Decision

2325-9965/16/$12.00 http://dx.doi.org/10.1037/dec0000073

Rejecting Outliers: Surprising Changes Do Not Always Improve Belief Updating

Alex Filipowicz, Derick Valadao, Britt Anderson, and James Danckert
University of Waterloo

An important human skill is the ability to update one's beliefs when they are no longer supported by the environment. Current models of dynamic decision-making suggest that more unexpected, or "surprising," events lead to quicker belief updating. The current article tests the ubiquity of the notion that surprising environmental changes are always positively related to updating. Using a novel task based on the game Plinko, we tracked participants' beliefs as they learned distributions of ball drops. At an unannounced point during the task, the distribution of ball drops changed and we computed how surprising these changes were relative to participants' beliefs and compared how this surprise factor influenced their ability to update their beliefs to reflect the change. We found that, consistent with current models, there were some situations in which belief updating was positively related to the surprise of a change. However, we also found a situation in which highly surprising changes were negatively related to updating—situations where participants tended to update less with increasingly surprising changes. This negative relationship seems due to participants' treating highly surprising events as "outliers" and choosing not to integrate them in their current beliefs. Our results provide a novel and more nuanced representation of the relationship between surprise and updating that should be considered in models of dynamic decision-making.

Keywords: belief updating, mental models, probabilistic learning, surprise

Supplemental materials: http://dx.doi.org/10.1037/dec0000073.supp

In Sir Arthur Conan Doyle's story *Silver Blaze,* Sherlock Holmes is faced with a mystery involving the disappearance of a prized race horse (Doyle, 1930/1892). The crucial detail in Holmes's solving of the case was that the stable dog did not bark the night the horse was taken away—so Holmes deduced that the person involved in the horse's disappearance must have been familiar to the dog, a fact that suggested an inside job. Holmes correctly identified that the *absence* of an expected event—the dog's barking—was an important factor to consider in solving the case, a fact that had been overlooked by other investigators.

This story poses interesting questions regarding the way humans process and learn from events that occur or do not occur in their environment. Indeed, individuals are regularly required to process large volumes of sensory information that often exceed their perceptual capacities (Barlow, 1961; Wei & Stocker, 2015). To make sense of this information, they build coherent "mental models" based on the frequency and prevalence of past experiences to guide the way they understand and interact with the world (Johnson-Laird, 2004; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Alex Filipowicz, Derick Valadao, Britt Anderson, and James Danckert, Department of Psychology, University of Waterloo.

This research was supported by Discovery Grant 261628-07 from the Natural Sciences and Engineering Research Council (NSERC) of Canada (to James Danckert), Canadian Institutes of Health Research Operating Grant 219972 (to James Danckert and Britt Anderson), and Ontario Graduate Scholarships and NSERC Alexander Graham Bell Canada Graduate Scholarships (to Alex Filipowicz and Derick Valadao). We thank Liat Koefler, Elliot Lee, Anna Pipkin, and Ryan Yeung for their assistance with data collection.

Correspondence concerning this article should be addressed to Alex Filipowicz, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada. E-mail: alsfilip@uwaterloo.ca

In addition to building mental models, an equally important skill is the ability to *update* one's models when faced with new information that is not explained by a current model (Collins & Koechlin, 2012; Danckert, Stöttinger, Quehl, & Anderson, 2012; Nassar, Wilson, Heasly, & Gold, 2010). Ideally, one's mental models should be updated whenever change occurs; however, as highlighted by the Sherlock Holmes example, detecting important events is not always particularly obvious—environmental changes can also be prompted by the *absence* of events. Detecting and efficiently utilizing such absences might prove more difficult than is responding to changes signaled by the *presence* of new events.

Research has demonstrated that individuals pay close attention to *surprising* information when judging probabilities (Fisk, 2002), and the concept of *surprise* plays an important role in current studies of updating (Mars et al., 2008; McGuire, Nassar, Gold, & Kable, 2014; Nassar et al., 2010; O'Reilly et al., 2013). In the context of learning and decision-making, surprise describes an unexpected and/or novel event, particularly one that is contrasted with another, more expected event (Teigen & Keren, 2003). Of importance, this definition implies that surprise is a subjective experience that depends on a person's current expectations. Thus, surprise can properly be measured only insofar as a person's prior expectations can be measured. Indeed, updating does not occur in a vacuum—it depends largely on the mental model an observer is using to interpret the environment (Collins & Koechlin, 2012; Filipowicz, Valadao, Anderson, & Danckert, 2014; Lee & Johnson-Laird, 2012; Stöttinger, Filipowicz, Danckert, & Anderson, 2014). One prominent challenge in measuring the influence of surprise on updating, therefore, is being able to accurately characterize the mental models an observer holds at any given moment.

Previous research measuring the effect of surprise on updating has generally approximated mental models from participant responses. These approximations are often obtained by building ideal observers (Mars et al., 2008; O'Reilly et al., 2013) or by fitting participants' responses to computational models (e.g., Bayesian change-point models; McGuire et al., 2014; Nassar et al., 2010). A measure of surprise is then obtained by measuring the discrepancy between these mental model approximations and the observations that the mental model is attempting to predict—the larger the discrepancy, the higher the calculated surprise of the event. This research has consistently found that participants update more quickly with increasing discrepancies between their predictions and current observations, suggesting that surprise is positively related to updating (McGuire et al., 2014; Nassar et al., 2010).

There are, however, some questions related to the ubiquity of this relationship. Does one always update when faced with surprising events? In some cases, individuals treat discrepant information with a sort of skepticism and discount it when building a representation of the environment (De Gardelle & Summerfield, 2011). For example, when attempting to classify an array of objects based on color, participants were found to base their responses more on coherent objects in the array and to reject the contribution of items that deviated strongly from the rest (De Gardelle & Summerfield, 2011). Although these rules have primarily been found in studies of human perception, some researchers have argued that these tendencies are also present in decision-making, leading one to sometimes treat highly surprising events as a type of "outlier" (Summerfield & Tsetsos, 2015). This suggests that rather than blindly integrating *any* surprising information, there may be situations in which one can be resistant to highly surprising changes.

The current study explores the relationship between surprise and updating in more detail. Using a task based on the game Plinko to accurately represent mental models, we exposed participants to distributions of events that changed at an unannounced point and varied in their level of surprise. We then used participants' responses to measure how the surprise of each change related to their ability to update. In contrast to the case in prior work, we did not find that updating was always positively related to the degree of surprise. Instead, we found some situations in which surprise and updating were negatively correlated, such that, rather than integrate highly surprising events, participants devalued them.

Method

Participants

Seventy-eight University of Waterloo undergraduates (54 female; mean age = 19.64 years, *SD* = 1.59 years) participated in our study in exchange for course credit.

Experimental setup

Participants were exposed to a computerized version of the game Plinko. In our game, participants saw that a red ball would fall through a pyramid of pegs and land in one of 40 possible slots located below the pegs (see Figure 1a).

The ball drops followed prespecified probability distributions that participants attempted to learn. Participants were informed that their goal was to accurately predict the likelihood that a ball would fall in any of the 40 slots on future trials. To represent their likelihood estimations, they drew bars under the slots with a computer mouse. It is important to note that these bars could be adjusted at the start of each trial, as participants saw new ball drops. These bars provided us with trial-by-trial probability distributions of participants' beliefs as the task progressed (see the online supplementary materials for the full procedure).

Measuring Accuracy

Performance was measured by computing how accurately participants managed to represent the computer's ball distribution with the bars they drew below each slot. Accuracy was calculated on every trial as the proportion of overlap between the participants' distribution

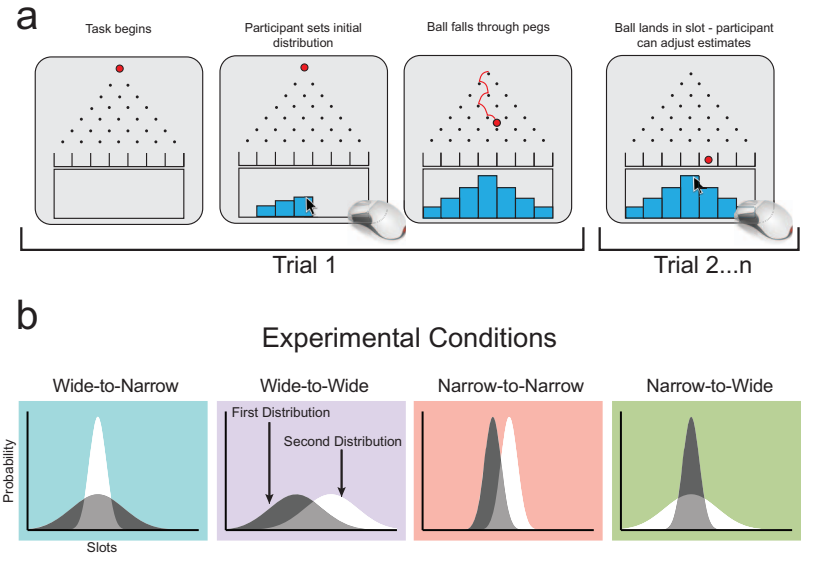


Figure 1. Plinko task environment and task conditions. Panel a: At the start of the task, participants were informed that a red ball would fall through a pyramid of pegs and land in one of 40 slots on each trial (note that only seven slots are pictured in this schematic). They were instructed to draw bars using the computer mouse to indicate how likely they believed the ball was to land in any of the 40 slots—with higher bars indicating an expectation of higher likelihood. They drew their first set of bars before seeing any ball drops and had the option of adjusting their bars at the start of each trial, but they were not required to do so. Panel b: Participants were assigned to one of four switch conditions. Participants saw a first distribution of 100 ball drops that was generated from either a wide or narrow Gaussian distribution. They were then switched to a second Gaussian distribution of 100 ball drops that either changed in mean while holding the variance constant (i.e., wide-to-wide and narrow-to-narrow conditions) or changed in variance while holding the mean constant (i.e., wide-to-narrow and narrow-to-wide). See the online article for the color version of this figure.

and the discrete distribution of ball drops they were attempting to estimate. We then fit a standard exponential learning curve to participants' accuracy scores over time, providing values that represented participants' asymptotic performance, starting accuracy, and learning rate (e.g., Estes, 1950; see Equation S2 in the online supplemental materials). Using these values, we characterized participants' performance by using the estimated starting accuracy to represent their starting accuracy value, learning rate to capture how quickly they reached their asymptote from their starting value, and their estimated accuracy on the last trial of the distribution they were estimating to indicate a final level of accuracy achieved (see the online supplementary materials for the full details).

Computing Surprise

To measure the surprise of a shift, we used measures taken from information theory to quantify the information, or surprise, that an event provides given a specific set of expectations. This method defines *surprise* as "the negative log probability" of an event occurring under a distribution of expectations (cf. Attneave, 1959; Shannon, 1948). This method has commonly been used to quantify surprise in learning tasks, primarily when participants' expectations are characterized as continuous probability distributions (Doya, Ishii, Pouget, & Rao, 2007; Mars et al., 2008; O'Reilly et al., 2013; Strange, Duggins, Penny, Dolan, & Friston, 2005). In our task, given that participants' distributions were both discrete and that participants were not required to draw bars under every slot (i.e., potentially leaving some slots with a value of 0), we could not compute a pure measure of negative log probability to characterize surprise. Instead, we used and compared two complementary measures of surprise to account for the discrete nature of participants' distributions.

The first measure involved a modification of participants' slot values to make them compatible for computing negative log probability. As a second measure, we used *weighted empirical log odds,* a proxy of log odds developed to compute odds ratios for discrete distributions (Cox & Snell, 1989; see the online supplemental materials).

Using these two measures to quantify surprise, we calculated a "surprise factor" S of a shift from any first distribution j to any second distribution k for each distribution shift in our task by summing the ratio of the surprise s of each slot i (computed using either negative log probability or weighted empirical log odds) of each of the two distributions as follows:

$$S_{jk} = \sum_{i=1}^{40} \frac{S_{ij}}{S_{ik}}. \quad (1)$$

It is important to note that our formula computes shifts that *include* unexpected events to be more surprising than are shifts that *omit* previously observed events. We used this quantification of surprise to generate our event distributions and to compare how surprising each shift was to each participant.

Experimental Conditions

All participants were exposed to a first Gaussian distribution of 100 ball drops, then switched to a second Gaussian distribution of 100 balls drops without any cues to indicate that a switch had occurred. The shifts were in the form of either a mean shift (i.e., the mean of the Gaussian was shifted, but variance was held constant) or a variance shift (i.e., the mean of the Gaussian was held constant, but variance changed). This produced four between-subjects distribution shift conditions: wide to wide mean shift (wide–wide), wide to narrow variance shift (wide–narrow), a narrow to narrow mean shift (narrow–narrow), and a narrow to wide variance shift (narrow–wide; see Figure 1b).

Although equivalent in their overlap, these distribution shifts varied in their calculated surprise factor. Using Equation 1, we computed the surprise factor for each shift condition using both our modified negative log probability (NLP) measure and our weighted empirical log odds (wElog) measure to compute a surprise value s for each slot i. As is evident in Table 1, both measures predict a similar trend for the surprise factors of each switch condition, with the highest surprise factor being predicted for the narrow–wide condition, midrange surprise for both mean shifts (narrow–narrow and wide–wide), and the lowest surprise for the wide–narrow condition.

Table 1
Calculated Surprise Values for Each Experimental Condition

Surprise measure	Narrow–Wide	Narrow–Narrow	Wide–Wide	Wide–Narrow
NLP	124.80	42.71	39.98	26.99
wElog[a]	31.51	44.86	52.00	60.02

Note. Narrow = Gaussian distribution of ball drops with a small standard deviation; Wide = Gaussian Distribution of ball drops with a larger standard deviation. NLP = negative log probability; wElog = weighted empirical log odds.
[a] Lower values indicate higher predicted surprise.

Results

Updating Is Worst for Low Surprise Shifts

We began by examining how updating accuracy differed between our different surprise conditions. We ran separate mixed factorial analyses of variance (ANOVAs) for each distribution participants were exposed to (first or second distribution), with trial accuracy as a dependent measure, trial number as a within-subject factor, and condition as between-subjects factor.

When examining performance between conditions in the first distribution, we found significant main effects of condition, $F(3, 74) = 5.852$, $MSE = 4.712$, $p < .002$, and trial number, $F(99, 7326) = 64.749$, $MSE = .015$, $p < .001$, and a Trial Number × Condition interaction, $F(99, 7326) = 3.227$, $MSE = .015$, $p < .001$, indicating that there were overall differences between groups in mean accuracy over the course of the first distribution and that the rate at which participants managed to learn the first distribution varied between switch conditions. Performance in the second distribution also yielded main effects of condition, $F(3, 74) = 10.41$, $MSE = 11.87$, $p < .001$, and trial number, $F(99, 7326) = 18.276$, $MSE = .038$, $p < .001$, but no Trial Number × Condition interaction, $F(99, 7326) = .918$, $MSE = .004$, $p = .837$, suggesting that although mean accuracy differed between switch conditions, their accuracy changed at a similar rate over the course of the second distribution (see Figure 2). When participants were exposed to both distributions, post hoc paired samples t tests indi-

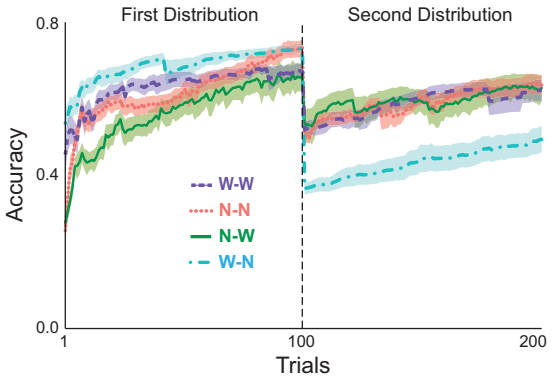


Figure 2. Accuracy performance for each surprise condition. Although all groups managed to learn the first distribution with equivalent accuracy, participants exposed to the low-surprise, wide–narrow shift (i.e., W-N; dotted-dashed line) finished the second distribution with the lowest accuracy of all four groups. There were no accuracy differences between participants in the medium-surprise, narrow–narrow and wide–wide conditions (i.e., N-N and W-W; dotted and dashed lines, respectively) and the high surprise, narrow–wide condition (i.e., N-W; solid line). The lines represent group means for each respective condition on each trial and shading represents ±1 standard error of the mean. See the online article for the color version of this figure.

cated that raw accuracy values increased from the first trial of each distribution (mean accuracy first distribution = .38, second distribution = .48) to the last trial (mean accuracy first distribution = .70, second distribution = .60; all ps $< .001$), indicating that participants were effectively learning the distributions they were observing. This effectively demonstrates that the changes participants made to their distributions, rather than being random, were directly related to the events they observed in the task environment.

We then used one-way ANOVAs and Tukey's honestly significant difference tests to examine how participants' performance fit parameters differed between our different switch conditions. In the first distribution, we found that start values differed between switch conditions, $F(3, 74) = 5.591$, $MSE = .181$, $p < .002$, with higher start accuracy in the wide–narrow condition of the first distribution than in both the narrow–narrow and narrow–wide conditions (all ps $< .04$) and higher start accuracy in the wide–wide condition than in the narrow–wide condition ($p < .04$). However, we found no differences between conditions when examining their learning rates, $F(3, 74) = .302$, $MSE = .005$, $p = .824$, or estimated last trial accuracy, $F(3, 74) = 1.953$, $MSE = .021$, $p = .128$. This indicates that although the groups differed in their starting accuracy, participants in all switch conditions learned the first distribution they were exposed to with a similar level of accuracy by the end of the 100 trials (see Figure 2 and for mean performance parameters see Table 2).

When comparing participants' performance parameters for the second distribution, we found differences in switch condition starting values, $F(3, 74) = 6.308$, $MSE = .117$, $p < .001$, and estimated last trial accuracy, $F(3, 74) = 5.811$, $MSE = .098$, $p < .002$. However, as indicated by the lack of Trial Number × Condition interaction in the overall ANOVA, we found no differences between switch condition learning rates, $F(3, 74) = 1.451$, $MSE = .006$, $p = .235$.

As predicted, the wide–narrow switch group, which had the lowest computed surprise, had the lowest overall estimated last trial accuracy when compared to all other switch groups (all ps $< .01$) and was also the group with the lowest overall start value (all ps $< .007$). Additionally, as expected, participants in the mean shift conditions (wide–wide, narrow–narrow) did not differ on any fit parameters (all ps $> .50$). However, contrary to expectations, participants in the high surprise condition (narrow–wide), although performing better than did participants in the low surprise condition, did not perform any better on any parameters than did participants exposed to mean shifts (all ps $> .74$; see Table 2).

Table 2
Participant Learning and Updating Performance Parameters

	First distribution			Second distribution		
Switch type	SV	LR	LTA	SV	LR	LTA
Narrow–Narrow	.35	.07	.71	.50	.05	.63
Wide–Wide	.46	.08	.67	.51	.02	.63
Narrow–Wide	.30	.05	.65	.51	.04	.62
Wide–Narrow	.51	.07	.72	.35	.01	.48

Note. Narrow = Gaussian distribution of ball drops with a small standard deviation; Wide = Gaussian Distribution of ball drops with a larger standard deviation. SV = start value; LR = learning rate; LTA = last trial accuracy.

High Surprise Shifts Do Not Always Lead to Better Updating

To understand why participants in our high surprise condition did not show any clear updating advantages over participants in our medium-surprise conditions, we computed parametric relations between surprise and updating accuracy to see whether these differed between our surprise conditions.

We began by examining how participants' relative surprise factor influenced their updating accuracy. We calculated two surprise factors, one using negative log probability (NLP) and one using weighted empirical log odds (wElog). These individual surprise factors were computed using the distributions participants had drawn after observing all 100 trials of the first distribution as the numerator in each equation (i.e., distribution j in Equation 1) and the second computer distribution they would be exposed to on the next 100 trials as the denominator (i.e., distribution k in Equation 1). We then compared participants' surprise factor with their estimated last trial accuracy on the second distribution as an estimate of how accurately they managed to update.

As is evident in Figure 3, participants with both the lowest and highest levels of estimated surprise seemed to perform more poorly than did participants with midrange surprise values. Nonlinear regressions comparing our two measures of surprise and estimated last trial accuracy on the second distribution found significant quadratic relationships when using both NLP ($b^2 = -.0004$, $b = .0029$), $t(75) = -5.458$, $p < .001$, $R^2 = .27$, and wElog ($b^2 = -.0006$, $b = .0058$), $t(75) = -5.575$, $p < .001$, $R^2 = .48$ (see Figure 3). This quadratic trend fit last trial accuracy better than did models using switch *magnitude* as a predictor (see the online supplemental materials).

The quadratic trend seemed driven by a negative relationship between surprise and accuracy in the narrow–wide switch condition compared to all other conditions. To test this, we performed two separate linear regressions for participants in our low- and middle-surprise groups (wide–wide, narrow–narrow, and wide–narrow) and for participants in our high-surprise group (narrow–wide). For the low- and middle-surprise groups, we found that higher surprise values were related to better estimated last trial accuracy for both our NLP measure ($b = .003$), $t(57) = 3.941$, $p < .001$, $R^2 = .20$, and wElog measure ($b = -.012$), $t(57) = -6.940$, $p < .001$, $R^2 = .45$. In contrast, a linear regression comparing surprise and estimated last trial accuracy in the narrow–wide condition demonstrated the opposite relationship, with higher surprise predicting lower estimated last trial accuracy: NLP: ($b = -.002$), $t(17) = -5.456$, $p < .001$, $R^2 = .62$; wElog: ($b = .022$), $t(17) = 5.206$, $p < .001$, $R^2 = .59$ (see the lower panels of Figure 3).

To better understand what was driving these condition differences, we took a closer look at how participants in each condition integrated varying levels of surprising information. The nature of our high-surprise shift was to increase the variance of the second distribution relative to the first, exposing participants to balls falling in previously unused slots. We therefore expected participants to see a higher number of surprising events when switched to the second distribution of the high-surprise condition, largely driven by balls' falling in slots under

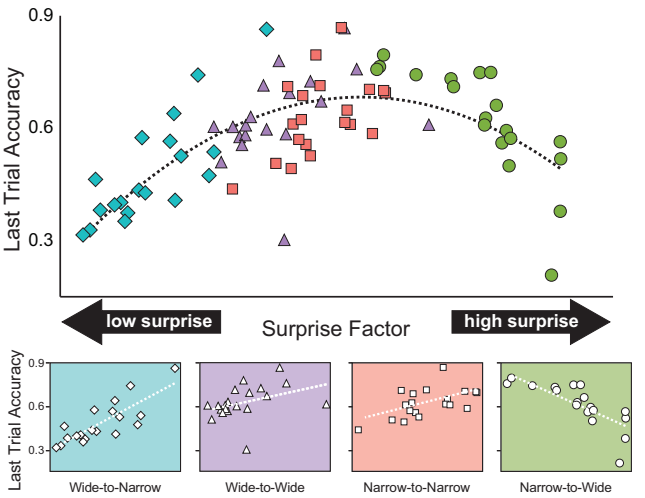


Figure 3. Surprise factor calculated using weighted empirical log odds and updating accuracy (i.e., estimated last trial accuracy on the second distribution of each switch condition) plotted for each participant. Updating accuracy was quadratically related to surprise factor, with both high and low surprise factor values predicting poor updating accuracy. This trend was due to a positive relationship between surprise factor and updating accuracy in the low- and medium-surprise conditions (wide-to-narrow, wide-to-wide, and narrow-to-narrow) and a negative relationship between surprise factor and updating accuracy in the highest surprise condition (narrow-to-wide). The dotted lines in the panels below the main figure represent the regression line for each individual surprise condition. See the online article for the color version of this figure.

which they had not drawn any bars (zero-probability slots) and would thus make larger and more frequent changes to their distributions. To examine this, we quantified the "change magnitude" on each trial as one minus the proportion of overlap between participants' distribution on the current trial and their distribution on the previous trial (with a value of zero indicating that participants had made no change to their distribution). To measure the frequency of changes, we also counted the number of instances in which participants made changes to their distributions (i.e., the number of instances where change magnitude greater than zero).

When learning the first distribution, participants' mean change magnitude did not differ between conditions (wide–narrow = .015, wide–wide = .031, narrow–narrow = .021, narrow–wide = .054), $F(3, 74) = .901$, $MSE = .006$, $p = .445$, and although participants first exposed to wide Gaussians made nominally more frequent changes to their distributions, the mean number of changes did not differ between conditions (wide–narrow = 41.6, wide–wide = 43.3, narrow–narrow = 28.5, narrow–wide = 27.8), $F(3, 74) = 2.15$, $MSE = 1{,}336.6$, $p = .101$.

When switched to a second distribution, as expected, the mean number of balls falling in zero-probability slots differed across conditions (wide–narrow = .42, wide–wide = 4.60, narrow–narrow = 2.35, narrow–wide = 20.32), $F(3, 74) = 10.85$, $MSE = 1{,}578.9$, $p < .001$, with participants in the narrow–wide condition experiencing more ball drops in zero-probability slots after a switch had occurred than did those in any of the other conditions (all ps $< .001$). However, although participants in our high-surprise condition experienced more surprising events after a computer switch, mean change magnitude did not significantly differ between conditions (wide–narrow = .006, wide–wide = .006, narrow–narrow = .009, narrow–wide = .053), $F(3, 74) = 1.513$, $MSE = .011$, $p = .218$, and participants in all conditions made the same average number of changes to their distribution (wide–narrow = 27.1, wide–wide = 28.6, narrow–narrow = 16.9, narrow–wide = 28.4), $F(3, 74) = .824$, $MSE = 626.1$, $p = .485$.

These results suggest that the negative correlation between surprise and updating observed in the narrow–wide condition could be due to participants' choosing not to integrate highly surprising events. If this were the case, one would expect that the *variance* of the last distributions drawn by participants should be narrower than the wide distribution presented to them after the switch. This was indeed the case. On average, participants' estimates had smaller standard deviations ($SD = 6.03$ slots) than the actual standard deviation of the discrete wide Gaussian distribution presented to them (computer $SD = 6.99$ slots), $t(18) = -2.10$, $SE = .46$, $p < .05$.

We wanted to see whether, following from this observation, this tendency to devalue surprising events generalized across all participants in our experiment—in other words, do participants in all four conditions, not just the narrow–wide condition, tend to discount surprising events? To do this, we first calculated the mode of participants' distributions on each trial by identifying the slot with the highest assigned probability value. On trials where participants had multiple bars with the same highest probability value, we used the mean position of these bars as a proxy for the mode. For each trial, we then calculated the absolute difference between participants' mode and the location of a ball drop on the same trial. Next, we calculated the proportion of changes made to their estimates of the underlying distribution when a ball fell at different distances from their mode. As is evident from Figure 4, participants made the fewest changes to their distributions when balls fell either near or far away from the mode of their distribution. We expected fewer changes at the mode, given that these events represent confirmatory evidence. We suggest that the few changes made when balls fell far from the mode represent discounting of outliers.

When we examined participants' postquestionnaire responses, we did not find any performance differences between participants who reported detecting a change to the ball distributions compared to those who had not (see the online supplemental materials). However, approximately one quarter of our participants reported that they made adjustments based on observations from the last few trials (in some cases, using the word *outlier* to refer to unexpected ball drops that they chose not to integrate in their estimates). To test this possibility, we used a mixed-effects logistic regression to measure the influence of factors that

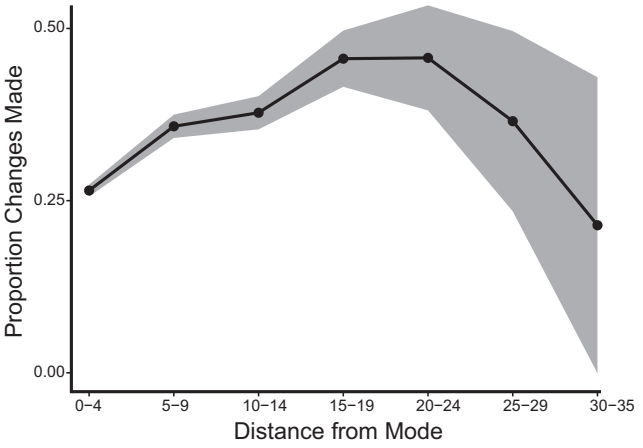


Figure 4. Proportion of changes participants made to their distributions in response to the absolute distance a ball fell away from their distribution mode. On each trial, participants' mode was identified as the slot with the highest estimated probability. The absolute distance was calculated as the distance a ball fell from the participants' mode on trial *t*, whereas the proportion of change was calculated using the changes participants made to their distributions on trial $t + 1$. The black dots represent mean proportion values, whereas shading represents ±95% confidence intervals.

contributed to likelihood that participants made changes to their distributions. The factors we included were participants' current trial and the surprise that participants had experienced on prior trials. We found that the surprise value from the immediately preceding trial ($n - 1$) had the greatest influence on the likelihood that participants would make an adjustment to their distribution. However, similar to participants' reports, trials as far as three trials back ($n - 3$) made additional, independent statistically significant contributions (see Table 3). When we performed the same logistic regression on each separate surprise condition, we found that this effect was primarily present in participants in our high-surprise condition, a condition where surprising events were more commonly observed, whereas participants in the low- and medium-surprise conditions relied mostly on the surprise from trial $n - 1$ (see Table 3).

Discussion

The goal of this study was to explore how the surprise of an environmental shift influences mental model updating. Our first result suggests that switches signaled by the *absence* of previously observed events (low surprise), although similar in magnitude, were learned less accurately than were shifts signaled primarily by the *presence* of new events. Additionally, we found situations in which, consistent with prior re-

Table 3
Influence of Surprise from Previous Trials on the Likelihood of Participant Updating on the Current Trial (n)

Predictor	All conditions	Wide–Narrow	Wide–Wide	Narrow–Narrow	Narrow–Wide
Trial	$-.009^{***}$	$-.010^{***}$	$-.009^{***}$	$-.010^{***}$	$-.007^{***}$
Surprise ($n - 1$)	$.067^{***}$	$.055^{***}$	$.045^{**}$	$.082^{***}$	$.074^{***}$
Surprise ($n - 2$)	$.007^{*}$	$.017^{*}$	.006	−.016	$.010^{*}$
Surprise ($n - 3$)	$.009^{**}$	.005	.003	.006	$.016^{***}$

Note. The values in the cells represent beta weights for each predictor. Narrow = Gaussian distribution of ball drops with a small standard deviation; Wide = Gaussian Distribution of ball drops with a larger standard deviation.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

search, updating was positively correlated with surprise: Our overall trial-by-trial analysis found that surprise predicted the likelihood that participants made changes to their distributions, and we also found that surprise was positively correlated with updating in our low- and medium-surprise conditions. Taken together, these results support the results of previous research that suggested that, under certain circumstances, surprising observations lead to more efficient updating (McGuire et al., 2014; Nassar et al., 2010).

However, we found, in addition to these positive correlations, the opposite trend in our high-surprise condition, where higher levels of surprise predicted poorer updating performance (see Figure 3). These results demonstrate that although surprise can play an important role in the updating process, highly surprising events do not always predict better updating.

We propose that the negative correlation we found between our high-surprise condition and updating may stem from a form of outlier devaluation, in which participants chose not to integrate highly surprising events into their estimates. As we outlined in the introduction, this idea is supported by studies demonstrating that participants tend to discount highly discrepant exemplars when categorizing events (De Gardelle & Summerfield, 2011; Summerfield & Tsetsos, 2015; Wei & Stocker, 2015). Additionally, participants in the highest surprise condition seemed to take more previous events into consideration when making changes to their distributions, rather than rely solely on the last trial seen. When participants were asked on the postexperimental questionnaire about the strategies they used to update their estimates, approximately one quarter of them reported waiting for the same event to occur a number of times in quick succession before committing to a change (some using the word *outlier* to describe unexpected events that they chose not to integrate).

This last finding is particularly important, because outlier devaluation is not a part of any current model of dynamic belief updating. Current models have suggested that *any* surprising event should increase individuals' propensity to change their beliefs (McGuire et al., 2014; Nassar et al., 2010; O'Reilly et al., 2013). These models were built to fit tightly controlled experimental environments, where changes are expected, they are very similar in their nature (e.g., consist primarily of mean shifts), and participants are encouraged to make changes to their beliefs on a trial-by-trial basis. One can see that when some of these constraints are removed, surprising information can be treated differently from what is predicted by these highly controlled environments.

Even though participants in our experiment were not made explicitly aware that changes would occur, we still saw some parallels with previous research. In our mean shift conditions, which most closely match prior work, surprise was positively related to updating. However, this was not the case in one of our variance conditions, suggesting that the *type* of change participants observe can have implications in the way they treat surprising information. We are not suggesting that *all* surprising information is devalued but merely that the type of change needs to be considered when attempting to measure the influence of surprise on updating.

One explanation for our results is that we focused on only the features we believed we needed, rather than encoding all events equally. Proponents of the "efficient coding hypothesis" have suggested that one weight perceptual events in proportion to the probability of their occurrence (Barlow, 1961; Wei & Stocker, 2015). In essence, we focused primarily on *modal* elements of a distribution. This hypothesis helps integrate our findings with the results of previous research. Participants exposed to changes signaled by mean shifts tended to update quickly, depending on the level of surprise prompted by the shift. However, in our variance shift conditions, although the *magnitude* of the distribution's change was similar to that of the mean shifts, the *modal* elements of each distribution remained the same. As a result, participants in these conditions either were worse overall or tended not to give as much weight to highly unexpected events. The efficient coding hypothesis provides a plausible explanation for our result and could potentially apply to the way in which individuals use information to inform their mental models (Summerfield & Tsetsos, 2015).

Taken together, our results provide new insights into the ways in which mental models influence one's ability to learn and integrate information from the environment. It is our hope that with a better understanding of the factors that

influence updating, we will be able help build more precise models of belief updating that will be applicable to a wider range of scenarios.

References

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. New York, NY: Holt.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology, 10*(3), e1001293. http://dx.doi.org/10.1371/journal.pbio.1001293

Cox, D., & Snell, E. (Eds.). (1989). *Analysis of binary data* (2nd ed.). New York, NY: Chapman and Hall.

Danckert, J., Stöttinger, E., Quehl, N., & Anderson, B. (2012). Right hemisphere brain damage impairs strategy updating. *Cerebral Cortex, 22,* 2745–2760. http://dx.doi.org/10.1093/cercor/bhr351

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

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *The Bayesian Brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Doyle, A. C. (1930). *The Complete Sherlock Holmes*. New York: Doubleday & Company, Inc. (Original work published 1892)

Estes, W. K. (1950). Towards a statistical theory of learning. *Psychological Review, 57,* 94–107. http://dx.doi.org/10.1037/h0058559

Filipowicz, A., Valadao, D., Anderson, B., & Danckert, J. (2014). Measuring the influence of prior beliefs on probabilistic estimations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 2198–2203). Austin, TX; Cognitive Science Society.

Fisk, J. E. (2002). Judgments under uncertainty: Representativeness or potential surprise? *British Journal of Psychology, 93,* 431–449. http://dx.doi.org/10.1348/000712602761381330

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7,* 185–207. http://dx.doi.org/10.3758/BF03212979

Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 179–212). Hove, Sussex, United Kingdom: Psychology Press.

Lee, N. Y. L., & Johnson-Laird, P. N. (2012). Strategic changes in problem solving. *Journal of Cognitive Psychology, 25,* 1–9. http://dx.doi.org/10.1080/20445911.2012.719021

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience, 28,* 12539–12545. http://dx.doi.org/10.1523/JNEUROSCI.2925-08.2008

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron, 84,* 870–881. http://dx.doi.org/10.1016/j.neuron.2014.10.013

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30,* 12366–12378. http://dx.doi.org/10.1523/JNEUROSCI.0822-10.2010

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 110*(38), E3660–E3669. http://dx.doi.org/10.1073/pnas.1305373110

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2,* 10. http://dx.doi.org/10.3389/neuro.11.010.2008

R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Available at http://www.R-project.org/

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioural sciences* (pp. 8602–8605). Amsterdam, the Netherlands: Pergamon Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423.

Stöttinger, E., Filipowicz, A., Danckert, J., & Anderson, B. (2014). The effects of prior learned strategies on updating an opponent's strategy in the rock, paper, scissors game. *Cognitive Science, 38,* 1482–1492. http://dx.doi.org/10.1111/cogs.12115

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2005). Information theory, novelty and hippocampal responses: Unpredicted or

unpredictable? *Neural Networks, 18,* 225–230. http://dx.doi.org/10.1016/j.neunet.2004.12.004

Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences, 19,* 27–34. http://dx.doi.org/10.1016/j.tics.2014.11.005

Teigen, K. H., & Keren, G. (2003). Surprises: low probabilities or high contrasts? *Cognition, 87,* 55–71. http://dx.doi.org/10.1016/S0

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011, March 11). How to grow a mind: Statistics, structure, and abstraction. *Science, 331,* 1279–1285. http://dx.doi.org/10.1126/science.1192788

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature neuroscience, 18,* 1509–1517.

Received March 21, 2016
Revision received October 19, 2016
Accepted November 16, 2016 ■

Supplemental Materials

Rejecting Outliers: Surprising Changes Do Not Always Improve Belief Updating

by A. Filipowicz et al., 2016, *Decision*

http://dx.doi.org/10.1037/dec0000073

TASK AND ANALYSIS SUPPLEMENTARY MATERIALS

Participant information

A total of 79 undergraduates participated in the current experiment. This number reflects the number of participants we were able to recruit over the course of two school semesters at the University of Waterloo. Data from one participant were excluded due to a computer failure, leaving a final total of 78 participants.

Task environment

We measured participant mental models using a computerized version of the game 'Plinko'[1]. The task environment was programmed in Python using the PsychoPy library (Pierce, 2009)[2]. During the task, a red ball fell through a pyramid of pegs into one of 40 possible slots. The pyramid consisted of 29 rows of black pegs that increased in number from the top to the bottom of the pyramid (i.e., the top row contained 1 peg and the bottom row contained 29 pegs). A rectangle was located below the 40 slots spanning their width. Participants were instructed to make their responses in this space (Figure 1a in main text).

Bars were drawn using the computer mouse: the height of the bars could be adjusted by holding down the left mouse button and dragging the cursor. The height of

1 Based on the game found on the American game show *The Price is Right* (http://www.thepriceisright.com).

2 A demo and code for the task can be obtained at https://osf.io/dwkie/

the bar would match the position of the cursor within the limits of the rectangle below the slots. Participants could also erase any single bar by right clicking with the cursor on the bar they wished to delete, or by clicking the backspace key to delete all bars on screen. The bars were not assigned any value; participants were simply told that taller bars represented a higher probability that a ball would fall in a slot, shorter bars a lower probability, and no bars represented zero probability. Participants were informed that they had the option of adjusting their bars at the start of every trial and that they had to have at least one bar on screen before proceeding with the trial. Once participants had indicated their likelihood estimates, they pressed the spacebar to proceed with the trial (Figure 1a).

After completing the task, participants completed a brief questionnaire, asking if they had noticed any structure to the locations of the ball drops, if they had noticed that the structure had changed at any point during the task, and to elaborate on any strategies they used to adjust their bars.

Experimental conditions

Sequences of ball drops were generated by randomly selecting integers between 1 and 40 from pre-specified probability density functions. The resulting sequences determined the slots in which the ball fell on each trial, with slot 1 representing the slot farthest to the left and slot 40 representing the slot farthest to the right.

During the experiment, participants were exposed to two distributions: a first distributions for 100 trials, then switched to a second distribution of 100 trials without any cues or other information to suggest that a switch had occurred. Participants were not informed that any switches would occur; they were simply instructed to have their bars

represent where they believed the ball would fall on future trials. Using this paradigm, participants were assigned to one of four switch conditions.

Participants began the task by being exposed to either a wide Gaussian (mean=17, SD=6), or a narrow Gaussian distribution (mean=17, SD=1.9). After 100 trials of this first distribution, participants were then switched either to a wide distribution (mean=25, SD=6), or a narrow Gaussian (mean=19.5, SD=1.9). This resulted in four between subject conditions: two mean shift conditions (narrow-narrow, and wide-wide), and two variance shift conditions (narrow-wide, and wide-narrow).

In all conditions, we kept the *magnitude* of each shift condition equivalent, while manipulating the amount of surprise expected from each shift. We define shift magnitude as the percent overlap between the first and second distribution of ball drops. We computed overlap using the same formula we used to compute participant accuracy (see equation S1 below). The overlap between the continuous distributions in each of these switches was nearly identical, with roughly 50% overlap between the first and second distributions. The discrete distributions we generated from these continuous distributions were also nearly identical in their overlap, with overlap between the first and second distributions ranging between 44% and 46%.

Measuring accuracy

Participant bars were normalized by dividing the height of each bar by the sum of the heights of all bars drawn on screen. This normalization process provided a discrete probability distribution on each trial for every participant.

Accuracy A was measured as the proportion of overlap between a participant's drawn distribution on any trial and the discrete distribution of ball drops they were being

presented with. The proportion of overlap was calculated by summing the minimum probability value x of every slot i between a participant's distribution P on any trial t and a computer's distribution C for any block of trials j:

$$A_t = \sum_{i=1}^{40} \min\left(P(x_{it}), C(x_{ij})\right) \qquad \text{(S1)}$$

Participant accuracy could range between 0 and 1, with 0 indicating no overlap between the participant's distribution and the computer's distribution for a particular block, and 1 indicating perfect overlap.[3]

Learning rates

Once participant accuracy was calculated on each trial, we fit a standard exponential learning curve to participant accuracy scores over time (e.g., Estes, 1950; Healthcote, Brown, & Mewhort, 2000; Ritter & Schooler, 2001):

$$\hat{A}_t = a_\infty - (a_\infty - a_0)e^{-\alpha t} \qquad \text{(S2)}$$

where t denotes the trial number, $\hat{A}_t$ a participant's estimated accuracy on trial t, a_0 a participant's starting accuracy, a_∞ asymptotic accuracy, and α a constant rate coefficient to capture how quickly participants reached their asymptote from their starting accuracy. We fit this function to each participant's accuracy scores using a nonlinear least squares

3 A potential alternative for measuring participant accuracy would be to measure the Kullback-Leibler divergence (D_{KL}) between participant and computer distributions (e.g., O'Reilly et al., 2013). A re-analysis of the participant performance trends using D_{KL} as a measure for accuracy revealed very similar results to those obtained using our overlap measure of accuracy. Given the similar trends reported by each measure, and that the conditions in our second experiment were designed to control for overlap rather than D_{KL}, we chose to use overlap as our metric of accuracy.

function in the R statistical package ('nls' function; R Core Team, 2014). Given that participant accuracy could only range between 0 and 1, we set the function's lower and upper limits for a participant's minimum starting accuracy and maximum asymptote value to 0 and 1 respectively. To characterize participant performance over time, we used the estimated starting accuracy to represent a participant's starting accuracy value, learning rate to capture how quickly they reached their asymptote from their starting value, and a participant's estimated accuracy on the last trial of the distribution they were estimating to indicate a final level of accuracy achieved.

We obtained the estimated last trial accuracy by computing a participant's accuracy using equation (S2), with the participant's 3 fit parameters (asymptote, starting accuracy, and learning rate), and the value for t set to the total number of trials in the distribution they were estimating. Since participants were exposed to each distribution for 100 trials, the value of t was always set to 100. We used this last parameter instead of the asymptote value because the asymptote reflects a participant's maximal estimated accuracy after an unspecified number of trials, not necessarily the estimated accuracy at the end of a finite block of trials. For example, a participant with a fitted starting value of .5, an asymptote of .95, and a learning rate of .01 would only reach asymptotic performance after more than 1000 trials, whereas their performance after 100 trials, the length of each of the distributions participants were asked to estimate in our experiments, would be a closer to .78.

Overall, the learning curves fit participant accuracy values quite well, with median R^2 values across all participants of .83. Importantly, curve fits did not differ

between conditions ($F(3,150) = .668$, $MSE = .077$, $p = .573$), indicating that our characterization of participant accuracy trends was equally accurate across all conditions.

Computing surprise

We based our measures of surprise on previous research, which characterizes the 'surprise' of an observation as the negative log probability of its occurrence. In our task, given that participants were not required to draw bars under all possible slots on any given trial, participant bar values could range between 1 (if they only had one bar on screen) and 0 (if they had no bars in a given slot). Given that the natural log of 0 is undefined, it becomes impossible to compute a pure measure of negative log probability for any events that land in slots with no probability. To circumvent this issue, we used and compared two measures of surprise meant specifically to account for the discrete nature of participant distributions.

Our first measure used a modification of participant slot probability values to make them compatible with a measure of negative log probability. We did this by replacing slot values equal to 0 with an arbitrarily small number $1x10^{-10}$, which approaches 0, but still has a natural logarithm.

As a second measure, we used *weighted empirical log odds* to compute surprise. The advantage of this measure is that it can be applied directly to discrete distributions, and does not require any modifications to participant responses. Although not entirely identical, odds and probabilities both provide information about likelihoods, and make predictions about how expected certain events are to occur.

To compute the *weighted* empirical log odds of a given probability value we first need to compute the empirical log odds (Elog). The Elog E of any probability p is defined as:

$$E_p = \ln\left(\frac{p+0.5}{1-p+0.5}\right) \qquad \text{(S3)}$$

If we know the frequency r of a specific event compared to the total number of events n that were used to compute p, E can also be defined as:

$$E_r = \ln\left(\frac{r+0.5}{n-r+0.5}\right) \qquad \text{(S4)}$$

Elog can be weighted depending on the number of observations n that are used to compute p. The variance V of Elog is defined as:

$$V = \frac{(n+1)(n+2)}{n(r+1)(n-r+1)} \qquad \text{(S5)}$$

of which the inverse can be used to compute a weight W:

$$W = \frac{1}{V} \qquad \text{(S6)}$$

The weighted empirical log odds $\hat{E}$ for any probability value in slot i can be computed as

$$\hat{E}_i = E_i \times W_i \qquad \text{(S7)}$$

Given that each distribution consisted of 100 ball drops, we calculated the weights for our wElog calculation using variance as expressed in equation (S5) with n=100 for each distribution.

Comparing predictors of last trial accuracy

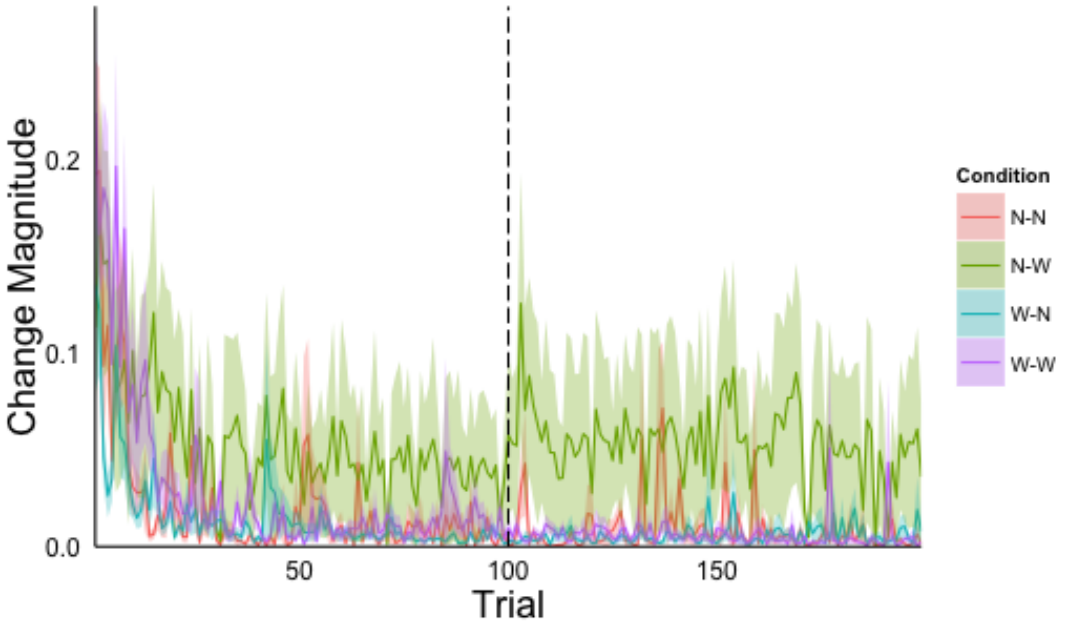
An alternative predictor of last trial accuracy could be that participants are updating in response to the magnitude of overlap between the distribution they drew on the last trial before the switch, and the computer distribution they would observe in the second half of the experiment. To test this, we compared our quadratic model using surprise factor as a predictor, with the best fitting model using switch magnitude as a predictor. Although magnitude of overlap was linearly related to last trial accuracy (b = -.64; $t(76)$ = -6.449, SE = .098, $p < .001$, R^2 = .35), our quadratic surprise factor model provided a significantly better fit ($F(1,75) = 20.223$, $p < .001$).

Questionnaire responses

A total of 71 participants responded to the post-experimental questionnaire asking if they had noticed a change during the experiment – although a nominally smaller proportion of participants first exposed to a wide Gaussian reported noticing that a change had occurred, a chi-square test of independence did not reveal any significant differences between the proportion of participants in each condition who detected whether or not a switch had occurred (proportion detected: wide-narrow = .53, wide-wide = .53, narrow-narrow = .85, narrow-wide = .76; $\chi^2(3, N = 71) = 6.714$, $p = .082$).

Additionally, a factorial ANOVA measuring the influence of Condition and Change Detection (detected, not detected) on updating accuracy found no main effect of Change Detection ($F(1,63) = .064$, $MSE = .001$, $p = .801$) and no Condition by Change Detection interaction ($F(3,63) = 1.094$, $MSE = .014$, $p = .358$). Additionally, the average surprise factor did not differ significantly based on whether or not participants had reported detecting a change (Mean wElog surprise factor: Detected = 48.62*, Not Detected = 53.72*; $t(38.34)=-1.729$, $SE = 2.802$, $p = .092$; *lower values indicate higher surprise).

Supplementary figures



Supplementary figure captions

Supplementary Figure 1. Mean trial-by-trial change magnitude between experimental conditions. One participant in the narrow-wide condition (P52) made large and frequent changes to their responses throughout the experiment, contributing to the seemingly high magnitude of change means displayed for the narrow-wide condition.

Additional References

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207. doi:10.3758/BF03212979

Peirce, J. W. (2009). Generating Stimuli for Neuroscience Using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. doi:10.3389/neuro.11.010.2008

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioural Sciences*, (pp. 8602–8605). Amsterdam: Pergamon.