Regime Change Detection –
Task factors as Drivers of Judgmental Effectiveness

Florian M. Federspiel & **Matthias Seifert**
Operations & Technology Area, IE Business School, Madrid, Spain

Lee Newman
School of Social & Behavioral Sciences, IE University & IE Business School, Madrid, Spain

Abstract:
We examine the decision maker's external environment as a predictor of her judgmental effectiveness in detecting regime changes. Relying on two experimental studies, our main task employed is one of regime-change detection over time. Our first study looks at the effect of different task factors (signal diagnosticity, transition probability, and signal length) on judgmental accuracy, particularly with regards to over- and under-reaction. We find evidence of systematic over- and under-reaction, further supporting what is known as the system-neglect hypothesis (Massey & Wu 2005). We also establish that signal length is a significant component in a regime change detection task and show that increasing amounts of information are related to increasing conservatism. Yet what has not received adequate attention in studies of regime change detection is the effect of the strength or extremeness of evidence. We hypothesize that signals that are highly representative of a change having occurred (a perception we anchor on streaks) are likely to lead to systematic over-reaction. Our second study examines the effect of signal streaks on judgmental effectiveness. We report a significant three-way interaction suggesting that streaks tend to govern over-reaction in regime change detection tasks.

Key words: judgmental forecasting, regime change; system neglect; streaks; under-reaction; over-reaction

Acknowledgements

This research project was generously supported by a grant from SAS and the International Institute of Forecasters.

1. Introduction

Detecting a fundamental change in regime is an important aspect of every day decision-making. Practical examples include a wide range of tasks from detecting the onset of a recession, to a structural change in demand, to a change in mood of one's partner. Each of those tasks essentially involves dealing with imperfect signals and incorporating those to make an inference about the true state of being, i.e. the current underlying regime or state. When detecting the onset of a recession, possible signals may relate to news from the housing market, employment data, and other indicators whereas the stability of the environment could be inferred upon based on past data about recessions (e.g. the general rate of occurrence, cycles). Naturally, given the uncertainties and complexity involved, most reactions to such signals will fall into the categories of under-reaction or over-reaction. Under-reaction may for example lead a decision maker to miss out on an important change in the market whereas over-reaction may lead to false alarms and premature action, e.g. regarding a market entrance. Rational as well as irrational reasoning may cause both over- and under-reaction. Aspects such as the possible cost and benefits of over- and under-reaction should certainly factor into a rational decision and make a universal judgment about the value or threat of over- and under-reaction impossible. Yet cognitive processes that affect how individuals, perhaps unwittingly, respond to certain features in change detection settings in ways that differ from normative prescriptions are potentially very harmful. Resulting errors in (biased) judgments may indeed harbor unfavorable and unintended consequences.

The study of regime change detection has for long involved varying environmental parameters in order to investigate why and how well individuals respond to changing conditions (e.g. see Chinnis and Peterson 1968, 1970). Recently, Massey and Wu (2005) have revived the stream on regime change detection and successfully unified previous inconsistent findings through their framework of system-neglect. The system-neglect hypothesis posits that subjects react primarily to signals they observe and somewhat disregard the underlying system that generates those signals, leading to relative over-reaction in stable systems and relative under-reaction in unstable systems. This framework manages to encompass previous findings and was systematically tested by Massey and Wu (2005) in a series of studies. Kremer et al. (2010) found further support for the stability of this pattern in time series forecasting tasks. In such task contexts, subjects tended to over-react to forecast errors in stable environments and under-react to forecast errors in less stable environments, reaffirming that subjects paid too little attention to the underlying environment of their task. Massey and Wu's (2005) system neglect hypothesis regarding over- and under-reaction in a non-stationary regime change detection setting is derived in large parts from earlier work by Griffin and Tversky (1992). In their studies on over- and under-confidence in stationary forecasting settings, Griffin and Tversky (1992) aimed at reconciling previous findings of pronounced conservatism (under-confidence) with the common over-confidence phenomenon. The occurrence of over-confidence and over-reaction in judgmental behavior had frequently been attributed to the representativeness heuristic (Kahneman and Tversky 1973a, 1973b). Subjects have been shown to predict outcomes of which the experienced signal is most representative. Yet representative signals do neither account for the potential level of noise in the signal nor for the underlying likelihood (i.e. do not relate to a relevant base rate of or take into account the sample size that was used for inference), thereby relying on a signal's representativeness may easily lead to over-reaction. On the other hand, the phenomenon of under-reaction has often been termed conservatism (Tversky and Kahneman 1974). Subjects have been shown to behave conservatively in evaluating the weight of information (e.g. sample size). Conservatism may thus lead to under-reaction. Under-reaction has also been attributed to potential problems regarding the proper aggregation of evidence with increasing information and the misperception of the impact of pieces of information (Rappoport and Summers 1973).

Griffin and Tversky (1992) reconciled these findings in a common framework, arguing that individuals focus primarily on aspects that are related to the strength or extremeness of evidence (as related to the representativeness heuristic) and generally much neglect the weight or credence of evidence. Whereas the weight of evidence can be thought of as an underlying base rate or sample size, the strength of evidence relates to the extremeness of a certain piece of evidence. Supported by a number of studies they showed that when weight is high and strength is low individuals suffer from under-confidence. To the contrary, in settings where strength is high and weight is low individuals display over-confidence (see Figure 1).

	Low weight	High weight
Low strength		*Under-confidence*
High strength	*Over-confidence*	

Figure 1

The notion of the interplay between the strength and weight of evidence has strongly influenced the inception of Massey and Wu's (2005) system-neglect hypothesis and the interpretation of its tests rests on its assumptions. Yet in experimentally studying the system neglect hypothesis, existing research by Massey and Wu (2005) has operationally focused on subjects' apparent disregard of the weight of evidence concept, which was varied through changing the characteristics of the signal generating system. Massey and Wu (2005) have varied two environmental parameters, namely diagnosticity and transition probability. Diagnosticity relates to the amount of information (indicative of a given regime) inherent in a signal (versus noise) whereas the parameter transition probability serves as a gauge of stability of the current regime. In their (2005) studies, subjects had for example to judge the probability of a switch from one regime, an urn filled with a certain distribution of red and blue balls, to a second (absorptive) regime, an urn with a symmetric (opposite) distribution of red and blue balls compared to the first urn. The diagnosticity parameter was operationalized as the relative proportion of red versus blue balls in the urns' distributions. A higher diagnosticity is given when the distributions approach extreme values (e.g. almost only red balls in the first urn and therefore almost only blue balls in the second urn). An alternative way to think of the diagnosticity parameter is as the strength of correlation between a signal and the urn it originates from.

When it comes to judging the probability of a switch from a first regime to a second, Massey and Wu (2005) argued that in settings where signal diagnosticity is high (i.e. precise signals) and the transition probability from the first to the second regime is high (i.e. an unstable first regime) the weight of evidence for a switch is the highest. Therefore, relative under-reaction when compared to all other settings should be expected. (The system-neglect hypothesis is silent about absolute levels of over- and under-reaction.) On the other hand, when signal diagnosticity is low and the transition probability is low, the weight of evidence pertaining to a switch is the lowest and thus most relative over-reaction should be seen in such a setting. Figure 2 shows this prediction graphically.

	Low transition probability	High transition probability
Low diagnosticity	*Over-reaction* *(lowest weight of evidence)*	
High diagnosticity		*Under-reaction* *(highest weight of evidence)*

Figure 2

However, what is missing from the current framework is the fact that decision makers are usually presented with several pieces of information at the same time when making a judgment. Previous research in the field of regime change detection has not investigated the impact of varying amounts of information at hand (Chinnis and Peterson 1968, 1970; Massey and Wu 2005; Kremer et al. 2010). Most recent studies (Massey and Wu 2005; Li et al. 2009; Kremer et al. 2010) focused only on single point sequences and others (Chinnis and Peterson 1968, 1970) held the amount of information constant throughout their studies. Varying the amount of information per signal introduces an important factor that much increases the ecological validity of the task. It further builds on previous research that has repeatedly linked increasing amounts of information (e.g. sample size) with increasing conservatism in likelihood judgments (Peterson et al. 1965, Slovic and Lichtenstein 1970). The amount of information at hand has equally been related to the concept of diagnosticity (e.g. Peterson et al. 1965; Slovic and Lichtenstein 1970, p. 71) as well as the weight of evidence (Griffin and Tversky 1992). Then, when making a judgment about whether a change from one regime to another has occurred, using signals for inference, signal length can be regarded as an analogue to the concept of sample size as discussed in previous research and thus the amount of information at hand. A signal that includes more information (in quantity) carries a higher weight of evidence and represents a higher level of diagnosticity to the decision maker. Therefore, with increasing amounts of information at hand when judging the likelihood of regime changes, relatively more under-reaction should be witnessed. Yet both whether this is the case and if so how strong the effect of varying amounts of information in a change detection setting is remain to be tested. Our first study sets out to investigate the role of varying amounts of information by varying signal length and its effect on under- and over-reaction in a regime change detection setting.

Furthermore, what previous research on the system-neglect hypothesis (Massey & Wu 2005, Li et al. 2009) has merely assumed as the weight of evidence's counterpart in determining behavior leading to over- or under-reaction is the effect of the strength or extremeness of evidence and its effect in a regime change detection setting. The concept of the strength of evidence has not been explicitly measured. In a regime change detection setting the strength of evidence may relate to the extremeness of a certain signal. Signals that are highly representative of a change having occurred are likely to lead to systematic over-reaction. An interaction between the effect of the weight and the strength of evidence further seems very likely and ought to be investigated for a better understanding of systematic patterns of over- and under-reaction in a non-stationary setting.

Massey and Wu (2005) have treated a diagnosticity parameter as part of a gauge for the weight of evidence. Yet we argue that the actual, realized (as opposed to probabilistic diagnosticity parameter that Massey and Wu (2005) used in creating sequences) distribution of signals could however be thought of as a gauge for the strength of evidence. One facet of the actual signal distribution that has previously received much attention outside of the regime change setting is the effect of streaks in data (Gillovich et al. 1985). Perceived streaks in perfectly random data relate to people's misperceptions of chance and yet point to streaks as being seen as representative of the underlying generating mechanism. For example, in one of Gillovich et al.'s (1985) study, individuals causally interpreted streaks in sports as "hot hands". A number of recent studies have implicitly investigated perceptions of randomness and the effect and perception of streaks. Reimers and Harvey (2011) showed once more that individuals seem to perceive positive autocorrelation in factually uncorrelated time series and Kremer et al. (2010) found support for a belief in illusionary trends akin to both the gambler's fallacy and the hot-hand effect in a time series forecasting task. Streaks may very well be perceived to be representative of a certain system and as such provoke a stronger reaction leading to over-reaction. It is then plausible to assume that streak perception plays a significant role in regime change detection since it is representative of the perceived strength of evidence concept which previous research has assumed to be the deciding behavioral influencer. Our second study sets out to directly test the effect of streak perception and its interplay with the other environmental background parameters such as transition probability and signal diagnosticity.

2. General experimental design

2.1 Experimental task

Our experimental task was based on Massey and Wu (2005) and required subjects to visually inspect 10 consecutive signals on a computer screen as well as to judge whether the signal originated from one of two possible "regimes". In particular, for each signal, subjects were asked to state their probability belief that the signal indicated a switch from one regime to another regime. The two regimes were represented by two urns filled with symmetric distributions of either more red or green balls (e.g. 60% red & 40% green in urn A, versus 40% red & 60% green in urn B) and for ease of interpretation these were labeled the red and green urn, respectively. For each game, we used a fixed probability (e.g. 5%) of a switch from the red to the green urn. Each signal consisted of several balls drawn from either the red or the green urn. The sequences always started in the red urn and were absorptive in case of a change (i.e. once a switch to the green urn had occurred, a switch back to the red urn was impossible). The absorptive characteristic of the green urn somewhat simplified the task which otherwise would likely become too cognitively challenging for subjects. An absorptive system is arguably also a realistic assumption for environments that generally have low likelihoods (e.g. 5%) of change from one state to another. If this restriction had not been imposed, the vast majority of randomly generated sequences would indeed not have exhibited more than one switch over 10 signals. The distributions and switch probabilities always remained constant throughout a given sequence. The two urns were always labeled and all the information (i.e. respective urn distributions, the probability of a switch with each new signal, and the full history of previous signals) was always shown to subjects during the task (see Figure 3 for an example screen shot). Each sequence to be judged was represented by a unique draw of balls from one of the urns and our experimental conditions related to variations in either the likelihood of a switch or the distribution of red and green balls in each of the two urns, or both. Consistent with previous studies, we referred to the ratio of red versus green ball as the diagnosticity of the two urns whereas we defined the likelihood of switching between the urns as the transition probability parameter.

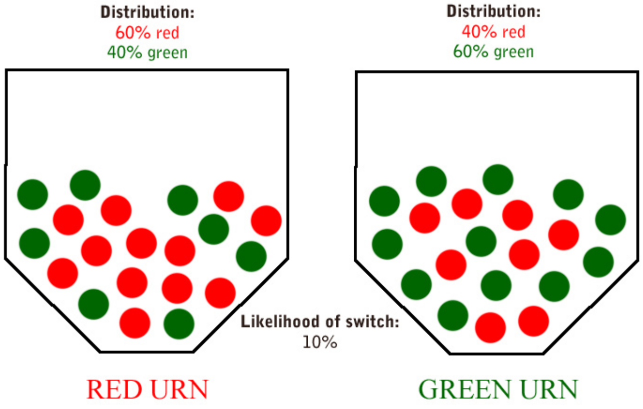


Figure 3

In order to generate a signal, we sampled a sequence of red or green balls from one of the urns. Each signal can be understood as the result of a Bernoulli process with uneven success probability reflecting the ratio of the two types of balls in each urn. As balls were drawn independently and replaced after each trial, the order of red and green balls in our signals should not have had any effect in judging its probability of having come from one regime or the other. Given this setup, we informed participants that all parameters would remain constant throughout a sequence. Our research design allowed us to analyze if subjects' judgments may have been influenced by their subjective perception of streaks among the presented signals. Specifically, we could achieve this by calculating the probability of a specific signal having come from one regime or the other, pretending that order mattered. A more detailed description of the operationalization of this parameter is described in part 4.2 below. As noted above, a high likelihood for a particular streak to have come from one regime was likely seen as representative of that regime, and thereby hypothesized to lead to over-reaction.

The task was computer-based, conducted via Microsoft Excel and created using Virtual Basic for Applications (VBA).

2.2 Payment formula

Subjects were paid based on their judgmental accuracy according to the same following symmetrical quadratic scoring formula as used by Massey & Wu (2005).

$$Payment\ per\ judgment = x - (2x * error^2)$$

Participants were paid a maximum of *x* per judgment and a minimum of -*x* depending on their accuracy in indicating whether the system had switched or not. The error term was calculated by comparing the accuracy of the probability judgment to the actual state of the system. For example, if a subject stated the probability of the current signal to come from the green urn (i.e. a switch is assumed to have occurred) to be 0.6 and indeed a switch had occurred, his error is 0.4 (= 1 – 0.6). Assuming for the same case that a switch had not occurred, the subject's error had been 0.6 (= 0.6 – 0). *X* varied between €0.05 in experiment 1 and $0.1125 in experiment 2. In experiment 2 subjects were additionally paid a flat fee of $1 regardless of their performance.

2.3 Normative model

Based on the information that could be extracted from the signals, Bayesian estimates could be computed that served as a normative benchmark to subjects' judgments. We were particularly interested in examining deviations from this Bayesian benchmark, which would indicate the degree of relative over- and under-reaction in subjects' probability estimates.

Let δ denote the diagnosticity level that stands for a given distribution of red and green balls in either the red (δ_R) or the green (δ_G) urn and δ_{r_R} and δ_{g_R} the respective probabilities of drawing a red or a green ball (in this example out of the red urn). Let *g* denote the number of green balls in a given signal and *r* denote the number of red balls in a given signal. *G* denotes the green urn. *R* denotes the red urn. H_t denotes the history of previous signals at *t* and τ denotes the transition probability parameter. As per Bayes Theorem the posterior likelihood that a switch to the green urn has occurred given the previous history of signals can be stated as

$$P(G|H)_t = \begin{cases} t = 0\,;\ \tau \\ t > 0\,;\ \dfrac{P(H|G)\,P(G)_{t-1}}{P(H|G)\,P(G)_{t-1} + P(H|R) * (1 - P(G)_{t-1})} \end{cases},$$

where

$$P(H|G)_t = {\delta_{g_G}}^{g_t} * {\delta_{r_G}}^{r_t},$$

and

$$P(H|R)_t = {\delta_{g_R}}^{g_t} * {\delta_{r_R}}^{r_t}.$$

The proper, normative Bayesian revision per signal can then be defined as

$$\Delta P(G|H)_t = P(G|H)_t - P(G|H)_{t-1}.$$

Further, let P_{e_t} denote the empirical probability judgment at *t*, then

$$\Delta P_{e_t} = P_{e_t} - P_{e_{t-1}},$$

defines the change in empirical judgments per given signal, which allows us to derive our variable of interest which is empirical under- or over-reaction as compared to the Bayesian standard. It can then simply be stated as

$$\Delta P_{e_t} - \Delta P(G|H)_t.$$

3. Experiment 1

The first experiment was carried out to check for the robustness of the system neglect hypothesis whilst extending the realism and ecological validity of the decision context by varying the amount of information (signal length) at hand.

3.1 Participants

Participants were students recruited on campus of a leading European business school for an experimental study advertised as related to decision making. A total of 41 participants completed the experiment. The average age was 30 years old (with a standard deviation of 6 years) and 41.5% were female.

3.2 Design

The experiment was administered as a within-subjects design. The task factors under investigation were three different levels of a diagnosticity parameter (the urns' distributions: 52/48, 60/40 and 68/32 – e.g. representing 68% red & 32% green in the red urn and vice versa in the green urn for the last case), two different probability levels of a switch (3.5%, 15%), and three different signal lengths (5, 8, or 11 balls). Sequences were randomly created using the respective parameters and sequences were presented in random order. Each condition consisted of 10 signals and each subject completed a total of 36 sequences resulting a total of 14,760 probability judgments.

3.3 Procedure

The experiment took place in a computer lab on campus. Subjects were assigned seats, briefed on how to access and run the experiment via Microsoft Excel and began the task at the same time. The task was not timed, subjects were free to stop at any time and to leave once finished. Once started, the experiment began with detailed instructions on the statistical process underlying the creation of the different sequences. (See appendix 1 for detailed instruction examples and screen shots.) Notably, subjects were explicitly told that all parameters were always constant per any given sequence. Subjects were told that their task was to estimate the probability that the last signal had come from the green urn (i.e. a switch having occurred). Subsequently subjects received detailed instructions on the payment formula and were told that their compensation was performance based. They were provided with a range of possible payouts as well as with a rough estimate of how long they should expect to complete the task. Thereafter, subjects went through a practice trial for which their performance was not recorded. The practice trial consisted of one complete sequence and was accompanied by instructions that repeated the nature of the statistical process underlying the sequences and guidelines on how to navigate the task. Only after completing the practice trial subjects proceeded to the actual 36 sequences. Further, with each new sequence, subjects were first shown a single sample signal stemming from the red urn. They were informed of their overall progress and payouts as proxy of judgmental accuracy after each sequence.

3.4 Results

On average the task took subjects 1h and 15 minutes to complete and the average payment was €11.94. All of the 14,760 judgments were used in the analysis. To investigate the statistical effects of diagnosticity, transition probability and signal length a repeated-measures ANOVA was run. To check for statistical effects across conditions, we averaged the differences between the respective empirical and Bayesian changes per signal (the average of $\Delta P_{e_t} - \Delta P(G|H)_t$ per condition). We found main effects of diagnosticity, $F(2, 162) = 117.33$, $p < 0.000$, $\eta_p^2 = 0.59$, transition probability, $F(1, 81) = 398.19$, $p < 0.000$, $\eta_p^2 = 0.83$, and length, $F(2, 162) = 34.07$, $p < 0.000$, $\eta_p^2 = 0.30$, all of which were in the predicted direction (see Figure 6 in appendix 2.1). We found interactions between diagnosticity and transition probability, $F(2, 162) = 67.08$, $p < 0.000$, $\eta_p^2 = 0.45$, diagnosticity and length, $F(4, 324) = 22.95$, $p < 0.000$, $\eta_p^2 = 0.22$, and transition probability and length, $F(2, 162) = 12.86$, $p < 0.000$, $\eta_p^2 = 0.14$ (see Figure 7 in appendix 2.2). We further found a significant three-way interaction between diagnosticity, transition probability and length, $F(4, 324) = 28.84$, $p < 0.000$, $\eta_p^2 = 0.26$ (see Figure 4).

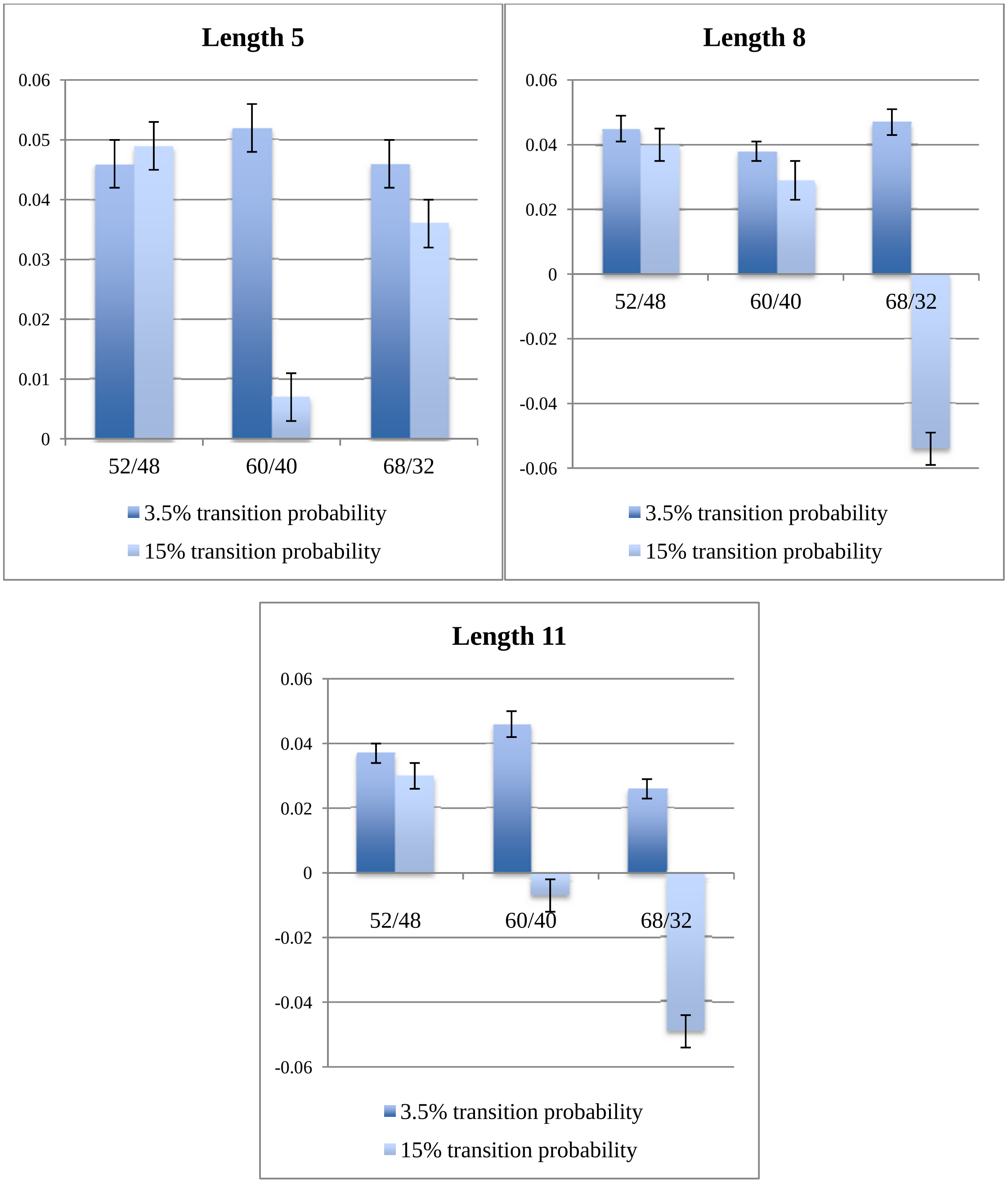


Figure 4 – Error bars represent one standard error

3.5 Discussion

As predicted, the amount of information available to subjects when making their judgments had a pronounced effect on its own and further in combination with the other two factors in determining relative over- and under-reaction. Our findings seem to confirm that the length of the signal in a regime change detection setting of this kind maps well with the notion of information at hand and previous findings that relate increasing amounts of information at hand to increasing conservatism (e.g. Slovic and Lichtenstein 1970). An increase in the amount of information to derive a judgment updating the likelihood of a switch can indeed be regarded as another manifestation of signal diagnosticity. As such, it ties into and potentially much alters the weight of evidence. In this sense, results from our first experiment further support the system-neglect hypothesis (Massey and Wu 2005), testing it in a novel, more complex and realistic setting. Subjects exhibited a pronounced disregard for changes in the weight of evidence offered to them when judging the likelihood of a regime change. We witnessed increasing conservatism with increasing levels of signal diagnosticity and with decreasing stability of the system (increasing transition probabilities). In relative terms, most over-reaction was exhibited in settings where the weight of evidence was lowest and most relative under-reaction was exhibited in settings where the weight of evidence was highest. Nonetheless, what until this point has not been explicitly tested is the effect of the strength of evidence in a regime change detection setting. It has been assumed as the counterpart to the weight of evidence, i.e. factors such as those discussed above, yet it has not been directly measured.

4. Experiment 2

Our second experiment aims to measure the effect of the strength of evidence and investigate its interplay with changes in the weight of evidence. As discussed, our measure for the strength of evidence are streaks within signals of either red or green balls, which may likely be perceived to be representative of the red or the green urn, respectively, and thus of a regime change in the case of the latter.

4.1 Participants

A total of 110 participants completed the online-based experiment. The mean age was 34 years old (with a standard deviation of 10 years) and 42.7% were female. Participation was location restricted to participants who were located in the USA. Amazon Turk provides access to a very large and very diverse base of subjects. In fact, Paolacci et al. (2010) as well as Buhrmester et al (2011) have shown evidence for its US subject base to be much more representative of the US population as a whole as compared to traditional university subject pools. Also, Horton et al. (2011) have replicated traditional experimental findings such as choice reversals through framing which provides further support for its validity as sampling source.

4.2 Design

The experiment was administered using a mixed design. The task factors under investigation were two different levels of a diagnosticity parameter (the urns' distribution: 60/40 and 68/32), two different probability levels of a switch (5% and 20%), both of which were administered as between-subject factors and two different streak probability ratios (high or low) that were administered as within-subject factor. Signal length was now held constant at 8 balls. For this experiment we first randomly created a number of sequences using the respective diagnosticity and transition probability parameters. We then chose one random sequence of each of the four parameter combinations and manually altered the location of red and green balls in each signal in order to create two derivate sequences, one of which maximized the occurrence of streaks of green balls and another minimized it. (Keep in mind that the system created all signals and sequences without memory, i.e. with replacement, and this thus does not matter for judging the likelihood of a regime change.) The eight different conditions were then split into four pairs of two that were randomly administered. Each condition again consisted of 10 signals and each subject completed two sequences, resulting in a total of 2,200 probability judgments. The streak probability ratios were computed by taking the order of the individual balls per signal into account. The variable was computed as the average ratio per sequence between the probability of a given streak in a signal, assuming that the signal was drawn from the green urn, and the probability of a given streak in a signal, assuming that the signal was drawn from the red urn. (See appendix 3 for a detailed account of the streak probability ratio factor.)

4.3 Procedure

The experiment took place online via Amazon's Mechanical Turk platform on which a link to the task was provided. The actual experiment was then carried out on Qualtrics in which the task was embedded. The instructions and the procedure were identical to the ones of the first experiment and differed only with respect to x in the payment formula and the number of overall sequences to be judged.

4.4 Results

The average completion time was 17 minutes. The average performance based pay was $1.24 and the average payment was $2.24 (including the $1 flat fee). To investigate the statistical effects of diagnosticity, transition probability and the streak probability ratio a mixed effects ANOVA was run. Diagnosticity and transition probability were both entered as between-subjects factors whereas the streak probability ratio was entered as a within-subject factor. To check for statistical effects across conditions, we averaged the differences between the respective empirical and Bayesian changes per signal (the average of $\Delta P_{e_t} - \Delta P(G|H)_t$ per condition). Once more, we found main effects of diagnosticity, F(1, 106) = 61.41, p < 0.000, η_p^2 = 0.37, and transition probability, F(1, 106) = 59.66, p < 0.000, η_p^2 = 0.36, both of which were in the predicted direction (see Figure 8 in appendix 2.3). We did not find a significant main effect for the streak probability ratio, F < 2. Moreover, we found a significant interaction between diagnosticity and transition probability, F(1, 106) = 11.96, p < 0.001, η_p^2 = 0.10 (see Figure 9 in appendix 2.4), and furthermore a significant three-way interaction between diagnosticity, transition probability and the streak probability ratio, F(1, 106) = 4.61, p < 0.034, η_p^2 = 0.04 (see Figure 5).

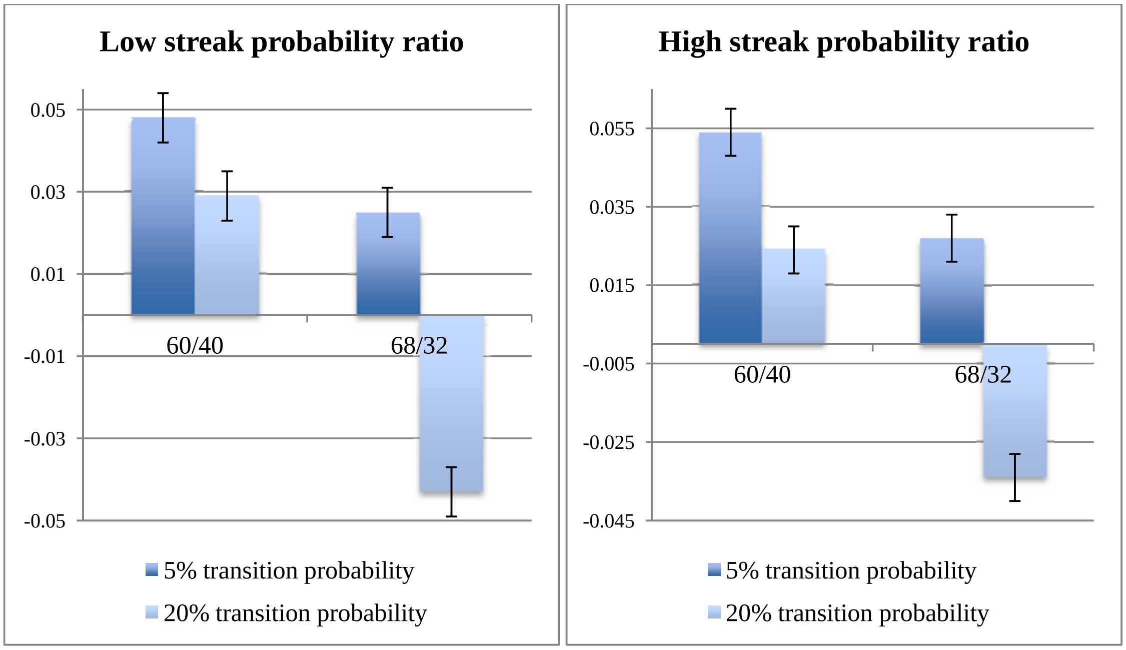


Figure 5 – Error bars represent one standard error

4.5 Discussion

Our second experiment has found further support for the effect of the weight of evidence, in that we replicated previous patterns of increasing relative conservatism with an increase in the weight of evidence. However, we did also find a significant interplay between the strength of evidence, measured by the streak probability ratio, and the weight of evidence. Although we failed to find a significant main effect of perceived streaks (whilst directionally it behaved as expected) it significantly increased relative over-reaction in situations of otherwise high weight of evidence (in our case highly unstable systems and high diagnosticity). Also, it needs to be noted that given our task design and our explicit instructions it was arguably much harder to find a significant belief in streaks. Subjects were clearly told that balls were sampled with replacement, which made the order of balls irrelevant when it came to judge the likelihood of a switch. Importantly, our task was also highly abstract and did not at all offer any rationale for a causal belief in human agency behind the creation of the sequences. Recent research on the belief in hot hands by Burns and Corpus (2004) found support for differences in the belief in hot hands to stem from how random subjects regarded the underlying system creating the sequence. For example, streaks in the case of basketball are likely regarded as less random than streaks in the case of a roulette wheel. Our evidence seems to indicate that our subjects still shared a causal perception that streaks were non-random and indicative of a certain system even though the instructions clearly indicated otherwise.

5. Conclusion

Our research ties into a long stream of research on regime change detection. We carried out two experiments, the first of which investigated the effect of varying amounts of information at hand when judging the likelihood of a regime change whereas the second investigated the effect of streaks on perceived likelihoods of regime changes. As in previous research, our subjects did not fully neglect relevant environmental parameters in judging the likelihoods of regime changes and yet notably deviated from the normative judgments of a perfectly Bayesian agent. Both of our studies have found further support for the system-neglect hypothesis (Massey and Wu 2005) and yet we have identified further relevant factors that significantly moderate relative over- and under-reaction in a regime change detection setting. We found novel support for both increasing relative conservatism with increasing amounts of information at hand as well as increasing relative over-reaction with an increase in apparent streaks in the data. Theoretically the amount of information strongly ties into the concept of the weight of evidence, whereas the apparent belief in streaks being representative of a regime change relates well to the concept of the strength or extremeness of evidence (Griffin and Tversky 1992). Both the belief in streaks and the effect of varying amounts of information are factors that are commonly involved in situations of regime change detection in practice. Whilst further factors remain to be investigated it is important to find ways to de-bias judgments in regime change detection settings since depending on the associated costs and benefits of over- and under-reaction the consequences of over- or under-reaction can be dire.

References

Buhrmester, M., Kwang, T., Gosling, S. D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data?. *Perspectives on Psychological Science*, 6, 1, 3-5.

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks: “Gambler’s fallacy” versus” hot hand “. *Psychonomic Bulletin & Review*, 11, 1, 179-184.

Chinnis, James O., Cameron R. Peterson (1968). Inference about a non-stationary process, *Journal of Experimental Psychology*, 77, 620–625.

Chinnis, James O., Cameron R. Peterson (1970). Non-stationary processes and conservative inference, *Journal of Experimental Psychology*, 84, 248–251.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 3, 295-314.

Griffin, Tversky (1992), The weighting of evidence and the determinants of confidence, *Cognitive Psychology*, 24, 411-435.

Horton, J. J., Rand, D. G., Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14, 3, 399-425.

Kahneman, D., Tversky, A. (1973a). On the Psychology of Prediction. *Psychological Review*, 80, 4, 237-251.

Kahneman, D., Tversky, A. (1973b). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.

Kremer, M., Moritz, B., Siemsen, E. (2011). Demand Forecasting Behavior: System Neglect and Change Detection. *Management Science*, 57, 10, 1827–1843.

Li, Y., Massey, C., Wu, G. (2009). Learning to Detect Change (January 30, 2009). Chicago Booth School of Business Research Paper No. 09-03. Available at SSRN: http://ssrn.com/abstract=1336724.

Massey, C., Wu, G. (2005). Detecting Regime Shifts: The Causes of Under- and Overreaction. *Management Science*, 51, 6, 932-947.

Paolacci, G., Chandler, J., Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5, 5, 411-419.

Rappoport and Summers (1973). *Human Judgment and Social Interaction*. Holt, Rinehart and Winston, Inc., USA.

Reimers, S., Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, 27, 4, 1196-1214.

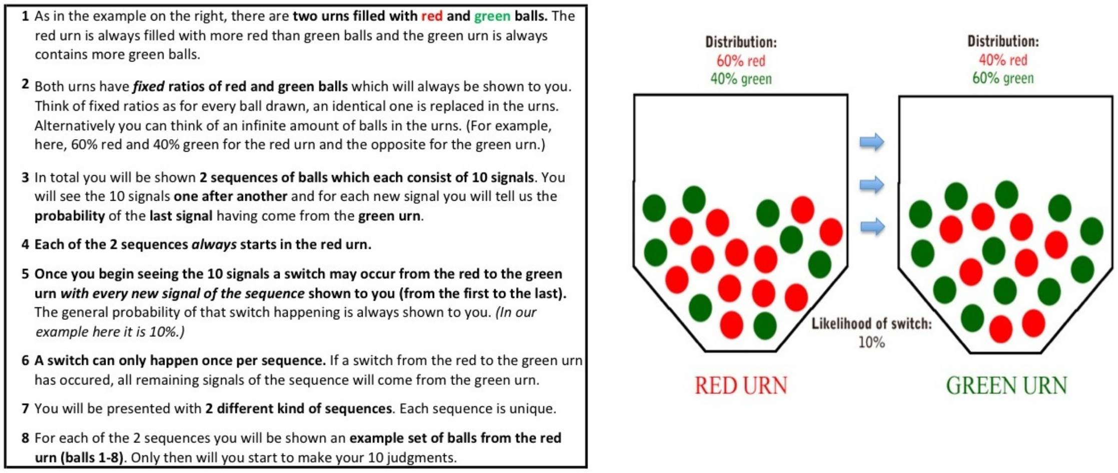
Slovic, P. Lichtenstein, S. (1970). *Comparison of Bayesian and regression approaches to the study of information processing in judgment*. Oregon Research Institute Research Monograph, 1970, Vol. 10, No. 1.

Tversky, A., Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 184, 1124–1131.

Appendices

1. Instructions and screen shots

1.1 Instruction example screen shots of experiment 2



1 As in the example on the right, there are **two urns filled with red and green balls.** The red urn is always filled with more red than green balls and the green urn is always contains more green balls.

2 Both urns have ***fixed* ratios of red and green balls** which will always be shown to you. Think of fixed ratios as for every ball drawn, an identical one is replaced in the urns. Alternatively you can think of an infinite amount of balls in the urns. (For example, here, 60% red and 40% green for the red urn and the opposite for the green urn.)

3 In total you will be shown **2 sequences of balls which each consist of 10 signals**. You will see the 10 signals **one after another** and for each new signal you will tell us the **probability** of the **last signal** having come from the **green urn**.

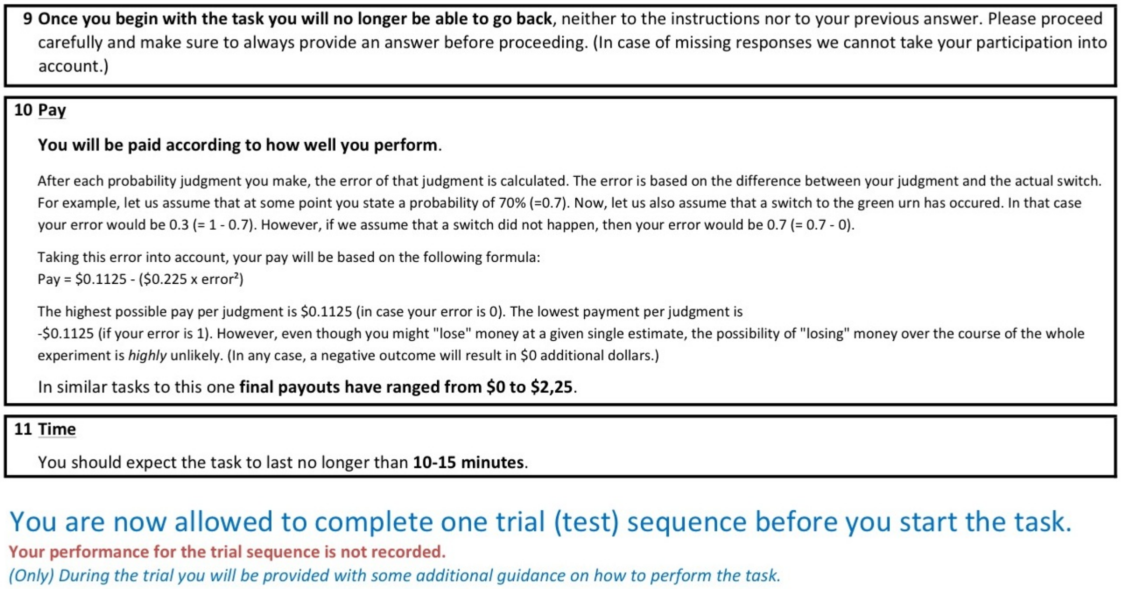
4 **Each of the 2 sequences *always* starts in the red urn.**

5 **Once you begin seeing the 10 signals a switch may occur from the red to the green urn *with every new signal of the sequence* shown to you (from the first to the last).** The general probability of that switch happening is always shown to you. *(In our example here it is 10%.)*

6 **A switch can only happen once per sequence.** If a switch from the red to the green urn has occured, all remaining signals of the sequence will come from the green urn.

7 You will be presented with **2 different kind of sequences**. Each sequence is unique.

8 For each of the 2 sequences you will be shown an **example set of balls from the red urn (balls 1-8)**. Only then will you start to make your 10 judgments.



9 **Once you begin with the task you will no longer be able to go back**, neither to the instructions nor to your previous answer. Please proceed carefully and make sure to always provide an answer before proceeding. (In case of missing responses we cannot take your participation into account.)

10 Pay

You will be paid according to how well you perform.

After each probability judgment you make, the error of that judgment is calculated. The error is based on the difference between your judgment and the actual switch. For example, let us assume that at some point you state a probability of 70% (=0.7). Now, let us also assume that a switch to the green urn has occured. In that case your error would be 0.3 (= 1 - 0.7). However, if we assume that a switch did not happen, then your error would be 0.7 (= 0.7 - 0).

Taking this error into account, your pay will be based on the following formula:
$\text{Pay} = \$0.1125 - (\$0.225 \text{ x error}^2)$

The highest possible pay per judgment is $0.1125 (in case your error is 0). The lowest payment per judgment is
-$0.1125 (if your error is 1). However, even though you might "lose" money at a given single estimate, the possibility of "losing" money over the course of the whole experiment is *highly* unlikely. (In any case, a negative outcome will result in $0 additional dollars.)

In similar tasks to this one **final payouts have ranged from $0 to $2,25**.

11 Time

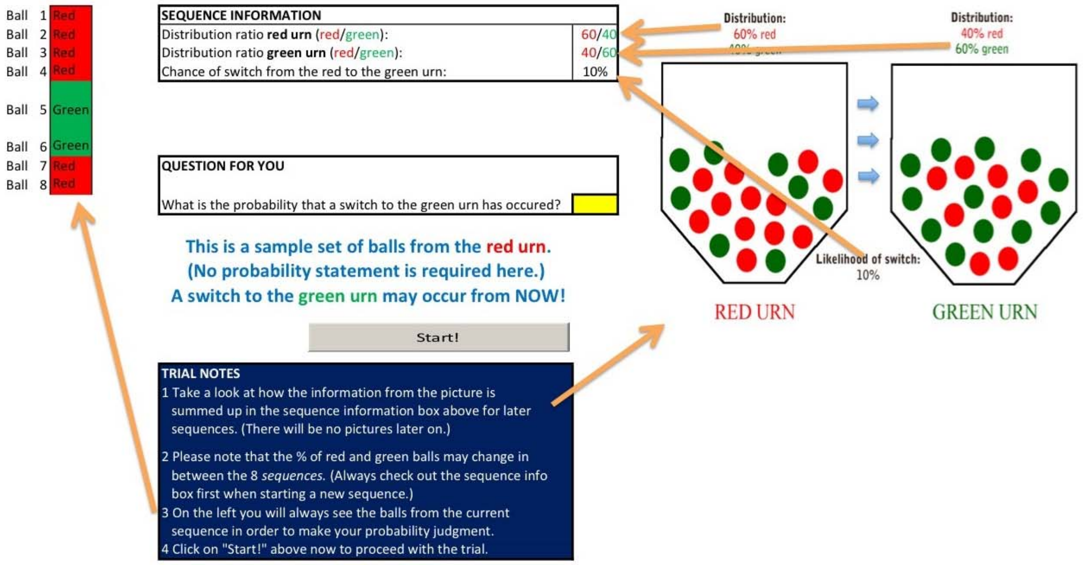
You should expect the task to last no longer than **10-15 minutes**.

You are now allowed to complete one trial (test) sequence before you start the task.

Your performance for the trial sequence is not recorded.

(Only) During the trial you will be provided with some additional guidance on how to perform the task.

1.2 Trial example screen shot of experiment 1



Ball 1 Red
Ball 2 Red
Ball 3 Red
Ball 4 Red
Ball 5 Green
Ball 6 Green
Ball 7 Red
Ball 8 Red

SEQUENCE INFORMATION	
Distribution ratio **red urn** (red/green):	60/40
Distribution ratio **green urn** (red/green):	40/60
Chance of switch from the red to the green urn:	10%

QUESTION FOR YOU

What is the probability that a switch to the green urn has occured?

This is a sample set of balls from the red urn.
(No probability statement is required here.)
A switch to the green urn may occur from NOW!

Start!

TRIAL NOTES

1 Take a look at how the information from the picture is summed up in the sequence information box above for later sequences. (There will be no pictures later on.)

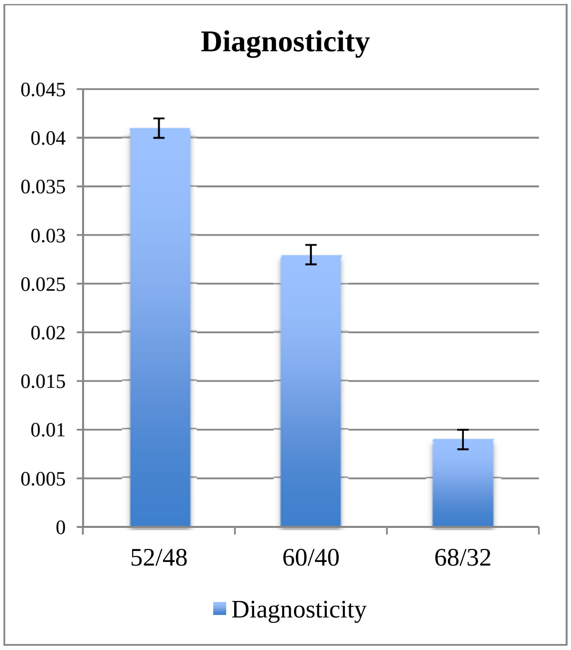
2 Please note that the % of red and green balls may change in between the 8 *sequences*. (Always check out the sequence info box first when starting a new sequence.)

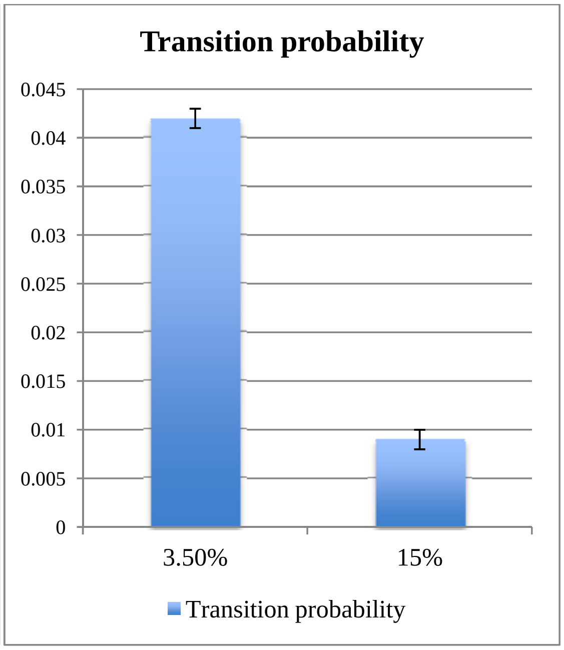
3 On the left you will always see the balls from the current sequence in order to make your probability judgment.

4 Click on "Start!" above now to proceed with the trial.

2. Estimated marginal means

2.1 Experiment 1: Main effects





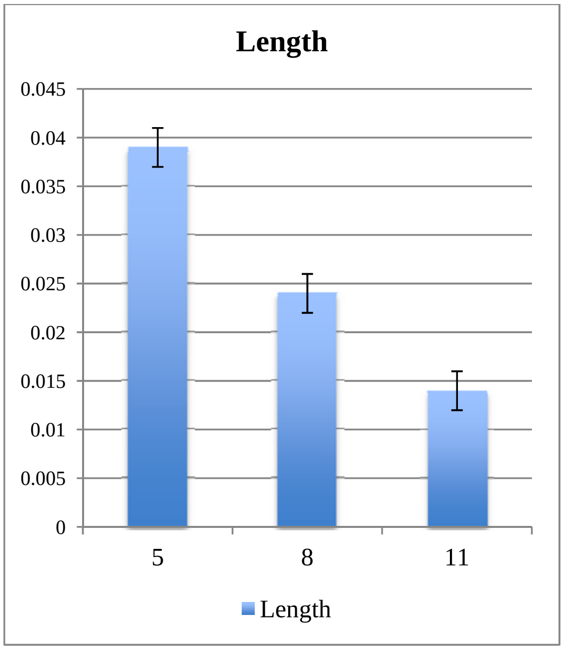
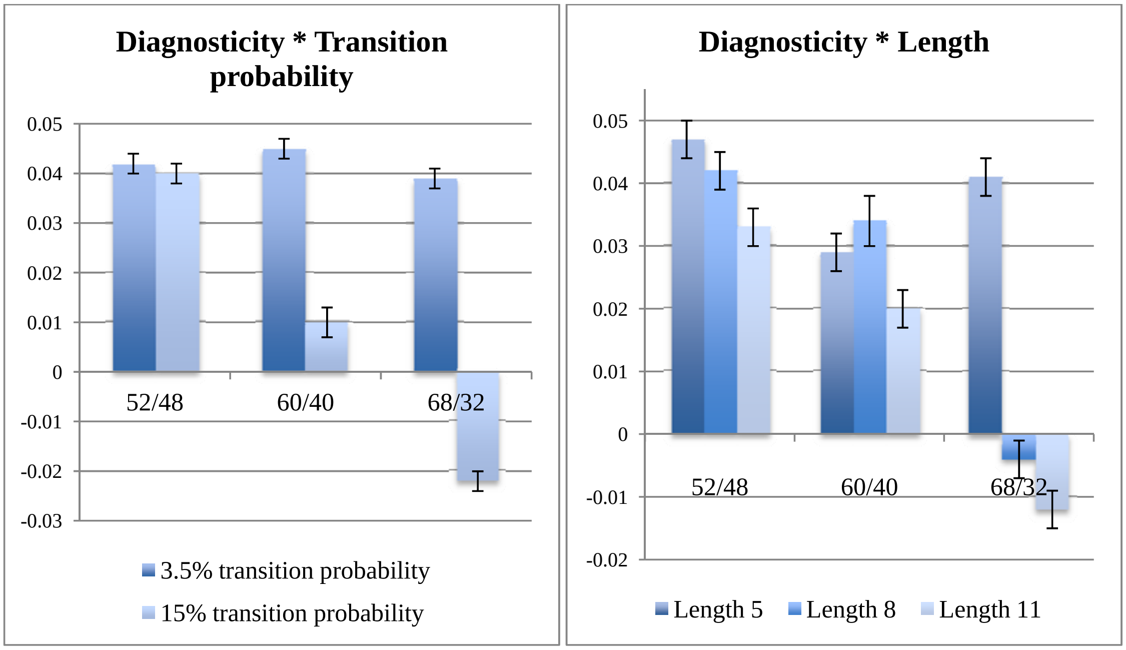


Figure 6 – Error bars represent one standard error

2.2 Experiment 1: Two-way interactions



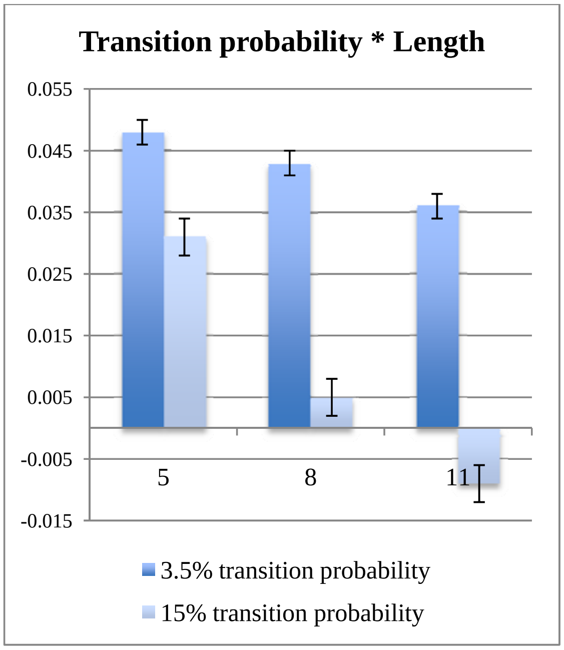
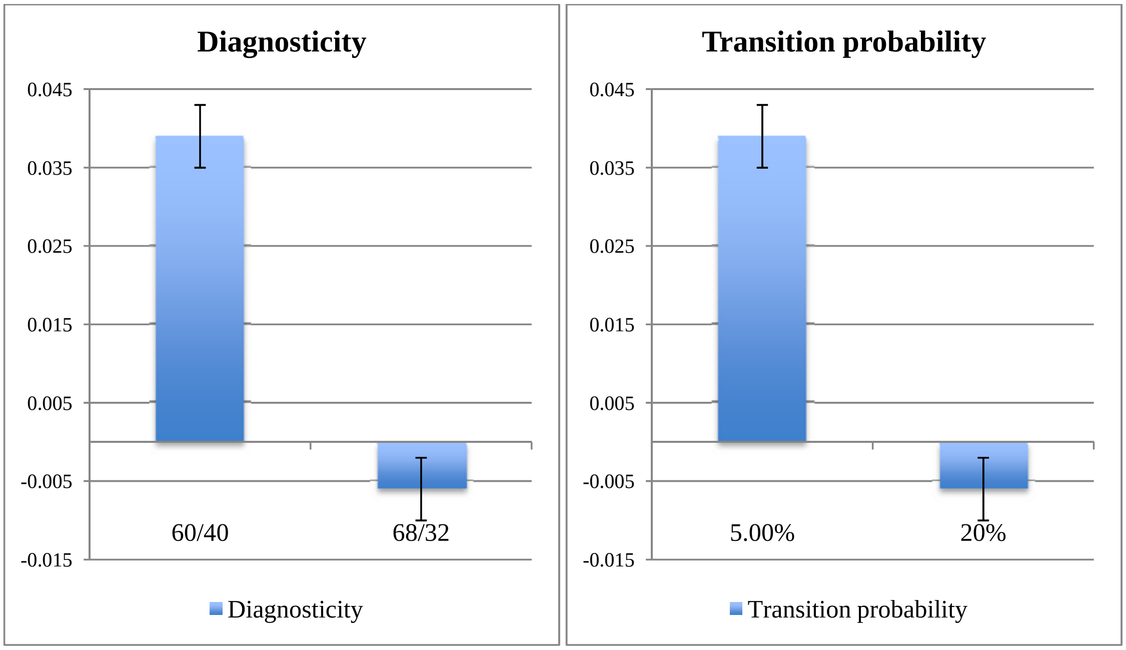


Figure 7 – Error bars represent one standard error

2.3 Experiment 2: Main effects



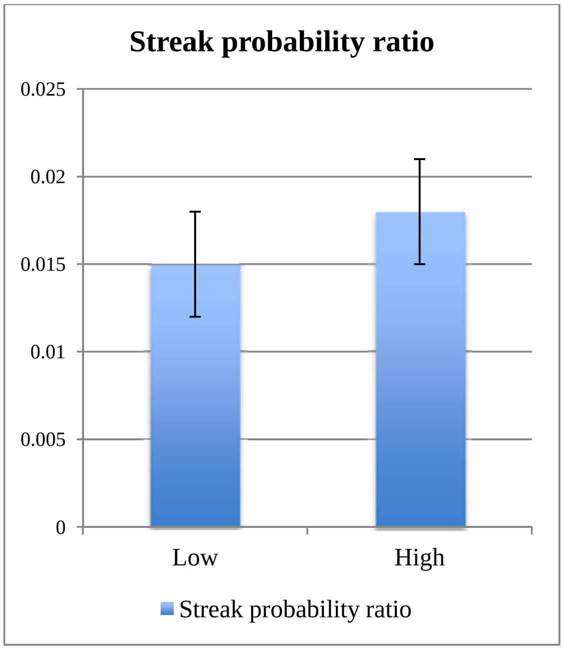
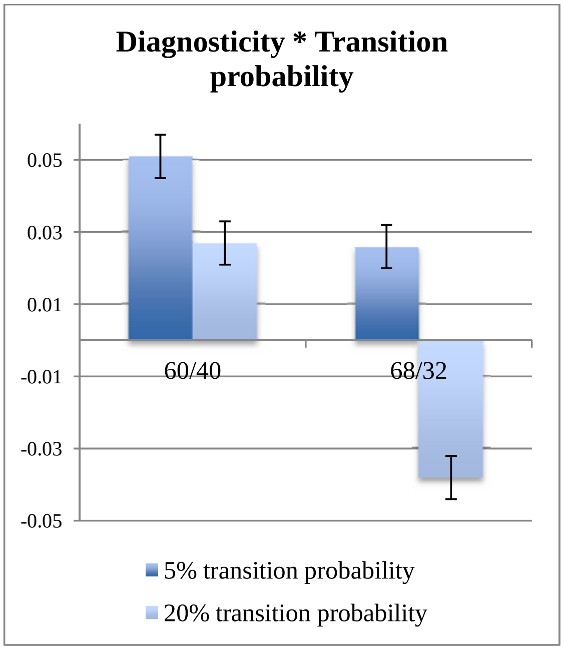
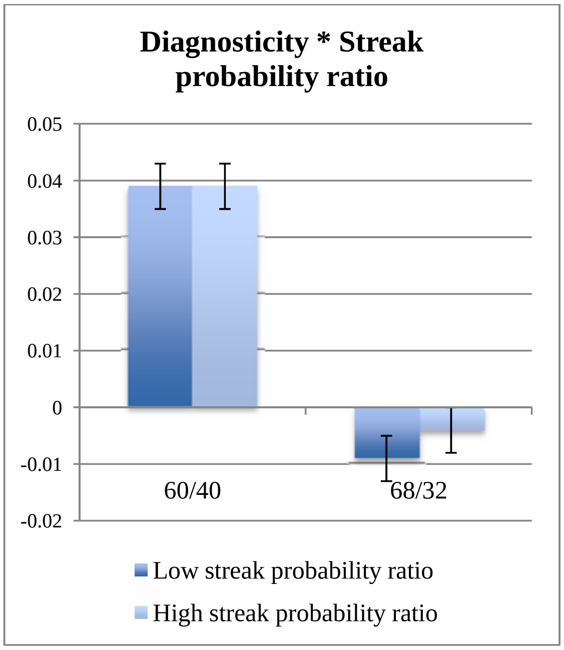


Figure 8 – Error bars represent one standard error

<u>2.4 Experiment 2: Two-way interactions</u>





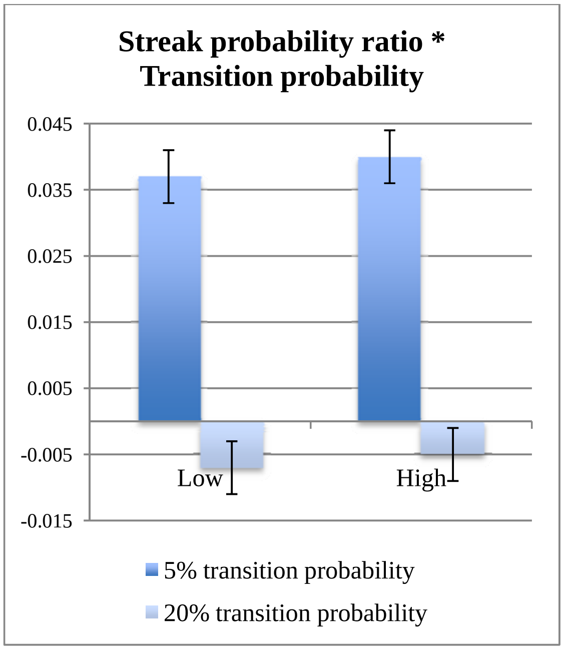


Figure 9 – Error bars represent one standard error

3. Variable computation

3.1 Streak probability ratio

We calculated the streak probability ratio factor by computing the ratio between (1) the probability of a given streak in a signal, assuming that the signal was drawn from the *green* urn, and (2) the probability of a given streak in a signal, assuming that the signal was drawn from the *red* urn. Even though the order of balls was normatively irrelevant when judging the likelihood of a regime change (balls were drawn with replacement) we were still able to compute such ratio for descriptive purposes taking into account the order of balls in a signal.

Consider the following example: In a signal of 8 balls the probability of seeing 5 green balls in a row (for example *r-g-g-g-g-g-r-r*) can be calculated by multiplying (1) the number of possible permutations of 5 green balls in a row divided by all possible permutations at length 8 by (2) the respective likelihood of seeing 5 green balls in a signal (e.g. 5 * 0.6, assuming that the balls have come from the green urn with a diagnosticity level of 40% red and 60% green balls).

We manipulated the order of balls in the previously at random crated sequences given certain diagnosticity and transition probability levels to create two sequences for each combination of the other factors. One of the two newly created sequences maximized the streak probability ratio (averaged over all signals per sequence) whereas the second sequence minimized the streak probability ratio. When the streak probability ratio was entered as a factor in our analysis, the former sequence represented the "*high*" level whereas the latter represented the "*low*" level. As such, this allowed us to tap into possible causal beliefs that associated apparent streaks (as opposed to diagnosticity levels) with a given underlying system.