How does prevalence shape errors in complex tasks?

Enkhbold Nyamsuren (e.nyamsuren@uva.nl)
Han L.J. van der Maas (h.l.j.vandermaas@uva.nl)
Department of Psychology, University of Amsterdam
Weesperplein 4, 1018 XA Amsterdam, Netherlands

Niels A. Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen,
Nijenborgh 9, 9747 AG Groningen, Netherlands

Abstract

This study shows that cause and types of errors in complex problem-solving tasks can be explained within a framework of the prevalence effect commonly studied only in simple visual search tasks. The explanation proposes that subjects make a series of probabilistic decisions aimed at balancing both speed and accuracy. Such decision is a complex process that relies not only on task instructions but also on cognitive biases established by the history of previous trials and progress of the current trial. We provide evidence based on both empirical data and cognitive modeling.

Keywords: problem-solving, cause of errors, prevalence, ACT-R

Introduction

Why and how do people make mistakes in complex problem-solving tasks? What are the primary cognitive mechanisms? We try to answer these questions using a computerized version of a board game SET[1]. Compared to typical laboratory tasks, SET is a more complex task requiring implicit and explicit strategies, coordination of bottom-up perceptual and top-down executive processes, making consecutive decisions and accumulation of evidence along several dimensions. Any of these components can be a source of errors. Despite a number of preceding studies focused on SET (Jacob & Hochstein, 2008; Mackey, Hill, Stone, & Bunge, 2011; Nyamsuren & Taatgen, 2013), none of them looked at the source of errors. However, the nature of errors can tell us a lot more about the process of problem solving than just the response times and accuracies. We employ a combination of empirical study based on Math Garden and cognitive modeling to tackle this problem. Math Garden (Klinkenberg, Straatemeier, & Van der Maas, 2011) is a web-based computer adaptive practice and monitoring system used by more than 2000 schools to train students' cognitive skills with serious games such as SET.

A SET trial starts with a number of cards dealt open (Figure 1). Each card is uniquely defined by a combination of four attributes: color, shape, shading and the number of shapes. Each attribute can have one of three distinct values. The goal is to find a unique combination of three cards, called a *set*, where values of each attribute are all same or all different. We refer to the number of different attributes in a set as the *set level*. For example, in Figure 1, a level 2 set is formed by three yellow cards. It has two same (color and shape) and two different (shading and number) attributes. In a level 4 set, all values of all attributes are different.

Jacob and Hochstein (2008) proposed that SET players use a dimension reduction strategy. They prefer to search for a set among cards that have the same attribute value thus effectively reducing the search space by one attribute dimension. For example, a subject may look for a set among cards of the same color. A later study (Nyamsuren, & Taatgen, 2013) confirmed Jacob and Hochstein's theory. Nyamsuren and Taatgen also found that dimension reduction is mostly used early in a trial. If dimension reduction strategy fails to find a set, subjects start searching for more dissimilar sets.

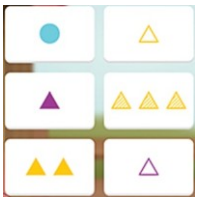

Figure 1: An example of a trial used in Math Garden. A level 2 set is formed by the yellow cards.

Experimental Results

The data was gathered in Math Garden between April 2014 and October 2014. It included 1374530 trials of 80 items (20 items per set level) played by 86964 subjects. Each item consisted of six cards and had exactly one set (e.g. Figure 1). A trial was terminated after a subject selected any combination of three cards. There was 30-seconds time limit per trial. Above sample does not include overtime trials or trials without proper responses (a subject can give up on a trial and request to shown an answer).

Accuracy and Response Time

The average accuracy[2] is around 70%. In 30% of the trials, subjects responded with wrong combinations of three cards (further referred as *triplets*). First, we study cause of errors

[1] SET is a game by Set Enterprises (http://www.setgame.com)

[2] Math Garden dynamically adjusts difficulty to maintain a 75% success rate. Therefore, relative accuracy is uninformative.

by analyzing response times (RT).

Confirmed by a linear regression carried out on trials' mean RT, Figure 2a shows that response times increase with set level for both correct and incorrect trials (β = 2.05, $t(156)$ = 20.2, p < .01). In correct trials, the increase is caused by two factors (Nyamsuren & Taatgen, 2013). Firstly, subjects tend to start a trial with search for a lower-level set and, if the set was not found, switch to search for higher-level sets. Secondly, it requires more effort to compare dissimilar attributes than similar attributes. It is likely that the same two factors are also responsible for RT increase in incorrect trials. Mean RT for correct trials is lower than mean RT for incorrect trials (β = -1.17, $t(156)$ = -4.2, p < .01). However, this difference in RT decreases as the set level increases (β = .35, $t(156)$ = 3.5, p < .01). Note that, for items with level 4 sets, mean RT for incorrect trials is lower than mean RT for correct trials.

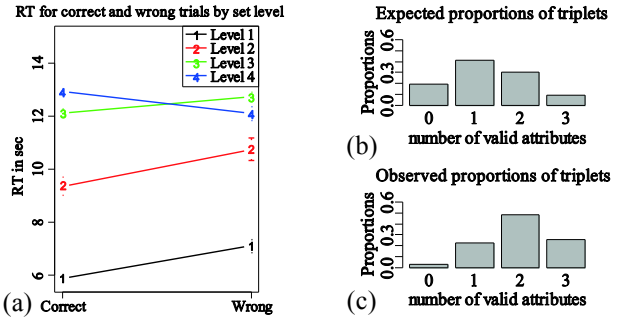

Figure 2: (a) Response times for correct and incorrect trials averaged by set level; Distributions of proportions of triplets by the number of valid attributes calculated from (b) all possible combinations of triplets in 80 items and (c) triplets provided as response by subjects.

Errors Based on Types of Triplets

Previous studies showed that perceptual aspects of SET have significant influence on subjects' decisions (e.g. Nyamsuren & Taatgen, 2013). Similarly, error types in SET may be affected by perceptual components of the task. In this section, we explore whether properties of a triplet defined by its combinations of attribute values affect subjects' decisions and error patterns.

Subsequent analyses concern incorrect trials where subjects responded with wrong triplets. Figure 2b shows a distribution of proportions of triplets by the number of valid attributes in a triplet. An attribute is valid if it follows the set rule and thus is either the same or different in all cards of the triplet. These proportions are calculated from all possible non-repeating combinations of triplets in all 80 items. They serve as a baseline. Figure 2c shows the same distribution, but with proportions calculated from wrong triplets provided as responses. According to Figure 2c, triplets with 2 or 3 valid attributes have significantly above chance probability of being chosen as a set. In other words, errors made by subjects are systematic and not random. More set-like triplets with higher number of valid attributes have higher probability of being incorrectly chosen as a set.

More importantly, there is a negative correlation between the number of valid attributes and RT. Errors with triplets with more valid attributes are made sooner than errors with triplets with fewer valid attributes. According to a linear regression analysis, RT decreases by 188 ms with each valid attribute in a triplet ($t(1518)$ = -2.26, p = .024).

Errors Based on Sameness and Difference

The previous section showed that the number of valid attributes in a triplet could have a significant impact on subjects' decisions. However, a valid attribute can be either same or different among three cards of the triplet. We found that sameness or difference of an attribute plays a substantial role in subjects' decisions.

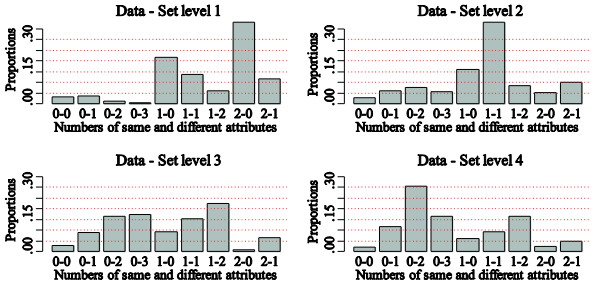

Figure 3: Distributions of proportions of errors types in trials categorized by set level. A wrong triplet consisting of three cards on the left of Figure 1 would give a 1-1 error type, since this answer contains one valid same attribute (shading) and one valid different attribute (color).

The following analysis concerns incorrect trials where subjects responded with wrong triplets. Figure 3 shows distributions of proportions of triplets with specific combinations of same and different valid attributes. The proportions were calculated separately for groups of trials of the same set levels. In trials with level 1 sets, most errors are made with triplets that had same valid attributes. For example, about 35% of all errors in level 1 trials involved triplets with two valid same attributes and no valid different attributes. The effect is completely opposite in trials with level 4 sets. In those trials, the most frequent errors involve triplets with different valid attributes. In fact, the gradual shift from sameness to difference can be observed in the distributions of proportions as set level increases. For levels 1 to 4, mean numbers of same attributes in wrong triplets are 1.4, 0.93, 0.61 and 0.46 against expected 0.67, 0.41, 0.33 and 0.21 if triplets were chosen randomly. Similar above chance preference was observed for different attributes in higher-level sets. Therefore, this shift likely represents a systematic shift in criterion against which subjects evaluate validity of attribute combinations.

Cause of Errors

An explanation of errors in SET can be derived from the prevalence effect. It is frequently observed in visual search

tasks where a target can be either present or absent. In low-prevalence conditions, subjects miss the target more often than in high-prevalence conditions (Wolfe, Horowitz, Van Wert, Kenner, Place, & Kibbi, 2007). Subjects do not explicitly try to speed up their responses using some time threshold. Instead, they adjust their internally estimated probability of a target being absent based on the sequence of previous trials (Ishibashi, Kita & Wolfe, 2012). This probability affects the decision on whether an object is a target or a distractor and the decision to quit the trial.

Within-trial Prevalence

With a proposal of within-trial prevalence, we assume that subject's internally estimated probability of finding a set changes during the progression of a trial. Subjects are aware that there is always one set present in each trial. Therefore, although probability of finding a set at the start of a trial is very low, it increases as a subject continues search and discards more distractor triplets. Wolfe and Van Wert (2010) proposed that the prevalence effect can be modeled via a drift diffusion model where decision is made when an evidence accumulation reaches a certain threshold. Similarly, we propose that subjects pick a triplet as a set when the accumulating probability reaches some threshold. During the trial, each discarded triplet increases probability of the next triplet being a set (green lines in Figure 4).

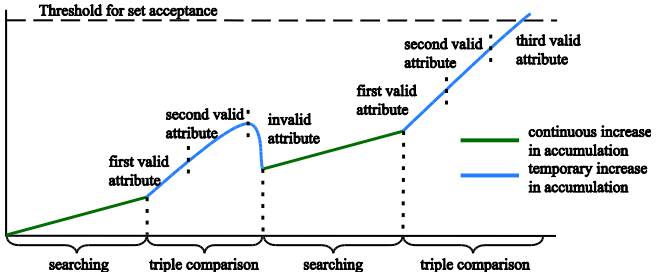

Figure 4: An evidence accumulation account of a within-trial prevalence in SET.

We know that the more set-like triplets are, the more likely they are to be chosen as a set. The effect can be explained with an assumption that an increasing similarity of a triplet to a set results in a temporary increase in within-trial probability. This process can be viewed as a large but temporary step-wise increase in accumulation caused by each new validated attribute (blue lines in Figure 4). Higher number of valid attributes will result in a larger increase in accumulation and a higher probability of exceeding the acceptance threshold. However, an attribute with invalid combination may negate the local boost in probability and results in the triplet being discarded as a potential set.

Even when a set is found early, the temporary increase in probability caused by four valid attributes is normally sufficient to exceed the threshold. On the other hand, in late trials, wrong triplets with few valid attributes will have a higher chance of exceeding the threshold due to constantly increasing probability. This process will result in incorrect trials having higher RT than correct trials (Figure 2a).

Finally, triplets with higher number of valid attributes may exceed the threshold sooner than triplets with fewer valid attributes explaining the negative correlation between RT and the number of valid attributes observed in the data.

Between-trial Prevalence

The prevalence effect also provides a framework for explaining why subjects shift from sameness to difference when validating attribute combinations. Here, changing prevalence of trials with particular set levels is a likely cause for such criterion shift. The adaptive algorithm in Math Garden ensures that next trial's difficulty is tailored to subject's skills. Therefore, new subjects start with easy trials with level 1 sets and are gradually introduced to more difficult trials. As a result, trials with level 1 and 4 sets initially have high and low prevalence respectively. However, as subjects gain more experience, prevalence of trials with level 1 and level 4 sets decreases and increases respectively causing subjects to shift their set acceptance criterion from similarity to dissimilarity. Based on data of 432 subjects who played at least 100 trials, proportions of trials with set levels 1 and 4 in first 25 trials are 63% and 7% respectively. For the fourth bin of 25 trials, the same proportions change to 18% and 28% respectively.

In terms of evidence accumulation account shown in Figure 4, different and same valid attributes make different contributions to the temporary increases in accumulation. In trials with level 1 sets, valid different attributes may not cause temporary increase in accumulation or may even have inhibitory effect on accumulation. However, as a subject is exposed more to trials with higher set levels, contributions of valid different attribute may gradually increase.

Threshold

The fact that the RT increases with set level indicates that the threshold is not the same among trials with different set levels. It is likely that subjects dynamically adjust their threshold whenever it is too low or too high, as in other visual search tasks. Chun and Wolfe (1996) showed that subjects' RT in target absent-present visual search tasks can be reproduced with a model using a dynamic threshold adjusted in a staircase manner. It was further suggested that RT in low- and high-prevalence search tasks can be modeled via adjustment of a quitting threshold (Wolfe & Van Wert, 2010). In a more recent visual foraging study, subjects adjusted in a staircase manner their probability of remaining on a patch depending on whether an instance of foraging was successful or not (Wolfe, 2013).

We draw an analogy from above examples and propose that subjects in SET are also adjusting set acceptance threshold in a staircase manner based on the result of the previous trial. After making a mistake, a subject may become more conservative and increase set acceptance threshold. The opposite will happen after a correct trial where the subject accepts a more liberal approach by

lowering threshold.

Cognitive Model

A cognitive model was used to formally test validity of the processes proposed in the preceding section. We have reused a model of a SET player developed in our earlier work. Due to space limit, we will describe only essential details of the model. The reader is referred to previous literature for a detailed description of the model (Nyamsuren & Taatgen, 2013). The model is based on ACT-R cognitive architecture (Anderson, 2007) that simulates functionality of essential cognitive resources such as declarative memory, working memory, the visual system and the production system. Within a model, task-related instructions are implemented as a set of production rules.

The overall strategy used by the model is simple. The model chooses a triplet and compares validity of four attributes one by one in a random order. This is done by having a production rule named '*compare*' repeatedly being called for each attribute. Only when all four attributes form valid combinations, the model chooses the triplet as a set ending a trial. When any attribute yields an invalid combination, the triplet is discarded and a new triplet is chosen. At the beginning of the trial, the model prefers triplets of cards having, at least, one common value (e.g. all green cards). Later in the trial, the model switches to triplets with cards that are more dissimilar.

The original model did not make mistakes. We have extended the model by implementing error-making mechanisms described in the preceding section. The next section describes those extensions.

Production Competition as a Cause of Errors

The original model took a conservative approach to set acceptance ensuring that all four attributes were valid in a triplet. The modified model adopts a more liberal approach and can accept a triplet as a set without validating all attributes. This is done by introducing a new production rule named '*valid-set*' that competes with the production rule '*compare*'. This process is shown in Figure 5. Given a triplet, the model can either validate an attribute in the triplet by calling '*compare*' production or accept the triplet as a set by calling '*valid-set*' production. A production rule with the highest utility value *U* is chosen.

Utility of '*valid-set*', $U(V)$, represents the accumulation shown in Figure 4 and indicates a probability of a triplet being a set. $U(V)$ is zero at a start of a trial but increases as the trial progresses according to $U(V) = 1/(T_t - T_c)$. T_t is the total number of unique triplets formed by six cards and equal to 20. T_c is the number of compared triplets, and $(T_t - T_c)$ is the number of remaining uncompared triplets.

$U(V)$ can temporary increase based on the number of validated attributes in a triplet. Given k validated attributes, probability of a set is $U(V) = 1/[(T_t - T_c) * P(k)]$, where $P(k)$ is a proportion of triplets that have k or more valid attributes. For each newly compared triplet, k is set to 0 but increases with each validated attribute in the triplet. It is unlikely that subjects can estimate $P(k)$ within each trial. However, it is probable that subject may able to learn prevalence of triplets with different values of k over many trials. Calculated from all unique triplets from all items, the proportions are $P(1) = .8$, $P(2) = .4$, $P(3) = .11$, and $P(4) = 1/(T_t - T_c)$ since there is only one set. Therefore, $U(V)$ increases with increasing k and is equal to 1 for $k = 4$ simulating the temporary increase in accumulation shown in Figure 4. Above proportions decrease as T_c increases.

$$T_{m_{i+1}} = \begin{cases} T_{m_i} - 6, & trial_i \text{ is correct} \\ T_{m_i} + 4, & trial_i \text{ is incorrect} \\ 12, & trial_i \text{ is overtime} \end{cases} \quad \text{(Eq. 1)}$$

The utility of '*compare*' production, $U(C)$, represents a threshold for set acceptance. The threshold for *i*-th trial is calculated as $U(C)_i = 1/(T_t - T_{m_i})$, where T_{m_i} is the minimum number of triplets to be compared in the trial. $U(C)$ remains the same during a trial, but T_m decreases or increases between trials according to the Eq. 1. For the next trial, $U(C)$ increases if the model makes a mistake by responding with a wrong triplet. If set is found, $U(C)$ decreases. If utility is too high and the model cannot find a set within a time limit then T_m is reset to 12, the minimum allowed value. This minimum value is set based on the assumption that subjects always have to perform some search. Numerical constants were fitted based on model simulations.

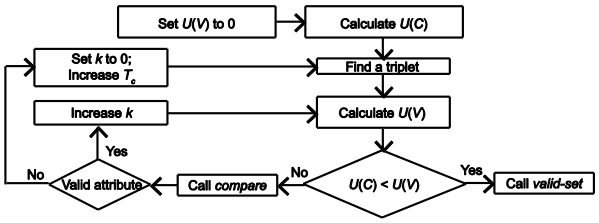

Figure 5: The competition between '*valid-set*' and '*compare*' productions is a cause of errors in the model.

Bias to similarity

Bias to similarity can occur at least at two decision points. First, bias can affect a choice of strategy. To replicate the effect of prevalence of trials with different set levels, the model was modified to be highly biased to the dimension reduction strategy while playing items with set level 1. However, this bias decreases with increasing set level following the decrease in observed proportions of trials with set level 1. Therefore, while playing items with set level 4, the model is more likely to use dissimilarity-based strategy.

Second, similarity bias can affect a decision whether a triplet is a set with subjects giving an initially higher weight to valid same attributes than to different attributes. This bias is simulated using two weights W_s and W_d that affect calculation of the value k: $k = W_s k_s + W_d k_d$. k_s and k_d are numbers of validated same and different attribute respectively. If the model is using the dimension reduction strategy then W_s and W_d are equal to 0.5 and -0.5 respectively. Otherwise, W_s and W_d are equal to -0.5 and 0.5

if the model is using a dissimilarity-based strategy.

Those changing weights represent shift in criterion for accepting a triplet as a set. More specifically, weights coupled with decreasing bias to the dimension reduction strategy simulate in the model a shift in set acceptance criterion from similarity to dissimilarity.

Simulation Results

80 items were divided into 10 blocks. Each block contained items of the same set level and same distance between set cards. The model was tested on 10000 trials in each block. The '*compare*' production had the minimum utility at the first trial of a block but was adjusted between trials of the same block.

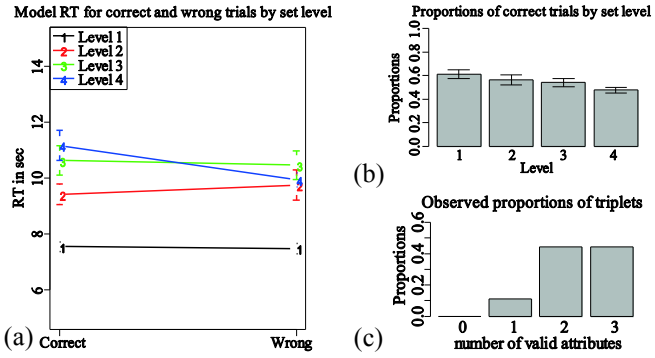

Figure 6: (a) Model's response time for correct and wrong trials of different set levels. (b) Model's accuracy by set level. (c) Proportions by the number of valid attributes of wrong triplets selected by the model.

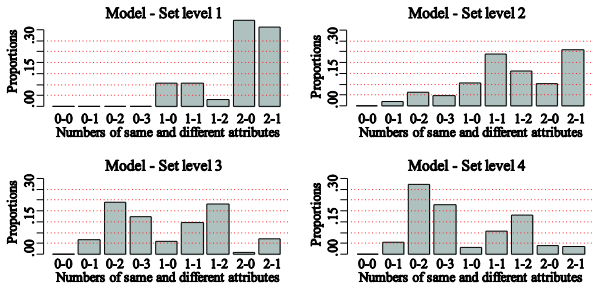

Figure 7: Distributions of proportions of errors made by the model in trials categorized by set level.

Figure 6a shows RT produced by the model. Similar to experimental data, model's RT increased as set level increased (β = 1.0, $t(156)$ = 7.5, p < .01). The main predictor for accuracy indicates that RT should be lower for correct trials (β = -.36, $t(156)$ = -.97, p = .3) with the interaction term indicating that this difference should decrease as set level increased (β = .2, $t(156)$ = -1.5, p = .14). However, both the main and interaction terms were not significant. The model was not able to simulate decreasing difference in RT between correct and incorrect trials observed in the empirical data (Figure 2a). Figure 6b shows model accuracy. The model predicts that the accuracy should decrease with set level. Most likely, this trend was not observed in subject data (Figure 2a) because Math Garden maintains 0.75 probability success by matching trial's difficulty to subject's skills. Figure 6c shows types of model-made errors defined by the number of valid attributes. Similar to subjects (Figure 2c), the model is more likely to make errors with triplets that have high number of two or three valid attributes. The increasing number of valid attributes also results in lower RT for incorrect trials (β = -1.1, $t(844)$ = -4.3, p < .01). However, the decrease of 1.1 seconds per valid attribute is much higher than 0.188 seconds observed in subjects' data.

Figure 7 shows distributions of proportions of incorrectly selected triplets with specific combinations of same and different valid attributes. The model's data closely resembles the empirical data shown in Figure 3. Correlations between the proportions in the empirical data and model data are $r(7)$ = .86, p < .01 for level 1, $r(7)$ = .84, p < .01 for level 2, $r(7)$ = .97, p < .01 for level 3, and $r(7)$ = .99, p < .01 for level 4. The model shows the same shift in criterion from similarity to dissimilarity in its decision of a triplet being a set. Overall, model simulations support our hypothesis that the prevalence effect, internally estimated probability of finding a set, can be a cause of errors.

Discussion and Conclusion

Speed-accuracy trade-off (Wickelgren, 1977) may also contribute to errors since Math Garden pressures subjects to finish a trial both quickly and accurately. However, it is unlikely to be the sole or even the main cause of errors in SET. First, the negative correlation between the number of valid attributes and RT is not easily explained by speeded responses. Second, a critical assumption behind the speed-accuracy trade-off is that errors should disappear if subjects are discouraged from giving fast responses. However, the reward system used in Math Garden severely punishes for fast incorrect responses making it more profitable to make slow correct responses. Therefore, the ideal strategy is either to give a correct response or let the time run out. The fact that subjects make early errors (Figure 2a) despite discouragement of fast responses violates the assumption behind the speed-accuracy trade-off. Wolfe at al. (2007) also explicitly differentiated the prevalence effect from speed-accuracy trade-off and showed that people resort to probabilistic decision making even in absence of a time pressure. The prevalence-based explanation assumes that that estimated probability causes changes in both RT and accuracy (Wolfe & Van Wert, 2010). Therefore, manipulations based on time should have little effect on estimated probability and, therefore, on accuracy explaining why subjects still made early errors in SET despite strong discouragement in Math Garden.

Therefore, a general question that has not been addressed in other studies is why a probabilistic decision is made despite an opportunity to verify their answers. We propose that it is due to an inherent nature of a human cognition to

pursue efficiency. Efficiency is achieved by minimizing the amount of cognitive effort to accomplish the task while still maintaining a reasonably high degree of success. This efficiency optimization is different from the common definition of optimization aimed at finding the optimal solution. Instead, in cognitive literature, efficient strategy is often referred to as heuristic (Gigerenzer & Brighton, 2009). Heuristics are simple strategies that do not guarantee absolute success rate but work most of the time. A necessity for heuristics is dictated by the framework of bounded rationality (Simon, 1972). It assumes that cognitive resources are limited and, therefore, processes utilizing the least amount of resources are favored even at the expense of accuracy. Note that a time pressure is not a required component for a formation or use of heuristics.

Above discussion suggests that the prevalence effect is a general phenomenon beyond simple visual search tasks commonly used to study the effect. Our study shows that it may play an important role in complex problem-solving tasks. For example, a similar effect is commonly observed in causal reasoning tasks. Griffiths, Sobel, Tenenbaum and Gopnik (2011) showed that subjects internally estimate Bayesian-like probabilities to judge causal relations between an effect and two possible causes. Although subjects were aware that it is possible for both options to independently cause the effect, their judgments were highly correlated with frequencies of both options causing the effect. The more prevalent option was not only likely to be classified as a cause but also decreased the probability of positive classification for the second option despite the independence of causes. Therefore, decision-making in causal task is not just frequency-based but a probabilistic process that incorporates frequency information.

Wolfe and Van Wert (2010) originally proposed that target prevalence in visual search can be modeled with a drift diffusion model with a changing starting point. Indeed, triplet comparison processes in Figure 4 can be viewed as a sequence of drift diffusion models where a consecutive model has a higher starting point than the previous one. However, Wolfe and Van Wert assumed that the starting point can only change between trials and not within trials. Our study shows that, in a complex task requiring several decisions, the starting point can and should change within a trial if there is a high expectation that the target is present. During the progression of visual search the estimated probability of finding a target should increase. This leads to our assumption that prevalence is not only a between-trial effect, but also can be observed within a trial in complex tasks such as SET.

ACT-R does not provide a suitable and standardized way to model an evidence accumulation process in the procedural system. In this study, we proposed that production rule's utility can change as a function of relevance to a changing context without the production rule being executed. It is not inconsistent with the existing utility learning mechanism, but adds an additional factor that influences utilities. As a proof of concept, the SET model used this mechanism to manipulate utility of a single production ruule during a trial to replicate a human behavior conventionally modeled with accumulation models. While we argue for the necessity for such mechanism, more studies are required for its implementation that is well integrated into ACT-R both theoretically and technically. Experimental data and the cognitive model can be downloaded from http://www.bcogs.net/models/

References

Anderson, J. R. (2007). *How can human mind occur in the physical universe?* New York: Oxford University Press.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, 30 (1), 39-78.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1 (1), 107-143.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. *Cognitive Science*, 35, 1407–1455.

Ishibashi, K., Kita, S., & Wolfe, J. M. (2012). The effects of local prevalence and explicit expectations on search termination times. *Attention, Perception, & Psychophysics*, 74 (1), 115-123.

Jacob, M., & Hochstein, S. (2008). Set recognition as a window to perceptual and cognitive processes. *Perception & Psychophysics*, 70 (7), 1165-1184.

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57 (2), 1813-1824.

Mackey, A. P., Hill, S. S., Stone, S. I., & Bunge, S. A. (2011). Differential effects of reasoning and speed training in children. *Developmental Science*, 14 (3), 582-590.

Nyamsuren, E., & Taatgen, N.A. (2013). Set as instance of a real-world visual-cognitive task. *Cognitive Science*, 37 (1), 146-175.

Simon, H. A. (1972). Theories of bounded rationality. Decision and organization, 1 (1), 161-176.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41 (1), 67-85.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, 13(3), 10.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, 136 (4), 623.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, 20 (2), 121-124.