Decision

© 2017 American Psychological Association
2325-9965/17/$12.00 http://dx.doi.org/10.1037/dec0000087

AQ: 1

Why We Should Quit While We're Ahead: When Do Averages Matter More Than Sums?

AQ: au
AQ: 2

Michael Brusovansky, Yonatan Vanunu, and Marius Usher
Tel-Aviv University

An enduring debate in decision-making and social cognition concerns the algorithm governing the formation of intuitive preferences and attitudes. Here we contrast 2 principles that are considered central to such judgments: averaging versus summation. Participants in 4 experiments were prompted to rely on their intuition when rating the Hall of Fame eligibility of basketball players, or their liking of athletes, lecturers or slot-machines, on the basis of rapid numerical sequences that represent performances, class feedback, or rewards. Experiment 1 showed that participants are sensitive to the sequences' averages, and prefer alternatives with high averages over those with high sums. Experiment 2 replicated these findings, and further showed that in a comparison between several models such as averaging, summation and the Peak-End heuristic, averaging type models account best for participants' preferences. Experiment 3 indicated that these evaluations are mediated by automatic/intuitive processes. Based on computational considerations we propose that the critical variable, determining whether preferences are more sensitive to sums or to averages, is the presentation and evaluation format: one by one versus grouped. This prediction is confirmed in Experiment 4.

AQ: 3

Keywords: intuitive preferences, averaging, summation, Peak-End heuristic, evaluation format

Supplemental materials: http://dx.doi.org/10.1037/dec0000087.supp

Intuitive preferences and attitudes can be viewed as "evaluative summaries" that determine our behavioral tendencies toward persons and objects and thus are an indispensable construct for understanding human judgment and decision-making (Allport, 1935; Fazio, 1989). Although much research has been dedicated to controlled information integration, viewed as a capacity constrained sequential adjustment of an estimate toward a criterion (Hogarth & Einhorn, 1992; see review in Juslin, 2015), a number of studies demonstrated that attitudes can also be formed in an *implicit* and automatic way (Schneider & Shiffrin, 1977). For example, in two seminal studies, Betsch and colleagues have demonstrated that human observers can form accurate attitudes toward alternatives (stocks characterized as sequences of returns), which are presented as distractors and without an explicit task of evaluation (Betsch, Kaufmann, Lindow, Plessner, & Hoffmann, 2006; Betsch, Plessner, Schwieren, & Gütig, 2001). Similar automatic mechanisms are indicated by studies of retrospective evaluations of affective episodes (Diener, Wirtz, & Oishi, 2001; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Redelmeier & Kahneman, 1996).

The principle that underlies such automatic evaluations is subject to debate. On the one hand, the information-integration theory has suggested that attitudes reflect an averaging principle (Anderson, 1981). A similar suggestion is the Peak-End rule—a special version of

AQ: 8

Michael Brusovansky, Yonatan Vanunu, and Marius Usher, School of Psychological Sciences and Sagol School of Neuroscience, Tel-Aviv University.

Yonatan Vanunu is now at School of Psychology, University of New South Wales.

This study was supported by ISF grant (743/12) and by (NSF-BSF-NIH Computational Neuroscience (CRCNS; 2014612) to Marius Usher. We thank Nira Liberman, Zohar Bronfman, Moshe Glickman, and Konstantinos Tsetsos for very helpful discussions and comments on the manuscript.

Correspondence concerning this article should be addressed to Michael Brusovansky, School of Psychological Sciences, Tel-Aviv University, Tel-Aviv, Israel. E-mail: michaelbrus25@gmail.com

an averaging model—which accounts for retrospective evaluations of affective episodes (Diener et al., 2001; Kahneman et al., 1993; Redelmeier & Kahneman, 1996). On the other hand, Betsch and colleagues have recently proposed that attitudes toward a stream of value-charged pieces of information about alternatives (stock returns) reflect a summative principle (Betsch et al., 2006, 2001).

It is important to note that both of these principles—averaging and summation—can be subsumed under the more general principle of *additive integration*, which is usually contrasted with the nonlinear integration that is required in normative accounts of probability judgment tasks (Juslin, 2015).[1] To illustrate the tension between the averaging and the summation principles, consider the case of Michael Jordan's comeback. Jordan is widely considered to be the greatest basketball player of all-time (National Basketball Association official website, 2015). He retired on the highest note possible, having just won his final championship and scoring the winning basket in his final game. His list of accomplishments speaks for itself: during the 13 years he wore the Chicago Bulls uniform, Jordan won the regular season's Most Valuable Player award 5 times, the NBA Finals Most Valuable Player award 6 times, 6 NBA championships, gained 12 All-Star selections and 10 All-NBA First Team selections. However, after spending 3 years away from the court, Jordan decided to make a comeback in the uniform of the Washington Wizards. Despite playing well for a 40-year-old, the two years he spent with the Wizards were not nearly as successful, prompting sports writer Bill Simmons to name it as one of the most depressing comebacks in NBA history (Simmons, 2010). However, none of Jordan's previous achievements were taken away during that comeback. If anything, Jordan still managed to add two additional All-Star selections and 3,000 points to his resume. The sum total of his achievements only got bigger. Obviously, this is a case in which the two principles diverge: whereas summation should predict an enhanced appreciation after the comeback, averaging predicts depreciation (Jordan's career averages dipped).[2]

The aim of this article is to examine the nature of the mechanism that mediates intuitive and automatic preferences (or attitudes) by contrasting between these two operating principles and examine the role of individual differences. Toward this aim we carried out four experiments, in which participants were exposed to rapid sequences of value-charged stimuli (numerical values of performance) about a number of alternatives and were asked to convey their liking of each, on an analog scale, by relying on their intuition. Previous studies have shown that instructing participants to adopt a certain mindset (Horstmann, Ahlgrimm, & Glöckner, 2009; Pham, Lee, & Stephen, 2012; Rusou, Zakay, & Usher, 2013; Usher, Russo, Weyers, Brauner, & Zakay, 2011) and using stringent time-constraints (Horstmann, Hausmann, & Ryf, 2010) are effective manipulations for inducing automatic/intuitive decision modes (but see Exp. 3, for additional validation).

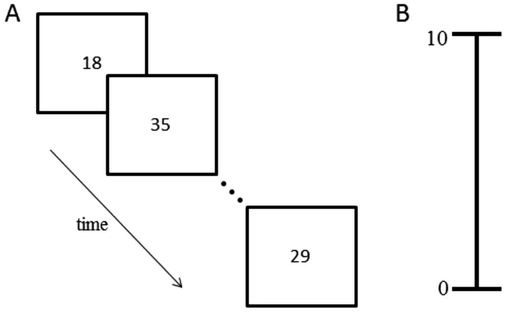
To anticipate our results, in our first three experiments we found that, in opposition to what we find when similar evaluations are made in an self-controlled format (see Discussion of Exp. 1–3), a clear domination of the averaging principle emerges, and the third experiment provide further support that the evaluations are indeed made under an automatic/intuitive mode. Based on these results and on neurocomputational considerations, we propose that a critical difference that determines whether intuitive preferences are dominated by averaging versus summation is the presentation and evaluation format (one-by-one vs. grouped). This prediction is confirmed in our fourth experiment.

Experiments 1 and 2

The experimental task was designed as a series of evaluations (30 decisions in Experiment 1 and 96 decisions in Experiment 2) about the eligibility of basketball players to the Hall Of Fame. Prior to each decision, the participants viewed a sequence of 6–12 numbers (ranging from 5 to 50), which represented a player's career, with each number corresponding to the average number of points he scored during

[1] Thus data showing additive integration (e,g., prior neglect) do not distinguish between them, as to do so one needs to compare attitudes for alternatives that differ in the number of samples.

[2] We use Jordan's example for illustrative purposes and do not claim that his evaluation change was actually based on intuitive (nonanalytical) judgments. In the experiments, however, we aim to provide conditions that engage intuitive preferences.



AQ: 9 *Figure 1.* (A) Time flow of sequence presentation in Exp. 1 and 2. (B) Evaluation scale.

F1 one season (Figure 1A). At the end of each sequence presentation the participants had to indicate how much they feel the player deserved to be inducted into the Basketball Hall Of Fame, by rating him on an analog attractiveness scale (Figure 1B).

In both experiments, we intended to engage intuitive judgments by instructing the participants to rely on their intuition and general impression of the numbers (Horstmann et al., 2009; Pham et al., 2012; Usher et al., 2011) and also by employing a fast presentation rate (Horstmann et al., 2010). The players which the participants had to rate (one per trial) were divided into two types: (a) regular players, testing participants' sensitivity to averages and to sequence length, and (b) critical players, contrasting between averages and sums by setting them in opposition. Whereas in Experiment 1 (which was aimed to test the basic paradigm) the additional seasons always came at the end of the sequence, in Experiment 2 (which was aimed to replicate the results with a larger sample and examine individual differences), we extended this design by inserting the additional seasons at three additional temporal locations. Experiment 2 also tested the sensitivity of intuitive evaluations to one additional factor: the temporal bias, which is an important component in the Peak-End heuristic (Kahneman et al., 1993).

Materials

A total of 30 players were presented in Exp. 1, 20 regular players and 10 critical players, each presented as a numerical sequence for evaluation in one experimental trial. The regular trials were constructed based on two factors, which were manipulated orthogonally: (a) sequence's average (10, 25 or 40) and (b) sequence's length (8, 10 or 12). Thus, the regular players averaged 10, 25 or 40 points per game throughout their careers, which lasted for 8, 10, or 12 seasons. The 10 critical players were constructed as pairs, in which one player had a short but successful career, in which he averaged 40 points per game in 9 seasons ("High-Average, Low-Sum" player), while the other player had the exact same first 9 seasons as the former, but with three additional seasons with diminished (though still above-average – 28 points per game) performance, for an overall average of 37 points per game over 12 seasons ("High-Sum, Low-Average" player; see Suppl. Information e.g., of critical trials). Thus, although the second player has lower career average, he dominates the first one in terms of total career achievements (higher sum) and therefore should be favored based on a summation principle.

In Exp. 2, the participants saw and evaluated a total of 96 players: 72 regular players and 24 critical players. The 72 regular players were constructed based on a factorial design with 3 factors: (a) sequence average (10, 20, 30, 40), (b) sequence length (6, 9, 12) and (c) temporal bias (primacy, balanced, recency); the numbers in the sequence were arranged so that either the larger numbers appeared in the first half (primacy profile), the second half (recency profile), or roughly equally (balanced profile), see supplemental information for additional details. The 24 critical players were constructed in six quadruplets. The first two players—the "High-Average" and the "High-Sum End" were similar to the critical players in Experiment 1 (this time the "High-Average" player averaged 40 points per game over 8 seasons and the "High-Sum End" player averaged 36 points per game over 11 seasons). The two new players in a quadruplet had the same career as the "High-Sum End" player, but the additional three seasons appeared either at the beginning of the player's career ("High-Sum Beginning") or at the middle of the career ("High-Sum Middle"); see suppl. information.

Method

Participants. Twelve students from Tel-Aviv University (7 females, age: 19–31, M = 24.3) participated in Experiment 1, and 29 students from Tel-Aviv University (20 females, age: 19–36, *M* = 23.6) participated in Experiment 2. All the participants received either course credit or payment, and all had normal or corrected to normal vision.

Procedure (Experiments 1 and 2). Participants were told that they are about to see sequences of the numbers that would be presented at a fast pace, and thus doing various calculations in the experiment is nearly impossible. They were therefore instructed to make the evaluations of the players intuitively, by relying on their general impression of the numbers (Rusou, Zakay, & Usher, 2016). In both experiments the participants were told that the maximum amount of points the players score is 50, and an "example player," which was not evaluated, was shown. Each trial presented one player, and the trials' order was randomized once, and then kept the same for all participants. When a player had to be evaluated, his "career" was presented sequentially. The numbers appeared in green color at the center of a black screen for 500 ms, followed by a blank screen for 100 ms before the next number of the sequence. The evaluation was done on evaluation sheets for Exp. 1 (see Figure 1) and on the computer for Exp. 2. Exp. 1 took approximately 15 min and Exp. 2 took approximately 30 min.

Analysis. First, the ratings of the regular trials in Experiments 1 and 2 were used to test if participants are sensitive to averages and to sequence length. We carried out a repeated measures ANOVA analysis on the regular players, with "average" and sequence length ("number of seasons") as within-participant variables (all factors and interactions were tested for the sphericity assumption, and when violated, a correction for the degrees of freedom was applied, using the Greenhouse-Geisser correction). Experiment 2 had an additional within-participant variable, the "temporal profile" (primacy/recency/balanced). Second, participants' ratings in the critical trials were contrasted to test the average versus the summation principles. In Exp. 1 the critical trials were contrasted via a paired *t* test, in Exp. 2 this was done via repeated-measures ANOVA. Finally, we examined individual differences in Exp. 2 by carrying out quantitative analysis that contrasts a number of competing models of intuitive evaluations, which are based on various combinations of the sequence properties, such as (a) average, (b) sum (or length of sequence), (c) last item, (d) peak-item.

Results and Discussion

Group-level results.

Regular trials. The average ratings of the regular candidates, as a function of sequence-average, and sequence length, are shown in Figure 2 for Exp. 1 and in Figure S1 (for Exp. 2). F2

In both experiments, the results indicate a strong effect of the average (Exp. 1: $F(2, 22) = 97.14$, $p < .001$; Exp. 2: $F(1.92, 53.70) = 539.13$, $p < .001$). The effect of sequence-length was less consistent. While no effect was found in Exp. 1, $F(2, 22) = 0.75$, $p = .49$, sequence-length was significant in Exp. 2, F(2,

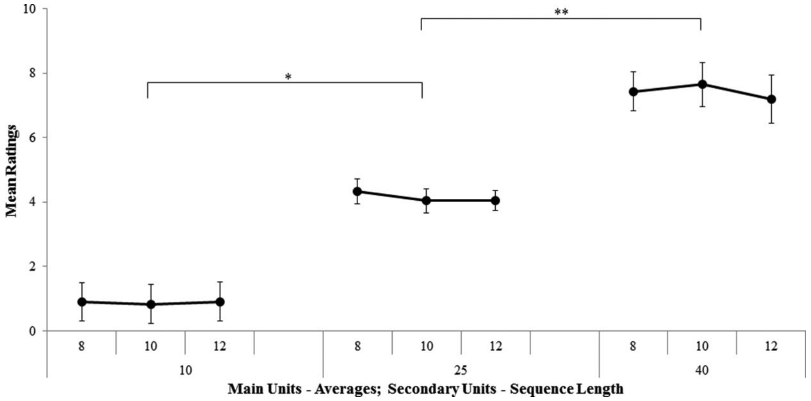


Figure 2. Mean ratings in Exp. 1 for the regular players, by average (10/25/40) and number of seasons (8/10/12). Error bars correspond to within-participant 95% confidence intervals. * $p < .05$, ** $p < .01$.

56) = 15.85, $p < .001$; this however was smaller in magnitude, compared with the effect of the average, as discussed below. Post hoc Tukey's tests revealed that that participants gave higher ratings to each successive increase in the players' points average (Exp. 1: $M_{10} = 0.88$, $SD_{10} = 1.02$; $M_{25} = 4.13$, $SD_{25} = 0.37$; $M_{40} = 7.43$, $SD_{40} = 1.11$; Exp. 2: $M_{10} = 1.31$, $SD_{10} = 0.61$; $M_{20} = 2.97$, $SD_{20} = 0.68$; $M_{30} = 5.11$, $SD_{30} = 0.66$; $M_{40} = 7.48$, $SD_{40} = 0.67$). Post hoc Tukey analyses on sequence length in Exp. 2 revealed that participants gave higher mean ratings to players who played 9 or 12 seasons than to players who played 6 seasons (no difference in the ratings between 9 seasons and 12 seasons; $M_6 = 3.96$, $SD_6 = 0.60$; $M_9 = 4.30$, $SD_9 = 0.32$; $M_{12} = 4.40$, $SD_{12} = 0.43$). No interaction between average and sequence length was found in Exp. 1, $F(4, 44) = 1.29$, $p = .29$, but was found in Exp. 2 (see Fig. S1): $F(4.42, 123.78) = 10.43$, $p < .001$. Post hoc Tukey analyses revealed that participants assigned higher ratings to players with 9 or 12 seasons than to players with 6 seasons—when the player's average was 30 or 40, but the number of seasons had no effect when the player's average was 10 or 20. Analysis of the players' temporal profiles in Exp. 2 revealed no main effect for the temporal profile, $F(1.63, 45.77) = 2.27$, $p = .123$, as well as no interaction between the temporal profile and the average, $F(4.08, 114.35) = 2.25$, $p = .066$, the length, $\mathrm{F}(4, 112) = 0.87$, $p = .483$, and no three-way interaction, $F(6.08, 170.15) = 1.26$, $p = .277$, see suppl. information.

Critical trials. As illustrated in Figure 3A, in Exp. 1 the "High-Average, Low-Sum" players received higher ratings on average ($M = 7.34$, $SD = 1.89$) than the "Low-Average, High-Sum" players ($M = 6.87$, $SD = 2.02$; $t(11) = 3.54$, $p = .005$), suggesting that participants' evaluations were more consistent with an averaging process rather than a summation one, supporting the "Jordan-effect." In Exp. 2, a main effect was found, $F(3, 84) = 3.57$, $p = .017$. Post hoc Tukey's tests revealed that participants gave the highest mean ratings to the "High-Average" players ($M = 7.31$, $SD = 0.84$), which were higher than the ratings given to the "High-Sum End" players ($M = 6.81$, $SD = 0.96$). The "High-Sum Beginning" players ($M = 7.04$, $SD = 0.77$) and the "High-Sum Middle" players ($M = 6.99$, $SD = 0.75$) received similar ratings, which were also not different from the "High-Average" players or the "High-Sum End" players (see Figure 3B).

The group-level results (Figure 2, Figure 3 and Fig. S1) indicate a similar pattern. Participants differentiated between the regular players based on their averages (an *increase-ratio*[3] of 1.64 (6.55/4) in Exp. 1 (6.55/4) and of 1.54 (6.17/4) in Exp. 2). The impact of sequence-length was less consistent (no effect in Exp. 1, and a very small effect, an *increase-ratio* of 0.22 (0.44/2) in Exp. 2); a summation model predicts *increase-ratios* of the same magnitude. The participants also preferred the "High-Average" player over the "High-Sum" player when the additional seasons were added in the end (see Figure 3), providing further support for the "Jordan-effect." However, the additional critical players added in Experiment 2 failed to provide a clear-cut support for either the Averaging principle (which predicted higher rating for the High-Average compared with High-Sum Beginning/Middle) or for the Peak-End heuristic (which predicted higher ratings for the latter two compared with the High-Sum End). In addition, the temporal bias factor was not significant in the regular players' analysis, suggesting that the impact of the temporal bias to the evaluation is smaller than that of the average, and subject to variability, which may involve individual differences. We thus carried out a quantitative analysis of individual participants, in which we contrasted a number of competing models that are based on various combinations of the sequence-average, number of seasons, the peak-season and the last-season.

Quantitative model comparison and individual differences. To examine which model best describes the participants' ratings, we subjected the participants' ratings in Exp. 2 to nine linear regression models, and estimated how well they account for the data. Our central aim was to contrast between summation and aver-

[3] We define the *increase-ratio* as the ratio of the increase in rating (e.g., in Exp. 2 the increase in ratings for the average was 7.48 − 1.31 = 6.17) divided by the ratio of increase between the lowest and the highest values of the independent variable (average or sequence length). For example, in both experiments, the average increased by a factor of 4 (40/10), and in Exp. 2, sequence-length increased by a factor of 2 (12/6). We divide the ratio of the corresponding ratings by these factors.

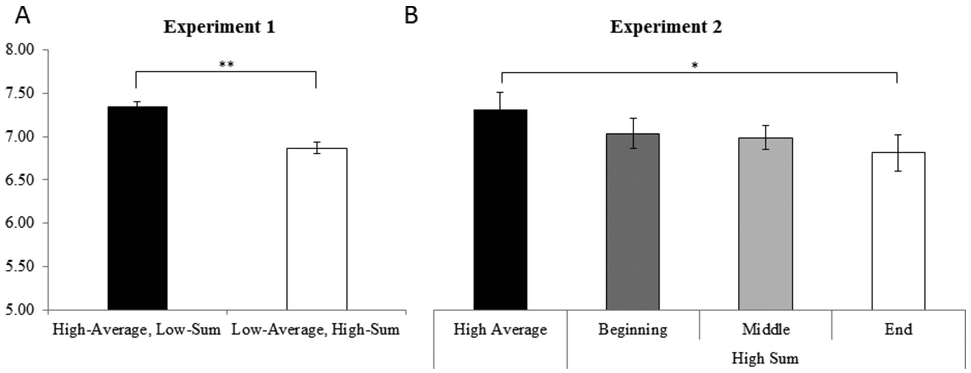


Figure 3. (A) Mean ratings given to the "High-Average, Low-Sum," and the "High-Sum, Low-Average" critical players in Exp. 1. (B) Mean ratings given to the critical players in Exp. 2. Error bars correspond to within-participant 95% confidence intervals. * $p < .05$, ** $p < .01$.

aging-based models and also to distinguish between averaging and Peak-End Heuristics. Therefore, the models were selected so that the first four are variants of the averaging principle: Whereas Model 1 is a pure averaging model, Models 2–4 include the average together with some additional factors that may be thought to contribute in addition to averaging, such as (a) the last value, (b) sequence length, *N*, (c) both of the above. The next three models are variants of the summation principle: Model 5 is a pure summation model, while Model 6 is a modified sum model, in which the summation is carried out relative to a reference—*R*: $Y = a_1 \times \Sigma (x_i - R) + b$ (when $R = 0$ the model acts as a regular summation model,[4] whereas for *R* in the midrange it predicts an interaction between average and *N*). Model 7 includes summation together with the end item, and Model 8 is the traditional Peak-End model (Kahneman et al., 1993). Note further that Models 3 and 4 are intermediate between averaging and summation, as they allow the number of elements to impact the evaluation, but depending on the *N*-coefficient this impact may be much smaller than expected from a summation model. The final Model 9 is the *anchor and adjustment* model (Hogarth & Einhorn, 1992), which is associated with controlled updating and was used to account for temporal order effects (see Table 1).

Fn4

T1

All the models were fitted to the data of each participant, by regressing the ratings that the participant produced in each trial, on the model-properties of that trial (e.g., Average and End-value, for Model 2). The regression was carried out in Matlab (by using the *LinearModel.fit* function), which produced the model coefficients that minimize the least square distance between model prediction and data and provides the likelihood of data given the model (at its best parameters).[5] This likelihood was used to compute the Bayesian Information Criterion ($BIC = -2 \cdot \ln(\hat{L}) + k \cdot \ln(n)$; Wasserman, 2000), which penalizes for extra free parameters. The aim of the analysis was to find, for each participant, which of these models best accounts for the ratings, using BIC (see Table S1 in suppl. information for individual BIC values for each model).

Fn5

AQ: 4

As illustrated in Table S1, the various averaging models provided the best BIC-fit for 21 of the 29 participants (~72%), with the pure averaging model (A) accounting best for 18 of them (~62%); see Fig. S2 in the Suppl., for the data fit of these participants in the filler trials. An averaging model which also includes the number of items in the sequence (A + N) provided the best fit for two of the participants, and an averaging model which includes the last item (A + E*) provided the best fit for one participant. While no subject was accounted best by

[4] The R value corresponds to an implicit reference, and we allowed the model the flexibility to assume that it can vary among the participants. For each participant we tested 41 different values of R (0-40), to find the value which provides the best fit—the highest explained variance.

[5] For Model 6, we also varied *R* on a grid (0-40, in steps of 1).

Table 1

AQ: 10 *The Nine Models Examined in Experiment 2*

Model name	Notation	Y: Dependent variable (evaluation)
1. Average	A	$Y = a_1{}^{*}\text{average} + b$
2. Average & End	A + E*	$Y = a_1{}^{*}\text{average} + a_2{}^{*}(\text{end-average}) + b$
3. Average & N	A + N	$Y = a_1{}^{*}\text{average} + a_2{}^{*}\text{number of items} + b$
4. Average & End & N	A + E*+N	$Y = a_1{}^{*}\text{average} + a_2{}^{*}(\text{end-average}) + a_3{}^{*}\text{number of items} + b$
5. Sum	S	$Y = a_1{}^{*}\text{sum} + b$
6. Sum*	S*	$Y = a_1{}^{*}\text{sum}^{*} + b;\ \text{sum}^{*} = \sum(x_i - R)$
7. Sum & End	S + E	$Y = a_1{}^{*}\text{sum} + a_2{}^{*}\text{end} + b$
8. Peak & End	P + E	$Y = \mathrm{a}_1{}^{*}\text{peak} + a_2{}^{*}\text{end} + b$
9. Anchor & Adjustment	An & Ad	$Y = a_1{}^{*}S^{*} + b;\quad S_i = S_{i-1} + \alpha^{*}S_{i-1}{}^{*}(X_i - R)\ for\ X_i \leq R$ $S_i = S_{i-1} + \beta^{*}(1 - S_{i-1})^{*}(X_i - R)\ for\ X_i > R$

Note. N = number of seasons; End = last item; Peak = largest item of each sequence; Sum* = summation relative to a reference. Because in our sequences the average and the End values are strongly correlated, we entered their difference in the regression, when the two appear together in the same model. S in Model 9 corresponds to the end of the iterative process at i = n (length of sequence).

the pure summation model, the summation relative to a reference model (S*) provided the best fit for 5 of the participants; see Fig. S2 in the suppl., for the data fit of these participants and for a demonstration that the models capture actual differences. Finally, the Peak-End model (P + E) provided the best fit for two of the remaining participants, while the anchor and adjustment model to only one.[6] Fn6

The results of the model comparison confirm the Group ANOVA results, by indicating the pure averaging model to be the main determinant of the intuitive evaluations. Recency also appears to have some influence on the evaluations, but this was observed for only 3 participants (two of which were accounted best by the Peak-End heuristic and one for whom the end item together with the average, was part of the best-fitting model). Finally, the modified summation relative to a reference accounted best for five participants.[7] Fn7

Experiment 3

The aim of Exp. 3 was to extend the average dominance effect with a different task-framing (preference for athletes competing in a six events track field contest), while also testing a property that is associated with automaticity and intuitive processing. To do so, the participants were informed that occasionally the "computer generates incorrect values," which are enclosed in a salient red square, and which they should ignore. This is based on the rationale that detecting and acting on negation is a distinguishing property of the rule based analytical/controlled system (e.g., Deutsch, Gawronski, & Strack, 2006; Gawronski & Bodenhausen, 2006). If the evaluations are mediated by an automatic/intuitive process, we should expect participants to be unable to filter or ignore values enclosed by red-squares, and those to-be-ignored values should receive equal weight as other values. For example, if the to-be ignored value has a lower value than the average of the others in the sequence, we expect that this will reduce the evaluation given to that athlete, despite the increase in the sequence's sum (a type of "Jordan-effect").

Method

Participants. Nineteen students from Tel-Aviv University (13 females, age: 19–26, $M = 23.4$) participated in the experiment, in exchange for course credit. All the participants had normal or corrected to normal vision and normal color vision.

Materials. The participants were asked to evaluate (on a scale of 0–10) performances of athletes competing in a six-event track & field

[6] If we consider differences in BIC values that are smaller than 1 to be a tie, the various Averaging models provide the best fit in 28 instances of 36 (78%, pure averaging model – 18/36), followed by summation relative to reference – 5/36, and the Peak-End model – 2/36.

[7] R values for these five participants were 13, 15, 16, 19, and 26, which for the majority are below the midrange of 25.

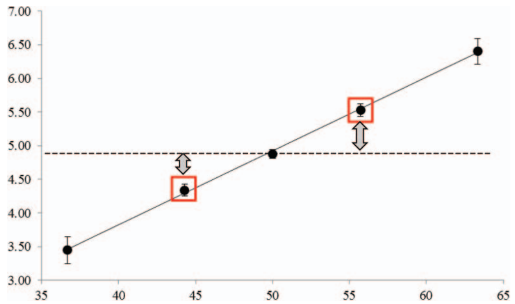


Figure 4. Mean ratings given to the different athletes' types in Experiment 4. The *x* axis represents the actual averages, including the "erroneous" values. Error bars correspond to within-participant 95% confidence intervals. The red-squares marks correspond to the "computer errors" conditions. The actual evaluations of those alternatives are on the regression line, and significantly differ from the evaluation given to the "50" alternatives (dashed line), which had the same average (excluding the "errors"). See the online article for the color version of this figure.

contest (100-m dash, long jump, etc.). The evaluations had to be based on the marks (range: 0–100) each athlete receives on the six events. The regular athletes (without computer errors) had averages of 36.7, 50, and of 63.3. The critical athletes had six values with an overall mean of 50, and the seventh (the computer error) chosen around 10 or 90, thus bringing the overall average to 44.3 or 55.7, respectively (See suppl. information and Table S2 for details).

Procedure. The procedure was similar to that of Experiments 1–2. When an athlete had to be evaluated, her marks were sequentially presented and the participant had to rate her before moving on to the next athlete. All the numbers appeared in green color at the center of a black screen for 500 ms, and were followed by a blank screen for 100 ms before the next number appeared. An additional seventh value, which appeared for 120 critical athletes, was enclosed inside a prominent red square and its location within the sequence was randomized. Participants were told in advance that they will encounter such "computer error" instances and that they should ignore them. Thus, each athlete had only six values that should influence her evaluation. At the start of the experiment participants saw examples of three athletes with average marks of 50/37/63. They were also told that the marks range from 0 to 100. After that participants completed 300 trials in six blocks of 50 trials each and with short breaks between the blocks. The whole procedure took approximately 60 min.

Analysis. A one-way repeated measures ANOVA was carried out, with average (36.7, 44.3, 50, 55.7, 63.3) as a within-participant variable.

Results

As shown in Figure 4, the average factor was highly significant, $F(1.30, 23.36) = 187.61$, $p < .001$. Post-Hoc Tukey's tests revealed that all the possible pairs of comparisons yielded significant results; that is, participants gave different ratings to each level of the "average" factor and in particular the critical athletes were rated differently from the ones with an average of 50, indicating that participants were not able to filter out the 7th ("incorrect") value. For the case in which the extravalue is lower than the average (i.e., the 44.3 sequence) the results provide further support for averaging compared with summation. Furthermore, the ratings were highly linear with the nominal average values that *include* the "computer error," showing that participants gave the same weight to the computer error as to the other values. This took place despite the fact that all the participants, when debriefed, reported seeing the values enclosed by the red-squares and expressed confidence that they ignored them.

Experiments 1–3: Discussion

The results of Experiments 1–3 provide strong support for the averaging principle, as a determinant of preferences between alternatives characterized by rapid numerical sequences (basketball players that are candidates for Hall of Fame and competing athletes). In a follow-up experiment (reported as Exp. 5, in the Suppl.) we show that the range of the average-prevalence effect extends to sequences of slot-machine rewards[8] (see also Exp. 4). In all these experiments the participants preferred alterna-

[8] In this experiment we find that participants are sensitive not only to the sequence average but also to its variance, with both determining the preference as suggested by a prominent economic model (risk-return; Markowitz, 1952; see also Weber, 2010).

tives with higher average and lower sums over alternatives with higher sums and lower average (an analog of the "Jordan-effect"). This conclusion is supported by the model comparison carried out in Experiment 2 on the individual participants. For the majority of the participants, the pure average was the dominating predictor for their evaluations, even when among the alternative models we included the Peak-End Heuristics (and some variants of the summation) which shares properties with the average.

As we presented the information at a fast rate (600 ms/sample) and as the instructions emphasized intuitive evaluations, we attribute the underlying process to one that is characteristic to automatic/intuitive processes (Horstmann et al., 2009, 2010). This interpretation is consistent with two supporting results. First, models of controlled (step-by-step) processing, such as the adjustment and anchoring model (Hogarth & Einhorn, 1992) provided a less good account of the data. Second the results from Exp. 3 showed that for the rapid sequential presentation we used, participants gave the same weight to to-be-ignored values as they gave to regular values. As the ability to act on negation is one of the characteristics of controlled processing (Deutsch et al., 2006; Gawronski & Bodenhausen, 2006), this supports an automatic/intuitive mechanism.

Although more research is needed to fully establish automaticity in our experimental setup, it is important to highlight that these results are remarkably different from those obtained in a *fully analytical* set up. In a control experiment for the Hall of Fame experiment, we asked the participants ($N = 29$) to evaluate candidates that correspond to the critical players in Exp. 1, using self-controlled evaluations, in which the alternatives were presented together in a table format and without time limits (see Control Experiment in suppl. information). The results were markedly different from those of Exp. 1–2. In strong contrast to our previous results, here most participants (21/29, $p = .012$; two-tailed binomial test for difference from 50%) ranked the "High-Sum, Low-Average" player as more eligible for induction to the Hall Of Fame, compared with his "High-Average, Low-Sum" counterpart (see suppl. information for details). This indicates that under the most optimal conditions, most (though not all) participants believe that alternatives that are equivalent to Jordan after his comeback should not be devalued relative to the precomeback status. Although it is beyond the scope of this paper to address the normativity of these principles (which requires a separate investigation), we believe that the results provides a lab analogue to the cold water experiments of Kahneman and colleagues (Kahneman et al., 1993; Redelmeier & Kahneman, 1996), in which the intuitive judgments favor the alternative that adds to total discomfort, but reduces its average.

While evidence that attitudes are determined by an averaging principle is not new (e.g., Anderson, 1981), the results differ from those obtained by Betsch et al. (2006, 2001), and which supported a summative evaluation principle. This is important since the Betsch studies, which highlighted the automaticity of the attitudes by presenting participants a single trial in which the sequences' values were framed as irrelevant distraction, were one of the major motivations of our investigation. Thus, if we are to make some progress in understanding the process involved, it is necessary to better understand the source of this difference. On the basis of computational considerations (discussed below) we suggest that one important factor is the mode of presentation and evaluation: one alternative at a time (in our experiments) versus all alternatives presented and rated together (in Betsch et al.'s studies). In our final experiment we test this prediction, but first we briefly outline its motivation.

A Neurocomputational Prediction: The Presentation/Evaluation Format

Establishing averaging as a mediating principle for intuitive preferences in Experiments 1–3 is only a first step in explaining the mediating process or mechanism. Unlike summation, which is naturally mediated by a variety of accumulator (or diffusion) models[9] (Brown & Heathcote, 2008; Busemeyer, 1985; Busemeyer & Townsend, 1993; de Gardelle & Summerfield, 2011; Forstmann, Ratcliff, & Wagenmak-

[9] Each alternative is associated with a neural accumulator that integrates the information samples that correspond to it. Although such models are usually (and naturally) employed in accounting for decisions, they can also generate "liking" ratings, if we assume that activations are systematically (albeit arbitrarily) converted onto an analog scale.

ers, 2016; Kiani, Corthell, & Shadlen, 2014; Stewart, 2009; Vickers, 1970; Zeigenfuse, Pleskac, & Liu, 2014), averaging appears to require a more complex process. As discussed by Betsch and colleagues (2001), averaging requires tracking both the sum and the number of samples for each alternative (and dividing accordingly, which is not very plausible for the rapid presentation conditions of our experiments in which the number of samples varies). The accumulator models described above, for example, are unable to extract averages as they cannot distinguish between the same activation that is the result of few large values or a larger number of smaller ones. A variant of such models based on leaky accumulators (Busemeyer & Townsend, 1993; Hogarth & Einhorn, 1992; Tsetsos, Chater, & Usher, 2012; Usher & McClelland, 2004; Yechiam, Busemeyer, Stout, & Bechara, 2005) can interpolate between a summation principle (at small *N*) and an averaging principle (at large *N*). Such models, however, predict temporal weights (i.e., order effects) that are not consistent with our data (e.g., Exp. 2).

Research on explicit numerical averaging, however, indicates that human participants are quite good at averaging sequences of two or three digit numbers, even at rapid presentation rates of 2 per seconds (Brezis, Bronfman, & Usher, 2015; Malmi & Samson, 1983) and has started to address the mediating mechanism. Malmi and Samson, for example, explicitly discussed two hypotheses regarding the averaging mechanisms: One is a *running-average* (a type of online updating with a decreasing weight for each new item; see also Hau, Pleskac, Kiefer, & Hertwig, 2008) that only maintains the running-average and discards distributional properties of the values sampled, and the other is a value-distribution account, which maintains the distribution of values (a type of histogram) and computes the average offline as its balance-point or "fulcrum" (p. 552) and they provided support for the latter. This idea was further developed in two neurocomputational studies, in which it was suggested that numerical averaging for rapid numerical sequencing is mediated by a population coding mechanism (Brezis, Bronfman, Jacoby, Lavidor, & Usher, 2016; Brezis et al., 2015; see also Jazayeri & Movshon, 2006) that operates an analog/intuitive or approximate pathway, in which Arabic numerals are rapidly and automatically translated from their digital code into a noisy quantity code on a "mental number line" (Dehaene, 1992, 2007; Dehaene, Molko, Cohen, & Wilson, 2004).

If indeed numerical averaging of rapid sequences is mediated by a population-averaging mechanism that operates on the representation of the sampled value distribution, one may expect that this mechanism will be easier to deploy in the one-by-one evaluation and presentation condition, which we used in our experiments, but more difficult in the grouped presentation conditions of Betsch et al. (2001), which require the maintenance of multiple distributions, one for each sequence presented in parallel, to generate distinct population averages (without being contaminated by the others).[10] Assuming task adaptivity (Payne, Bettman, & Johnson, 1988; Tsetsos et al., 2016), we predict that in the grouped presentation/evaluation condition, most participants employ an accumulator-based mechanism that results in preferences that are dominated by the summation principle (Betsch et al., 2006, 2001), or exhibit a balance between the two. Indeed, this is a result we obtained in a preliminary experiment (M. Brusovansky, MA thesis), in which we presented four sequences of values in a randomized order and required the participants to make the evaluations for each at the end of the presentation as in Betsch et al. (2006, 2001). We test this prediction is Exp. 4, which presents each participant the same sequences for evaluation, once in a one-by-one format (session-1) and once in a grouped format (session-2, counterbalanced).

Experiment 4

To test this prediction and to extend the previous results to new domains, we carried out an experiment, in which each participant viewed the same numerical sequences (now framed as possible outcomes of various slot-machines (4a), or as ratings given by students to lecturers (4b)) and made liking ratings under two conditions: *one-by-one*, versus *grouped*.

[10] This is similar to the limitation that participants have in maintaining more than one signal detection criterion when trials of different difficulties are randomly mixed within a block (Gorea & Sagi, 2000).

Method

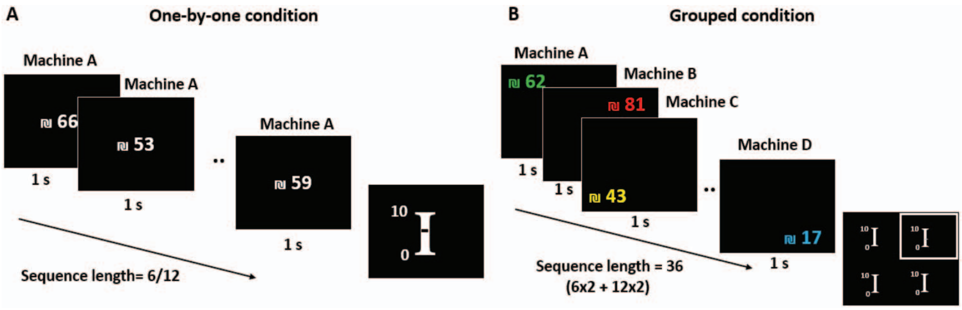
Participants. Twenty-six students (21 females, age: 19–38 years, M = 22.5) from Tel Aviv University participated in two sessions of the study, in exchange for credit. All participants had normal or corrected vision. One participant in the grouped session of Exp. 4b only completed 2/3 of the trials. Removing this participant did not affect the results.

Materials. The experiment was conducted over two sessions, which were run one week apart (order was counterbalanced), once in a one-by-one and once in a grouped format. Each session included two parts (4a and 4b), which were administered as different blocks. Apart from the format factor, *part 4a* (slot-machines) consisted of a factorial 2 × 2 design of average (40/56) and sequence-length (6/12). *Part 4b* consisted of filler and critical trials, similar to Exp. 1, with one critical sequence corresponding to a "High-Average, Low-Sum" condition (6 items with an average of 65), while the other corresponding to a "Low-Average, High-Sum" condition (6 items with an average of 65 and additional 3 items with an average of 50, which were placed randomly in the sequence). "Filler" alternatives with a low average (40) and a moderate set size (7 items) were included, to make the "critical" sequences less conspicuous. In all the sequences the numerical values were in the range 1–99 and were drawn from a Gaussian distribution with an average determined by the factorial design and a fixed *SD* of 13.

Procedure. In the *one-by-one evaluation* condition, the sequences were presented separately, one per trial, as illustrated in Figure 5 (left panel). In Exp. 4a, the task was framed as involving outcomes of slot machines and each trial corresponded to a single slot-machine, whose rewards were displayed sequentially at the center of the screen, with a presentation rate of 1/sec. Following each trial, participants were asked to indicate their "liking" of the slot-machine by using the mouse cursor on a continuous (0–10) scale. Each session started with a practice phase presenting eight slot-machines. Following the completion of Exp. 4a, participants received instruction for Exp. 4b, in which the task was framed as an evaluation of lecturers based on rating scores given by their students in previous years. Overall, in part 4a 96 slot-machines were displayed during the test phase (24 slot-machines for each of the four average × set size conditions), and in part 4b 60 lecturers (15 pairs of 30 critical alternatives and 30 fillers).

F5

In the *grouped evaluation* condition (Exp. 4a) participants were told that they would see different "rooms" (one per trial), each one with four slot-machines (see Figure 5, right panel). The four slot-machines presented in each trial corresponded to the 4 average × set size conditions (i.e., there were two sequences with six values and two with 12 values). To facilitate differentiation between the four slot-machines in each "room," each machine was presented in



ONLINE COLOR

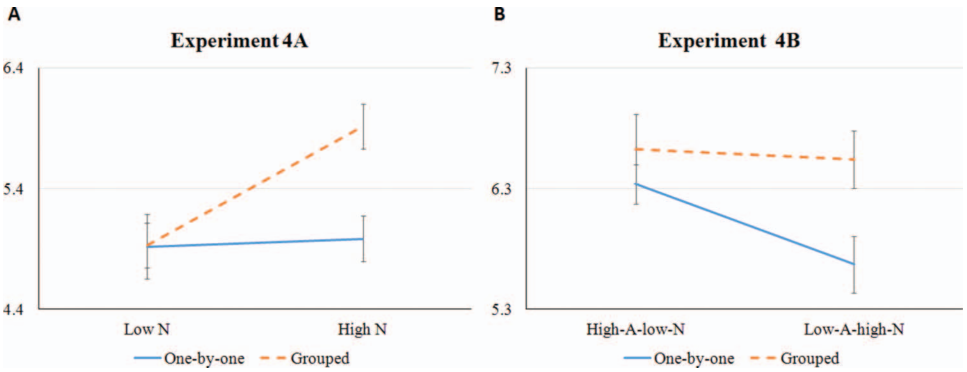
Figure 5. Exp. 4, one-by-one (A) and grouped (B) evaluation presentations of numerical sequences, followed by evaluation scales. In (A) the sequences were presented in the middle of the screen, while in (B) they were presented in the four quadrants (two sequences with six and two with 12 values, each with a different color) in a pseudorandom order with the constraint of never displaying two outcomes successively from the same sequence. See the online article for the color version of this figure.

a different corner of the screen, and had a unique color (red, green, blue or yellow; see Figure 5 right panel). Each trial presented a sequence of 36 (6 × 2 + 12 × 2) outcomes, in a random order, subject to the constraint of never displaying two outcomes successively from the same slot-machine (screen-corner). At the end of each trial, four visual analogue scales ranging from 0 to 10 were displayed, one in each quadrant, and participants were asked to evaluate the corresponding slot-machine. The order of the machines' evaluations was not constrained, and participants were free to choose which machine they would prefer to evaluate first (a rectangle around each scale, colored either red, green, blue, or yellow, indicated which of the scales was currently operated). The session started with a short practice phase, in which two rooms were presented. After completing Exp. 4a participants received instructions for Exp. 4b, in which a similar procedure was executed, but now the participants were instructed that they would view different "courses" taught by four different lecturers (two of the four were a critical pair and two were fillers; participants were not informed of the critical alternatives). Each trial presented a sequence of numerical values corresponding to the ratings the 4 lecturers received in past years, after which participants were asked to evaluate (on a scale of 0–10) how much they wish to study with each lecturer.

Results

Exp. 4a. We carried out a repeated measures ANOVA with average (low vs. high), set size (small vs. large) and evaluation format (one-by-one vs. grouped) as within-participant variables. Consistent with results from Experiments 1–3, the effect of average was highly significant (high averages: $M = 6.22$, $SD = 0.13$; low averages: $M = 4.16$, $SD = 0.15$); $F(1, 25) = 97.00$; $p < .0001$). The predicted interaction between evaluation formats and sequence-length was also highly significant, $F(1, 25) = 15.33$; $p < .0005$. Post hoc Tukey's tests indicated that participants gave higher rating to large sequence-length in the grouped condition ($p < .0005$), but (as in our previous experiments) exhibited no sensitivity to sequence-length in the one-by-one condition ($p = .999$).

Exp. 4b. We carried out a repeated measures ANOVA solely on the critical sequences' data, with average-sum ("Low-Average, High-Sum" vs. "High-Average, Low-Sum") and evaluation format (one-by-one vs. grouped) as within-participant variables. Main effects of evaluation formats ($F(1, 25) = 7.34$; $p = .012$) and of average-sum ($F(1, 25) = 19.15$; $p < .001$) were found, but critically, we found the predicted interaction between evaluation formats and average-sum ($F(1, 25) = 13.92$; $p < .001$; see Figure 6B). A F6 post hoc Tukey's test revealed that in the one-by-one condition, participants prefer the "High-



ONLINE COLOR

Figure 6. (A) The slot-machines' average ratings in Exp. 4a, showing an interaction between presentation format and set size collapsed across average. (B) The lecturers' average ratings in the critical sequences in Exp. 4b, showing an interaction between presentation formats and average-sum conditions. Error bars correspond to within-participant 95% confidence intervals. See the online article for the color version of this figure.

Average, Low-Sum" alternatives to the "Low-Average, High-Sum" one (a replication of our "Jordan-effect"; $p < .001$), but no difference in preference between the two was found in the grouped condition ($p = .867$).

General Discussion

The results of Exp. 4 confirm the predicted presentation/evaluation-format effect. In Exp. 4a we find that while increasing sequence-length had no effect on the evaluations in the one-by-one condition (see Figure 6A, blue line), it did increase the evaluations in the grouped condition (see Figure 6A, red dashed line), consistent with a transition from an averaging to an accumulator model. In Exp. 4b, we replicated the "Jordan-effect" in the one-by-one condition (see Figure 6B, blue line), extending it to a different domain. However, we find that this effect disappears in the grouped condition (see Figure 6B, red dashed line), where both the average and the sequence-length appear to influence the evaluations, consistent with the results reported by Betsch et al. (2006, 2001).[11] Fn11 Moreover, this preference change was demonstrated in the same group of participants, who viewed the same alternatives under the two conditions. This presentation/evaluation format effect is somewhat reminiscent of the joint versus separate evaluation effect in description-based multiattribute decisions (Hsee, 1996). While in the latter, the choice bias is the result of differential difficulty to evaluate one of the two attributes (as participants lack knowledge of the expected range of values on that attribute), the effect which we report here takes place in experience-based decisions, in which the range of values is specified.

We have proposed that this presentation/evaluation-format effect is the result of an adaptational change in the preference construction mechanism. While under the one-by-one condition, participants rely (mostly) on a population-averaging mechanism, in the grouped condition they rely (mostly) on an accumulator (or sequential sampling) mechanism. While the latter is a standard assumption in many models of preference construction (e.g., Busemeyer, 1985; Busemeyer & Townsend, 1993; Stewart, 2009; Tsetsos et al., 2012, 2016; Usher & McClelland, 2004; Zeigenfuse et al., 2014), the population-averaging mechanism was used less in the decision making literature. Note, however, that such a mechanism was proposed to account for performance in tasks of explicit numerical averaging (Malmi & Samson, 1983; Brezis et al., 2016; Brezis et al., 2015). Furthermore, such a mechanism was proposed to account for the impressive ability of human observers in the evaluation of summary statistics of perceptual stimuli (Alvarez & Oliva, 2008; Ariely, 2001; Bronfman, Brezis, Jacobson, & Usher, 2014; Chong & Treisman, 2003, 2005; de Gardelle & Summerfield, 2011). While perceptual summary studies involve the averaging of perceptual stimuli (size, color, orientation, etc.), here we suggest that a similar mechanism may apply to intuitive evaluations of numerical values, an idea consistent with Kahneman's suggestion that intuition operates at the interface between perception and cognition (Kahneman, 2003), and with recent suggestion that perceptual and preference-based decisions processes share cognitive machinery (Busemeyer, Jessup, Johnson, & Townsend, 2006; Summerfield & Tsetsos, 2012).

Here we have argued that the prevalence of the accumulator type mechanism in the decision-making literature results from its dominant reliance on methods that present groups of alternatives for choice or for grouped evaluations, which generate a representational bottleneck of maintaining multiple value-representations without mixing them. This challenge, however, is not posed in the one-by-one evaluation condition, for which the population-averaging is an efficient mechanism that allows an automatic (and effortless) way to compute an approximate average. Furthermore, population-averaging allows the (approximate) estimation of higher order statistics, such as the variance, opening the way for a mechanism to implement a well-known economic theory of risk preference: the *risk-return* model (Markowitz, 1952; see also Weber, 2010). According to this model, risk preferences are generated by a tradeoff between the estimated average and variance of an alternative (see Exp. 5 in the Suppl. for an illustration of this model to alternatives made of numerical sequences). Here we suggest that risk biases will only follow the risk-return model (whose signature is that risk biases do not depend on the average) under conditions of one-by-one evalua-

[11] Our data indicate individual differences in the tendency of the participants to deploy an accumulator or an averaging mechanism.

tion and that they will conform to an accumulator type model in grouped evaluation conditions (Vanunu, Pachur & Usher, 2017). AQ: 5

Based on these results, we propose that the preference construction mechanism is adaptationally contingent on the type of presentation and evaluation format, with a population-averaging mechanism favoring one-by-one conditions that results in average-based preferences and risk-return type biases (see Suppl.), and with an accumulator type mechanism favoring grouped decisions (Tsetsos et al., 2012, 2016; Vanunu et al., 2017; Zeigenfuse et al., 2014). AQ: 6 Although this characterization is framed at the group-level and focuses on the effects on task-framing, the results also indicate important individual differences (see Table S1 in the Suppl.). Thus, in addition to task-contingency, the preference mechanism is also subject to individual differences; some individuals may have a tendency to rely on one type of mechanism more than on the other in both conditions. Future research will be needed to investigate the nature of these individual differences, their interaction with the task demand and their dependence on cognitive resources (e,g., WM-capacity, or resistance to memory interference).

If correct, this distinction between one-by-one and group evaluations and decisions may have an important impact on the quality of daily decisions. Although most research has focused on choices between pairs of alternatives, and demonstrated marked deviations from normative principles, some of which resulting from attentional biases (Shafir, 1993; Tsetsos et al., 2012, 2016; Zeigenfuse et al., 2014), it is quite possible that reliance on a one-by-one evaluation mode (as we do here and as done in some older studies on attitudes; Anderson, 1981) may uncover a reduction of such deviations (but see Hsee, 1996 for cases of multiattribute decisions from description). If this is the case, one may prefer, in order to reduce biases in choices between consumer products (e.g., clothes or cellphones), to examine and evaluate each in isolation, rather than in parallel, as is often the case when we shop.

To conclude, we suggest that averaging and summation are principles that characterize complementary neural mechanisms, which participants employ in attitude formation, and whose prevalence depends on the task complexity. Future research is required to validate this hypothesis and to test whether reliance on these mechanisms is adaptive to the nature of the task. While the normative preferences in the tasks we used here were intentionally ambiguous, future studies can manipulate the framing to favor either summation or averaging in terms of the objective reward.

References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worchester, MA: Clark University Press.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398. http://dx.doi.org/10.1111/j.1467-9280.2008.02098.x

Anderson, N. H. (1981). Integration theory applied to cognitive responses and attitudes. In R. E. Petty, T. M. Ostrom, & T. C. Brock (Eds.), *Cognitive responses in persuasion* (pp. 361–397). Hillsdale, NJ: Erlbaum.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162. http://dx.doi.org/10.1111/1467-9280.00327

Betsch, T., Kaufmann, M., Lindow, F., Plessner, H., & Hoffmann, K. (2006). Different principles of information integration in implicit and explicit attitude formation. *European Journal of Social Psychology, 36,* 887–905. http://dx.doi.org/10.1002/ejsp.328

Betsch, T., Plessner, H., Schwieren, C., & Gütig, R. (2001). I like it but I don't know why: A value-account approach to implicit attitude formation. *Personality and Social Psychology Bulletin, 27,* 242–253. http://dx.doi.org/10.1177/0146167201272009

Brezis, N., Bronfman, Z. Z., Jacoby, N., Lavidor, M., & Usher, M. (2016). Transcranial direct current stimulation over the parietal cortex improves approximate numerical averaging. *Journal of Cognitive Neuroscience, 28,* 1700–1713. http://dx.doi.org/10.1162/jocn_a_00991

Brezis, N., Bronfman, Z. Z., & Usher, M. (2015). Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Scientific Reports, 5,* 10415. http://dx.doi.org/10.1038/srep10415

Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We see more than we can report: "cost free" color phenomenality outside focal attention. *Psychological Science, 25,* 1394–1403. http://dx.doi.org/10.1177/0956797614532656

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178. http://dx.doi.org/10.1016/j.cogpsych.2007.12.002

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 538–564. http://dx.doi.org/10.1037/0278-7393.11.3.538

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks, 19,* 1047–1058. http://dx.doi.org/10.1016/j.neunet.2006.05.043

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459. http://dx.doi.org/10.1037/0033-295X.100.3.432

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404. http://dx.doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900. http://dx.doi.org/10.1016/j.visres.2004.10.004

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 13341–13346. http://dx.doi.org/10.1073/pnas.1104517108

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44,* 1–42. http://dx.doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S. (2007). A few steps toward a science of mental life. *Mind, Brain and Education, 1,* 28–47. http://dx.doi.org/10.1111/j.1751-228X.2007.00003.x

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology, 14,* 218–224. http://dx.doi.org/10.1016/j.conb.2004.03.008

Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91,* 385–405. http://dx.doi.org/10.1037/0022-3514.91.3.385

Diener, E., Wirtz, D., & Oishi, S. (2001). End effects of rated life quality: The James Dean Effect. *Psychological Science, 12,* 124–128. http://dx.doi.org/10.1111/1467-9280.00321

Fazio, R. H. (1989). On the power and functionality of attitudes: The role of attitude accessibility. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function. The third Ohio State University volume on attitudes and persuasion* (pp. 153–179). Hillsdale, NJ: Erlbaum.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67,* 641–666. http://dx.doi.org/10.1146/annurev-psych-122414-033645

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692–731. http://dx.doi.org/10.1037/0033-2909.132.5.692

Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 12380–12384. http://dx.doi.org/10.1073/pnas.97.22.12380

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518. http://dx.doi.org/10.1002/bdm.598

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J

Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making, 4,* 335–354.

Horstmann, N., Hausmann, D., & Ryf, S. (2010). Methods for inducing intuitive and deliberate processing modes. In A. Glöckner & C. Witteman (Eds.), *Foundations for tracing intuition: Challenges and methods* (pp. 219–237). New York, NY: Psychology Press.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67,* 247–257. http://dx.doi.org/10.1006/obhd.1996.0077

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9,* 690–696. http://dx.doi.org/10.1038/nn1691

Juslin, P. (2015). Controlled information integration and Bayesian inference. *Frontiers in Psychology, 6,* 70.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697–720. http://dx.doi.org/10.1037/0003-066X.58.9.697

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4,* 401–405. http://dx.doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84,* 1329–1342. http://dx.doi.org/10.1016/j.neuron.2014.12.015

Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. *Journal of Verbal Learning & Verbal Behavior, 22,* 547–559. http://dx.doi.org/10.1016/S0022-5371(83)90337-7

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7,* 77–91.

National Basketball Association official website. (2015). *NBA History Legends profile: Michael Jordan*. Retrieved May 8th, 2016, from http://www.nba.com/history/legends/michael-jordan

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552. http://dx.doi.org/10.1037/0278-7393.14.3.534

Pham, M. T., Lee, L., & Stephen, A. T. (2012). Feeling the future: The emotional oracle effect. *Journal of Consumer Research, 39,* 461–477. http://dx.doi.org/10.1086/663823

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66,* 3–8. http://dx.doi.org/10.1016/0304-3959(96)02994-6

Rusou, Z., Zakay, D., & Usher, M. (2013). Pitting intuitive and analytical thinking against each other: The case of transitivity. *Psychonomic Bulletin & Review, 20,* 608–614. http://dx.doi.org/10.3758/s13423-013-0382-7

Rusou, Z., Zakay, D., & Usher, M. (2016). Intuitive number evaluation is not affected by information processing load. In J. I. Kantola, T. Barath, S. Nazir, & T. Andre (Eds.), *Advances in human factors, business management, training and education* (pp. 135–148). New York, NY: Springer.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66. http://dx.doi.org/10.1037/0033-295X.84.1.1

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21,* 546–556. http://dx.doi.org/10.3758/BF03197186

Simmons, B. (2010). *The book of basketball: The NBA according to the Sports Guy*. New York, NY: Ballantine Books.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 62,* 1041–1062. http://dx.doi.org/10.1080/17470210902747112

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *frontiers in Neuroscience, 6*(70). http://dx.doi.org/10.3389/fnins.2012.00070

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 9659–9664. http://dx.doi.org/10.1073/pnas.1119569109

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 3102–3107. http://dx.doi.org/10.1073/pnas.1519157113

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111,* 757–769. http://dx.doi.org/10.1037/0033-295X.111.3.757

Usher, M., Russo, Z., Weyers, M., Brauner, R., & Zakay, D. (2011). The impact of mode of thought in complex decisions: Intuitive decisions are better. *frontiers in Psychology, 2*(37). http://dx.doi.org/10.3389/fpsyg.2011.00037

Vanunu, Y., Pachur, T. & Usher, M. (2017). *Constructing preference from rapid sequential samples: the impact of evaluation format on risk attitudes*. Manuscript submitted for publication. AQ: 7

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13,* 37–58. http://dx.doi.org/10.1080/00140137008931117

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44,* 92–107. http://dx.doi.org/10.1006/jmps.1999.1278

Weber, E. U. (2010). Risk attitude and preference. *WIREs Cognitive Science, 1,* 79–88. http://dx.doi.org/10.1002/wcs.5

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16,* 973–978. http://dx.doi.org/10.1111/j.1467-9280.2005.01646.x

Zeigenfuse, M. D., Pleskac, T. J., & Liu, T. (2014). Rapid decisions from experience. *Cognition, 131,* 181–194. http://dx.doi.org/10.1016/j.cognition.2013.12.012

Received November 2, 2016
Revision received February 16, 2017
Accepted May 12, 2017 ■

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQau—Please confirm the given-names and surnames are identified properly by the colors.
🟥= Given-Name, 🟦= Surname
The colors are for proofing purposes only. The colors will not appear online or in print.

AQ1—Author: This article has been lightly edited for grammar, style, and usage. Please compare against your original document and make changes on these paged proofs. If no change is required in response to a question, please write “OK as set” in the margin.

AQ2—Author: Please be sure to provide the name of the department(s) with which you and your coauthors are affiliated at your respective institutes if you have not already done so. If you are affiliated with a governmental department, business, hospital, clinic, VA center, or other nonuniversity-based institute, please provide the city and U.S. state (or the city, province, and country) in which the institute is based. Departments should be listed in the author footnote only, not the byline. If you or your coauthors have changed affiliations since the article was written, please include a separate note indicating the new department/affiliation: [author’s name] is now at [affiliation].

AQ3—Author: Please supply three to five key words.

AQ4—Author: Please review the typeset table carefully against your original table to verify accuracy of editing and typesetting, provide a title for the table, and cite the table in the article text.

AQ5—Author: Please provide year of publication for Vanunu et al or, if not yet accepted for publication, year of last revision.

AQ6—Author: Please provide year of publication for Vanunu et al or, if not yet accepted for publication, year of last revision.

AQ7—Author: If reference Vanunu et al. has been accepted for publication, please provide all available publication information; if it has not, provide date (year) of last revision.

AQ8—Author: Please confirm that all authors’ institutional affiliations (including city/state/country locations) and correspondence information are correct as shown in the affiliations footnote.

AQ9—Author: Please confirm figure(s) and legend(s) as set.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQ10—Author: Please provide table citation in the text.