A Model of the "Guilty Knowledge Effect:" Dual Processes in Recognition

Travis L. Seymour (NOGARD@Umich.Edu)
Department of Psychology; 525 East University
Ann Arbor, MI 48109-1109 USA

Colleen M. Seifert (SEIFERT@Umich.Edu)
Department of Psychology; 525 East University
Ann Arbor, MI 48109-1109 USA

Abstract

Recent alternatives to the polygraph-based Guilty Knowledge Test by (Farwell & Donchin, 1991; Seymour, Mosmann, & Seifert 1997) raise important questions about automatic access to knowledge already in memory. Despite subjects' intentions, "guilty" knowledge in memory can be detected because its automatic access interferes with other recognition tasks (Seymour, et al., 1997). To account for this finding, we present a model based on classic models of recognition (e.g. Kintch 1970; Anderson & Bower 1972). We posit that 'recognition' is a dual process involving a *familiarity* component where recent occurrence is quickly assessed, and a slower *source resolution* component, where the source of the familiar information is identified. Our model of the Guilty Knowledge Effect can account for patterns of response time and accuracy used to measure access to guilty knowledge (Seymour, et al., 1997). We also explain why strategies used to mask the Guilty Knowledge Effect are likely to fail given constraints on the recognition process, and discuss potentially successful strategies suggested by the model.

Introduction

In most cases, information regarding the contents of another person's memory is filtered through that person's wants, needs and biases. Individuals may deny or alter reports of what information they have in memory. In order to determine whether someone does possess particular privileged knowledge, a polygraph-based measure called the Guilty Knowledge Test (e.g. Lykken, 1981) was devised. In this method, "lying" is presumed when a suspect's responses to crime-related information consistently result in higher levels of arousal compared to control questions. This arousal is detected by physiological measures such as galvanic skin response and heart rate. The subsequent pattern of results from the polygraph are then used to infer the emotional impact of a suspect's responses. Though such polygraph-based tests are admissible in many states (provided that both sides consent), there are numerous studies questioning their reliability and validity (e.g., Bashore & Rapp, 1993; Furedy & Heslegrave, 1988).

As an alternative to the Guilty Knowledge Test, Farwell & Donchin (1991) set out to determine whether a suspect possesses crime-related information by measuring knowledge activation in memory rather than physiological response. Their assumption was that in order to answer the question "Do you know the color of the getaway car," guilty suspects must automatically access the target information before answering, regardless of whether they intended to admit to having the knowledge. Innocent suspects, however, would not have this knowledge and therefore it could not be accessed. Using the P300 (P3) component of evoked related potentials (ERPs) measured during an Old/New recognition task, Farwell & Donchin were able to reliably distinguish subjects possessing guilty knowledge and those without such knowledge. Because their paradigm measured recognition of information in memory, and because ERPs are thought to be difficult to manipulate, Farwell and Donchin propose that their higher reliability and accuracy (90% "guilty" and 85% "innocent" classification) make their method a superior alternative to the polygraph.

The Guilty Knowledge Effect

A variation of this methodology using response times (RT) rather than ERPs, Seymour, Mosmann and Seifert (1997) showed that a better differentiation between "innocent" and "guilty" subjects ("guilty" 90% and "innocent" 100% correct classification) could be attained without the use of ERPs. Based on Farwell & Donchin's (1991) paradigm, Seymour, et al. (1997) brought subjects into the laboratory and then asked them to "commit" a crime: Log into an email account on one university computer and send a message to a student suspected of computer fraud. Subjects learned 6 pieces of information in order to carry out their "assignment," which included the name of a street on campus, name of a person to login as, some identifying article of clothing, a "mission" name, a file folder to ask for and the name of the documents within. After memorizing this information, subjects logged in and sent the message as directed (a mockup of the university interface provided a realistic experience, though no message was actually sent).

Next, in an ostensibly unrelated task, subjects were asked to memorize a Target List of items very similar to the list they had learned for the mock crime. Subjects then participated in a List Priming Task where items were presented on a computer screen and subjects asked to make "old/new" judgements by pressing one of two response keys. One sixth of the items presented were Target Items from the list most recently learned, and required an "old" response. Two thirds of the items presented were Irrelevant Items (Non-Target), to which subjects were to make a "new" response. Unknown to the subject was a third category of stimuli that

occurred on 1/6 of the trials. These stimuli were, during Guilty blocks, items from the mock crime the subject had committed and, during Innocent blocks, items from a mock crime for which the subject had no knowledge. Items from this Probe stimulus category were to be responded to as Irrelevant ("new") items.

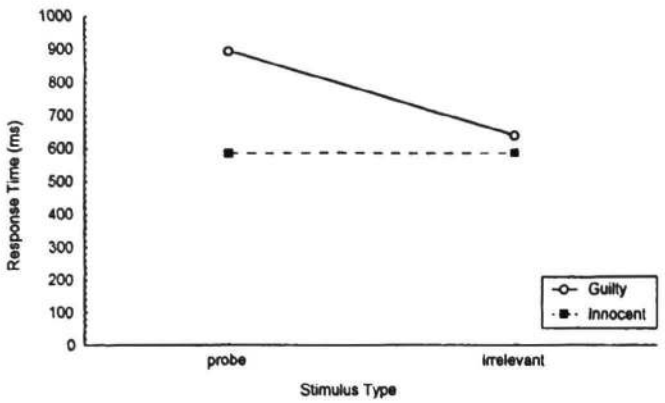

Figure 1. RT results from Seymour et al. (1997).

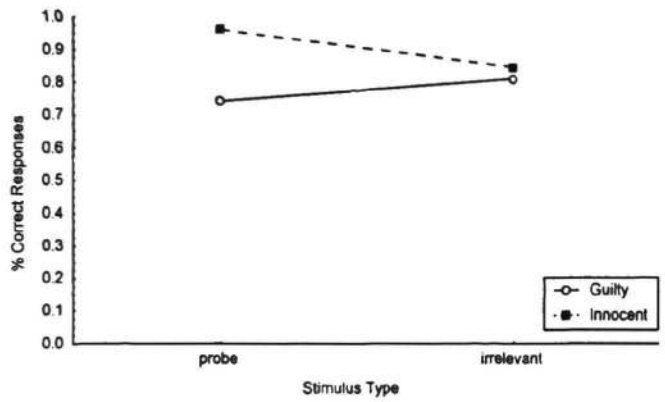

Figure 2. Accuracy results from Seymour et al. (1997).

Figure 1 shows the standard RT result for the critical comparison between Probe and Irrelevant items (both were "new" judgements). As predicted, the only reliable RT differences were between Guilty-Probe Items and Innocent-Probe Items. This "Guilty Knowledge Effect" was also supported with the accuracy data (see Figure 2), where only accuracy for Probe items differed as a function of Guilt. This pattern suggests that subjects had no trouble rejecting Innocent-Probe items as Irrelevant while there was considerable difficulty rejecting Guilty-Probe items as Irrelevant.

In sum, the recognition of items as related to the earlier crime interfered with subjects' performance on the "old/new" task: they were both slower and less accurate if items referred to the crime.

In other experiments, subjects were informed about the nature of the testing, and even given specific instructions about how to "beat" the test by minimizing interference from the crime words. However, in several different studies, no subject was able to mask the longer response times associated with crime related words. The deadline procedure employed (around 1000ms) forced subjects to respond too quickly to allow strategic manipulation of their responses (Ratcliff & McKoon, 1981). Even subjects motivated to foil the test were unable to do so.

Given these results, our goal was to posit a theory of the recognition processes which lead to the Guilty Knowledge Effect. This model could then be used to account for how and when conscious strategies can affect the activation of related knowledge in memory.

Models of the Recognition Process

The recognition process in human memory has been characterized by Kintsch (1970), and refined by Anderson & Bower (1972). Kintsch's model of recognition essentially described a single process of signal detection (Tanner & Swets, 1954; Green & Swets, 1966), where familiarity is judged in the context of "old" and "new" item distributions using a positive D-prime (difference between each distribution's mean activation) and Beta (subjects' decision criterion indicating their threshold for "old" responses).

Kintsch hypothesized that study of the test list of items causes an increase in the activation associated with those items in memory. The set of distractors was hypothesized to have a similarly-distributed set of activation levels, with a mean activation lower than for the test list. The subject sets a decision threshold for the degree of activation required to say "old," and the activation of a stimulus on each trial is used to make the "old/new" judgement. Subjects are not only good at making such distinctions, but perform nearly as well when required to distinguish among items from several target lists in an "old/new" paradigm (Anderson & Bower, 1972). To account for this finding, Anderson and Bower posited that instead of a single general activation, items are associated with specific elements of the study context. The "old/new" judgment is then mediated by the degree to which the test stimuli are associated with specific elements of the target context (i.e., relative to some particular target list).

More recently, models of recognition have posited two separate processes, where both "know" and "remember" judgements are made (Mander, 1980; Rajaram, 1993; Tulving, 1985; but see Hirshman & Henzler, 1998 for a single process account). "Know" judgements are based on familiarity alone: fast, automatic judgements that an item was on the study list, but without a recollection of its actual occurrence. "Remember" judgements are slower and more deliberate, producing explicit memories of having seen the specific item on the study list in question. It is this dual process of recognition that we propose underlies recognition in the guilty knowledge paradigm (Seymour et al, 1997).

A Model of the Guilty Knowledge Effect

Figure 3 depicts, our model of the processes involved in the Guilty Knowledge Effect in terms of standard recognition memory processes. The two most important stages of the model are the Familiarity and Context stages. In our model, the familiarity judgement ("Know") is described by the standard signal detection model, with the addition of an

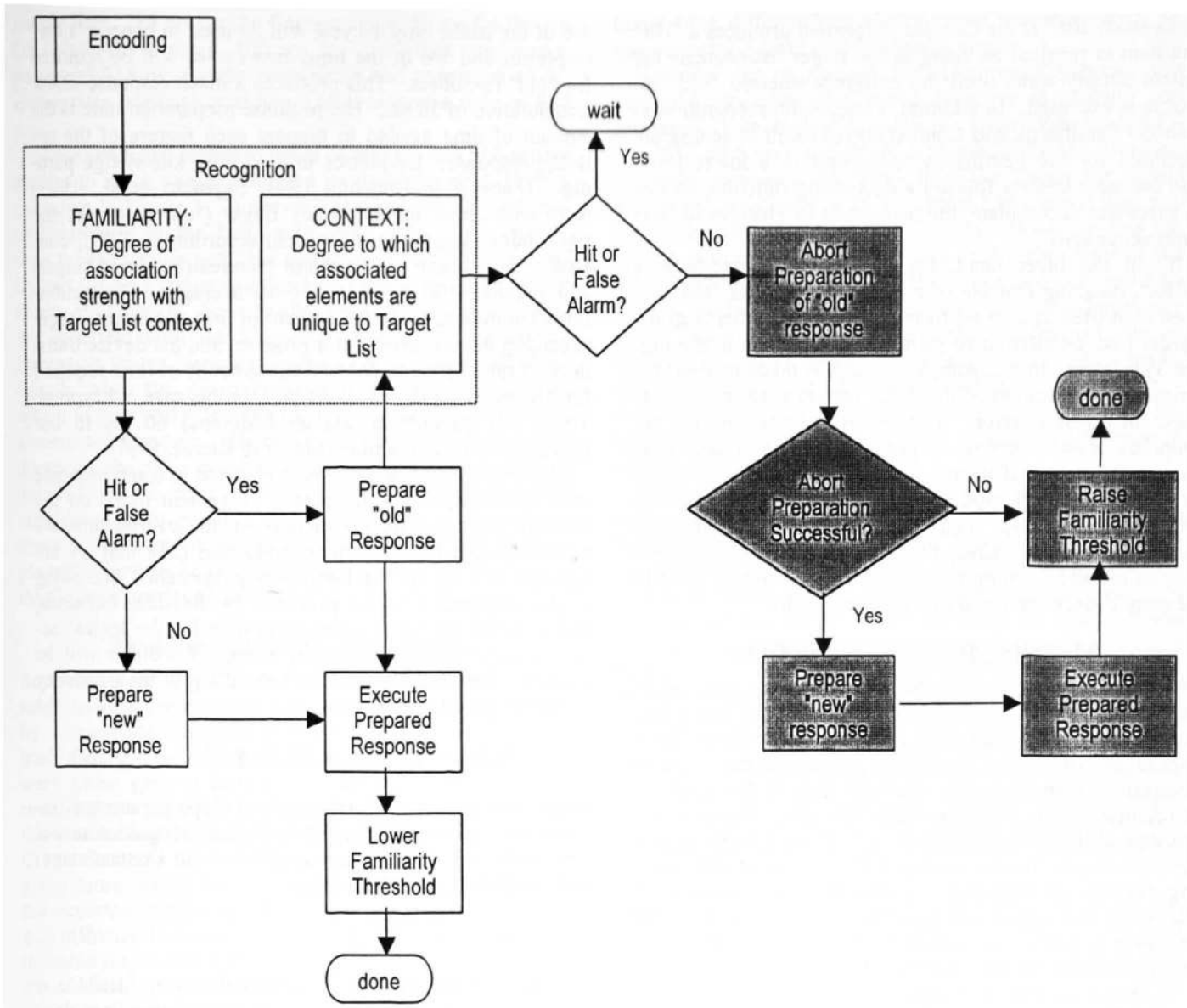

Figure 3. Model of Guilty Knowledge Effect.

evaluation of the degree to which the test item is associated with the study list context (as proposed by Anderson and Bower, 1972). This is accomplished without identifying a specific source context. After our Familiarity stage, while Response Preparation begins, the Context stage commences. The Context stage is also described by a signal detection process where the differentiation is between a composite of contextual elements associate with the test item and a composite of contextual elements associated with the study context. This process requires resolving the contextual associative links of the test item; consequently, this judgement requires more time when the target and distractor contexts are similar. We believe that it is only at this second stage that list-discrimination can occur.

Combining these components, we can walk through the model during the recognition judgement that produces the Guilty Knowledge Effect. On each trial of the List Priming Task described above, subjects are presented with a stimulus item. After encoding, this stimulus is subject to a familiarity judgement, which decides whether an "old" or a "new" response will be prepared. Because the activation of "old" and "new" items are assumed to be equally distributed with similar variances, the judgement made in the Familiarity stage is fallible (Green & Swets, 1966; Kintch, 1970; Klatzky, 1980; Johnson, 1996; Tanner & Swets, 1954) (may produce either a Hit or False Alarm). Its completion initiates the Response Preparation process, and once completed, its output motor program will be automatically executed. Concurrently, while preparing an "old" response, the Context stage checks whether the "Know" judgement that has been made is in fact a "Remember" judgement. That is, the Context stage checks to see if the test item comes from the list (context) in question -- the Target list -- or from some other list having similar items and context; for example, the

crime-study list. If the Context judgement produces a "Hit" (the item is recalled as being in the target list context) the system simply waits until the currently selected "old" response is executed. In addition, a successful recognition in both the Familiarity and Context stages result in setting the threshold for the Familiarity judgement to a lower level. The decrease in Beta follows a decreasing function, so that as successes accumulate, the magnitude of changes to Beta approaches zero.

If, on the other hand, the Context stage produces a "Miss", meaning that the stimulus just judged as "old" has been identified as coming from some context other than the Target List, an alternative path (the shaded path in the Figure 3) is taken. In this path, an attempt is made to abort the erroneous preparation of the "old" response already underway, since it is in error. If the abort is successful, the appropriate "new" response is prepared and eventually executed. Otherwise, if Response Preparation is too far along or if Response Execution has begun, the erroneous response will not be prevented. As a result of traversing this alternate path, the RT will increase. Finally, the Familiarity threshold will be raised according to a decreasing function in order to be more conservative and avoid further errors.

Modeling Response Time Data

The main predictions from this model are that, because of the visual and contextual similarity between the Target List (study) and the Probe List (crime), Guilty-Probe items should lead to positive Familiarity judgements and negative Context judgements -- the alternate path in the model -- much more often than any other item type. Furthermore, because of the stochastic nature of both the Familiarity and Context stages, the Abort stage will be initiated with varying degrees of Response Preparation already completed. Therefore, the system will sometimes fail to abort the "old" response in time, and should lead to more errors during Guilty-Probe trials than other trial types.

In order to capture the general RT result we used the Executive Process Interactive Control (EPIC) model of human multiple-task performance (see Meyer & Kieras (1997) for an overview) to determine parameters for the response selection and motor stages of our task. Table 1 shows the relevant EPIC response execution parameters in milliseconds.

Table 1: EPIC Response Execution Parameters.

Parameter Name	Mean Value
Cognitive processor cycle duration	50 ms
Response preparation time per feature	50 ms
Action-initiation time	50 ms
Device transduction time	10 ms

The first parameter, cycle duration is essentially the duration of a single decision cycle and is necessary to determine the response selection time. For simplicity, we will posit a response selection process for our "old/new" judgements without repetition effects (due to the fact that "old" responses are only correct on 1/6 of the trials). Therefore, on 5/6 of the trials, only 1 cycle will be used to choose "new" responses and 1/6 of the time, two cycles will be required for "old" responses. This produces a mean response selection duration of 58 ms. The response preparation time is the amount of time needed to prepare each feature of the selected response. Responses in the guilty knowledge paradigm (Farwell & Donchin, 1991; Seymour et al., 1997) were with either the left-index finger ("old") or with the right-index finger ("new"), which, according to EPIC, consists of programming two motor features (hand and finger) and requires 100 milliseconds on average. The action-initiation time refers to the amount of time required to begin executing the prepared motor program and the device transduction time refers to the average amount of time required for the recording device to register, in this case, a keypress. These two parameters add an additional 60 ms to our movement production time (Meyer & Kieras, 1997).

Meyer & Schvaneveldt (1976) reported average response time for a two-choice familiarity judgement to be on the order of 550 ms. By subtracting out the overall time we have indicated for movement production (218 ms) we are left with 332 ms for the familiarity judgement. According to the recognition model proposed by Reichle, Pollatsek, Fisher & Rayner (1998), the mean time for the lexical access (similar to our Familiarity stage) of word n will be equal to a linear function of the natural log of the frequency of that word. More formally:

$$t(f_n) = f_b - (f_m \bullet \ln(freq_n)) \qquad (1)$$

Where b and m are the intercept and slope parameters, respectively. A second component of their recognition model represents lexical completion (similar to our Context stage) and is a constant multiple of $t(f_n)$:

$$t(lc_n) = \Delta \bullet t(f_n) \qquad (2)$$

Where Δ is a fixed parameter greater than zero. Reichle et al. found that a Δ of 0.65 produced a reasonable fit to their recognition data, and when we substitute the 332 ms derived for our familiarity judgements into the $t(f_n)$ term of Equation 2, our Context stage requires approximately 218 ms.

The remaining stages "Abort Preparation - Successful" and "Hit Or False Alarm" are single decision stages, and are assumed to take one cognitive cycle (50ms each). The final stage, "Abort Preparation Of 'Yes' Response," is presumed to take two decision cycles if successful (the first checks if an abort is possible and the second effects the abort) and one decision cycle if unsuccessful, or 100 ms and 50 ms respectively.

With these latencies assumed in our model, responses to three of the item types -- Guilty-Irrelevant, Innocent-Irrelevant and Innocent-Probe items -- should take approximately 600 ms on average. This estimate represents the simplest path through the model, requiring 332 ms for the Familiarity Stage, 50 ms for the "Hit Or False Alarm" stage, 218 ms for the "Response Preparation" and "Response Execution" stages together. However, for Guilty-Probe responses where the "Abort Preparation" stage is successful, the path is composed as follows: 600 ms for the

simple path, 218 ms for the Context stage, 50 ms for the Hit Or False Alarm stage, 100 ms for the Abort Preparation stage and 218 ms for the preparation and execution of the appropriate response, yielding a total response time of 968 ms.

These predicted response times of 600 ms and 968 ms are similar to mean values observed by Seymour et al. (1997), as shown in Figure 1. However, we point out that not only will variation in the familiarity stage affect the response time for the simple path, but variation in both the Familiarity and Context stages will produce variability in the RT for Guilty-Probe trials. Specifically, because the Context stage is concurrent with the Response Selection and Response Execution stages, variation in the Context stage will affect exactly when the Abort Preparation process is attempted, which in turn will effect the likelihood and time to produce a corrected response. Therefore, although the critical path durations are close to those reported by Seymour et al., a detailed analysis must take stage variability and overlap into consideration to provide appropriate ranges of response times. However, in broad strokes, this analysis paints a plausible picture of the recognition processes involved in the Guilty Knowledge Effect.

Modeling Strategic Differences

Seymour et al. (1997), conducted 2 experiments where subjects were given instructions about how to "beat" the test by suppressing response differences on the Guilty-Probe trials (i.e., appear "innocent"). In one experiment, subjects were given general instructions about the task, including revealing the presence of the Probe stimulus category and the relationship between the List Priming Task and the Crime Scenario. In a second experiment, subjects were given more detailed information about the task, including the expected patterns of results and warnings about strategies that would lead to failure or easy detection (e.g., giving the same response to both "old" and "new" items, or missing the deadline). Subjects were also given an incentive to appear "innocent" during Guilty trials: They were told that the experiment would end sooner if they succeeded in masking their knowledge of the crime. Because of the repetitive nature of the task, subjects found this incentive quite compelling.

The results showed that, in both experiments, subjects were unable to produce response times to Guilty-Probe items that looked similar to responses to Irrelevant items. In fact, Guilty Knowledge Effect observed was nearly identical in RT and magnitude to uninformed subjects. The difficulty of strategically manipulating response times under 800ms has been used to characterize processes that occur automatically (Ratcliff & McKoon, 1981). Because appearing "innocent" during Guilty-Probe trials involves producing RTs of approximately 600 ms and requires strategic processes which are not available for at least 800ms, it is not surprising that such attempts are unsuccessful. Similarly, the model we have outlined involves mostly automatic processes with stages supporting strategic intervention arising only after the bulk of processing has already occurred and thereby adding additional time to the response. This additional time differentiates Guilty-Probe responses from responses arising from only automatic stages.

However, the recognition model proposed suggests how particular strategies may prove effective in altering performance. One method to mask the Guilty Knowledge Effect is to set a very conservative threshold for the Familiarity stage, reasoning that items from the Target list should have more association with the Target List context than will any distractors, including Probe items. However, this approach is made more difficult due to the similarity of the Target and Probe items, and of the Crime Scenario and Target List study contexts. The high degree of similarity may make Target and Guilty-Probe items indistinguishable on the basis of context association strength alone. Any measures that increase the distinctiveness of the two contexts -- such as a long time delay between the crime context and the target study context -- will facilitate systematic difference in context association.

Another strategy involves suspending the Response Execution stage until the source context for the test items has been verified in the Context stage. This strategy amounts to avoiding the Guilty Knowledge Effect by simply making sure to accurately classify Guilty-Probe items as "new." In Seymour et al., (1997) mean accuracy for Guilty-Probe items is considerably improved when subjects are motivated to appear "innocent." However, their response times still reliably reveal their difficulty with Guilty-Probe items. In general, any strategy that involves judging familiarity in advance of verifying that the item is from the Target context will, necessitate additional time to halt and replace the prepared response already underway. These additional stages (marked in gray in Figure 3) in Guilty-Probe trials will foil attempts to respond as in Irrelevant and Innocent-Probe trials.

One strategy suggested by the model is to avoid the initiation of the Response Preparation and Response Execution stages until the source of the stimulus item has been verified in the Context stage. In this way, there is no need to abort an erroneous motor program when a "False Alarm" is detected in the Context stage. This strategy would, using logic from the previous section, produce response times on the order of 767ms for all stimulus types and therefore typically occur before the response deadline. Though these processes appear to proceed automatically, it may be possible to avoid initiating them until the Context assessment has been completed. Whether or not subjects can actually use this strategy is an empirical question, though it is clear that it was not spontaneously used in Seymour et al. (1997).

Conclusion

We have proposed a model of the Guilty Knowledge Effect, where correct rejection of items related to a prior crime takes longer than for other, irrelevant items (Farwell & Donchin, 1991; Seymour et al., 1997). This model builds upon an existing theory of human performance (Meyer & Kieras, 1997) and prior theories of recognition (Anderson & Bower 1972; Kintch 1970; Mandler, 1980; Rajaram, 1993; Tulving, 1985). This model predicts response time difference that closely correspond to observed response times (Seymour, et al., 1997). We have also shown that the model

rules out certain strategies aimed at attenuating the Guilty Knowledge Effect, and suggests potentially more successful ones. Because it can account for absolute and relative differences in mean response time for Guilty compared to Innocent subjects, the model suggests a promising direction for theories of memory recognition.

Acknowledgments

We would like to thank Andrea Mosmann, Bill Gehring, Gail McKoon and Roger Ratcliff for their help and comments on this research.

References

Anderson, J. R. & Bower, G. H. (1972). Configurational properties in sentence memory. *Journal of Verbal Learning and Verbal Behavior, 11*, 594-605.

Bashore, T. R., & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin, 113(1)*, 3-22.

Farwell, L. A. & Donchin, E. (1991). The Truth will out: Interrogative Polygraphy ("Lie Detection") with event-related brain potentials. *Psychophysiology, 28(5)*, 531-547.

Furedy, J. J. & Heslegrave, R. J. (1988). Validity of the lie detector: A psychophysiological perspective. *Criminal Justice & Behavior, 15(2)*, 219-246.

Green, D., & Swets, J. (1966). Signal detection theory and psychophysics. New York: John Wiley & Sons.

Hirshman E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9(1)*, 61-65.

Johnson, M. K. (1996). Some problems with the process-dissociation approach to memory. *Journal of Experimental Psychology: General, 125(2)*, 181-194.

Kintsch, W. (1970). *Learning, memory and conceptual processes*. New York: Wiley.

Klatzky, R. (1980). *Human Memory* (pp. 237-263). San Francisco: W.H. Freeman.

Lykken, D. T. (1981). *A Tremor In The Blood: Uses and Abuses of the Lie Detector*. New York: McGraw Hill.

Mandler, G. (1980). Recognizing: The judgement of previous occurrence. *Psychological Review, 87*, 252-271.

Meyer, D. E., & Schvaneveldt, R. W., (1976). Meaning, memory structure, and mental processes. In Cofer, C. (Ed.) *The structure of human memory* (pp. 55-89). San Francisco: W. H. Freeman.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104*, 3-65.

Reichle, E. Pollatsek, A., Fisher, & D., Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105(1)*, 125-157.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21*, 89-102.

Ratcliff, R., & McKoon, G. (1981). Automatic and strategic priming in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*, 204-215.

Seymour, T. L., Mosmann, A. L., & Seifert, C. M. (1997). Using Reaction Time to Measure "Guilty Knowledge". *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 1044). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40*, 385-398.