P. Grice, 1975), and scale-norming (D. Kahneman & T. D. Miller, 1986). The impact of these 2 factors are examined in 4 parallel studies that manipulate these factors orthogonally. The major finding is that response modes that reduce reliance on conversational norms or standard of comparison also increase the attention that participants pay to duration.

Intertemporal choices—decisions with consequences that extend over time—are both common and important. Whether to save money or splurge, diet or indulge, devote oneself to learning a foreign language or indulge in a sitcom are a few examples of the myriad intertemporal choices that most people confront on a daily basis. The outcome of these decisions have momentous consequences, not only for individuals but, as Smith (1976) pointed out in *An Inquiry Into the Nature and Causes of the Wealth of Nations,* for whole societies. Not surprisingly, then, the topic of intertemporal choice has attracted considerable attention from empirical researchers in diverse disciplines.

Until recently, however, research on intertemporal choice has been curiously limited; it has dealt almost exclusively with single, discrete outcomes (e.g., pellets delivered to a rat). As embodied in the dominant discounted utility model (DU), it was assumed that the findings involving simple outcomes could be extrapolated in a simple fashion to more complex *sequences of outcomes.* That is, DU assumes that the value of a sequence of outcomes corresponds to the sum of the discounted values of its component parts. Recent research has challenged this assumption by showing that people care about the temporal relationships between outcomes in a way that is not predicted by DU (Loewenstein & Prelec, 1993). These findings are important because most real-world intertemporal choices are not between individual, discrete outcomes but between sequences of outcomes. For example, restaurant meals typically have multiple courses, vacations have multiple days (and possibly locations), jobs extend over time and have ups and downs, and even one's daily commute may consist of a string of episodes (e.g., easy suburban driving, highway congestion, search for a parking space, etc.). So choices between meals, vacations, or jobs almost always entail choices between sequences of experiences.

The most consistent finding from the research on preferences for sequences is that people prefer sequences of experiences that improve over time. Consider, for example, four dental treatments spaced over a week for which the intensity of pain either increases {2, 3, 4, 5} or decreases {5, 4, 3, 2}. Although both sequences deliver the same total amount of discomfort, most people prefer the sequence of decreasing pain (see Ariely, 1998; Chapman, 2000). Note that time discounting predicts the opposite—that people would prefer to experience the best (or least bad) outcome first and the worst outcome last. Preferences for improving sequences have been demonstrated in many domains, such as monetary payments (Loewenstein & Sicherman, 1991), life experiences such as vacations (Loewenstein & Prelec, 1991, 1993), emotional episodes (Fredrickson & Kahneman, 1993; Varey & Kahneman, 1992), TV advertisements (Baumgartner, Sujan, & Padgett, 1997), queuing experiences (Carmon & Kahneman, 1996), pain (Ariely, 1998; Ariely & Carmon, 2000), discomfort (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Ariely & Zauberman, 2000;

Dan Ariely, Sloan School of Management, Massachusetts Institute of Technology; George Loewenstein, Department of Social and Decision Sciences, Carnegie Mellon University and Center for Advanced Study in Behavioral Sciences, Palo Alto, California.

We are grateful for financial support from the Sloan School of Management and Carnegie Mellon University, Integrated Study of the Human Dimensions of Global Change (National Science Foundation Grant SBR-9521914). This article was written while George Loewenstein was on sabbatical at the Center for Advanced Study in the Behavioral Sciences; George Loewenstein's sabbatical research was supported by National Science Foundation Grant SBR-960123 to the Center for Advanced Study in Behavioral Sciences.

We are also grateful for comments and suggestions from Shane Frederick, Colin Camerer, Drazen Prelec, Richard Thaler, Charles Schreiber, and Daniel Kahneman. The usual caveats apply.

More information about the authors can be obtained by visiting their Web sites at www.mit.edu/ariely/www or www.hss.cmu.edu/departments/sds/faculty/loewenstein.htm.

Correspondence concerning this article should be addressed to Dan Ariely, Massachusetts Institute of Technology, 38 Memorial Drive, E56-329, Cambridge, Massachusetts 02142, or to George Loewenstein, Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890. Electronic mail may be sent to ariely@mit.edu or gl20@andrew.cmu.edu.

Schreiber & Kahneman, 2000), medical outcomes and treatments (Chapman, 2000; Redelmeier & Kahneman, 1996; Katz, Redelmeier, & Kahneman, 1997), gambling (Ross & Simonson, 1991), and academic performance (Hsee & Abelson, 1991). In addition, Hsee and his colleagues (Hsee & Abelson, 1991; Hsee, Abelson, & Salovey, 1991; Hsee, Salovey, & Abelson, 1994) have found that a sequence's rate of improvement or "velocity" affects its overall evaluation.

Improvement is not the only attribute that has been examined. For example, some research suggests that people care about the *spread* of a sequence—how evenly good parts and bad parts are distributed over time (Loewenstein & Prelec, 1993). Other work suggests that the continuous nature of the sequence, whether it is perceived as a continuum or as a series of discrete events, changes the way it is evaluated (Ariely & Zauberman, 2000). There is also considerable evidence that people care about the peak and final value of sequences (Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Schreiber & Kahneman, 2000).

All of these findings (spread, partitioning, peak, end, improvement) involve features of sequences that people care about that, according to DU, they should not care about. In contrast, Kahneman and coauthors (Kahneman, Wakker, & Sarin, 1997; Schreiber & Kahneman, 2000) have recently argued that there is one feature of sequences that DU predicts they should care about but which they do not (or not as much as DU predicts): the duration of the sequence. DU predicts that the total utility of a constant sequence of pleasure or pain is exactly proportional to the duration of the sequence (ignoring considerations of time discounting). Kahneman and coauthors suggested in earlier articles (e.g., Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Redelmeier & Kahneman, 1996) that people ignore or severely underweight duration (which they referred to as *duration neglect*). In later articles, they demonstrated that people do not evaluate sequences in the multiplicative fashion predicted by DU—a phenomenon they label an *additive duration effect* (Schreiber & Kahneman, 2000). The additive duration effect means that people do care at least weakly about duration but that their concern for duration does not depend on the intensity of the stimuli whose duration is varied. DU, in contrast, predicts that the impact of the duration of an experience depends on its intensity; it predicts, for example, that people should care much more about how long a 110-V shock lasts than they care about how long a 10-V shock lasts. The additive duration effect would imply that people's aversion to extending the shock does not depend on the intensity of the shock, which, if true, could lead to extremely suboptimal decision-making behavior.

Our focus in this article is not on the question of whether people neglect duration, either globally or in the additive sense, but on factors that influence the weight that decision makers place on duration. As Kahneman and his coauthors acknowledge, people's concern for duration is unlikely to be fixed across situations but is likely to be greater in some situations than in others.

We examine two factors that, on the basis of prior research, we expected to affect the weight that people would place on duration. First, we predict that the weight placed on duration would depend on the nature of the evaluation that people are asked to make and specifically whether people are asked to rate the desirability of different sequences or make decisions about them (e.g., price them or choose between them). Second, we predict that the weight placed on duration would depend on whether people evaluate sequences one at a time or in an explicitly comparative fashion. After reviewing past literature on the role of duration in evaluations of sequences, we turn to a discussion of the theoretical considerations that led us to focus on these two factors. Our empirical analysis in the following section examines the impact of both of these factors. We end with a discussion of normative issues regarding the role of duration in encoding and decision making.

Summary of Past Findings on the Importance of Duration

Varey and Kahneman (1992) were the first to draw attention to the problem of duration neglect. They presented participants with hypothetical experiences that differed in duration and in intensity-pattern over time and asked them to provide a global evaluation of each experience on a 0–100 scale. Varey and Kahneman found that ratings of these experiences were primarily based on the maximum and final intensities of the experiences, with little weight on duration. They also observed violations of monotonicity. For example, participants rated the overall pain in the hypothetical sequence {2, 5, 8} as worse than the overall pain in the sequence {2, 5, 8, 4}. (Large numbers in the sequence refer to greater pain than small numbers.)

Whereas this first study (Varey & Kahneman, 1992) was *prospective,* in the sense that research participants evaluated sequences that they had not previously experienced, subsequent investigations of the impact of duration have all been *retrospective,* meaning that participants evaluate sequences to which they have been previously exposed. By examining retrospective, as opposed to prospective, evaluation of sequences, previous research has focused on how people remember and encode past experiences rather than on how they choose experiences.

The first of these investigations (Fredrickson & Kahneman, 1993) focused on the role of duration in overall retrospective ratings of affective episodes. In the first study reported in that article, the authors showed participants either long and short movie clips that were pleasant (e.g., a puppy playing) or unpleasant (e.g., an amputation) and asked them to provide a global rating of the pleasantness or unpleasantness of the experience. Fredrickson and Kahneman's results showed that after accounting for the maximum and final intensities, duration had little impact on these participants' overall evaluations. A second study reported in the same article used rankings instead of ratings. Participants ranked sequences of pleasant or unpleasant films in order of overall pleasantness or unpleasantness. Again providing support for duration neglect, after accounting for the peak and end intensities, rankings of pleasant and unpleasant film clips were unrelated to the duration of the clips.

Redelmeier and Kahneman (1996) conducted a study with patients who underwent colonoscopy or lithotripsy. Patients were asked to report the total pain they experienced on a 10-point scale. The treatments in their study varied substantially in the amount of time they took (4–67 min for colonoscopy and 18–51 min for lithotripsy). Nevertheless, the results showed no significant correlation between the duration of the procedure and its retrospective evaluation. A similar neglect of duration emerged when Re-

delmeier and Kahneman asked for the physicians' retrospective evaluations of the patients' pain.[1]

Three other studies obtained mixed support for duration neglect with aversive stimuli. Schreiber and Kahneman (2000) found that longer unpleasant sounds were evaluated as worse than shorter sounds. The effect of duration in their experiments was apparent when examining the effect of duration as the only independent variable and also after accounting for the maximum and ending intensities of the sequences. Schreiber and Kahneman pointed out that although the effect of duration was substantial, it was not multiplicative, which they referred to as an additive duration effect. Finally, although they did not test the idea directly, Schreiber and Kahneman suggested that duration was salient in their experiments because participants evaluated multiple sounds of varying durations.

Ariely (1998) compared the impact of duration on ratings of stimuli that did not change much over time (*constant*) and stimuli that had substantial changes in their magnitude over time (*patterned*). His experiments, which involved real pain (produced by either a heat probe or pressure applied to the finger), revealed that although the ratings of constant sequences were not affected by changing their duration, the ratings of patterned sequences were sensitive to their duration. His research thus suggests that attention to duration may depend, in part, on the specific nature of the sequences being evaluated.

Another factor that seems to influence attention to duration is attentional focus. In a study that illustrated the importance of focus of attention, Rinot and Zakay (1999) had some participants evaluate the overall annoyance they experienced from each sound in a series of annoying sounds and had others evaluate both the overall annoyance and the duration of each experience. The overall evaluations of participants in the latter condition were more sensitive to duration.

In addition to relying mostly on retrospective judgments, the studies just reviewed used mostly (but not exclusively) ratings as the dependent measure. Two other studies—one involving sequences of discomfort from cold water (Kahneman et al., 1993) and the other involving sequences of aversive noise (Schreiber & Kahneman, 2000)—used, in addition to ratings, dependent measures involving choice. In both studies, participants chose between specific sequences of aversive sensations (cold water and noise) and dominated sequences that were created by adding a mildly aversive segment—that is, an extra period of mildly cold water or an extra period of moderately loud sound. In the cold water study (Kahneman et al., 1993), participants experienced two trials of cold water discomfort, one short and one long. In the short trial, participants immersed their hand in mildly cold water (14°C/57°F) for 60 s. In the long trial, they immersed their hand in the same cold water (14°C/57°F) for 60 s followed by 30 s at a slightly more comfortable temperature (15°C/59°F). When participants later chose which of the two trials to repeat, a significant majority (69%) opted to repeat the long trial. In the aversive noise study (Schreiber & Kahneman, 2000), participants were exposed sequentially to two sequences of aversive noise that were identical except that one sequence added a period of mildly aversive noise at the end. Despite the fact that the shorter sequence dominated the one with noise added to the end, the majority of participants violated dominance for several of the stimulus pairs—that is, they chose a sequence that contained the other sequence plus some discomfort.

Kahneman and his coauthors (Kahneman et al., 1993; Schreiber & Kahneman, 2000) concluded from the fact that participants chose to repeat the longer, dominated sequence that they must not be attending to duration. However, alternative conclusions are plausible, such as that participants put considerable weight on end level or on final slope. In fact, all that can be concluded from dominance violations is that participants do not base their overall evaluations of sequences on the integral (sum) of pleasure or pain. Any deviation from integration—such as giving special weight to peak, end, final slope, or any other specific feature of the sequence—can produce violations of dominance, regardless of whether participants do or do not attend to duration. Consider, for example, the sequences of pain {2, 5, 8} and {2, 5, 8, 4} in which the former dominates the latter. An individual who based his or her overall value of a sequence on the sum plus the final slope,

$$v(x_1, x_2, \ldots x_n) = \Sigma x_i + (x_n - x_{n-1}),$$

is fully sensitive to duration but would nevertheless value the former at 18 and the latter at 15 and would thus prefer the latter. Thus, violations of dominance do not, in and of themselves, point to duration neglect. Neglect of duration could, of course, produce or contribute to violations of dominance in the same way that disproportionate weighting of final slope or end level could, but duration neglect is not a necessary condition for dominance violations to occur. In these studies, therefore, it is possible that participants cared a lot about duration but that their concern for duration was overwhelmed by either a preference for improvement or (closely related) a preference for ending on a good note. To test for duration neglect, per se, requires a systematic investigation of the impact of duration on evaluations of sequences as compared with one or more other sequence features (e.g., peak, end, slope, average intensity).

In sum, support for complete duration neglect is mixed. Among those studies that have systematically manipulated duration, several have failed to observe any impact of duration on ratings of hedonic stimuli. However, some studies have observed some impact of duration—on rankings of aversive sequences of stimuli (Schreiber & Kahneman, 2000), ratings of patterned stimuli (Ariely, 1998), and ratings of aversive sequences of stimuli—when duration was also estimated (Rinot & Zakay, 1999). In addition, two studies involving choice have documented violations of dominance that could be, but are not necessarily, attributable to duration neglect.

Whether people do or do not take duration into account in retrospective evaluations or the degree to which they do so may, however, be inherently unproductive questions to explore. In fact, there appear to be some situations in which people place little weight (possibly zero) on duration and others in which they care about it a lot. A more fruitful line of inquiry, therefore, may be to investigate when and why respondents show much or little concern for duration. Next, we detail two factors that we predict would

[1] In an extension of the Redelmeier and Kahneman study (reported in Kahneman et al., 1997), some patients were assigned to a condition in which the procedure was extended by leaving the colonoscope in place for about a minute after the completion of the clinical examination. The authors report that the prolongation of the colonoscopy produced a significant improvement in the global evaluations of the procedure.

affect the concern that respondents showed for duration: the type of evaluation being made and whether the evaluation is or is not explicitly comparative.

Mechanisms That Could Affect the Weight Placed on Duration

Type of Evaluation Goal (Rating vs. Decision)

As noted in the review of past research on duration neglect, many studies examining the impact of duration have used ratings as a primary dependent measure. The implicit argument in this work—indeed the motivation for conducting it—has been that if people neglect duration in ratings of extended episodes, they are also likely to do so when they make choices between such episodes. There are good reasons to question whether this assumption is correct. Like people who are engaged in ordinary conversations, research participants naturally assume that the answers they provide to the questions they are asked will serve some kind of purpose (Clark & Schober, 1991). Different response methods may incorporate duration to a different extent in part because the purpose to which the evaluations are likely to be put is different or is expected to be different. Ratings of experiences are generally used for one of two purposes: (a) to communicate preferences to other persons and (b) to encode one's own preferences for use in future decision making. For either of these purposes, ignoring duration may be perfectly appropriate.

Communicating Preferences to Others

Grice (1975, 1987) proposed that conversationalists attempt (and believe that their partners will try) to make their utterances relate to commonly recognized goals (the maxim of relation). Supporting this assertion, research on conversational norms has shown that speakers alter their communications on the basis of their partner's goals in a conversation (Clark & Wilkes-Gibbs, 1986; Russell & Schober, 1999). The idea that people try to provide conversational partners with information that is relevant to their goals may help to explain why people's ratings of the overall global goodness or badness of sequences might not take their duration into account.

There are several possible ways to interpret the request to provide a global rating of a sequence. Under many, if not most, of these interpretations, duration neglect is entirely appropriate. Suppose, for example, that a colleague asked you, "Overall, how would you rate your recent trip to the Grand Canyon?" In this situation, the appropriate response is probably one that does not incorporate the duration of your recent trip. You would most likely mislead your colleague if you tried to factor the duration of your visit into your response, that is, to rate it more extremely simply because you spent a long time there (except insofar as duration affected your average momentary pleasure from the visit). The typical reason for being asked a question of this type is that the questioner is evaluating the desirability of a visit to the canyon. The most useful answer, which would be in line with the questioner's expectations, is to give some type of average rating of your visit that does not encode duration. If you responded "wonderful" because you had spent a full 2 weeks of slightly better-than-average days at the canyon, the questioner would be severely misled. The same would be true if you responded "awful" because you had spent only one, albeit spectacular, day at the canyon. The questioner may not know how long you spent there, and you are unlikely to know how long the questioner plans to spend there if he or she does visit. Indeed, the questioner may use your answer to the question, in part, to decide how long to spend there.[2]

The implication of conversational norms for laboratory research on duration neglect is that the insensitivity to duration often observed in summary evaluations of sequences may reflect participants' interpretation of the question they are being asked, rather than an actual lack of concern for duration.

In all the experiments discussed here, care was taken to use wording that would elicit overall global evaluations that sum the experience over time (using wording such as "global evaluation of how bad the overall experience is," "the total amount of pain," "maximize the overall pleasantness of the experience").[3] Even if people understand what is being asked for, however, "total amount of pain" may be an alien concept for participants who are not used to summing pain over time. Asking for total quantities of experiences, such as sleep or calorie intake, is perfectly normal, but in other cases, the concept of a total is less well defined. For example, most people would find it difficult to answer a question such as, "what was the total volume of the rock concert?" or "what was the total windiness this morning?" Variables such as badness (Varey & Kahneman, 1992), pleasure (Fredrickson & Kahneman, 1993), and pain (Redelmeier & Kahneman, 1996) lie somewhere between these two extremes. Reporting the total pain one experiences over an interval is noncustomary, although it is conceivable that people could do it if they were asked to do so. In all of these cases, participants could plausibly interpret the question they are being asked as one that does not call for an explicit incorporation of duration. Given the ambiguity over what the question is calling for, duration neglect in overall ratings may be sensible and not necessarily indicative of the concern that participants would show for duration if they were actually making decisions involving sequences of experiences.

---

[2] Arguing that duration neglect may sometimes be sensible does not mean that duration should be neglected in all settings. Childbirth, jail sentences, and waiting in line are some of the experiences that are readily and naturally described in terms of their duration. In fact, when asking, "How bad was your wait in line at the supermarket?" the questioner is implicitly asking, "How long did you have to wait?" perhaps with minor allowances for the conditions under which the waiting occurred. (For an interesting discussion of queuing experiences, see Carmon & Kahneman, 1996.)

[3] In the second study by Fredrickson and Kahneman (1993), participants were instructed to rank sequences to help the experimenters select clips for inclusion in a future study. For pleasant (or unpleasant) sequences, participants were asked to "MAXIMIZE [or MINIMIZE] the overall pleasantness [or unpleasantness] of the experience of viewing the pleasant [or unpleasant] videotape that we make." On the basis of the instructions they were given, participants may have thought that neglect of duration was appropriate. They may well have believed that it would be most useful for the experimenters to have a measure of the average pleasantness or unpleasantness of the film clips that ignored duration because the researchers would make their own decisions about the duration of each clip that would be included in the final experiment.

Encoding Preferences for Use in Future Decision Making

An analogous argument applies to situations in which one rates an extended episode for use as an input into one's own future decision making. For purposes of future decision making (i.e., deciding whether to repeat a past experience), it is almost certainly more useful to code summary measures of desirability that do not include duration. Such duration-free evaluations allow the decision maker to take the planned duration of future episode into account as he or she wants to at the time of making the decision. If duration were encoded into the stored representation of desirability and the decision maker was deciding whether to experience a new episode of different duration from the one already experienced, judging the new sequence would require the decision maker to partial out the effect of duration from his or her judgment of the initial evaluation and then combine the new duration with it. Such an adjustment requires storage of an additional piece of information (the duration of the original episode) and is, in practice, difficult to perform. Again, as is true for communication among people, duration neglect may be sensible when individuals are trying to encode the goodness or badness of a sequence with the intention of using this information as an input into future decisions. When these evaluations are used for making future decisions, however, it is quite plausible that decision makers will take duration into account.

Duration in Choice

Although duration neglect may be a necessary aspect of conversational efficiency, it is not sensible to completely neglect duration in choices between future sequences of outcomes. A 2-hr dental procedure is unambiguously worse than a similar procedure that ends after 10 min. A person who made the mistake of treating them the same (and ignoring duration in general) would lead a much worse life. Indeed, it is difficult to think of any case of a choice between temporally extended outcomes in which complete duration neglect would be appropriate. Moreover, as Kahneman et al. (1997) pointed out, ignoring duration in choice can lead to violations of normatively compelling principles of choices—most prominently dominance. Numerous studies involving choices between sequences have, in fact, found that decision makers pay robust attention to sequence duration. For example, Read and Loewenstein (1999) elicited people's willingness to experience different durations of coldpressor pain in the future in exchange for payment. Participants' willingness to accept pain in exchange for payment (WTA) in Read and Loewenstein's experiments depended strongly and monotonically on the duration of the pain they would be exposed to. Hoeffler and Ariely (1999) examined how people learn trade-offs between attributes (duration, intensity, and money) over time. Their results clearly showed that participants traded off duration against the other two attributes. Ariely, Loewenstein, and Prelec (1999) elicited participants' willingness to listen to loud noises of varying duration in exchange for payment. Again, WTA depended on duration in a highly systematic fashion. Our first central prediction, therefore, contrasts the role of duration in ratings and decisions: There will be greater sensitivity to duration when participants make decisions about sequences than when they rate them for encoding or communication purposes (Prediction 1).

Separate Versus Comparative Valuation

A second (and closely related) influence on the weighting of duration is whether sequences are evaluated comparatively—by explicitly comparing them with one another—or separately (i.e., one at a time). There is a substantial literature documenting dramatic differences in preference between these two modes of evaluation (Hsee, Loewenstein, Blount, & Bazerman, 1999; Nowlis & Simonson, 1997; Tversky, 1969).

Evaluability

One specific cause of such divergences is what Hsee and colleagues (1996; Hsee et al., 1999) call the evaluability effect: When judging items separately (i.e., one at a time), attributes that are not easily judged independently are given little weight. However, when the same items are judged in an environment that facilitates comparison to other items, respondents place much greater weight on the same attributes. For example, in one study reported by Hsee (1996), participants were asked how much salary they would be willing to pay to two job candidates who differed in their experience with the programming language they would be using and also differed on their undergraduate GPA. One candidate had a higher GPA, whereas the other was more experienced with the programming language. In the *joint evaluation* condition, participants were presented with the information on the two candidates side by side. In the *separate evaluation* condition, participants were presented with the information on only one of the candidates. The results revealed a significant preference reversal as a function of whether the information was presented jointly or separately. In the joint evaluation, willingness-to-pay salaries were higher for the candidate who had more programming experience. In the separate evaluation, willingness-to-pay salaries were higher for the candidate with the higher GPA. Hsee et al. (1999) attribute this effect (and a variety of related effects documented in their review) to the fact that programming experience is an attribute that is difficult to evaluate (people don't know what a good or bad amount of experience would be) and hence receives much lower weight when alternatives are evaluated separately.

Expanding these findings to the issue of duration neglect suggests that (if duration is not easily judged separately), judgments of a single experience without a referent (such as ratings) will show little concern for duration, whereas comparative judgments (in cases in which there is a standard of comparison) will show much higher concern for duration. Indeed, there is some suggestive evidence that people have difficulty evaluating duration as an attribute. Herrnstein, Loewenstein, Prelec, and Vaughan (1993) conducted a series of studies to test Herrnstein's "melioration" theory of choice. The theory applies to situations in which people make choices between alternatives and in which the choices they make have internalities—they affect the quality of alternatives they will face in the future. The concept of melioration refers to the assertion that people ignore such internalities. In a series of studies (Herrnstein et al., 1993), participants experienced a sequence of trials in which they chose between two buttons that caused coins to drop from a hopper and accumulate as earnings. Choosing one button always led to a higher immediate payoff but also caused the value of coins from both hoppers to decline to such an extent that total earnings would be lower. This was the meliorating choice—

the choice people would naturally make if they ignored the effects of their choices on coin value. In some studies, making the meliorating choice caused the value of subsequent coins to decline; in other studies, making the meliorating choice led to an increase in the time delay prior to each drop of a coin (which also led to a decline in total payoffs that was equivalent to the declining coin value conditions). Consistent with the notion that people have difficulty evaluating duration, participants were much more likely to meliorate—that is, to ignore the internality—in the coin delay condition than in the declining coin value condition.

Scale-Norming

A second important cause of discrepancies between comparative and separate evaluation is automatic scale-norming. In their presentation of *norm theory,* Kahneman and Miller (1986) showed that virtually all evaluations automatically evoke some type of norm of comparison—even those that are not explicitly comparative. In rating the Grand Canyon visit, for example, one will likely compare it to other vacation trips he or she took, or even, perhaps, to other trips he or she took to the southwestern United States. In contrast, one is unlikely to compare the Grand Canyon trip to the average restaurant dinner or game of squash. The same principle applies to duration. When one evaluates a particular morning's commute, he or she is unlikely to evaluate it relative to a recent cross-country drive. Duration is one of many variables that people use to classify stimuli for purposes of scale-norming.[4]

Applied to retrospective evaluations of sequences, scale-norming could be an important factor contributing to duration neglect. If participants are asked to rate a series of sequences that differ in duration, it is possible that they will norm each sequence against sequences of similar duration. If they do so, they will exhibit duration neglect, regardless of whether they take duration into account when making decisions about sequences. Thus, if participants are asked to rate sequences of similar duration and duration is only manipulated in a between-subject manner, duration neglect would seem to be virtually inevitable. Duration neglect is also likely to be observed if participants rate experiences that differ in duration but also differ on some other important dimension. For example, if participants rated a pleasurable 2-week vacation and an unpleasant 1-week work trip, they would probably norm the vacation against other vacations and the work trip against other work trips. Duration would then be neglected, even though it was explicitly manipulated in a within-subject design.

Automatic scale-norming of this type would be much less likely, and attention to duration commensurately more likely, if people compared experiences that are similar on most dimensions other than duration. In such a situation, duration would be highly salient, and it would most likely be taken into account. Indeed, Schreiber and Kahneman (2000) have suggested that once participants experience multiple episodes, they begin to rely more heavily on the experience's duration in their judgments (see also Rinot & Zakay, 1999; Ariely & Carmon, 2000). The preceding discussion points to a second major prediction: There will be greater sensitivity to duration when participants engage in evaluations that involve explicit comparisons between sequences than when they evaluate sequences one at a time (Prediction 2).

Why Other Features of Sequences, Such as Patterns, May Not Be Influenced by Evaluation Goals and Comparison Standards

What about other features of sequences, such as their peak, end, and slope? Should we expect these features to differ as a function of these two factors—that is, ratings versus decisions and separate versus comparative evaluation? Again, the answer may lie, in part, in conversational norms and norms of evaluation. In some cases, such features are an inherent, immutable aspect of a sequence. For example, movies provide a specific sequence of affect, and hikes and white-water rafting trips typically provide a relatively invariant sequence of terrain and excitement. In such cases, it is consistent with conversational norms to incorporate these features into one's evaluation. Thus, in recommending a film to another person, it would be a mistake for the recommender to ignore the fact that the movie's happy ending left him or her feeling exhilarated. Evaluative norming also does not normally imply a neglect of these other sequences. Most people don't rate films or wilderness excursions relative to other films or excursions that have similar peaks, ends, or temporal patterns of affect.

The situation changes somewhat for extended experiences consisting of components that do not have an inherent temporal order. For example, consider a 4-day visit to the Grand Canyon in which it rained for either the first 2 or last 2 days. Most people would choose to experience the rain on the first 2 days, consistent with the widespread preference for improving sequences. If the purpose of a rating is to provide a recommendation, however, it would be less normatively desirable to rate such an improving vacation as more desirable because the person requesting the rating may not be aware of the specific weather pattern that prevailed and can certainly not predict the weather that will prevail on his or her prospective visit. It is not clear, however, whether people, in practice, factor out improvement and features that result from the ordering of changeable sequence components. As Schwarz (1996) and Schwarz and Clore (1983) have shown for judgments of subjective well-being, people's tendency to remove such transient influences depends in part on whether the features are made salient. For example, a person would be more likely to factor weather-induced improvement out of his or her ratings of a Grand Canyon visit if he or she was first asked to report on the weather during the trip. These considerations led to our making a third prediction: The impact of sequence pattern (e.g., increasing vs. decreasing) on evaluations will be relatively invariant across rat-

[4] This probably makes perfect sense. To truly take duration into account—by multiplying the utility of each experience by its duration—would imply that one should give any experience a rating of zero. Consider, for example, someone trying to describe the amazingly good day he or she just had on a scale from 0 (*very bad day*) to 100 (*a wonderful day*). In trying to take duration into account, this person will realize that he or she could be asked in the future to say how good were his or her last two days, last week, last month, or even whole lifetime (e.g., when on the verge of death). Given that this person will want to obey the multiplicative criterion (which is equivalent to taking the integral of pleasure or pain), the maximum ratings he or she can give a day is 100 over the maximum amount of days he or she expects to live. In our case (because we expect to live to 90), this would be $100/(364 \times 90) = 0.00305$.

ings versus choice and comparative versus one-at-a-time evaluations (Prediction 3).

We tested these predictions in four parallel experiments in which participants evaluated sequences of aversive noise. The four experiments differed in the type of evaluation that was elicited from participants. To test the first prediction, we designed two of the four experiments to use ratings as a dependent measure and the other two to use measures that involved decisions. To test the second prediction, we had participants in two of the four experiments (crossed orthogonally with the first manipulation) evaluate sequences separately; in the other two experiments, participants evaluated sequences in a comparative fashion, relative to a standard sequence. The four experiments can thus be viewed as composing the four cells of a 2 × 2 factorial design.

In the first experiment, we adopted the one-at-a-time ratings of sequences method used in most prior research: Participants rated the "overall annoyance" of each sequence. In the second experiment, we introduced the element of decision while retaining one-at-a-time evaluation by having participants evaluate willingness to accept monetary compensation for listening to each sequence again. The third experiment involved comparative evaluation without decisions by having participants rate sequences of sound relative to a fixed standard sequence that was identical for all participants and constant across all the trials in the experiment. The fourth experiment involved both comparative evaluation and decision; after exposure to each sequence, participants decided whether they preferred to re-experience that sequence or to experience the standard sequence (the same standard used in the third experiment).

Experiments

Method

Participants

Participants in Experiments 1 (separate ratings) and 4 (choice) were Duke University undergraduates. Participants in Experiments 2 (WTA) and 3 (rating relative to standard) were Massachusetts Institute of Technology undergraduates. Although our main findings relate to comparisons across studies, it is very unlikely that these arise from the differences in participant population, particularly in light of the fact that, as predicted, the greatest differences in attention to duration were observed between Experiments 1 and 4, which were conducted with the same participant population.

Stimuli

The stimuli used in all four experiments were annoying sounds. We selected annoying sounds as stimuli because they have two desirable properties: (a) they permit delivery of many stimuli to a single participant, unlike, for example, cold water, and, (b) they show little or no adaptation over time (Ariely & Zauberman, 2000), which means that subjective stimulus levels correspond closely to objective levels, with little effect of prior exposure.[5] This property is especially important for work on sequences because with adaptation, preference for improvement could be due to the fact that adaptation to adverse early stimuli renders later stimuli less noxious.

To generate the stimuli, we used a tone-generating application (SoundEdit, 1997) and created a 16-bit triangular wave in a frequency of 3000 Hz, 10% amplitude. These sounds were delivered via a computer sound card (Crystal 3D 16bit, IBM). The different intensity levels were created by starting with a single base sound and systematically manipulating its intensity between 50% and 80% (corresponding approximately to 60–80 dB). All stimuli sounded like a high-pitched scream, similar to the broadcasting warning signal. All four experiments used 27 stimuli (see Table 1), which were grouped into two clusters: *constant* and *patterned* (see Figure 1). Constant stimuli did not change in intensity over time. Constant stimuli were presented in three durations (10 s, 15 s, and 22.5 s) and five different intensity levels (which we henceforth refer to as Levels 1 to 5), making a total of 15 different stimuli. Patterned stimuli did change in intensity over time. Patterned stimuli included four specific temporal trajectories (*up, down, up and down,* and *down and up*), each presented in three durations (10 s, 15 s, and 22.5 s), making a total of 12 different stimuli. For a description of the different stimuli, see Figure 1.

Table 1
A Numerical Description of the 27 Annoying Stimuli Used in All Four Experiments

Stimulus name	Pattern	Intensity level			
		Starting	Middle	Final	Mean
Up	Patterned	1	3	5	3
Down	Patterned	5	3	1	3
Up and down	Patterned	1	5	1	3
Down and up	Patterned	5	1	5	3
Constant-1	Constant	1	1	1	1
Constant-2	Constant	2	2	2	2
Constant-3	Constant	3	3	3	3
Constant-4	Constant	4	4	4	4
Constant-5	Constant	5	5	5	5

Note. Each of the 9 stimuli described here was presented in three durations (10 s, 15 s, and 22.5 s), making a total of 27 stimuli.

The four experiments differed in the procedure and dependent measures that participants used to evaluate the 27 different stimuli. Because the main hypotheses involve comparisons across the four experiments, we present the methods from all four studies before presenting results from any of them.

Common Elements

Participants sat in front of a computer and wore headphones. To introduce them to the sounds, we first presented them with sample sounds that spanned the whole range of the stimuli from the weakest to the most extreme constant sounds, as well as the up and the down sounds. After participants had indicated that this was an acceptable range for them to continue with the experiment, we gave them specific instructions (depending on the specific experiment in which they were participating). After completing the experiment, participants were debriefed, paid, and thanked for their participation.

Table 2 summarizes the four experiments in a way that highlights their connection to the two major predictions discussed in the previous section.

Experiment 1

Traditional Method: Separate Ratings of Sound Sequences (Separate Ratings Experiment). After the initial introduction and instructions, participants received each of the 27 stimuli in a random order. On each of

[5] There is a long history of formal and informal observations about the close relation between loudness and annoyance. For example, Stevens (1975, p. 69) commented on the similarity of results obtained when participants are asked to match noises for loudness, noisiness, or annoyance.

the 27 trials, participants were presented with a screen that asked them if they were ready to experience the next sound. Once they answered positively to this question, the trial proceeded and one of the sounds was played. After the sound terminated, participants were asked to rate the sound by answering the question, "Overall, how annoying was it?" and were asked to respond on a 100-point scale with 0 being *not annoying at all* and 100 being *very annoying*.[6]

Experiment 2

Willingness to Repeat Sound Sequences in Exchange for Payment (WTA Experiment). After the initial introduction and instructions, participants were told that the experiment consisted of two parts. In Part 1, they would hear a series of annoying sounds and, after each one, would state the lowest price they would demand as a compensation for hearing the sound again. Participants were told that in Stage 2 of the experiment, the computer would randomly generate a price for each sound. If their stated price was lower than this price, then they would be exposed to the sound again and get paid accordingly. If their stated price was higher, then they would not be exposed to the sound again and not get paid. Participants received each of the 27 stimuli in a random order. On each of the 27 trials, participants were presented with a screen that asked them if they were ready to experience the next sound. Once they answered positively to this question, the trial proceeded, one of the sounds was played, and they stated the minimum price for which they would be willing to listen to the same sound again in Stage 2 of the experiment. We set up the experiment in this way to maintain a similarity in the amount of experience and sounds that participants experienced across all four experiments. During Stage 1 of the experiment, participants made real decisions that they believed had immediate implications for their near future; but, after making these 27 responses, participants were spared a repetition of the annoying stimuli.

Experiment 3

Rating of Sound Sequences Relative to a Fixed Standard (Rating-Relative-to-a-Standard Experiment). After the initial introduction and instructions, participants were asked to listen repeatedly (eight times) to a sound that was labeled the *standard stimulus*. Participants were told to listen carefully to this sound in order to become familiar and remember it for future judgments. This standard stimulus was always constant at a level of three (the midpoint of the range) with a duration of 15 s (the intermediate value of the three stimulus durations). The results suggest that this procedure was successful in helping participants remember the standard sound. The mean rating of the standard was 50.556, which was not significantly different than the desired mean of 50, $t(44) = 0.403$, $p = 0.688$. The implied choice proportions were 49%, which were also not statistically different than the expected 50%, $t(44) = -0.15$, $p = 0.88$. After they became familiar with the standard, participants received each of the 27 stimuli in a random order. On each of the 27 trials, participants were presented with a screen that asked if they were ready to experience the next sound. Once they answered this question positively, the trial proceeded and one of the sounds was played. After listening to each of the sequences, participants were asked, "Overall, how annoying was the sound you just heard compared with the standard?" This annoyance rating was done on a 100-point scale in which 0 meant that the sequence was not annoying at all, 50 meant that the annoyance was equivalent to that of the standard sequence, and 100 meant that the sequence was very annoying.

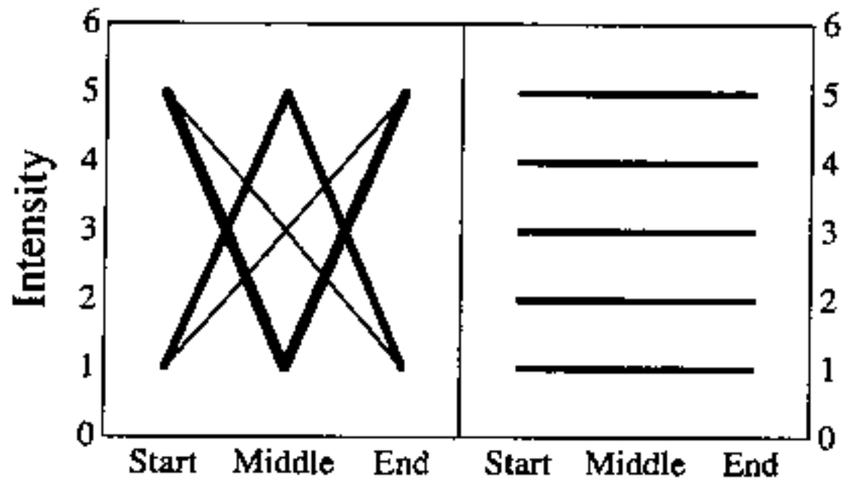

Figure 1. Left panel: A schematic illustration of the patterned stimuli (down, up and down, down and up, up). Right panel: A schematic illustration of the constant stimuli. Each of the 9 stimuli described here was presented in three durations (10 s, 15 s, and 22.5 s), making a total of 27 stimuli.

Table 2
Summary of Four Experiments

Evaluation	Rating	Decision
Separate	Experiment 1 (separate ratings)	Experiment 2 (WTA)
Comparative	Experiment 3 (rating relative to standard)	Experiment 4 (choice)

Note. WTA = willingness to accept pain in exchange for payment.

Experiment 4

Choice Between Sound Sequences (Choice Experiment). After the initial introduction and instructions, participants were asked to listen, learn, and become familiar with the standard stimulus (the same stimulus as used in Experiment 3). After becoming familiar with the standard, participants were told that the experiment had two parts to it. In Part 1 of the experiment, they would get a new annoying sound and upon its termination would be asked to choose whether, in Stage 2 of the experiment, they would prefer to experience the stimulus they just experienced or the standard stimulus. Participants were also told that after making several such choices, they would participate in a second stage of the experiment in which they would be exposed to all the sounds they had chosen in the first stage. We set up the experiment in this way to maintain uniformity in participants' exposure to sounds across the four experiments. During Stage 1 of the experiment, participants made real choices that they believed had immediate implications for their near future; however, after making these 27 responses, participants were spared a repetition of the annoying stimuli.

On each choice occasion, participants saw a scale on the computer screen with a probe at its middle. The scale was anchored on the left by the statement *Absolutely certain I prefer the standard* and on the right by the statement *Absolutely certain I prefer the new sound.* Participants moved the probe by pressing on the right and left arrow keys. They were not allowed to keep the probe at the middle (indifference). The instructions for using this "graded-choice" method emphasized the two aspects of the response. Participants were instructed that their decision to move the probe either left or right from the center of the scale would completely determine the choice outcome they would experience in Stage 2 of the experiment. They were told that distance of movement on the scale should reflect their confidence in a particular choice.

On each of the 27 trials, participants were presented with a screen that asked them if they were ready to experience the next sound. When they

[6] In a different setting, we asked 30 of the participants whether their response would have changed if we had replaced "Overall" with "Total." Twenty-eight of the 30 participants said they would not have changed their responses, and the 2 who changed their opinion gave slightly lower responses.

responded positively, the trial proceeded and one of the sounds was played. After the sound terminated, the graded-choice scale appeared on the screen and participants used it to express their preferences. To help the participants remember the standard, yet not present it too many times, we gave participants the standard stimulus as a reminder on every 6th trial.

To facilitate comparison with the annoyance ratings from the previous studies, we express, when presenting the results, both choices and confidence judgment on scales in which high numbers indicate preference for the standard sequence, that is, in which high numbers indicate that participants disliked a particular sequence. We encoded two dependent measures: *choices*, which were binary variables coded as 0 when the participant chose the focal sequence and coded as 1 when the participant chose the standard sequence, and *graded choices*, which also ranged from 0 (meaning that participants expressed complete confidence in their choice of the focal sequence) to 100 (meaning that participants expressed complete confidence in their choice of the standard sequence).

Results

Impact of Duration (Predictions 1 and 2)

Figure 2 presents the evaluations of the sequences as a function of duration, separately for the four experiments. These evaluations included (a) the raw 0–100 responses in Experiments 1 and 3, (b) the WTA evaluations (which ranged between \$0.01 and \$5.60 with *SD* of 0.28) from Experiment 2, and (c) the 0–100 graded choices from Experiment 4. The first row includes both patterned and constant stimuli, the second row includes only constant stimuli, and the third row includes only patterned stimuli.

Visual inspection of the figure suggests that, consistent with Prediction 1, duration had a greater impact on evaluations in the WTA experiment (in which participants engaged in decisions) than in the separate ratings experiment. Likewise, it appears that, consistent with Prediction 2, duration had a greater impact on evaluations in the rating-relative-to-standard experiment than in the separate ratings experiment. Consistent with the idea that both considerations are important, the impact of duration appears to be greatest in the choice experiment.

To provide a more rigorous test of the relative impact of duration across the four experiments, in each experiment we regressed each participant's evaluations against duration. The results are summarized in the top half of Table 3, which compares the impact of duration on responses in the four experiments on the basis of

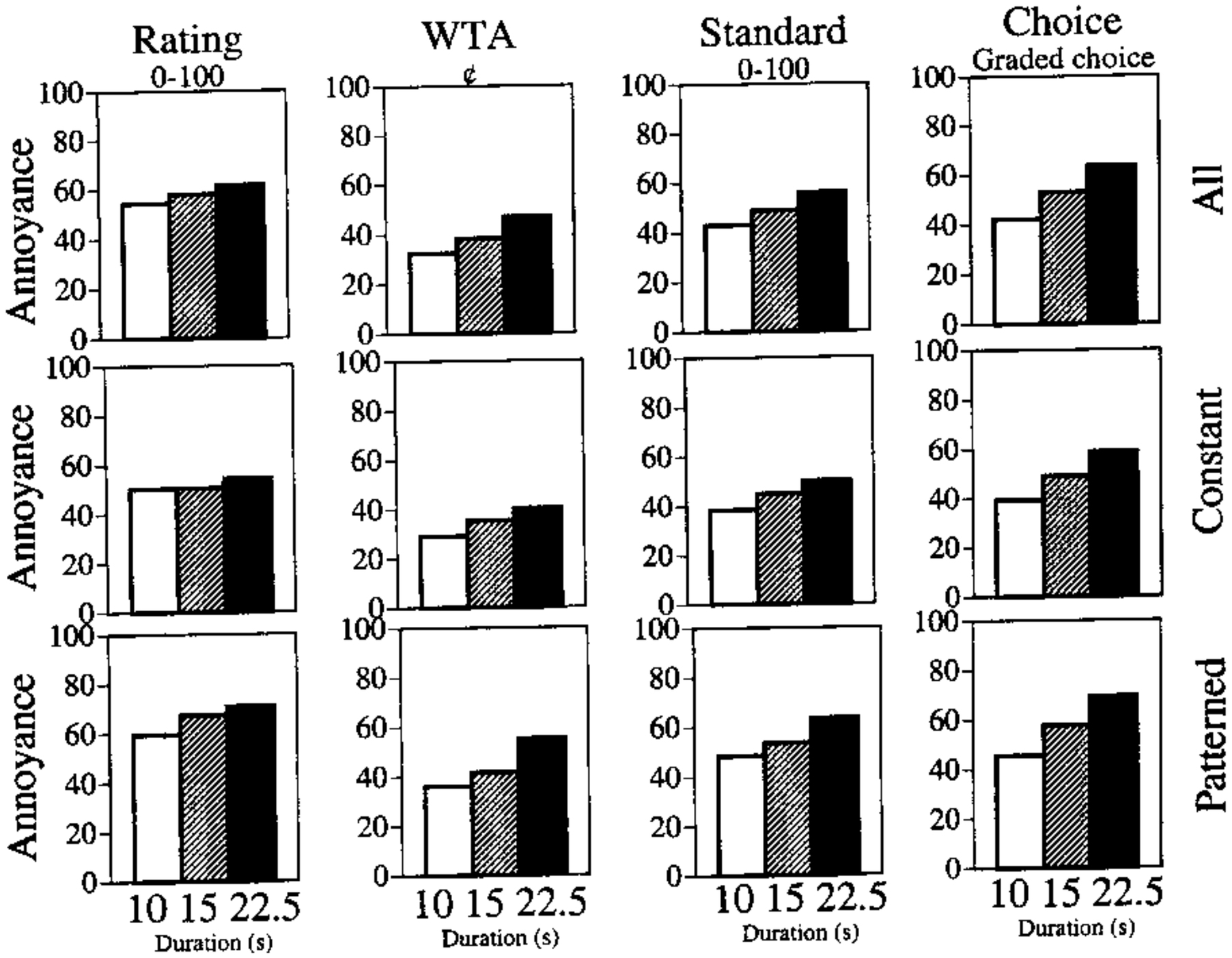

Figure 2. Mean overall annoyance in the four experiments, plotted separately for each experiment and each duration. For the choice experiment, the measure plotted is the continuous certainty measure. For all other experiments, the measure plotted is the original measure. In the top panel, both the constant and patterned stimuli are plotted. In the middle panel only the constant stimuli are plotted, and in the bottom panel only the patterned stimuli are plotted. Rating = separate ratings; WTA = willingness to accept pain in exchange for payment; Standard = rating relative to standard.

Table 3
Parameters From the Regression Analysis

Parameters	Expt. 1 (separate ratings)	Expt. 2 (WTA)	Expt. 3 (rating relative to standard)	Expt. 4 (choice)
Continuous response scale				
Standardized coefficients	.122	.334	.266	.304
R^2	.027	.131	.085	.108
Choice and pseudochoice				
Coefficient/*SE*	.435	1.171	1.093	1.245
R^2	.022	.069	.064	.069

Note. All models were analyzed by running separate regressions for each participant and comparing the magnitude of the model parameters across participants. The top two rows present the parameters from the regression models done on the continuous response scale. The bottom two rows present the parameters from the logistic regression models done on choice (and pseudochoice, which is defined as a binary outcome of the continuous choice response). Expt. = Experiment; WTA = willingness to accept pain in exchange for payment.

standardized regression coefficients (first row) and R^2 (percentage of variance explained by duration). Both comparisons reinforce the visual impression in Figure 2 that the separate ratings experiment is an outlier in terms of the small impact of duration on evaluations. In all of the other conditions, participants displayed substantial concern for duration. However, the regression analyses do not support the visual impression that concern for duration is greater in the choice experiment (Experiment 4) than in Experiments 2 and 3. These conclusions were substantiated in a 2 × 2 analysis of variance (ANOVA) in which participants' individual standardized regression coefficients were analyzed as a function of whether the evaluations were susceptible to issues of evaluation goals (Experiments 1 and 3 were susceptible; Experiments 2 and 4 were not) and whether the evaluations were susceptible to issues of standards of comparisons (Experiments 1 and 2 were susceptible; Experiments 3 and 4 were not). This analysis of the standardized regression coefficients yielded a significant main effect for evaluation goals, $F(1, 221) = 38.18$, $p < 0.001$, and a significant main effect for standards of comparisons, $F(1, 221) = 6.43$, $p = 0.011$. As implied by Figure 2 and Table 3, there was also a significant two-way interaction, in which the only experiment that yielded reduced weight to duration was the separate ratings experiment, $F(1, 221) = 28.19$, $p < 0.001$.

Note that in the between-experiment comparisons just presented, the choice experiment was at a disadvantage because in that experiment, the main response was binary but converted into a 0–100 scale by using participants' expressions of confidence in their choice. Another way to make comparisons across experiments is to compare the choices in Experiment 4 to pseudochoices in the other experiments created by coding whether each sequence was evaluated higher or lower than the standard sequence. (In cases of ties, we alternated between coding a sequence as preferred and inferior). Figure 3 shows the results from such comparisons. In each of the graphs, the .5 level separates sequences that were, on average, evaluated as superior or inferior to the standard sequence. Again, the separate ratings experiment appears to be an outlier in terms of the low impact of duration on evaluations, although even in this experiment, duration has an effect; visual inspection suggests that concern for duration was greatest in the choice experiment. The data in the bottom half of Table 3, which presents logistic regression coefficients and pseudo R^2s from regressions of duration on choice separately for each of the experiments, reinforces these conclusions.

To assess the specific effects of ratings versus choice and of separate versus comparative evaluation, a 2 × 2 ANOVA was conducted in which participants' individual standardized coefficients (coefficients/*SE*) were analyzed as a function of whether the evaluations were susceptible to issues of evaluation goals (Experiments 1 and 3 were susceptible; Experiments 2 and 4 were not) and whether the evaluations were susceptible to issues of standards of comparisons (Experiments 1 and 2 were susceptible; Experiments 3 and 4 were not). The analysis of the standardized logistic regression coefficients yielded a significant main effect for evaluation goals, $F(1, 221) = 9.13$, $p < 0.003$, and a significant main effect for standards of comparisons, $F(1, 221) = 5.28$, $p = 0.022$. As implied by Figure 3 and Table 3, there was also a significant two-way interaction, in which the only experiment that yielded reduced weight to duration was the separate ratings experiment, $F(1, 221) = 14.45$, $p < 0.001$.

Duration, it can be seen, had a greater impact on evaluations in Experiments 2, 3, and 4 than in Experiment 1. However, this does not necessarily mean that the absolute impact of duration on judgments was large in these experiments. What does it mean for the impact of duration to be large or small? If intensity were on a ratio scale, then it might be possible to compare, for example, the relative impact of doubling intensity versus doubling duration, but intensity is not a ratio scale. Thus, the best we can do is to compare the impact of duration with the impact of other sequence features that were manipulated in the experiments. Note that the results of this analysis depend critically on the relative range of manipulated sequence features. Thus, for example, intensity would look more important if the range of intensity were greater in the experiment.

To examine the impact of duration relative to other sequence features, we ran separate regressions for each participant, regressing the evaluations against the peak, ending value, and duration of each of the 27 sequences. The results, which are presented in Table 4, once again show that duration has the least impact on evaluations in the separate ratings experiment. In the WTA, rating-relative-to-standard, and choice experiments, the impact of duration is of roughly similar magnitude to the impact of peak and end. In the separate ratings experiment, however, the impact of duration is much smaller than that of peak and end. Note that peak and end are only two of many different possible ways of summarizing the nonduration features of the sequences. We ran similar regressions using ending slope and mean value as explanatory variables rather than peak and end and obtained very similar results (see Table 4).

Visual examination of Figures 2 and 3 also show that there is greater concern for duration with patterned stimuli than with constant stimuli (see Ariely, 1998). To test this claim, we performed a 2 × 3 × 4 ANOVA (Patterned/Constant × Duration × Experiment). The three-way interaction was significant, $F(8, 440) = 2.842$, $p = 0.004$, suggesting that indeed across the four experiments the effect of duration was higher for the patterned stimuli than for the constant stimuli. However, when inspecting this effect separately for the four experiments, a somewhat differ-

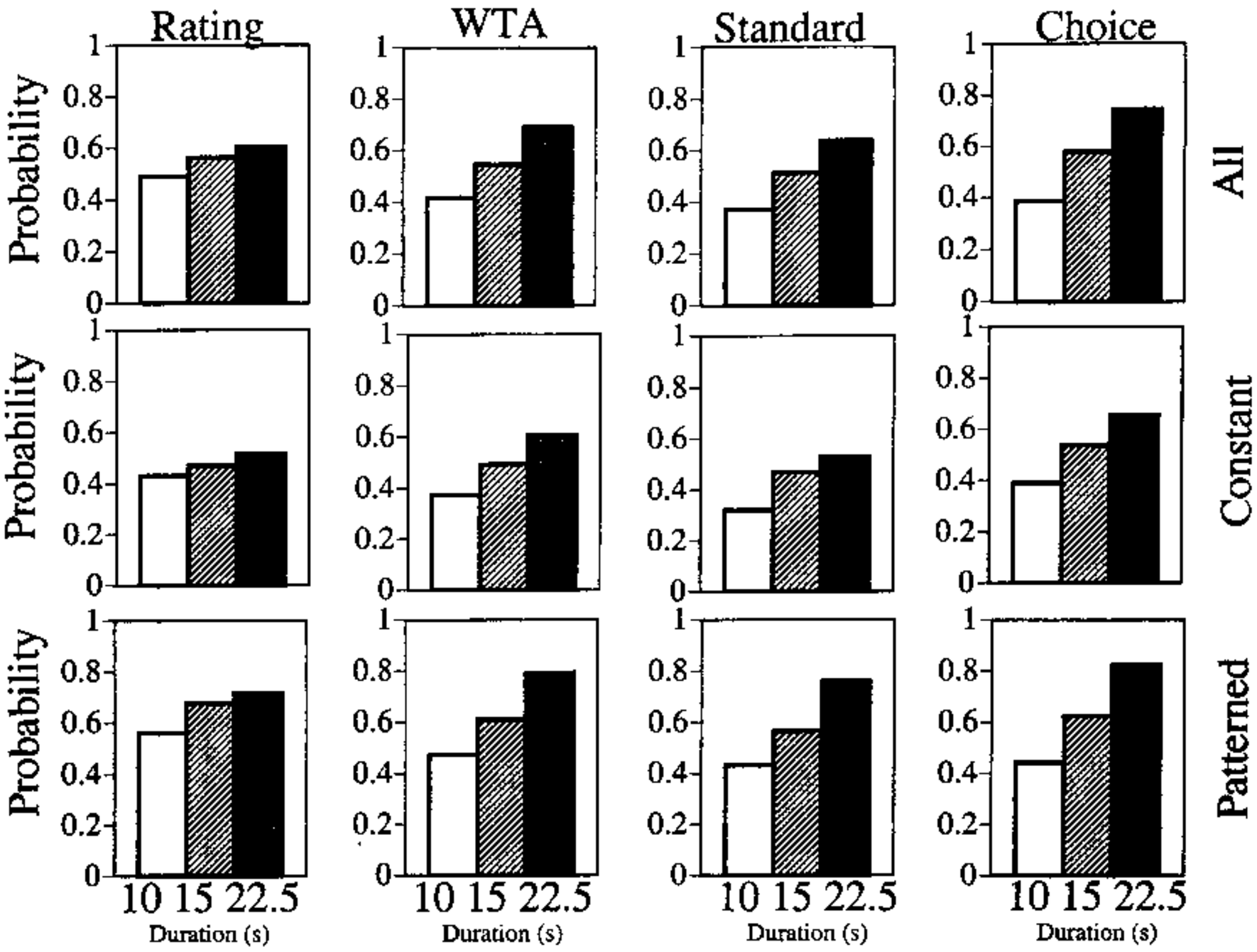

Figure 3. Probability of selection of the standard sequence in the four experiments, plotted separately for each experiment and each duration. For the choice experiment, the measure plotted is the choice respondents made during the experiment. For the other three experiments (rating relative to standard, WTA, and separate ratings), the measure plotted is a pseudochoice that converts the rating compared with the rating of the standard. In the top panel, both the constant and patterned stimuli are plotted. In the middle panel only the constant stimuli are plotted, and in the bottom panel only the patterned stimuli are plotted. Rating = separate ratings; WTA = willingness to accept pain in exchange for payment; Standard = rating relative to standard.

Table 4
Mean Standardized Regression Coefficients (and Standard Errors) and R^2s From Individual Participant Regressions

	Experiment			
Sequence features	Separate ratings	WTA	Ratings relative to standard	Choice
Peak	.437 (.032)	.416 (.029)	.494 (.031)	.357 (.032)
End	.394 (.030)	.206 (.035)	.299 (.035)	.345 (.036)
Duration	.122 (.016)	.334 (.021)	.266 (.018)	.304 (.019)
R^2	.63	.52	.65	.56
Final slope	.237 (.018)	.148 (.027)	.212 (.023)	.194 (.027)
Mean objective intensity	.703 (.017)	.538 (.030)	.687 (.015)	.640 (.016)
Duration	.122 (.016)	.334 (.021)	.266 (.018)	.304 (.019)
R^2	.61	.51	.64	.60

Note. The top half of the table shows the results of regressing evaluations on peak, end, and duration. The bottom half shows the results of regressing evaluations on final slope, mean, and duration. WTA = willingness to accept pain in exchange for payment.

ent picture emerges. The two-way interaction for the separate rating, $F(2, 88) = 7.17$, $p = 0.001$, rating-relative-to-standard, $F(2, 88) = 4.85$, $p = 0.01$, and WTA, $F(2, 88) = 6.18$, $p = 0.003$, experiments were significant but the two-way interaction for the choice experiment was not, $F(2, 88) = 1.44$, $p = 0.24$. That is, the effect of duration was similar for both patterned and constant stimuli in the choice experiment.

Pattern Preferences

Figure 4 presents mean evaluations of the four patterned stimuli: up, down, up–down and down–up, separately for each of the four experiments. Consistent with past findings (Ariely, 1998; Loewenstein & Prelec, 1993), sequences that ended with an improving slope (down and up–down) were rated more favorably than those that ended with a worsening trend of noise (up and down–up). Supporting our third prediction, elimination of conversational norms (Experiment 2), scale-norming (Experiment 3), or both (Experiment 4) did not change the taste for improvement significantly. A more detailed inspection of the preference for improvement reveals that in the WTA experiment, this tendency was a bit lower than in the other three experiments, but it was highly significant in all cases.

Finally, it is worth noting that the choice experiment was biased against showing robust attention to duration. To better understand this point, imagine a participant who cares a lot about duration and who is evaluating two annoying sounds, one lasting 10 s and the other lasting 15 s. When evaluating these sounds on a continuous scale, the participant evaluates the first sound as 20 (on a scale from 0–100) and the second sound as 40—thus showing his or her concern for duration. Now if the same participant was to evaluate the same stimuli using a choice method, the results would be somewhat different. In this case, the participant is asked to choose whether he or she wants to experience each of the two sounds or the standard stimuli. If the participant evaluated the standard to be more aversive than both stimuli, he or she would choose never to listen to the standard, and if the participant evaluated the standard to be less aversive than both stimuli, he or she would always choose to listen to the standard. In other words, unless the aversiveness of the standard is somewhere between the aversiveness of the focal stimuli, the results will show insensitivity to duration. The data in Figure 5 support this idea by showing that in the experiments in which there was a continuous response scale (such as the separate ratings experiment), participants showed sensitivity to duration at each level. However, in the choice experiment, sensitivity to duration was low for stimuli that are plotted at the top

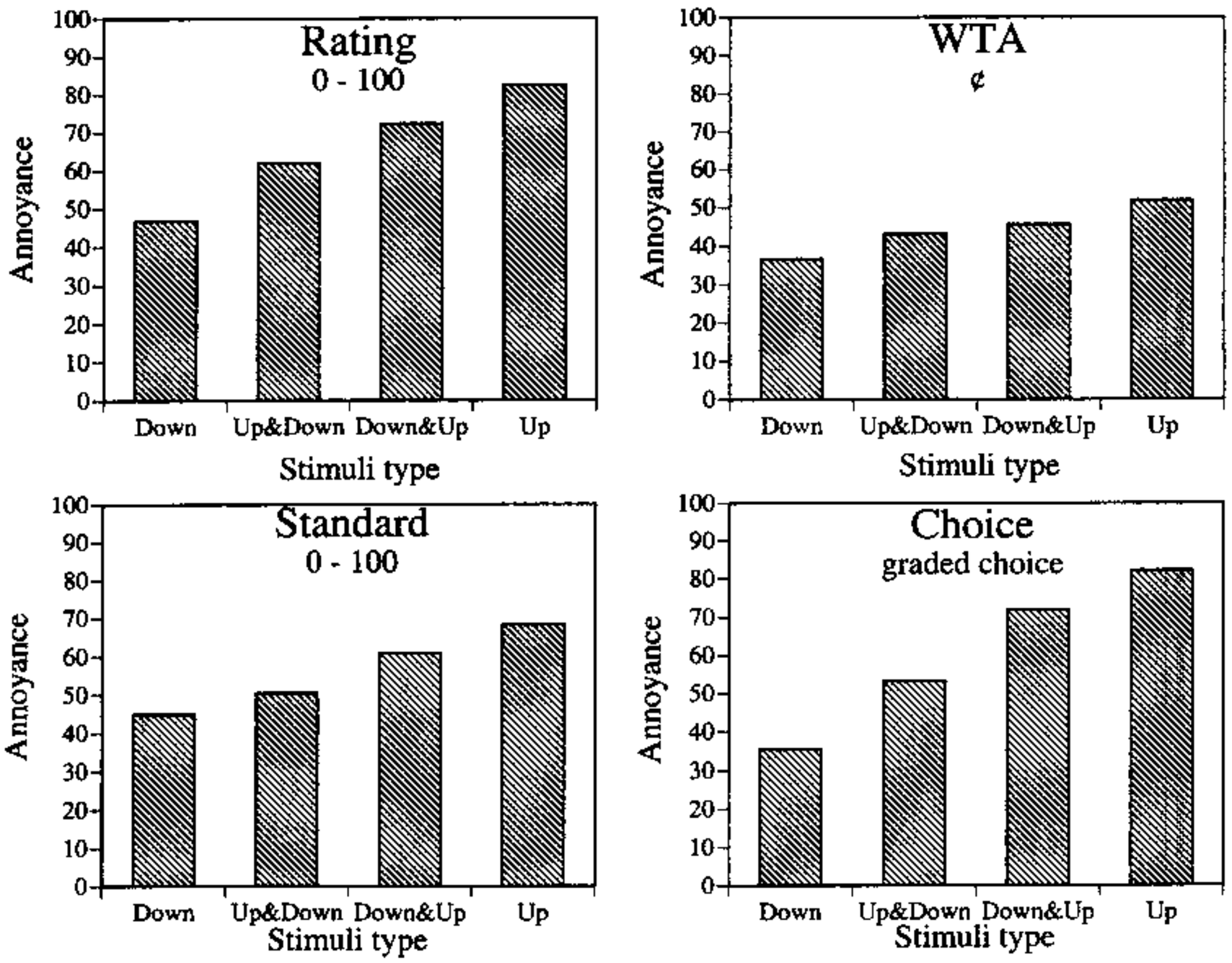

Figure 4. Mean overall annoyance in the four experiments, plotted separately for each experiment and each pattern. For the choice experiment, the measure plotted is the continuous certainty measure. For all other experiments, the measure plotted is the original measure. Rating = separate ratings; WTA = willingness to accept pain in exchange for payment; Standard = rating relative to standard.

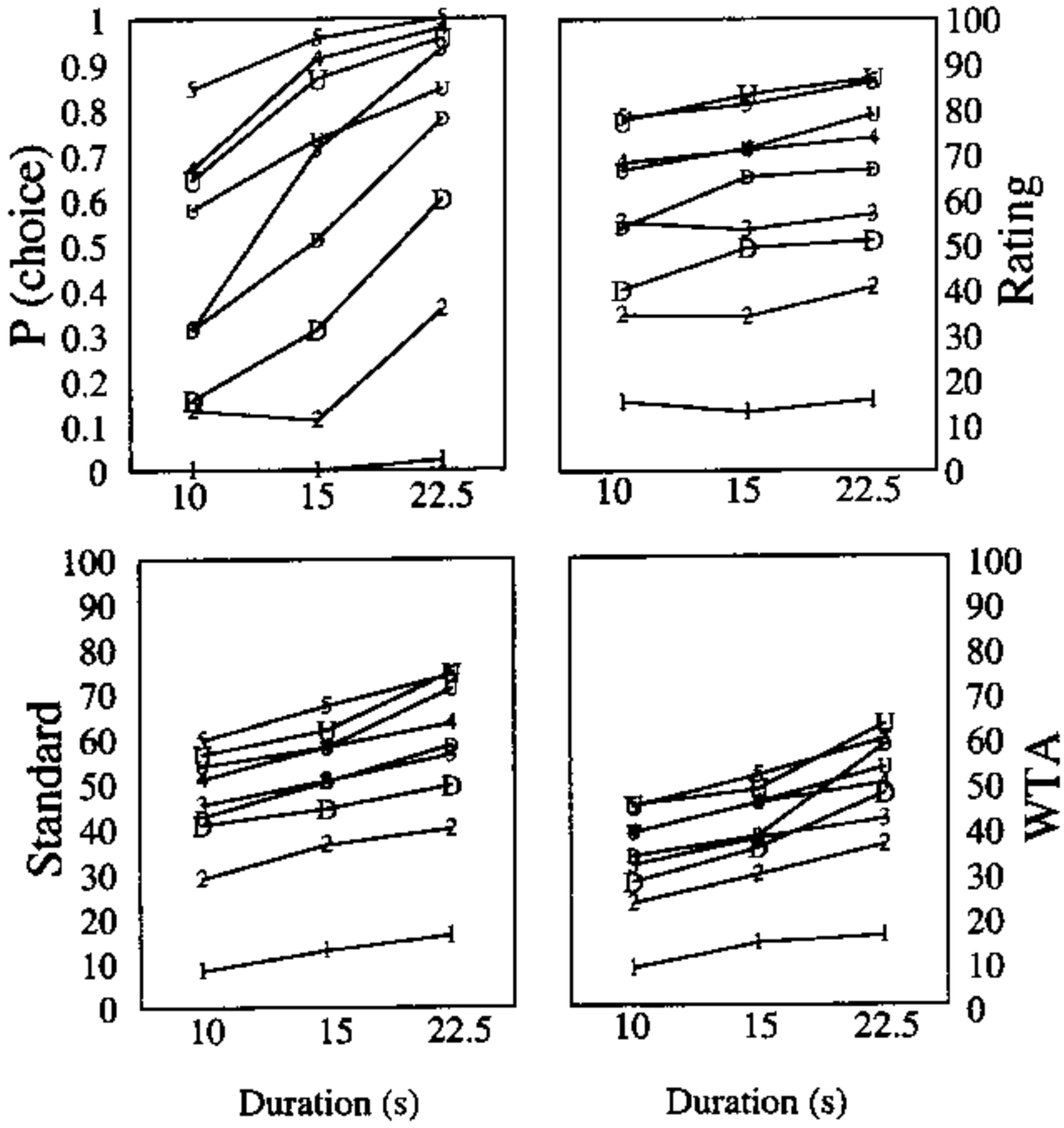

Figure 5. Overall annoyance, plotted separately for each experiment, each pattern, and each duration. The constant patterns (1–5) are noted by their number. The up and down patterns are noted by a large *U* and *D*, respectively, and the down and up and up and down patterns are noted by a small *U* and *D*, respectively. For the choice experiment, the measure plotted is the choice respondents made during the experiment. For the other three experiments (rating relative to standard, WTA, and separate ratings), the measure plotted is the original responses. P = probability; Rating = separate ratings; Standard = rating relative to standard; WTA = willingness to accept pain in exchange for payment.

(highly annoying stimuli) or at the bottom (not very annoying stimuli). The highest sensitivity in the choice experiment is shown for stimuli that are in the middle range—those that can be traded off with the standard stimulus. Another interesting point relates to the curvature shown in the choice experiment. Very annoying sounds show sensitivity at the low durations and a decreased sensitivity at the high durations, whereas not very annoying sounds show low sensitivity at the low durations and an increased sensitivity at the high durations. The combination of the overall change in sensitivity across the range, and the curvature in the responses, strengthens our belief that participants indeed traded off duration and intensity.

Discussion

The central goal of the current article was to explore some mechanisms that we expected to affect the weight placed on duration in evaluations of sequences of outcomes. To explore these mechanisms, we examined the impact of different evaluation methods on the role of duration in evaluations. We used a range of elicitation procedures: ratings as traditionally used, ratings of sequences relative to a well specified standard sequence, willingness to pay, and a graded-choice procedure in which participants repeatedly chose between the standard sequence and each of the 27 focal sequences. Comparing these four evaluation methods, we see that the ratings procedure commonly used in previous research elicited the least sensitivity to duration. This pattern was observed when participants' evaluations of sequences were treated as continuous variables and also when they were converted into a binary variable that designated preference relative to the standard sequence. The pattern of results was evident when the four experiments were compared on the basis of mean evaluations of the different stimuli and also on the basis of a variety of different measures designed to capture concern for duration: standardized regression coefficients and R^2s from regressions of continuous evaluations on duration and standardized regression coefficients and pseudo-R^2s from a logistic regression of the binary preference variable on duration. Moving from unanchored ratings to decisions (either WTA or pairwise choice) or from unanchored ratings to