Probabilistic Modelling of Cognitive Processes in the Remember-Know Paradigm

Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem
Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von Dipl.-Psych. Christine Sattler
geboren am 11.08.1974 in Hamburg
Gutachter

1. Prof. Dr. Thorsten Meiser, Marburg
2. Prof. Dr. Klaus Rothermund, Jena

Tag des Kolloquiums: 30.10.2007
Acknowledgments

I am much obliged to my supervisor Thorsten Meiser, who accompanied this project through all its phases. I thank Maya Machunsky for helpful comments on chapters of this thesis.

I am grateful to Wolfgang Prinz for the possibility to finish this thesis in Leipzig.

In particular, I thank Alexandra Löffler and Thomas Schrader for their friendship and emotional support during these years.
# Contents

1  Introduction ................................................................................................................. 7

2  Explaining Remember-Know Judgments ........................................................................ 12  
   2.1  Distinctiveness-Fluency Framework ...................................................................... 12 
   2.2  Multiple Memory Systems ................................................................................... 14 
   2.3  Unidimensional Signal-Detection Model ............................................................. 16 

3  Critique ....................................................................................................................... 20  
   3.1  Old-New Response Bias ...................................................................................... 22 
   3.2  Recognition Memory Performance ..................................................................... 26 

4  Problem Formulation and Research Goal .............................................................. 28 

5  Probabilistic Modelling of the Remember-Know-(Guess) Paradigm ......................... 30  
   5.1  Multinomial Processing Tree Models .................................................................... 30 
   5.2  Measurement Models for the Remember-Know Paradigm .................................... 34  
      5.2.1  Task-Oriented Model ................................................................................... 34 
      5.2.2  Multi-Memory Model .................................................................................. 37 
   5.3  Measurement Models for the Remember-Know-Guess Paradigm .......................... 39  
      5.3.1  Task-Oriented Model ................................................................................... 39 
      5.3.2  Multi-memory model ................................................................................... 42 

6  Evaluation of the Measurement Models ................................................................... 45  
   6.1  Strategy of Model Evaluation .............................................................................. 45 
   6.2  Remember-Know Paradigm ................................................................................. 48  
      6.2.1  Database ....................................................................................................... 48 
      6.2.2  Results and Discussion ............................................................................... 49 
      6.2.3  Conclusion ................................................................................................... 57
<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3 Remember-Know-Guess Paradigm</td>
</tr>
<tr>
<td>6.3.1 Database</td>
</tr>
<tr>
<td>6.3.2 Results and Discussion</td>
</tr>
<tr>
<td>6.3.3 Conclusion</td>
</tr>
<tr>
<td>7 Experiments Manipulating Old-New Response Bias</td>
</tr>
<tr>
<td>7.1 Experiment 1</td>
</tr>
<tr>
<td>7.1.1 Method</td>
</tr>
<tr>
<td>7.1.2 Results and Discussion</td>
</tr>
<tr>
<td>7.1.3 Summary</td>
</tr>
<tr>
<td>7.2 Experiment 2</td>
</tr>
<tr>
<td>7.2.1 Method</td>
</tr>
<tr>
<td>7.2.2 Results and Discussion</td>
</tr>
<tr>
<td>7.2.3 Summary</td>
</tr>
<tr>
<td>7.3 Discussion of Experiments 1 and 2</td>
</tr>
<tr>
<td>7.4 Evaluation of the Measurement Models with the Extended Database</td>
</tr>
<tr>
<td>7.4.1 Remember-Know Paradigm</td>
</tr>
<tr>
<td>7.4.2 Remember-Know-Guess Paradigm</td>
</tr>
<tr>
<td>7.4.3 Conclusion</td>
</tr>
<tr>
<td>8 Experiments Manipulating Recognition Memory Performance</td>
</tr>
<tr>
<td>8.1 Experiment 3</td>
</tr>
<tr>
<td>8.1.1 Method</td>
</tr>
<tr>
<td>8.1.2 Results and Discussion</td>
</tr>
<tr>
<td>8.2 Experiment 4</td>
</tr>
<tr>
<td>8.2.1 Method</td>
</tr>
<tr>
<td>8.2.2 Results and Discussion</td>
</tr>
<tr>
<td>8.3 Discussion of Experiments 3 and 4</td>
</tr>
</tbody>
</table>
9 General Discussion ........................................................................................................... 111

  9.1 Overview ................................................................................................................ 111

  9.2 Remember-Know Judgments and the Retrieval of Context Features................ 115

  9.3 Judgment Processes in the Remember-Know Paradigm .................................... 120

  9.4 Remember-Know Judgments and Recognition Memory Performance .......... 126

  9.5 Conclusion ............................................................................................................. 130

References ....................................................................................................................... 132

Appendix .......................................................................................................................... 141

  A. Model Equations and Proofs of Identifiability ..................................................... 141
      A.1 Task-Oriented Models for the Remember-Know Paradigm ......................... 141
      A.2 Task-Oriented Models for the Remember-Know-Guess Paradigm ............ 142

  B. Remember-Know-(Guess) Instructions ............................................................... 146
      B.1 Experiment 1 ................................................................................................. 146
      B.2 Experiment 2 ................................................................................................. 147
      B.3 Experiment 3 ................................................................................................. 148
      B.4 Experiment 4 ................................................................................................. 149

Summary ......................................................................................................................... 151

Zusammenfassung .......................................................................................................... 153

Ehrenwörtliche Erklärung .............................................................................................. 156
1 Introduction

The remember-know paradigm investigates the subjective experience that accompanies retrieving information from memory. The occasion for retrieving information from memory is typically created by a recognition memory test, which consists of a mixture of items that have been presented in a previous study phase (i.e., targets) and items that have not been presented in the study phase (i.e., distractors). For each test item that participants report to recognise from the study phase, they are asked to classify their retrieval experience as remember versus know. The term remember denotes the ability to mentally re-experience a previous event, here the occurrence of the item in the study phase. That is, one is able to become consciously aware again of one or more aspects of the past episode of studying the item. For example, one may recollect perceptual features of the item, one's thoughts or feelings, or something else that co-occurred with the item presentation. In contrast, the term know means that the item is confidently recognised, but does not bring back to mind any aspect of the item's prior occurrence. That is, the item simply appears familiar in the absence of a specific recollection.

Since the remember-know paradigm was introduced by Tulving (1985b), it has been used in numerous recognition memory experiments (for reviews, see Dunn, 2004; Gardiner & Java, 1993a, 1993b; Gardiner & Richardson-Klavehn, 2000; Rajaram, 1999; Rajaram & Roediger, 1997). Even though it is likewise possible to ask for remember-know judgments in free recall or cued recall tests, this has hardly ever been done (for exceptions see Hamilton & Rajaram, 2003; Tulving, 1985b).

Research with the remember-know paradigm has been conducted in pursuit of different overarching goals. In one approach, the remember-know paradigm is used in order "to understand the nature of the subjective experience through a first-person account" (Rajaram, 1999, p. 261) and the ultimate goal is to explain "the nature of retrieval experience per se" (Rajaram, 1999, p. 262). In another approach to the remember-know paradigm, the ultimate goal is to explain recognition memory performance rather than retrieval experience. For example, Gardiner and Java (1990) stated:

The main purpose of the present research was to provide further tests of the "dual-component hypothesis", that is, the hypothesis that these measures of consciousness [i.e., remember and know response rates] reflect qualitatively
distinct components of memory performance, rather than some unitary dimension. (p. 24)

That is, in this approach remember-know judgments are used as a tool for isolating two hypothetical bases of recognition memory performance. A comprehensive review of dual-process models of recognition memory performance and of measurement methods used for isolating the postulated processes is given by Yonelinas (2002). While dual-process models are very popular for explaining recognition memory performance, single-process models (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) constitute an important alternative. These models suggest that all items in a recognition memory test can be ordered along a single dimension which captures the strength of evidence available from memory in favour of their occurrence in the study phase, and suggest a mechanism how the strength-of-evidence values of the test items are generated. Importantly, it has been claimed that the data obtained with the remember-know paradigm are compatible with single-process models (e.g., Donaldson, 1996; Dunn, 2004; Hirshman & Master, 1997). If this claim is correct, the dual-process interpretation of remember-know responses advanced by others (e.g., Gardiner, 1988; Gardiner & Parkin, 1990; Rajaram, 1993; Tulving, 1985b) is not cogent.

Regardless of whether the research interest is directed at retrieval experience per se or at retrieval experience as a tool for isolating different bases of recognition memory performance, data analyses in the remember-know paradigm have typically been based on the rates of remember and know responses, with the focus on remember and know responses to targets. A large body of empirical evidence shows experimental dissociations between remember and know response rates to targets. Some experimental manipulations affect remember response rates, but not know response rates, whereas others affect know response rates, but not remember response rates. Other variables produce opposite effects on remember and know response rates. These patterns of results amount to a double dissociation of remember and know response rates.

Experimental manipulations that have been found to selectively affect the remember response rate include low as opposed to high word frequency (Gardiner & Java, 1990; Kinoshita, 1995; Strack & Förster, 1995), full versus divided attention at study (Gardiner & Parkin, 1990; Parkin, Gardiner, & Rosser, 1995), or generating as opposed to reading items at study (Dewhurst & Hitch, 1999; Gardiner, 1988; but see Wippich, 1992). Selective effects on the know response rate have been reported for modality match as
opposed to mismatch across study and test following a highly perceptual orienting task (Gregg & Gardiner, 1994), for an unattended additional presentation of the item between study and test (Mäntylä & Raudsepp, 1996), and for different priming manipulations at test (Rajaram, 1993; Rajaram & Geraci, 2000). Compared to test items that were preceded by an unrelated prime, a selective increase in the know response rate was observed for test items preceded by an identical prime (Rajaram, 1993) or by a semantically related prime (Rajaram & Geraci, 2000). Examples of variables that have been found to produce opposite effects on remember and know response rates complete this illustrative enumeration. Nonwords received fewer remember responses than words, but more know responses (Gardiner & Java, 1990). Test words that were studied as pictures received more remember responses than test items that were studied as words, but fewer know responses (Dewhurst & Conway, 1994; Rajaram, 1993). Highly imageable words received more remember responses than poorly imageable words, but fewer know responses (Dewhurst & Conway, 1994).

Various theoretical frameworks have been suggested to account for this kind of findings. In an endeavour to explain "the nature of retrieval experience per se" (Rajaram, 1996, p. 262) by capturing the common gist of those variables that influence remember and know response rates, respectively, Rajaram (1996, 1998, 1999) proposed the distinctiveness-fluency framework. According to this taxonomy, remember judgments are affected by variables that provide distinctive or salient information, whereas know judgments are influenced by variables that affect the fluency or ease of information processing.

In a different approach that started from a theoretical conception of the organisation of memory, Tulving (1985b, 1993) suggested that remember and know responses reflect retrieval from two different memory systems, episodic and semantic memory. In particular, he assumed that retrieving information about a personally experienced event from episodic memory is accompanied by a distinctive state of awareness that is called autonoetic awareness and captured by remember responses. If no relevant information can be retrieved from episodic memory, it may nevertheless be possible to retrieve relevant information from semantic memory. If information is only retrieved from semantic memory, it is accompanied by a different state of awareness that is called noetic awareness and captured by know responses. That is, according to this conception, the contribution of the episodic memory system to recognition memory performance is indicated by remember
responses, whereas the additional contribution to recognition memory performance produced by the semantic memory system is indicated by know responses. In other words, the concept of multiple memory systems belongs to the class of dual-process models of recognition memory performance, and retrieval experience provides a tool for isolating the memory systems.

However, as mentioned above, other theorists have claimed that the remember-know data can be explained by assuming a single strength-of-evidence continuum as the only basis of recognition memory performance (e.g., Donaldson, 1996; Dunn, 2004; Hirshman & Master, 1997). This unidimensional signal-detection account implies that remember-know judgments are equivalent to confidence ratings, such that the remember-know paradigm cannot yield insights over and above conventional confidence rating experiments. Accordingly, from this perspective, it is not of interest to investigate retrieval experience.

It is noteworthy that both approaches that do consider it worthwhile to investigate retrieval experience, either in order to isolate different theoretical constructs that are assumed to jointly produce recognition memory performance or in order to understand the retrieval experience per se, draw on on the same measures of retrieval experience, even though these approaches imply partially conflicting requirements for a suitable measure. If the goal is to isolate different theoretical constructs that simultaneously mediate a state of retrieval experience and contribute to recognition memory performance, it suggests itself to employ measures that capture these aspects simultaneously. At the theoretical level, such measures represent the assumed constructs, episodic and semantic memory. At the descriptive level, such measures are two-in-one measures of retrieval experience and a component of recognition memory performance. Because an increase in overall recognition performance is mediated by a higher contribution of at least one memory system, measures of these memory systems are not independent of overall recognition memory performance. In contrast, if the goal is to investigate retrieval experience per se, a pure measure of retrieval experience is indispensable, which is required to be independent of recognition memory performance.

On the basis of arguments and empirical evidence that suggest that remember-know response rates are not suited as measures of retrieval experience for either purpose, this thesis investigates alternative measures and demonstrates empirically that the traditional
The analysis of remember and know response rates can lead to erroneous conclusions about the effect of experimental manipulations.

The following chapter 2 introduces the explanations of remember-know judgments in more detail. Chapter 3 is devoted to the issue of whether the rates of remember and know responses are suitable for investigating the retrieval experience per se or for isolating theoretical constructs that jointly produce recognition memory performance. Considerations of requirements for suitable measures for either purpose on the one hand and of features of remember and know response rates on the other hand, in conjunction with a review of empirical evidence, lead to the conclusion that remember and know response rates are not suited for either purpose. The problems persist if participants are allowed to classify their retrieval experience as guessing in addition to remembering and knowing, an addition that yields the remember-know-guess paradigm. Therefore, the goal of the present thesis was to find suitable measures of retrieval experience, as pointed out in chapter 4.

The alternative measures that are investigated in this thesis are formulated within multinomial processing tree models. After an overview of the methodology of multinomial modelling, measurement models for the remember-know paradigm and the remember-know-guess paradigm are introduced in chapter 5. Chapter 6 presents a formal evaluation of these measurement models by means of reanalyses of published datasets. While the reanalyses suggested one of the measurement models for the remember-know and remember-know-guess paradigm as most promising, neither model proved to be globally valid. Therefore, as reported in chapter 7, the database was extended by two new experiments, in which a manipulation of old-new response bias was crossed with whether the response option guess was available for classifying retrieval experience. Drawing on the extended database, one measurement model for the remember-know paradigm turned out to be globally valid. As described in chapter 8, this measurement model was used to demonstrate empirically that the traditional analysis of remember and know response rates can lead to erroneous conclusions about the effect of independent variables on retrieval experience. Participants studied items with full versus divided attention in Experiment 3, and Experiment 4 compared semantic to phonemic processing. A general discussion in chapter 9 concludes the thesis.
2 Explaining Remember-Know Judgments

2.1 Distinctiveness-Fluency Framework

The distinctiveness-fluency framework (Rajaram, 1996, 1998, 1999; Rajaram & Geraci, 2000) aims to explain retrieval experience by capturing the common gist of those variables that influence remember and know response rates, respectively. In Rajaram's (1998) own words:

The distinctiveness-fluency framework is drawn from previous accounts of memory performance (see Hunt & McDaniel, 1993; Hunt & Mitchell, 1982; Jacoby & Dallas, 1981; Luo, 1993) and is applied here to account for the retrieval experience. In other words, this framework is designed to predict and explain why different experimental manipulations produce different states of subjective experience at retrieval. Within this framework, processing of salient or distinctive attributes of the stimuli, which may be conceptual or perceptual in nature, leads to the experience of remembering at retrieval. In contrast, the retrieval experience of knowing is influenced by the fluency with which perceptual or conceptual information is processed. (p. 72)

Rajaram took care to make it very clear that her aim was to explain "the nature of retrieval experience per se" (Rajaram, 1999, p. 262), but not recognition memory performance: "As a point of clarification, it should be noted that the distinctiveness-fluency framework is not designed to explain how these two classes of underlying processes combine to give rise to overall memory performance" (Rajaram, 1996, p. 74).

An earlier line of research that used retrieval experience as a tool for isolating different hypothetical bases of recognition memory performance had used the distinction between conceptual and perceptual processing as a taxonomy of experimental variables that influence remember and know response rates, respectively (e.g., Gardiner & Parkin, 1990; Rajaram, 1993). Specifically, the idea that guided early research with the remember-know paradigm was that conceptual processing selectively influences remember responses, whereas perceptual processing selectively influences know responses.

In a direct test of the claim that remember responses are not influenced by perceptual manipulations, Rajaram (1996) found evidence to the contrary. In two experiments, she
used line drawings of objects as stimuli. At test, the drawings were presented either in the same versus a different size (Experiment 2) or with the same versus reversed left-right orientation (Experiment 3). Changing these perceptual features from study to test reduced the remember response rate as compared to the perceptually identical condition. In an effort to account for these results as well as extant data, Rajaram (1996) put forward the distinctiveness-fluency framework. She argued that manipulations that had previously been shown to increase the remember response rate, like performing a semantic rather than a phonemic orienting task (e.g., Gardiner, 1988, Experiment 1) or generating the target word at study in response to a semantic cue rather than simply reading the cue-target pair (e.g., Gardiner, 1988, Experiment 2), not only differ with regard to the predominantly conceptual or perceptual character of the processing induced at study, but also increase the probability that distinctive or salient features of the study items are processed. Likewise, she reasoned that manipulations that had been shown to selectively increase the know response rate, like masked repetition priming at test (Rajaram, 1993) or modality match across study and test following adverse study conditions which by virtue of the task design minimised the probability of non-perceptual processing (Gregg & Gardiner, 1994), can reasonably be assumed to increase the processing fluency.

Rajaram (1998) examined the distinctiveness component of the distinctiveness-fluency framework by contrasting stimuli that had a priori been determined as more versus less perceptually distinctive. As predicted, orthographically distinctive words (e.g., subpoena) selectively received more remember responses than orthographically common words (e.g., cookie). Importantly, orthographically distinctive and common words were matched for word frequency, which is known to selectively influence remember responses as well (Gardiner & Java, 1990; Kinoshita, 1995; Strack & Förster, 1995). In a related study, Dewhurst and Hitch (1999) increased the distinctiveness of encoding events by using orienting tasks that required more cognitive effort. Participants simply read the study items aloud or had to generate them from easy or difficult anagrams, which differed in the number of misplaced letters out of five. The remember response rate increased from read items through easy anagrams to difficult anagrams, whereas the know response rate was not affected, in line with the predictions of the distinctiveness-fluency framework.

Rajaram and Geraci (2000) examined the fluency component of the distinctiveness-fluency framework by preceding each test item with a prime that was either semantically related or unrelated to the test item. As predicted, this conceptual manipulation of
processing fluency selectively affected the know response rate to targets and distractors. This finding complements earlier evidence that preceding test items with a masked presentation of the same versus an unrelated word selectively affected the know response rate to targets and distractors (Rajaram, 1993). That is, both a perceptual and a conceptual manipulation of processing fluency selectively affected know responses.

Altogether, the available evidence is largely consistent with the suggestion of the distinctiveness-fluency framework that the processing of distinctive or salient aspects at study influences the remember response rate, whereas the fluency of processing at test influences the know response rate. However, note that the effect of distinctiveness manipulations is not always limited to the remember response rate, but may be accompanied by a change of the know response rate in the opposite direction. For example, highly imageable words as opposed to poorly imageable words received more remember responses and fewer know responses (Dewhurst & Conway, 1994).

### 2.2 Multiple Memory Systems

According to Tulving (1985b, 1993), remember and know responses reflect retrieval from episodic and semantic memory, respectively. These are two of five memory systems which form the components of Tulving's model of the overall organisation of human memory (e.g., Schacter & Tulving, 1994, Tulving, 1984, 1985a, 1985b, 1995). A memory system is conceived as a neurocognitive system that serves specialised functions, thereby making possible performance on a large class of tasks. Among the properties of a memory system are the kind of information it processes, rules of operation, and neural substrates. The following paragraphs specify how the episodic and semantic memory systems are conceived in Tulving's model.

The crucial feature of the episodic memory system is that it makes possible mental time travelling (Tulving, 1985b, 1993, 2002; Wheeler, Stuss, & Tulving, 1997). The episodic memory system stores information about personally experienced events. Retrieval from episodic memory entails the subjective experience of mentally re-experiencing a past event – or travelling back in subjective time to a previous episode – where one is aware of the event as a veridical part of one's own past existence. This particular state of awareness, which necessarily accompanies retrieval from episodic memory, is called autonoetic awareness.
In contrast to the episodic memory, the semantic memory system deals with general factual knowledge of the world (Tulving, 1985b, 1995). Though the name suggests otherwise, it is not tied to language or meaning, but can contain anything that can be represented propositionally. Retrieval from semantic memory is accompanied by noetic awareness. Thereby, semantic memory makes it possible to cognitively operate on aspects of the world that are not currently present.

Importantly, not only the episodic memory, but also the semantic memory can process information about personally experienced events (Tulving, 1985b). Specifically, according to the serial-parallel-independent conception (Tulving, 1995), which is concerned with the relations among memory systems, successful processing by the semantic memory system is a necessary prerequisite for encoding information into the episodic memory system. Given successful encoding, information pertaining to a particular episode can be stored in both memory systems in parallel, where each memory system stores a different kind of information according to its function. Retrieval of information from one memory system is independent of retrieving the corresponding information from the other memory system. Thus, in case of episodic memory retrieval failure, it may nevertheless be possible to retrieve information pertinent to the target episode from semantic memory, which is accompanied by noetic awareness instead of autonoetic awareness.

Tulving (1985b) derived from his model of memory organisation the suggestion to use the subjective experience that accompanies memory retrieval, captured by binary remember-know responses, to draw inferences about the extent to which episodic memory is involved in performing a given experimental task. In particular, remember responses in recognition memory experiments are assumed to reflect autonoetic awareness of the episode of the item's occurrence in the study list, thereby indicating retrieval from episodic memory. Conversely, know responses are assumed to reflect noetic awareness of the item, thereby indicating retrieval from semantic memory but not episodic memory.

That is, the remember and know response rates are assumed to represent the theoretical constructs of episodic and semantic memory, respectively. Importantly, these two memory systems not only mediate a characteristic state of awareness, but also jointly produce the recognition memory performance. The contribution of the episodic memory system to recognition memory performance is indicated by remember responses, whereas the additional contribution to overall recognition memory performance produced by the semantic memory system is indicated by know responses. Therefore, the concept of
multiple memory systems belongs to the class of dual-process models of recognition memory performance.

If remember and know responses reflect functionally different memory systems, it should be possible to selectively affect retrieval from the episodic memory system, but not the semantic memory system and vice versa, which would be reflected in experimental dissociations of remember and know response rates. Such experimental dissociations have indeed been observed, as reported in previous sections of this thesis, and Gardiner and colleagues (e.g., Gardiner, 1988; Gardiner & Java, 1990; Gardiner & Richardson-Klavehn, 2000) interpreted them as evidence in support of multiple memory systems.

### 2.3 Unidimensional Signal-Detection Model

According to the unidimensional signal-detection model account (e.g., Donaldson, 1996; Dunn, 2004; Hirshman & Master, 1997), remember and know responses reflect different levels of confidence rather than distinct states of awareness and are based on the strength of a memory signal along a single continuum plus a decision process.

The unidimensional signal-detection model account makes three fundamental assumptions (see Dunn, 2004; Hirshman & Master, 1997). First, all items in a recognition memory test can be ordered along a single dimension which captures the strength of evidence available from memory. Second, for each class of items exists a particular distribution of strength-of-evidence values. On average, the strength of evidence is larger for targets than for distractors, but the distributions overlap. Therefore, a particular strength value does not unambiguously indicate whether a test item is old or new, which necessitates a decision process. However, the ability to discriminate between targets and distractors (i.e., the recognition memory performance) is uniquely determined by the degree of overlap between the strength-of-evidence distributions and independent of the decision process. Third, the response is determined by placing two criteria on the strength-of-evidence axis. If an item's strength value falls below the more lenient criterion, the item is called new, else it is called old. If the strength value falls between the two criteria, the item receives a know response. If the strength value exceeds the stricter criterion, it is called remember. That is, remember responses are considered as equivalent to highly confident recognition responses.

Experimental dissociations between remember and know responses have been taken as evidence in support of dual-process accounts, as discussed above. However, Hirshman and
Master (1997) and Dunn (2004) demonstrated that these same experimental dissociations can be produced by varying one or more parameters of the unidimensional signal-detection model with equal-variance normal distributions, namely the degree of overlap between the target and distractor distributions and the positions of the response criteria on the strength-of-evidence axis. Thus, experimental dissociations do not necessitate a dual-process account.

Some researchers directly compared remember-know judgments to confidence ratings by asking different groups of participants to qualify their old responses by sure-unsure judgments or remember-know judgments in otherwise identical experiments. For example, Rajaram (1993) showed that in the remember-know paradigm, masked repetition priming compared to masked priming with an unrelated word increased the know response rate to targets and distractors, whereas the remember response rate was not affected. In contrast, both sure and unsure response rates were higher after masked repetition priming, significantly for distractors, descriptively for targets. Rajaram (1993) inferred from these patterns of results, which differed as a function of response set, that remember-know responses are not equivalent to confidence judgments and cannot be explained in terms of a single strength-of-evidence dimension. However, Dunn (2004) demonstrated that this inference is not warranted. He fitted a unidimensional signal-detection model to the data, which incorporated the restriction that recognition memory performance does not differ as a function of response set. This model successfully reconstructed the observed pattern of results. For each priming condition, the order of response criteria was (from lenient to strict) unsure, know, sure, remember. Thus, this model-based analysis suggests that given remember-know instructions, participants simply apply stricter response criteria to the same underlying strength-of-evidence dimension than given the response options sure-unsure. This is a reasonable finding in that the remember-know instructions stressed that a know response required being certain of recognising the item.

A critical test of the unidimensional signal-detection model interpretation of the remember-know paradigm derives from the fact that according to the model's assumptions, the recognition memory performance is independent of the bias in the decision process, which solely affects criterion placement. That is, if remember and know judgments result from applying different response criteria to the same strength-of-evidence dimension, bias-free estimates of recognition memory performance should be equal when calculated from remember response rates and when calculated from remember plus know response rates. In
various meta-analyses, using different datasets and different measures of recognition memory performance, the respective authors arrived at contradictory conclusions (Donaldson, 1996; Dunn, 2004; Gardiner & Gregg, 1997; Gardiner, Ramponi, & Richardson-Klavehn, 2002; Macmillan, Rotello, & Verde, 2005). The largest data-base was applied to this question by Dunn (2004). He found that $d'$, a bias-free measure of recognition memory performance given equal-variance normal distributions, was virtually identical when based on remember response rates and remember plus know response rates, where a scatterplot showed no systematic deviations. A frequently used alternative measure of recognition memory performance is $A'$. Dunn (2004) corroborated previous findings that $A'_r$ based on remember response rates is slightly, but consistently smaller than $A'_{r+k}$ based on remember plus know response rates. At first glance, this contradicts the finding for $d'$. However, Dunn (2004) showed that these findings can be reconciled. Given equal-variance normal distributions, $A'$ is influenced by response bias in such a way that the unidimensional signal-detection model actually predicts $A'_r < A'_{r+k}$ rather than $A'_r = A'_{r+k}$, in line with the empirical finding. That is, assuming equal-variance normal distributions, $A'$ is not a valid measure of recognition memory performance and the unidimensional signal-detection model passes the test.

Rotello, Macmillan, and Reeder (2004) devised another test of the unidimensional signal-detection model, which focuses on the slope of the receiver operating characteristic in $z$-coordinates, $z$ROC. A theoretical $z$ROC for a pair of target and distractor distributions is obtained by varying the position of the response criterion along the strength-of-evidence axis. Each criterion placement (i.e., level of confidence) generates a particular pair of hit rate (i.e., rate of correct old responses to targets) and false-alarm rate (i.e., rate of incorrect old responses to distractors). The $z$ROC is obtained by transforming hit and false-alarm rates into $z$-scores and plotting $z$-hit rate as a function of $z$-false-alarm rate. For normal strength-of-evidence distributions, the $z$ROC is linear. Importantly, the slope of the $z$ROC is equal to the ratio of the standard deviations of the distractor and target distributions. Rotello et al. (2004) summarised a large body of recognition memory experiments that varied response bias by asking for confidence ratings or by means of experimental manipulations and concluded that the slope of the $z$ROC is typically about 0.8, which indicates that the target distribution has a larger variance than the distractor distribution. Rotello et al. (2004) reasoned that if responses in remember-know and non-remember-know experiments are generated by the same mechanism of placing response criteria on a
single strength-of-evidence dimension, then the slope of the \( z \)ROC should be equal for both classes of experiments, irrespective of response set. Contrary to this prediction, a meta-analysis showed that the distribution of remember-know \( z \)ROC slopes has a larger central tendency and a larger variance than the distribution of non-remember-know slopes. Hence, Rotello et al. (2004) concluded that remember-know responses are not equivalent to confidence judgments, but reflect some other source of evidence in addition to or instead of a single strength-of-evidence dimension. However, Wixted and Stretch (2004) pointed out that this conclusion may be invalid because of procedural differences other than response set between remember-know and non-remember-know experiments, which they suspect to differentially affect the \( z \)ROC slope. In order to control for this potential confound, they looked at experiments that asked for remember-know judgments versus confidence ratings under otherwise identical conditions. Interestingly, across this small set of experiments, the mean \( z \)ROC slope was close to 0.8 for both remember-know and non-remember-know datasets.

To summarise, the unidimensional signal-detection model of recognition memory suggests that a single strength-of-evidence dimension forms the basis of remember-know judgments, old-new judgments and confidence ratings alike. Based on a comprehensive review of the remember-know literature, Dunn (2004) concluded that the remember-know data are compatible with a unidimensional signal-detection model if one assumes normal distributions with equal variance. In line with this assumption, Rotello et al. (2004) found a mean remember-know \( z \)ROC slope of 1.0. However, unidimensional signal-detection model analyses of recognition memory experiments without remember-know judgments indicated that typically, the target distribution has a larger variance than the distractor distribution. Wixted and Stretch (2004) suggested that this difference of \( z \)ROC slopes might reflect a confound. Else, it would imply that the assumption of a single strength-of-evidence dimension cannot simultaneously explain both sets of results, which would call the unidimensional signal-detection account into question.
3 Critique

The theoretical explanations of remember-know judgments reviewed in the previous chapter offer very different views on the remember-know paradigm. From the perspective of the unidimensional signal-detection model, remember-know judgments are equivalent to confidence judgments. This implies that the remember-know paradigm cannot yield insights over and above conventional confidence rating experiments. Accordingly, analyses of the remember-know paradigm with the unidimensional signal-detection model do not involve a measure of retrieval experience, but proceed in terms of measures of recognition memory performance and response bias. In contrast, both the multiple-memory systems model and the distinctiveness-fluency account assume that remember-know judgments are not equivalent to confidence ratings, but capture a distinct phenomenon. Analyses of retrieval experience are typically based on the rates of remember and know responses.

It is noteworthy that both models that consider it worthwhile to investigate retrieval experience, either in order to isolate different theoretical constructs that are assumed to jointly produce the recognition memory performance or in order to understand the retrieval experience per se, draw on the same measures, although these approaches imply partially conflicting requirements for a suitable measure. If the goal is to isolate different theoretical constructs that simultaneously mediate a state of retrieval experience and contribute to recognition memory performance, it is desirable to employ measures that capture these aspects simultaneously. At the theoretical level, such measures represent the assumed constructs, episodic and semantic memory. At the descriptive level, such measures are two-in-one measures of retrieval experience and of a component of recognition memory performance. In contrast, if the goal is to investigate retrieval experience per se, a pure measure of retrieval experience is required.

Are the remember and know response rates suitable for either purpose? Remember and know response rates are obtained by asking participants to qualify their old responses as remember versus know. However, it is textbook knowledge that the rate of old responses in recognition memory tests varies as a function of both the ability to discriminate targets from distractors and the bias in the old-new decision (e.g., Buchner & Brandt, 2002; Neath & Surprenant, 2003). This suggests that remember and know response rates, which are derived by partitioning the old response rate, vary as a function of these factors as well.
Depending on whether remember-know response rates are required to be pure measures of retrieval experience or to be two-in-one measures of retrieval experience and a component of recognition memory performance, at least one of these expected covariations would constitute an objectionable contamination.

A pure measure of retrieval experience is required to be sensitive to changes of retrieval experience, but to be invariant to changes of recognition memory performance or old-new response bias. The use of remember and know response rates in order to measure retrieval experience per se neglects the possibility of differences between experimental conditions with regard to recognition memory performance and old-new response bias. If such differences exist, they are expected to influence remember and know response rates, as argued above. Given the invariance requirement, such an influence would constitute an objectionable contamination.

In contrast, from the perspective of the multiple-memory systems model (Tulving, 1985b), a covariation of remember-know response rates with recognition memory performance would not constitute an objectionable contamination, but rather a natural consequence of the presumed mechanisms of recognition memory. As outlined before, each memory system simultaneously mediates a state of retrieval experience and contributes to recognition memory performance. Remember-know response rates are used to represent these memory systems and hence simultaneously measure both aspects that are mediated by a memory system. An increase in overall recognition memory performance can only be achieved by means of greater retrieval success from at least one of these memory systems. Therefore, measures that represent these memory systems are not independent of overall recognition memory performance, contrary to what is required of pure measures of retrieval experience. Hence, from the perspective of the multiple-memory systems model, a covariation of remember-know response rates with overall recognition memory performance simply reflects the fact that remember-know response rates are two-in-one measures of retrieval experience and of a component of recognition memory performance and does not constitute a contamination.

However, two-in-one measures of retrieval experience and a component of recognition memory performance are required to be invariant to changes of old-new response bias. Because it is well-established that experimental manipulations of old-new response bias do not affect the ability to discriminate between targets and distractors, it is reasonable to assume that such manipulations do not affect the ability to retrieve information from the
episodic and semantic memory system, either, which jointly produce the recognition memory performance. Accordingly, an effect of an experimental manipulation of old-new response bias on two-in-one measures of retrieval experience and a component of recognition memory performance constitutes a contamination, indicating that the measures are not suited to represent the assumed theoretical constructs, episodic and semantic memory.\(^1\)

Note that a *pure* measure of retrieval experience, which is required to be invariant to changes of old-new response bias, can *not* be required to be invariant to experimental manipulations of old-new response bias, because it is possible that the manipulation affects the *true* retrieval experience in addition to the old-new response bias. In this case, it would not be warranted to speak of a contamination. Only if the retrieval experience is assumed to indicate a theoretical construct that simultaneously mediates the ability to discriminate between targets and distractors is it warranted to assume that this ability to discriminate between targets and distractors, and hence the true retrieval experience, is not affected by a manipulation of old-new response bias. Therefore, only two-in-one measures of retrieval experience and a component of recognition memory performance are required to be invariant to experimental manipulations of old-new response bias. The empirical evidence with regard to remember-know response rates is reviewed in the following section.

### 3.1 Old-New Response Bias

The influence of old-new response bias on remember and know response rates was first investigated experimentally by Strack and Förster (1995). They argued that judgmental strategies would be more likely to influence the response to a test item that does not elicit details of the study episode. Accordingly, they predicted that an experimental manipulation of old-new response bias would affect know responses, but not remember responses. They manipulated old-new response bias by informing different groups of participants just prior to the recognition memory test that the percentage of targets among the test items was 30\% or 50\%, when in fact it was 50\% in both conditions. Varying the baserate information

\(^1\) Although in principle, constant recognition memory performance could occur if an increased contribution of one memory system to recognition memory performance is offset by a decreased contribution of the other memory system, there is no reason to suspect that a manipulation of old-new response bias, which is well-established not to affect the ability to discriminate between targets and distractors, should affect the ability to retrieve information from either memory system.
successfully manipulated old-new response bias, as evidenced by a standard measure of old-new response bias calculated from old responses to targets and distractors. As predicted by Strack and Förster (1995), the manipulation of baserate information affected the know response rate, but not the remember response rate to both targets and distractors. That is, the know response rate, but not the remember response rate was contaminated with old-new response bias.

Hirshman and Henzler (1998) reasoned that a stronger manipulation of old-new response bias should affect not only the know, but also the remember response rate. Indeed, informing different groups of participants that the percentage of targets in the test was 30% or 70% (with a constant true percentage of 50%) affected not only the know response rate, but also the remember response rate to both targets and distractors. This shows that not only the know response rate, but also the remember response rate may be contaminated with old-new response bias.

This evidence was extended and corroborated by further experiments that employed other experimental manipulations of old-new response bias. For example, Postma (1999) explicitly instructed different groups of participants to respond conservatively versus liberally in the old-new decision. Standard measures of old-new response bias calculated from old response rates served as manipulation check and showed that the instruction successfully influenced old-new response bias. The remember response rates to targets and the know response rates to distractors differed as a function of instruction, supporting the notion that both remember and know response rates may be contaminated with old-new response bias.

Erdfelder (2000) varied pay-off matrices, which is an established manipulation of old-new response bias (e.g., Banks, 1969; Smith, 1970). A pay-off matrix assigns rewards to both types of correct responses, namely correct old responses to targets (i.e., hits) and correct new responses to distractors (i.e., correct rejections), and penalties to both types of false responses, namely old responses to distractors (i.e., false alarms) and new responses to targets (i.e., misses). If an old response is associated with a high reward if it is correct, but only a low penalty if it is wrong and at the same time a new response is associated with a low reward if correct and a high penalty if wrong, it is rational to prefer old responses over new responses and have a liberal old-new response bias. Conversely, a conservative bias is rational when high and low values are assigned to the opposite types of correct and wrong responses. Erdfelder (2000) constructed two pay-off matrices along these lines and
offered book rewards for the highest scores in each group as an incentive. This manipulation selectively affected the remember response rate to distractors.

Finally, two experiments manipulated old-new response bias by varying the test procedure (Eldridge, Sarfatti, & Knowlton, 2002; Hicks & Marsh, 1999). Typically, participants in a remember-know experiment are asked to make an old-new judgment to each test item, and if they respond old, they are subsequently asked for a remember-know judgment. Alternatively, they may be asked for a judgment only once by simultaneously presenting the response options remember, know and new. These test arrangements are called two-step and one-step test procedure, respectively. Standard measures of old-new response bias calculated from old response rates indicated that as predicted, the old-new response bias was more liberal with the one-step test procedure in both studies. With regard to remember and know response rates, Hicks and Marsh (1999) found for the one-step test procedure a higher remember response rate to targets and distractors and a higher know responses rate to distractors. Eldridge et al. (2002) replicated the increase in remember response rate to targets and know response rate to distractors with the one-step test procedure, but did not find an effect on the remember response rate to distractors. However, the two experiments converge in showing that both remember and know response rates may be contaminated with old-new response bias.

In sum, various experiments with the remember-know paradigm that manipulated old-new response bias by different means corroborate the notion that both remember and know response rates may be contaminated with old-new response bias and are hence not suited as measures of retrieval experience that represent retrieval from episodic and semantic memory, respectively.

Gardiner and colleagues (Gardiner, Java, & Richardson-Klavehn, 1996; Gardiner, Richardson-Klavehn, & Ramponi, 1997) suggested that a contamination of remember and know response rates with old-new response bias can be prevented by providing the additional response option guess for qualifying the old responses. This addition yields the remember-know-guess paradigm. This suggestion is based on the rationale that an experimental condition that aims to induce a liberal response bias, for example by informing participants that the percentage of targets in the test is 70%, might make it impossible for participants to obey both the response-bias and remember-know instructions at the same time. This dilemma occurs if participants do not encounter enough items that they could call remember or know in line with the remember-know instructions in order to
approach a perceived rate of 70% old responses. If they choose to attain the liberal response bias nevertheless, they are forced to disobey the remember-know instructions. However, this constellation can be avoided by adding guess responses as a third type of old responses, which can be used to attain the desired level of old responses without violating the remember-know-guess instructions for qualifying old responses. In principle, adding a third response option for qualifying old responses has the potential to eliminate the contamination of remember-know response rates with old-new response bias because it abolishes the restriction that remember and know response rates add up to the old response rate.

Numerous experiments have used the remember-know-guess paradigm in the hope of obtaining uncontaminated remember-know response rates (see Gardiner et al., 2002, for an overview), but I know of only three experiments that examined this claim empirically (Eldridge et al., 2002; Erdfelder, 2000; Gardiner et al., 1997).\(^2\) The remember-know-guess paradigm is successful if experimental manipulations of old-new response bias selectively affect the guess response rate, but neither the remember nor the know response rate.

Gardiner et al. (1997) replicated the experiment by Strack and Förster (1995), which demonstrated a contamination of the know response rate, with the additional response option guess. Indeed, neither the remember nor the know response rates to targets or distractors varied between an alleged percentage of 30% or 50% targets in the test, but only the rate of guess responses to distractors. This result demonstrates empirically that the remember-know-guess paradigm may successfully prevent contamination of remember-know response rates with old-new response bias.

Erdfelder (2000) varied pay-off matrices. This manipulation of old-new response bias affected the guess response rate to distractors and, furthermore, the remember response rate to targets and distractors as well as the know response rate to targets. Eldridge et al. (2002) compared the two-step and one-step test procedures. This manipulation affected not only the guess response rate to distractors, but also the know response rate to distractors. That is, both experiments failed to replicate invariant remember-know response rates across the levels of an old-new response bias manipulation (Gardiner et al., 1997). Rather, they

\(^2\) Xu and Bellezza (2001) crossed the manipulation of percentage of targets in the test with whether the response option guess was available for classifying the retrieval experience. However, they did not report simple effects of percentage of targets separately for the response option conditions.
demonstrate that remember-know response rates may be contaminated with old-new response bias in spite of the additional response option guess.

The finding that remember-know response rates may vary with old-new response bias, regardless of whether or not it is possible to qualify old responses as guess, indicates that they are unsuitable as measures of retrieval experience that represent retrieval from episodic and semantic memory, respectively.

As mentioned before, in contrast to measures that represent retrieval from episodic and semantic memory, a *pure* measure of retrieval experience, though it is required to be invariant to changes of old-new response bias, cannot be required to be invariant to experimental manipulations of old-new response bias, because it is possible that the manipulation affects the *true* retrieval experience in addition to the old-new response bias. In this case, it would not be warranted to speak of a contamination. Therefore, the reported evidence that remember-know response rates vary across levels of manipulations of old-new response bias cannot be taken as conclusive evidence that remember-know response rates are unsuitable as pure measures of retrieval experience. However, as remember-know response rates are obtained by partitioning the old response rate, and the old response rate in recognition memory tests varies as a function of old-new response bias (as well as recognition memory performance), there is reason to doubt the suitability of remember-know response rates as pure measures of retrieval experience. Accordingly, it is not appropriate to interpret these findings as evidence of effects on retrieval experience per se, either. Possible effects of response-bias manipulations on retrieval experience per se can only reasonably be examined with measures that by virtue of their construction do not a priori raise the suspicion of being susceptible to contamination with old-new response bias. The following section is devoted to the second invariance requirement that applies to pure measures of retrieval experience.

### 3.2 Recognition Memory Performance

A pure measure of retrieval experience is required to be invariant to changes of recognition memory performance. As remember and know response rates are obtained by partitioning the old response rate and the old response rate in recognition memory tests varies as a function of the ability to discriminate targets from distractors (as well as old-new response bias), remember and know response rates are expected to vary as a function of recognition memory performance as well. Thus, recognition memory performance
constitutes a potential source of contamination of the traditional measures of retrieval experience in both the remember-know and the remember-know-guess paradigm.

Take the levels-of-processing effect as an example. It is well-known that recognition memory performance is better after semantic processing than after phonemic processing (e.g., Craik & Tulving, 1975; Jacoby & Dallas, 1981). Does this manipulation also affect retrieval experience? Gardiner (1988) was the first to address this issue with the remember-know paradigm. As an incidental orienting task, different groups of participants had to write down a semantic associate or a rhyme for each study word. Gardiner (1988) found a higher remember response rate after semantic processing than after phonemic processing. The know response rates to targets, which were considerably lower, did not differ between conditions. Remember and know response rates to distractors were extremely low (none above .02). Gardiner (1988) concluded from the selective effect on the remember response rate to targets that the levels-of-processing manipulation selectively affected retrieval experience of the kind that he termed recollective experience. However, the observed pattern of results is equally compatible with a selective effect of the levels-of-processing manipulation on recognition memory performance, where neither the rate of truly recognised items judged as remember nor the rate of items that are not recognised, but nevertheless guessed to be old and given a remember response differ between conditions. In this scenario, retrieval experience does not change across conditions in that a constant proportion of truly recognised targets receives a remember response and a lower, constant proportion of items guessed to be old receives a remember response. In this scenario, the old-new response bias does not differ, either, but is equally conservative in both conditions. This alternative interpretation of an alleged effect on retrieval experience in terms of a selective effect on recognition memory performance illustrates the ambiguity inherent in the use of remember and know response rates for measuring retrieval experience per se.
4 Problem Formulation and Research Goal

The remember-know-(guess) paradigm investigates the subjective experience that accompanies retrieving information from memory. Regardless of whether the overarching research interest is directed at retrieval experience per se (Rajaram, 1996) or at retrieval experience as a tool for isolating different theoretical constructs that jointly produce the recognition memory performance (Tulving, 1985b), data analyses have typically been based on the rates of remember and know responses as measures of retrieval experience. However, the arguments and empirical evidence presented in the previous chapter lead to the conclusion that remember-know response rates are not suitable as measures for either purpose, regardless of whether the response option guess is available for classifying the retrieval experience or not. Hence, remember-know response rates do not constitute a suitable basis for evaluating whether the distinctiveness-fluency framework and/or the multiple-memory systems model provide adequate explanations of retrieval experience as captured by the remember-know-(guess) paradigm.

The present thesis examines alternative measures of retrieval experience. Specifically, the first research goal is to evaluate the suitability of these alternative measures. If the suitability of a measure can be established, the second goal is to use this measure in order to demonstrate empirically that the traditional analysis of remember and know response rates can lead to erroneous conclusions about the effect of independent variables on retrieval experience. In doing so, the present thesis provides tests of both the distinctiveness-fluency framework and the multiple-memory systems model of retrieval experience.

The alternative measures of retrieval experience that are examined in this thesis are formulated within multinomial processing tree models, which allow one to measure cognitive events assumed to underlie the observable responses and to test whether a suggested model fits the data. In an attempt to provide pure measures of retrieval experience, the task-oriented measurement models start by specifying measures of recognition memory performance and old-new response bias. These are supplemented by two measures of retrieval experience, one of which is specified conditionally on true recognition, the other conditionally on old responses due to guessing. By virtue of their construction, these measures can be expected to be pure measures of retrieval experience. In contrast, the multi-memory measurement models implement Tulving’s (1985b) model of
multiple memory systems by specifying parameters that reflect retrieval from episodic and semantic memory. Additionally, these measurement models allow for the possibility that remember and know responses may result not only from memory retrieval, but also from guessing, which is captured with separate parameters. These modelling approaches are applied to both the remember-know paradigm and the remember-know-guess paradigm.

The measurement models are evaluated by examining whether model parameters that reflect recognition memory performance or components thereof are invariant to experimental manipulations of old-new response bias. A reanalysis of published datasets was followed up by two new experiments that crossed a manipulation of old-new response bias with whether the response option guess was available for classifying retrieval experience. Drawing on the extended database, one measurement model for the remember-know paradigm proved to be globally valid. This measurement model was used in order to demonstrate empirically that the traditional analysis of remember and know response rates can lead to erroneous conclusions about the effect of independent variables on retrieval experience. Participants studied items with full versus divided attention in Experiment 3, and Experiment 4 compared semantic to phonemic processing.
5 Probabilistic Modelling of the Remember-Know-(Guess) Paradigm

The measurement models that are examined in this thesis are formulated as multinomial processing tree models. Before delineating the psychological substance of the models for the remember-know-(guess) paradigm, the methodology of multinomial modelling is reviewed.

5.1 Multinomial Processing Tree Models

Multinomial processing tree (MPT) models are a class of probabilistic models for categorical data that allow the measurement of latent cognitive events hypothesised to underlie observable responses (Batchelder & Riefer, 1999; Erdfelder, 2000; Riefer & Batchelder, 1988).

Assume an experiment with \( J \) mutually exclusive and jointly exhaustive response categories \( C_j, j = 1, \ldots, J \). That is, any observation falls into one and only one response category. Let parameter \( p_j \) denote the probability that an observation falls into category \( C_j \). Because the response categories are mutually exclusive and jointly exhaustive, \( \sum_{j=1}^{J} p_j = 1 \). Hence, of the \( J \) parameters \( p_j \) that represent category probabilities, \( J-1 \) are functionally independent.

Assume that a random sample of \( N \) independent and identically distributed observations is drawn. Let \( \mathbf{Y} = (Y_1, \ldots, Y_J) \) be a random vector whose values \( \mathbf{y} = (y_1, \ldots, y_j, \ldots, y_J) \) denote the count of observations in each category \( C_j \), with \( \sum_{j=1}^{J} y_j = N \). Then the distribution of \( \mathbf{Y} \) is the multinomial distribution defined by the probability function

\[
P(y_1, \ldots, y_J; p_1, \ldots, p_J) = \frac{N!}{y_1! \cdots y_J!} \cdot p_1^{y_1} \cdots p_J^{y_J}.
\]

(5.1)

If an experiment contains more than one experimental condition or item type, an extension is required. Assume \( K \) experimental conditions, \( k = 1, \ldots, K \), where each condition has a separate system of \( J_k \) mutually exclusive and jointly exhaustive response categories. Let \( C_{jk} \) denote response category \( j \) in condition \( k \), and let \( p_{jk} \) denote the
probability that an observation falls into response category $C_{jk}$, with $\sum_{j=1}^{J_k} p_{jk} = 1$ for each condition.

Assume that for each condition an independent sample of $N_k$ independent and identically distributed observations is drawn. For each condition, let $Y_k = (Y_{1k},...,Y_{jk},...,Y_{J_k})$ be a random vector whose values $y_k = (y_{1k},...,y_{jk},...,y_{J_k})$ denote the count of observations in each category $C_{jk}$, with $\sum_{j=1}^{J_k} y_{jk} = N_k$. Finally, assume that each $Y_k$ has an independent multinomial distribution. Then the joint random vector $Y$ across all $K$ conditions has a product multinomial distribution defined by the probability function

$$P(y_{11},...,y_{J_K}; p_{11},...,p_{J_K}) = \prod_{k=1}^{K} \frac{N_k!}{y_{1k}! \cdots y_{J_k}!} \cdot p_{jk}^{y_{jk}}$$

(5.2)

This joint multinomial model is the most general statistical model for the situation under consideration ($K$ independent samples of independent, identically distributed observations). To reiterate, the parameters $p_{jk}$ represent category probabilities. Of these $\sum_{k=1}^{K} J_k$ category probabilities in total, $M = \sum_{k=1}^{K} (J_k - 1)$ are functionally independent.

MPT models reparameterise the product multinomial distribution with a vector $\theta = (\theta_1,...,\theta_s,...,\theta_S)$ of $S$ functionally independent model parameters. That is, each category probability parameter is defined as a function of the model parameters, $p_{jk} = f(\theta)$, in a system of model equations. Each model parameter $\theta_i$ represents the probability of a discrete cognitive event. Because each model parameter reflects a probability, it must be contained in the closed interval $[0,1]$. An observable response is generated by a combination or succession of cognitive events, which may be represented graphically as a branch in a tree that terminates in a particular response category. Different branches may terminate in the same response category, where each branch stands for a different possible cause of such a response. To obtain the model equation for response category $C_{jk}$, $p_{jk} = f(\theta)$, one multiplies all parameters in a given branch and sums across all branches that terminate in response category $C_{jk}$. Thus, MPT models formalise psychological assumptions about cognitive events that underlie the observable responses.
problems: estimating parameters, determining whether the model fits the data, and testing specific hypotheses about model parameters.

An important concept in the context of determining the values of the model parameters is *global identifiability*. Let $p(\theta)$ denote the vector of category probabilities as a function of the model’s parameter vector $\theta$. The MPT model is globally identifiable if and only if $p(\theta) = p(\theta^*)$ implies $\theta = \theta^*$ for all $\theta, \theta^*$. That is, an MPT model is globally identifiable if and only if any vector of category probabilities can be generated by one vector of model parameters at most. In other words, if the vector of category probabilities can be produced by the model at all, there exist unique parameter values. A necessary condition for global identifiability is that there are no more functionally independent model parameters than functionally independent category probabilities, $S \leq M$. The following discussion assumes globally identifiable MPT models.

Parameter values are estimated with maximum likelihood (ML) methods. The ML estimator $\hat{\theta}$ of the parameter vector $\theta$ is the best asymptotically normal estimator of $\theta$ (Erdfelder, 2000). That is, for $N \to \infty$, it has the following desirable properties: First, $\hat{\theta}$ is asymptotically unbiased, $E(\hat{\theta}) = \theta$. Second, $\hat{\theta}$ is asymptotically efficient, that is, its variance is no larger than that of any other asymptotically unbiased estimator. Finally, if suitably standardised, it has an asymptotic standard normal distribution, which can be used for constructing asymptotic confidence intervals. Assuming certain regularity conditions (Erdfelder, 2000), the ML estimate $\hat{\theta}$ of $\theta$ is the value that maximises the likelihood function

$$L(\theta; y) = \prod_{k=1}^{K} \frac{N!}{y_{1k}! \cdots y_{J_k}!} \cdot p_{1k}(\theta)^{y_{1k}} \cdots p_{J_k}(\theta)^{y_{J_k}}.$$  \hfill (5.3)

Note that the likelihood function $L(\theta; y)$ is a function of the parameter vector $\theta$ for a given vector of observed response frequencies $y$, whereas the probability function introduced above is a function of $y$ for a given vector of category probabilities $p$.

Computer programs for calculating ML estimates for MPT models employ the expectation-maximisation algorithm (Erdfelder, 2000; Hu & Batchelder, 1994). This is a general iterative method for obtaining ML estimates for such statistical models that can be conceptualized as models for incomplete data. With multinomial processing tree models, the data on branch frequencies are incomplete: The summed frequencies across all
branches terminating in a particular category are given by the observed category frequencies, but the frequencies of individual branches are missing. The expectation-maximisation algorithm starts with initial (random or user-specified) values of the model parameters. In the expectation step, expected branch frequencies given current parameter values and observed category frequencies are calculated. In the maximisation step, these branch frequencies are inserted into closed-form expressions to calculate the ML estimates of model parameters. This procedure is reiterated until the change in parameter estimates from one cycle to the next is sufficiently small.

Parameter estimates for a given MPT model can only be interpreted meaningfully if the model fits the data, that is, if the category probabilities \( p \) can indeed be produced by the model parameters \( \theta \). If an MPT model has fewer functionally independent model parameters than there are functionally independent category probabilities (i.e., \( S < M \)), the goodness of fit of the model can be tested with the log-likelihood ratio statistic \( G^2 \),

\[
G^2(\theta; y) = -2 \cdot \sum_{k=1}^{K} \sum_{j=1}^{J_k} y_{jk} \cdot \ln \left( \frac{N_k \cdot p_{kj}(\theta)}{y_{kj}} \right),
\]

which is asymptotically distributed according to a \( \chi^2 \)-distribution with \( M - S \) degrees of freedom. If the model is fully parameterised, \( S = M \), it is only possible to conclude that the model fits the data if the expected frequencies for a given number of observations and a particular vector of parameter values, \( N_k \cdot p_{kj}(\theta) \), are identical to the observed frequencies \( y_{kj} \), which implies \( G^2 = 0 \).

Hypotheses about the effect of experimental conditions on model parameters may be analysed by imposing restrictions on the model parameters. Such restrictions reduce the number of functionally independent model parameters from \( S \) to \( S' \), thereby creating a submodel with parameter vector \( \theta' \). With regard to statistical analysis, two cases need to be distinguished. First, if \( S < M \) for the initial model \( M_S \), the parameter hypothesis can be tested by assessing the difference in the goodness of fit between the initial model \( M_S \) and the submodel \( M_{S'} \) that is due to the parameter restriction. This difference can be analysed with the conditional log-likelihood ratio test statistic

\[
\Delta G^2 = G^2(M_{S'}) - G^2(M_S),
\]

(5.5)
which has a $\chi^2$-distribution with $S-S'$ degrees of freedom. Second, if a fully parameterised initial model fits the data, which implies $G^2 = 0$, the parameter hypothesis can be tested by assessing the goodness of fit of the submodel with the log-likelihood ratio statistic.

What can be done if one is for theoretical reasons interested in an overparameterised model, $S > M$, that necessarily lacks global identifiability? A way out is to add a second experimental condition, thereby doubling the number of functionally independent response categories, in a way that does not require doubling the number of functionally independent model parameters. This can be achieved by using an experimental manipulation that allows for the additional assumption that one or more model parameters are constant across this manipulation, such that introducing these equality restrictions yields a testable model with fewer functionally independent model parameters than response categories. If the restricted model holds, one can safely conclude that the theoretically interesting model structure is adequate.

In short, MPT models are probabilistic models whose parameters represent hypothetical cognitive events. They are useful measurement tools that allow for statistical testing and have been applied successfully in many areas of cognitive psychology (see Batchelder & Riefer, 1999, and Erdfelder, 2000, for reviews). In the next sections, several MPT models for the remember-know paradigm and its remember-know-guess variant are described.

### 5.2 Measurement Models for the Remember-Know Paradigm

#### 5.2.1 Task-Oriented Model

The task-oriented measurement model depicted in Figure 5.1 is derived from a pragmatic approach to constructing pure measures of retrieval experience: The typical two-step remember-know procedure supplements a standard old-new recognition test with a second judgment about the nature of the retrieval experience. Analogously, a standard measurement model for recognition memory (Snodgrass & Corwin, 1988) is used to describe the old-new decision, and conditional measures of retrieval experience are added that describe the additional remember-know decision in a way that is not confounded with the old-new decision. This approach was suggested by Meiser and Bröder (2002), who used it to extend a measurement model for multidimensional source memory to encompass remember-know judgments.
Figure 5.1  Processing tree representation of the two-high threshold task-oriented model for the remember-know paradigm. The one-high threshold task-oriented model results from setting \( D = 0 \) in the lower tree. \( D \) = probability of recognising a target as old and of identifying a distractor as new; \( R \) = probability of a remember response to a truly recognised target; \( b \) = probability of guessing that an item is old; \( R^* \) = probability of a remember response to an item guessed to be old.

The upper tree of Figure 5.1 illustrates that a target is truly recognised as old with probability \( D \). A truly recognised target is classified as remembered with probability \( R \). With the complementary probability \( 1 - R \) it obtains a know response. If a target is not recognised, which happens with probability \( 1 - D \), it may nevertheless be judged old on the basis of guessing processes, which happens with probability \( b \). In that case, too, a remember-know judgment is required. The probability of a remember response to an item guessed to be old is specified by parameter \( R^* \). A know response is given with the complementary probability \( 1 - R^* \). Finally, a target is called new if it is neither recognised nor guessed to be old. As depicted in the lower tree of Figure 5.1, a distractor is truly identified as new with probability \( D \). If the identification of the distractor fails, the same
cognitive events are assumed as in the case of failing to recognise a target. That is, guessing results in a remember response with probability \( b \cdot R^* \), in a know response with probability \( b \cdot (1 - R^*) \), and in a new response with probability \( 1 - b \).

Note that the task-oriented model contains two parameters that specify the conditional probabilities of remember judgments in two different cases: \( R \) denotes the probability of a remember response given true recognition of a target, whereas \( R^* \) denotes the probability of a remember response given that a test item is guessed to be old, regardless of its status as target or distractor. By virtue of their construction, these measures can be expected to be pure measures of retrieval experience. Furthermore, recognition memory performance is measured with parameter \( D \) and response bias in the old-new decision is measured with parameter \( b \).

This model is a two-high threshold (2HT) model of the old-new decision because it assumes two thresholds that define three discrete memory states, namely recognition as old, identification as new, and uncertainty. The thresholds are called high thresholds because each threshold can only be crossed by items of one type. In particular, a target can only exceed the old recognition threshold; if it does, it is called old. Analogously, a distractor can only exceed the new identification threshold; if it does, it is called new. An item that does not cross the respective high threshold is in the state of uncertainty. It is called old or new on the basis of guessing. The standard 2HT model includes the additional assumption that targets and distractors exceed the respective high threshold with equal probability (Snodgrass & Corwin, 1988). This equality assumption is reflected in the task-oriented model in the specification of a single parameter \( D \) that refers to both targets and distractors. Dropping the assumption that distractors can be identified as new, thereby abolishing the memory state of identification as new and the high threshold that refers to distractors, results in a one-high threshold (1HT) model. Formally, the 1HT task-oriented model is a submodel of the 2HT task-oriented model that is obtained by imposing the restriction \( D = 0 \) in the lower tree of Figure 5.1. The model equations and proofs of identifiability for the 2HT task-oriented model and the 1HT task-oriented model can be found in Appendix A.1.

Note that the model parameter \( D \) is simply a measure of recognition memory performance. This specification does not imply any theoretical commitment how the recognition memory performance is achieved, that is, whether exceeding the old
recognition threshold can be accomplished by means of one or more memory processes. Rather, it simply reflects the pragmatic approach to constructing pure measures of retrieval experience.

### 5.2.2 Multi-Memory Model

The multi-memory measurement model for the remember-know paradigm (Erdfelder, 2000; Xu & Bellezza, 2001) implements Tulving's (1985b) model of multiple memory systems. Two parameters represent retrieval from the episodic and semantic memory system, respectively. In addition, the measurement model allows remember and know responses to result not only from memory retrieval, but also from guessing that an item is old rather than new.

The multi-memory model as suggested by Erdfelder (2000) is depicted in Figure 5.2. At test, both memory systems are searched. In case of a target, illustrated in the upper tree of Figure 5.2, the item is successfully retrieved from episodic memory with probability $r$, and retrieval from episodic memory always leads to a remember response. If, with the complementary probability $1-r$, the item is not retrieved from the episodic memory system, it may still be retrieved from the semantic memory system with conditional probability $k$, which entails a know response. If, with probability $(1-r) \cdot (1-k)$, a target is not retrieved from either memory system, the response is based on guessing. Guessing is captured by two parameters. The parameter $g_r$ denotes the probability of responding remember given that the item is not retrieved from either memory system. The parameter $g_k$ denotes the probability of responding know given that the item is not retrieved from either memory system and not called remember by guessing. With the complementary probability $1-g_k$ the item is called new by guessing.

The lower tree in Figure 5.2 illustrates the cognitive events assumed to underlie responses to distractors. Erdfelder (2000) assumed that a distractor, due to not having been presented in the study phase, can never be retrieved from either memory system, reflected in $r=k=0$. Furthermore, based on research by Strack and Bless (1994), he assumed that such retrieval failure may be diagnostic, thus affording the identification of the item as a distractor by means of metamnestic inference with probability $d$. If, with probability $1-d$, the distractor is not identified, the same guessing processes as in the case of failing to retrieve a target are assumed to take place.
Figure 5.2 Processing tree representation of the two-high threshold multi-memory model for the remember-know paradigm. The one-high threshold multi-memory model results from setting \( d = 0 \) in the lower tree. \( r \) = probability of retrieving a target from episodic memory; \( k \) = probability of retrieving a target from semantic memory, but not episodic memory; \( g_r \) = probability of guessing that an item is old and calling it remember; \( g_k \) = probability of guessing that an item is old and calling it know; \( d \) = probability of identifying a distractor.

Thus, the multi-memory model contains two parameters that represent retrieval from different memory systems: Parameter \( r \) represents the probability of successful retrieval from episodic memory, and parameter \( k \) represents the conditional probability of successful retrieval from semantic memory, given retrieval failure for the episodic memory system. According to Tulving’s (1985b) conception, retrieving information from a memory system has two aspects: It is accompanied by a characteristic state of retrieval experience, and it contributes to recognition memory performance. The parameters \( r \) and \( k \) capture both aspects, that is, they are two-in-one measures of retrieval experience and of a component of recognition memory performance. Additionally, Erdfelder (2000) assumed that a third
component contributes to recognition memory performance, namely identifying distractors as new by means of metamnestic inference, which is captured by parameter \( d \). Besides these parameters that measure components of recognition performance, the model specifies two parameters that capture response biases, namely the conditional probabilities \( g_r \) and \( g_k \) of guessing remember and know, respectively.

With regard to the old-new decision, this model is a two-high threshold model: Targets exceed the old recognition threshold with probability \( r + (1 - r) \cdot k \), and distractors exceed the new identification threshold with probability \( d \). Note that unlike the standard 2HT model and the 2HT task-oriented model, the 2HT multi-memory model does not assume that targets and distractors exceed their respective thresholds with equal probability. Imposing the restriction \( d = 0 \) yields a submodel with a single high threshold. The 1HT multi-memory model was suggested by Xu and Bellezza (2001).

Note that the 2HT multi-memory model has more model parameters than there are functionally independent response category probabilities in a single experimental condition. Therefore, this model is not identifiable for a single experimental condition, but can only provide parameter estimates and be tested in conjunction with equality restrictions across different experimental conditions.

5.3 Measurement Models for the Remember-Know-Guess Paradigm

Extending the remember-know paradigm by the option to classify the retrieval experience as guessing yields the remember-know-guess paradigm. How can the MPT models described in the previous section be adapted to account for the use of this additional response category?

5.3.1 Task-Oriented Model

First consider the task-oriented modelling approach, which reflects the typical two-step experimental procedure by extending a standard measurement model that describes the old-new decision by conditional measures that describe the judgment of the retrieval experience. As discussed above, the model for the old-new decision involves assuming the discrete memory states recognition as old, uncertainty and, for the 2HT model, identification as new. Accordingly, modelling the judgment of retrieval experience is reduced to deciding which of the response options for classifying retrieval experience are
eligible given a certain memory state. The mapping of memory states to responses is very straightforward with the options remember, know, and new, but less obvious if the response option guess is added. The reasoning that underlies the 2HT task-oriented model for the remember-know-guess paradigm depicted in Figure 5.3 is developed in the following paragraphs.

To begin, note that in the standard two-step test procedure, the response option guess is available to the participants only to qualify old responses, but not new responses. Therefore, the response option guess does not denote a neutral point between old and new equivalent to complete indecision, but has a tendency towards old rather new. Thus, the memory state of identification as new, which refers to distractors, is logically compatible only with responding new, but not with any of the other response options, all of which imply oldness. Therefore, this part of the task-oriented model remains unchanged (see lower tree of Figure 5.3).

Next, consider the memory state of uncertainty, which may occur for targets and distractors and in which a response is based on guessing. Does this allow any response but guess? Given the tendency of the guess option as typically used in the remember-know-guess paradigm towards old rather than new, the uncertain memory state should be compatible with responding not only guess, but also new. But how about responding remember and know, which – according to the instruction – imply certainty as opposed to uncertainty? This question can be answered empirically by looking at the responses to distractors in studies that employ the remember-know-guess paradigm: Not all old responses to distractors are specified as guess. Rather, there are substantive rates of remember and know responses to distractors even if it is possible to explicitly report guessing (cf. Table 1 in the review by Gardiner et al., 2002). Therefore, the memory state of uncertainty must be compatible with responding remember and know apart from guess or new. Accordingly, the task-oriented MPT model is adapted by adding a new parameter $g^*$ that denotes the probability of responding guess given that the test item is indeed guessed to be old. The parameter $R^*$ again denotes the probability of responding remember rather than know given the expanded condition that the item is actually guessed to be old, but not called guess.

---

3 For a discussion of how to map memory states to confidence judgments on rating scales, see Erdfelder and Bucher (1998) and Malmberg (2002).
Figure 5.3 Processing tree representation of the two-high threshold task-oriented model for the remember-know-guess paradigm which allows for guess responses to truly recognised targets. Setting $g = 0$ in the upper tree precludes guess responses to truly recognised targets. The one-high threshold task-oriented models result from setting $D = 0$ in the lower tree. $D$ = probability of recognising a target as old and of identifying a distractor as new; $g$ = probability of a guess response to a recognised target; $R$ = probability of a remember response to a truly recognised target; $b$ = probability of guessing that an item is old; $g^*$ = probability of a guess response to an item guessed to be old; $R^*$ = probability of a remember response to an item guessed to be old.

Finally, consider the memory state of recognition as old. As discussed before, the task-oriented MPT model makes no theoretical commitment how a target may achieve this
state. Therefore, there is no a priori theoretical reason that precludes or demands the possibility of guess responses to truly recognised targets. Next, consider potential implications of task instruction and procedure. There are three response options that refer to the quality of the retrieval experience – remember, know, and guess. Considering that the difference between remembering and knowing does not refer to recognition confidence, participants in the state of recognition as old may well use the response option guess to indicate a lower degree of confidence as compared to the retrieval experiences of remembering or knowing. However, this suggested use is by no means mandatory, and the use of the response option guess may well be restricted to the memory state of uncertainty. Making allowances for this ambiguous state of affairs, two different adaptations of the task-oriented model are suggested and evaluated empirically, thus allowing for a data-driven answer to this issue. The more general model is depicted in the upper tree of Figure 5.3. A parameter $g$ has been introduced, which denotes the probability of a guess response given true recognition of a target. The parameter $R$ again denotes the probability of responding remember rather than know given true recognition and, in addition, that the response is not guess. Imposing the restriction $g = 0$ formalises the assumption that true recognition precludes responding guess.

Finally, 1HT models for the remember-know-guess paradigm result from dropping the assumption that distractors can be identified as new, formalised by imposing the restriction $D = 0$ in the lower tree of Figure 5.3. Analogous to the 2HT models for the remember-know-guess paradigm, the 1HT models may either allow for guess responses to truly recognised targets by leaving parameter $g$ unrestrained, or preclude guess responses to truly recognised targets by setting $g = 0$.

In total, applying the task-oriented modelling approach to the remember-know-guess paradigm resulted in four different MPT models that were evaluated empirically. The model equations and proofs of identifiability can be found in Appendix A.2.

### 5.3.2 Multi-memory model

Figure 5.4 depicts the 2HT multi-memory model for the remember-know-guess paradigm. The memory state of recognition as old can only be achieved by retrieval from one of two memory systems, each of which entails a distinct state of retrieval experience captured by remember and know responses, respectively. Accordingly, a target retrieved
Figure 5.4 Processing tree representation of the two-high threshold multi-memory model for the remember-know-guess paradigm. The one-high threshold multi-memory model results from setting $d = 0$ in the lower tree. $r$ = probability of retrieving a target from episodic memory; $k$ = probability of retrieving a target from semantic memory, but not episodic memory; $g_r$ = probability of guessing that an item is old and calling it remember; $g_k$ = probability of guessing that an item is old and calling it know; $g_g$ = probability of guessing that an item is old and calling it guess; $d$ = probability of identifying a distractor.

from memory never receives a guess response. Retrieval failure for both memory systems leads to the memory state of uncertainty, in which any of the response options remember, know, guess and new can be chosen on the basis of guessing, as discussed in the previous section. Accordingly, a further guessing parameter $g_g$ has been added that specifies the conditional probability of responding guess given that the target is not retrieved from either memory system and not called remember or know by guessing. Distractors that are not
identified as new but remain in the uncertain state are again assumed to be responded to on the basis of the same guessing processes as targets that cannot be retrieved from memory. The 2HT multi-memory model for the remember-know-guess paradigm assumes that distractor items are identified as distractors with probability $d$ (Erdfelder, 2000), whereas the 1HT multi-memory model assumes that this never happens, $d = 0$ (Xu & Bellezza, 2001).
6 Evaluation of the Measurement Models

Remember and know response rates, which are the traditional measures of retrieval experience employed in the remember-know-(guess) paradigm have been criticised because they are susceptible to confounding retrieval experience with old-new response bias. In response to this shortcoming, several MPT models have been suggested for each paradigm that explicitly provide for the possibility of differences in old-new response bias. Is any of these models a suitable measurement tool, and how do the various models fare in comparison? In this chapter, I outline a strategy for addressing these issues and derive some preliminary answers from a model-based reanalysis of published data.

6.1 Strategy of Model Evaluation

An MPT model is a suitable measurement tool for the remember-know-(guess) paradigm if it satisfies the validity criterion that model parameters that represent cognitive events pertaining to the accuracy of the recognition memory performance be invariant to experimental manipulations of old-new response bias. In the following, such parameters are referred to as memory parameters. All models derived from the task-oriented approach contain a single memory parameter $D$ that reflects the recognition memory performance. On the other hand, each multi-memory model specifies separate parameters that reflect retrieval from episodic memory, $r$, versus retrieval from semantic memory, $k$, given retrieval failure from episodic memory. Unlike the 1HT multi-memory models for the remember-know and remember-know-guess paradigms, the 2HT multi-memory models assume that distractors are truly identified as new with a certain probability represented by parameter $d$. This parameter reflects a component that contributes to recognition memory performance and is thus a third memory parameter.

The validity criterion of invariance to old-new response bias may be implemented by equating memory parameters across two or more experimental conditions that are expected to differ in old-new response bias. An MPT model is valid if it fits the empirical data with this equality restriction, that is, if the restricted MPT model may be retained. As discussed above, the goodness of fit can be assessed with the log-likelihood ratio statistic $G^2$, which is asymptotically $\chi^2$-distributed with degrees of freedom determined by the difference between the numbers of functionally independent response category probabilities and model parameters. Only with equality restrictions imposed on the memory parameters has
every model at least one degree of freedom, that is, only then can the model fit be tested statistically.

In addition, it is of interest to compare the performance of the models restricted according to the requirements of the validity criterion in order to find out whether one of the MPT models is a particularly suitable measurement model for the remember-know-(guess) paradigm. This comparison must balance a model’s goodness of fit against its complexity as indicated by the number of functionally independent model parameters (e.g., Wickens, 1998). Note that the models for each set of response options are not nested within a common model hierarchy. Therefore, it is impossible to employ conditional log-likelihood ratio tests for the model comparison. However, one can resort to information criteria like Akaike’s Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which afford comparisons between non-nested as well as nested models and provide a ranking of all models under consideration by quantifying the evidence in favour of each model (Burnham & Anderson, 2002, 2004; Kuha, 2004, 2005; Weakliem, 2004). While AIC and BIC share this common goal, they were derived from different rationales and may arrive at different rankings of the candidate models. In the next paragraphs, I shortly outline the basic ideas underlying AIC and BIC, respectively.

Fundamental to the approach underlying Akaike’s Information Criterion (Burnham & Anderson 2002, 2004; Kuha, 2004, 2005) is the notion that any model is a simplification of reality and that no model ever comprises full reality. Hence, the aim is not to identify a “true” model, but to find the model that best approximates the structural information contained in the data, given the number of observations. Information is defined as Kullback-Leibler discrepancy, which is the information lost when a particular model is used to approximate full reality. This is a fundamental quantity in information theory. AIC was derived as an estimator of relative expected Kullback-Leibler discrepancy. It is given by

\[
\text{AIC} = -2 \cdot \ln (L(\hat{\theta}; y)) + 2 \cdot S, \tag{6.1}
\]

where \(L(\hat{\theta}; y)\) denotes the maximum likelihood estimate for the observed data \(y\) and \(S\) denotes the number of functionally independent model parameters. The AIC values calculated according to this equation contain arbitrary constants, which can be eliminated by subtracting the minimal AIC value in the set of candidate models from the AIC value of each candidate model, thereby making the relevant information more easily visible: The
best model obtains $\Delta \text{AIC} = 0$, and the larger $\Delta \text{AIC}$, the less the model is supported by the data (Burnham & Anderson, 2002; 2004).

The Bayesian Information Criterion was derived from a Bayesian approach to model comparison (Kuha, 2004; 2005; Weakliem, 2004). A fundamental quantity in this framework is the Bayes Factor $p(y \mid M_1)/p(y \mid M_2)$, which is a measure of the evidence provided by the observed data $y$ in favour of model $M_1$ rather than model $M_2$ being the true model. A monotonic transformation of the Bayes Factor is approximated by the difference $\Delta \text{BIC}$ between the BIC values of the models $M_1$ and $M_2$. BIC itself is given by

$$\text{BIC} = -2 \cdot \ln (L(\hat{\theta}; y)) + S \cdot \log(N),$$

where $L(\hat{\theta}; y)$ denotes the maximum likelihood estimate for the observed data $y$, $S$ denotes the number of functionally independent model parameters, and $N$ denotes the number of observations. Analogous information to that provided by $\Delta \text{AIC}$ can be obtained by comparing each candidate model with the model in the set that is optimal according to this criterion. This is implemented by subtracting the minimal BIC value in the set from the BIC value of each candidate model. Thus, the best model obtains $\Delta \text{BIC} = 0$, and the larger $\Delta \text{BIC}$, the less the model is supported by the data.

AIC and BIC share the same structure. Their identical first term, $-2 \cdot \ln (L(\hat{\theta}; y))$, indicates how well a model fits the data. The second term penalises model complexity as indicated by the number of functionally independent model parameters, $S$. Thus, both information criteria implement the principle of parsimony. Comparing the penalty terms of AIC, $2 \cdot S$, and BIC, $S \cdot \log(N)$, reveals that the BIC penalises model complexity more strictly than the AIC if the number of observations exceeds eight. Hence, the BIC tends to prefer models with fewer parameters than the AIC.

In summary, the aim of evaluating the various MPT models suggested for the remember-know and remember-know-guess paradigms is to find out whether one or more of these models is a suitable measurement tool and how the candidate models fare in comparison. With regard to the first question, a model is a suitable measurement tool if it satisfies the validity criterion that model parameters pertaining to the accuracy of the recognition memory performance be invariant to manipulations of old-new response bias. This is indicated by the fit of the MPT models with equality restrictions imposed on their respective memory parameters to empirical data. Second, the suitability of the models can
be compared by means of the information criteria AIC and BIC, which balance a model’s goodness of fit against its complexity as indicated by the number of functionally independent model parameters. The next sections describe the results of evaluating the MPT models by means of a reanalysis of published datasets, starting with data from the classic remember-know paradigm without the response option guess.

6.2 Remember-Know Paradigm

6.2.1 Database

Searching the literature yielded seven experiments in which the old-new response bias was manipulated (cf. chapter 3). A range of different independent variables was employed, namely manipulating information about the (actually constant) percentage of targets in the recognition test (Hirshman & Henzler, 1998; Strack & Förster, 1995); the actual percentage of targets in the recognition test (Xu & Bellezza, 2001); explicit instructions to respond conservatively versus liberally (Postma, 1999); pay-off matrices (Erdfelder, 2000); and the test procedure of having participants report their decision for each item in two successive steps (old-new response, then in case of an old response remember-know response) or simultaneously in one step with the response options remember-know-new (Eldridge et al., 2002, Experiment 1; Hicks & Marsh, 1999).

In three of these experiments (Erdfelder, 2000; Hicks & Marsh, 1999; Hirshman & Henzler, 1998), the manipulation of old-new response bias was crossed with a manipulation known to affect recognition memory performance. Here, it is possible to evaluate the effect of manipulating the old-new response bias separately at each level of recognition memory performance. However, in two of the experiments (Erdfelder, 2000; Hirshman & Henzler, 1998), there is only one set of distractors common to both levels of expected recognition memory performance, rendering it impossible to obtain separate estimates of response rates to distractors at each level of expected recognition memory performance. This makes it meaningless to apply the 2HT task-oriented model to pairs of old-new response bias conditions at either level of expected memory performance for the following reason: The 2HT task-oriented model assumes that the probability of truly recognising a target as old, $D_{\text{old}}$, is equal to the probability of truly identifying a distractor as new, $D_{\text{new}}$. Therefore, the model contains a single parameter $D = D_{\text{old}} = D_{\text{new}}$ that represents recognition memory performance. This is a well-established assumption of 2HT models of recognition memory beyond the remember-know paradigm (e.g., Snodgrass &
Corwin, 1988). If an independent variable is expected to affect recognition memory performance, one expects $D_{\text{better}} > D_{\text{worse}}$, where the subscripts denote the levels of the independent variable related to better versus worse recognition memory performance, respectively. $D_{\text{better}} > D_{\text{worse}}$ implies (because of the model assumption $D_{\text{old}} = D_{\text{new}}$) that $D_{\text{old; better}} = D_{\text{new; better}} > D_{\text{new; worse}} = D_{\text{old; worse}}$. However, if an experiment contains only a single set of distractors shared by conditions that are expected to differ with regard to recognition memory performance, the design enforces $D_{\text{new; better}} = D_{\text{new; worse}} = D_{\text{new}}$. Hence, applying the 2HT task-oriented model implies $D_{\text{old; better}} = D_{\text{new; better}} = D_{\text{new; worse}} = D_{\text{old; worse}}$, in short $D_{\text{better}} = D_{\text{worse}}$. This is obviously nonsensical. Therefore, the 2HT task-oriented model is only meaningfully applicable if the experimental design allows for separate estimates of $D_{\text{new}}$ for each condition. Note that the other candidate models (i.e., 1HT task-oriented model, 2HT and 1HT multi-memory model) are applicable regardless of whether separate distractor sets are available in each condition.\textsuperscript{4} In total, the 2HT task-oriented model can meaningfully be applied to 7 datasets, whereas the other models can be applied to 11 datasets. If the observed frequencies for each response category summed across participants were not available, they were reconstructed by multiplying the mean response rates with the number of test items and participants per condition.\textsuperscript{5}

### 6.2.2 Results and Discussion

**Null-hypothesis testing.** Table 6.1 displays the results of fitting the four MPT models to each dataset with their respective memory parameters restricted to be equal across the levels of the old-new response bias manipulation.\textsuperscript{6} As outlined above, a model is a suitable measurement tool if it may be retained with these equality restrictions as assessed by the log-likelihood ratio statistic $G^2$. The significance level for each model test was set to $\alpha = .05$. The power to detect model violations of small effect size $\omega = .10$ (Cohen, 1988)

---

\textsuperscript{4} The 2HT multi-memory model, while assuming that distractors are truly identified as new with probability $d$, does not assume that this probability is equal to the probability $r + (1-r) \cdot k$ of retrieving a target from at least one memory system, thereby truly recognising it as old. Finally, the 1HT multi-memory and task-oriented models assume that distractors cannot be truly identified as new.

\textsuperscript{5} I am obliged to Francis Bellezza and Albert Postma for providing me with their original frequency data.

\textsuperscript{6} The multinomial analyses were conducted with the programs MBT (Hu, 1993) and HMMTree (Stahl & Klauer, 2005, in press).
<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>df</th>
<th>$G^2$</th>
<th>p</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ABIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eldridge, Sarfatti, &amp; Knowlton, 2002, Experiment 1 ($N = 2,593$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>43.96</td>
<td>0.000</td>
<td>3658.75</td>
<td>40.03</td>
<td>3693.91</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>33.37</td>
<td>0.000</td>
<td>3650.16</td>
<td>31.44</td>
<td>3691.19</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>18.97</td>
<td>0.000</td>
<td>3635.76</td>
<td>17.04</td>
<td>3676.79</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>1.93</td>
<td>0.165</td>
<td>3618.72</td>
<td></td>
<td>3659.74</td>
</tr>
<tr>
<td>Erdfelder, 2000, deep processing ($N = 4,764$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>0.03</td>
<td>0.986</td>
<td>8939.65</td>
<td>0.00</td>
<td>8978.46</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>0.03</td>
<td>0.860</td>
<td>8941.65</td>
<td>2.00</td>
<td>8986.93</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>0.02</td>
<td>0.899</td>
<td>8941.64</td>
<td>1.99</td>
<td>8986.92</td>
</tr>
<tr>
<td>Erdfelder, 2000, shallow processing ($N = 4,763$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>0.05</td>
<td>0.976</td>
<td>8695.75</td>
<td>0.00</td>
<td>8734.56</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>0.05</td>
<td>0.820</td>
<td>8697.76</td>
<td>2.00</td>
<td>8743.04</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>0.02</td>
<td>0.875</td>
<td>8697.73</td>
<td>1.98</td>
<td>8743.01</td>
</tr>
<tr>
<td>Hicks &amp; Marsh, 1999, 4.5s ($N = 7,841$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>8.91</td>
<td>0.012</td>
<td>12054.15</td>
<td>3.95</td>
<td>12095.95</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>2.97</td>
<td>0.085</td>
<td>12050.21</td>
<td>0.00</td>
<td>12098.98</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>6.58</td>
<td>0.010</td>
<td>12053.82</td>
<td>3.61</td>
<td>12102.59</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>3.12</td>
<td>0.077</td>
<td>12050.36</td>
<td>0.15</td>
<td>12099.13</td>
</tr>
<tr>
<td>Hicks &amp; Marsh, 1999, 1s ($N = 7,840$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>44.52</td>
<td>0.000</td>
<td>13958.49</td>
<td>39.78</td>
<td>14000.29</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>33.72</td>
<td>0.000</td>
<td>13949.69</td>
<td>30.98</td>
<td>13998.46</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>2.74</td>
<td>0.098</td>
<td>13918.71</td>
<td>0.00</td>
<td>13967.48</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>4.17</td>
<td>0.041</td>
<td>13920.14</td>
<td>1.43</td>
<td>13968.90</td>
</tr>
<tr>
<td>Hirshman &amp; Henzler, 1998, 2s ($N = 3,841$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>4.42</td>
<td>0.110</td>
<td>6669.01</td>
<td>0.15</td>
<td>6706.53</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>3.11</td>
<td>0.078</td>
<td>6669.70</td>
<td>0.84</td>
<td>6713.48</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>2.27</td>
<td>0.132</td>
<td>6668.86</td>
<td>0.00</td>
<td>6712.64</td>
</tr>
<tr>
<td>Hirshman &amp; Henzler, 1998, 0.5s ($N = 3,840$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td></td>
<td>6</td>
<td>2</td>
<td>8.71</td>
<td>0.013</td>
<td>6385.14</td>
<td>3.02</td>
<td>6422.66</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td></td>
<td>7</td>
<td>1</td>
<td>3.70</td>
<td>0.054</td>
<td>6382.13</td>
<td>0.00</td>
<td>6425.90</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td></td>
<td>7</td>
<td>1</td>
<td>5.77</td>
<td>0.016</td>
<td>6384.20</td>
<td>2.07</td>
<td>6427.97</td>
</tr>
</tbody>
</table>

(Continued)
### Evaluation of the Measurement Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>θ</th>
<th>df</th>
<th>$G^2$</th>
<th>$p$</th>
<th>AIC</th>
<th>ΔAIC</th>
<th>BIC</th>
<th>ΔBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postma, 1999 ($N = 2,880$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory rk</td>
<td>6</td>
<td>2</td>
<td>24.45</td>
<td>0.00</td>
<td>4136.45</td>
<td>22.44</td>
<td>4172.24</td>
<td>16.47</td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory rkd</td>
<td>7</td>
<td>1</td>
<td>22.52</td>
<td>0.00</td>
<td>4136.53</td>
<td>22.51</td>
<td>4178.28</td>
<td>22.51</td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>12.00</td>
<td>0.001</td>
<td>4126.00</td>
<td>11.99</td>
<td>4167.76</td>
<td>11.99</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>0.01</td>
<td>0.922</td>
<td>4114.01</td>
<td><strong>0.00</strong></td>
<td>4155.77</td>
<td><strong>0.00</strong></td>
<td></td>
</tr>
<tr>
<td>Strack &amp; Förster, 1995, low word frequency ($N = 1,150$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory rk</td>
<td>6</td>
<td>2</td>
<td>4.93</td>
<td>0.085</td>
<td>1922.02</td>
<td>2.86</td>
<td>1952.30</td>
<td><strong>0.00</strong></td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory rkd</td>
<td>7</td>
<td>1</td>
<td>4.00</td>
<td>0.046</td>
<td>1923.08</td>
<td>3.93</td>
<td>1958.42</td>
<td>6.11</td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>3.43</td>
<td>0.064</td>
<td>1922.51</td>
<td>3.36</td>
<td>1957.85</td>
<td>5.54</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>0.07</td>
<td>0.795</td>
<td>1919.15</td>
<td><strong>0.00</strong></td>
<td>1954.49</td>
<td>2.18</td>
<td></td>
</tr>
<tr>
<td>Strack &amp; Förster, 1995, high word frequency ($N = 1,150$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory rk</td>
<td>6</td>
<td>2</td>
<td>2.76</td>
<td>0.252</td>
<td>1795.71</td>
<td>0.72</td>
<td>1826.00</td>
<td><strong>0.00</strong></td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory rkd</td>
<td>7</td>
<td>1</td>
<td>1.92</td>
<td>0.166</td>
<td>1796.88</td>
<td>1.89</td>
<td>1832.21</td>
<td>6.21</td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>2.09</td>
<td>0.148</td>
<td>1797.05</td>
<td>2.06</td>
<td>1832.38</td>
<td>6.38</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented D</td>
<td>7</td>
<td>1</td>
<td>0.04</td>
<td>0.851</td>
<td>1794.99</td>
<td><strong>0.00</strong></td>
<td>1830.32</td>
<td>4.32</td>
<td></td>
</tr>
<tr>
<td>Xu &amp; Bellezza, 2001 ($N = 14,800$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory rk</td>
<td>8</td>
<td>4</td>
<td>131.64</td>
<td>0.000</td>
<td>27812.68</td>
<td>109.26</td>
<td>27891.92</td>
<td>98.65</td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory rkd</td>
<td>9</td>
<td>3</td>
<td>21.08</td>
<td>0.000</td>
<td>27704.12</td>
<td>0.70</td>
<td>27793.27</td>
<td><strong>0.00</strong></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented D</td>
<td>10</td>
<td>2</td>
<td>111.50</td>
<td>0.000</td>
<td>27796.54</td>
<td>93.12</td>
<td>27895.59</td>
<td>102.32</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented D</td>
<td>10</td>
<td>2</td>
<td>18.38</td>
<td>0.000</td>
<td>27703.42</td>
<td><strong>0.00</strong></td>
<td>27802.47</td>
<td>9.20</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column θ displays the number of functionally independent model parameters given these restrictions.

The 1HT multi-memory model was rejected for 6 of 11 datasets. To assess whether this finding has any significance for the global validity of the 1HT multi-memory model, it is useful to consider the 11 datasets as a sample and employ a binomial test of the null hypothesis that in the population of datasets with experimental manipulations of old-new response bias, the probability $\pi$ of rejecting the model is $\pi \leq .05$. This reflects the expected value of the relative frequency of incorrectly rejecting a valid model, i.e. committing a Type I error, when employing the significance level of $\alpha = .05$ in each model test. Hence, $H_0: \pi \leq .05$ exceeded 99% for all datasets apart from Strack and Förster (1995). The power calculations were conducted with the program GPOWER (Erdfelder, Faul, & Buchner, 1996; Faul & Erdfelder, 1992).

---

7 The power calculations were conducted with the program GPOWER (Erdfelder, Faul, & Buchner, 1996; Faul & Erdfelder, 1992).
represents the hypothesis that the model is a globally valid measurement model. $H_0: \pi \leq .05$ was also tested at $\alpha = .05$. The probability of incorrectly rejecting the 1HT multi-memory model in at least 6 of 11 datasets is $p < .0001$, indicating that the 1HT multi-memory model is not globally valid. Both the 2HT multi-memory model and the 1HT task-oriented model were rejected for 5 of 11 datasets. The probability of incorrectly rejecting a model in at least 5 of 11 datasets is $p = .0001$, indicating that neither the 2HT multi-memory model nor the 1HT task-oriented model is globally valid. The last candidate model, the 2HT task-oriented model, can meaningfully be applied to only 7 of the 11 datasets for the reasons given above. For 2 of 7 datasets, the 2HT task-oriented model was rejected. The probability of incorrectly rejecting a model in at least 2 of 7 datasets is $p = .0444$. Hence, $H_0: \pi \leq .05$ is to be rejected at the significance level of $\alpha = .05$ and there is reason to doubt the global validity of the 2HT task-oriented model, too, if not quite as overpowering as for the other models. Thus, according to this reanalysis, none of the models can be trusted to be a globally valid measurement model.

Further insights can be gleaned from considering how the model rejections are distributed across the datasets. The dataset by Xu and Bellezza (2001) is the only one for which all candidate models were rejected. Its number of observations, $N = 14,800$, is by far the largest of the datasets considered, leading to a disproportionate power to detect even minute model violations. However, the memory parameter estimates per condition suggest that the differences may be relevant. For each of the other 10 datasets, at least one model could be retained. In particular, at least one task-oriented model could be retained for all datasets but one (Hirshman & Henzler, 1998, 0.5s), for which the 1HT task-oriented model was rejected and the 2HT task-oriented model could not meaningfully be applied. On the other hand, for three datasets both the 2HT and the 1HT multi-memory models were rejected (Eldridge et al., 2002, Experiment 1; Hicks & Marsh, 1999, 1s; Postma, 1999). In these datasets that are incompatible with the multi-memory models, old-new response bias was manipulated by means of the explicit instruction to respond conservatively versus liberally in the old-new decision (Postma, 1999) or by contrasting the standard 2-step test procedure with a 1-step test procedure (Eldridge et al., 2002, Experiment 1; Hicks & Marsh, 1999, 1s). However, in the third dataset that varied the test procedure, the easier condition of Hicks and Marsh (1999) with 4.5s presentation duration, the 2HT multi-memory model could be retained as well as the 2HT task-oriented model, whereas both 1HT models were rejected.
AIC. Further comparisons of the performance of the MPT models are afforded by information criteria. Table 6.1 shows the results. According to the AIC, the 2HT task-oriented model is the best model for five of the seven datasets to which it is meaningfully applicable. These are the datasets from Eldridge et al. (2002, Experiment 1) and Postma (1999), for which all other candidate models were rejected by the log-likelihood ratio test with \( \alpha = .05 \); both datasets from Strack and Förster (1995), for which at least three candidate models could be retained; and the dataset from Xu and Bellezza (2001), for which all candidate models were rejected. In the two remaining datasets (Hicks & Marsh, 1999), the 2HT task-oriented model was ranked second-best by the AIC. In the easier condition with 4.5s presentation duration, the AIC value of the 2HT task-oriented model is practically tied with that of the 2HT multi-memory model. In the more difficult condition with 1s presentation duration, the 1HT task-oriented model was ranked best by the AIC, the only model that could be retained by the log-likelihood ratio test. In sum, the model comparisons by means of the AIC corroborate and extend the findings obtained with null-hypothesis testing by providing further evidence that the 2HT task-oriented model is the most suitable measurement tool among the candidate models.

BIC. Turning to the BIC, two different cases are apparent for the datasets with two levels of old-new response bias. First, for the datasets for which both multi-memory models were rejected, the BIC converges with the AIC on the same best model, the task-oriented model with either two high thresholds (Eldridge et al., 2002, Experiment 1; Postma, 1999) or one high threshold (Hicks & Marsh, 1999, 1s). Else, the 1HT multi-memory model is optimal according to the BIC. Note that with the equality restrictions required by the validity criterion that memory parameters be invariant to manipulations of old-new response bias, this model has one parameter fewer than the other candidate models. In considering this discrepancy, note that if AIC and BIC do not converge on the same candidate model, it is to be expected that the BIC prefers a model with fewer parameters than the AIC, because the BIC penalises the number of model parameters more strictly. The BIC uses a penalty term, \( S \cdot \log(N) \), that depends not only on the number of functionally independent model parameters \( S \), but also on the number of observations \( N \). As the numbers of observations in the datasets under consideration are very large, the BIC has a strong tendency to prefer more parsimonious models.

In this context, it is important to note that in contrast to the 1HT and 2HT multi-memory models, the 1HT and 2HT task-oriented models may well waste parsimony if restricted
according to the validity criterion. The validity criterion demands only that the parameter $D$, which represents recognition memory performance, be constant across levels of old-new response bias manipulations. Retrieval experience for items truly recognised as old, captured by the model parameter $R$, may or may not be affected by manipulations of old-new response bias. That is, the task-oriented models’ validity is not impaired if the parameter $R$ is affected by an experimental manipulation intended and established to affect the old-new response bias. However, if a manipulation of old-new response bias does not affect the retrieval experience for truly recognised items, allowing for $R_{\text{conservative}} = R_{\text{liberal}}$, the task-oriented models without this restriction unnecessarily spend one model parameter on specifying separate parameters for both conditions, thus wasting parsimony that the task-oriented models’ structure would permit to achieve. Hence, it is interesting to extend the model comparison by considering the task-oriented models with the additional restriction $R_{\text{conservative}} = R_{\text{liberal}}$, not required to hold for the models to be valid. Note that the 1HT task-oriented model with equality restrictions imposed on both $D$ and $R$ is data-equivalent to the 1HT multi-memory model with equality restrictions imposed on $r$ and $k$.

Table 6.2 displays the results of the extended model comparison including the task-oriented models with equality restrictions on $D$ and $R$. Three different cases are apparent.
<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>df</th>
<th>G^2</th>
<th>p</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ΔBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erdfelder, 2000, shallow processing (N = 4,763)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>DR</td>
<td>6</td>
<td>2</td>
<td>0.05</td>
<td>0.976</td>
<td>8695.75</td>
<td><strong>0.00</strong></td>
<td>8734.56</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>0.05</td>
<td>0.820</td>
<td>8697.76</td>
<td>2.00</td>
<td>8743.04</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>rkd</td>
<td>7</td>
<td>1</td>
<td>0.02</td>
<td>0.875</td>
<td>8697.73</td>
<td>1.98</td>
<td>8743.01</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>0.05</td>
<td>0.820</td>
<td>8697.76</td>
<td>2.00</td>
<td>8743.04</td>
</tr>
<tr>
<td>Hicks &amp; Marsh, 1999, 4.5s (N = 7,841)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>6</td>
<td>2</td>
<td>8.91</td>
<td>0.012</td>
<td>12054.15</td>
<td>3.95</td>
<td>12095.95</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>5.47</td>
<td>0.065</td>
<td>12050.71</td>
<td>0.51</td>
<td>12092.51</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>6.58</td>
<td>0.010</td>
<td>12053.82</td>
<td>3.61</td>
<td>12102.59</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>3.12</td>
<td>0.077</td>
<td>12050.36</td>
<td>0.15</td>
<td>12099.13</td>
</tr>
<tr>
<td>Hicks &amp; Marsh, 1999, 1s (N = 7,840)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>6</td>
<td>2</td>
<td>44.52</td>
<td>0.000</td>
<td>13958.49</td>
<td>39.78</td>
<td>14000.29</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>36.67</td>
<td>0.000</td>
<td>13950.64</td>
<td>31.93</td>
<td>13992.44</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>33.72</td>
<td>0.000</td>
<td>13949.69</td>
<td>30.98</td>
<td>13998.46</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>2.74</td>
<td>0.098</td>
<td>13918.71</td>
<td><strong>0.00</strong></td>
<td>13967.48</td>
</tr>
<tr>
<td>Hirshman &amp; Henzler, 1998, 2s (N = 3,841)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>6</td>
<td>2</td>
<td>4.42</td>
<td>0.110</td>
<td>6669.01</td>
<td>0.15</td>
<td>6706.53</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>3.11</td>
<td>0.078</td>
<td>6669.70</td>
<td>0.84</td>
<td>6713.48</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>2.27</td>
<td>0.132</td>
<td>6668.86</td>
<td><strong>0.00</strong></td>
<td>6712.64</td>
</tr>
<tr>
<td>Hirshman &amp; Henzler, 1998, 0.5s (N = 3,840)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>6</td>
<td>2</td>
<td>8.71</td>
<td>0.013</td>
<td>6385.14</td>
<td>3.02</td>
<td>6422.66</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>3.70</td>
<td>0.054</td>
<td>6382.13</td>
<td><strong>0.00</strong></td>
<td>6425.90</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>5.77</td>
<td>0.016</td>
<td>6384.20</td>
<td>2.07</td>
<td>6427.97</td>
</tr>
<tr>
<td>Postma, 1999 (N = 2,880)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>6</td>
<td>2</td>
<td>24.45</td>
<td>0.000</td>
<td>4136.45</td>
<td>22.44</td>
<td>4172.24</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>7</td>
<td>1</td>
<td>26.29</td>
<td>0.000</td>
<td>4138.29</td>
<td>24.28</td>
<td>4174.09</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>22.52</td>
<td>0.000</td>
<td>4136.53</td>
<td>22.51</td>
<td>4178.28</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>12.00</td>
<td>0.001</td>
<td>4126.00</td>
<td>11.99</td>
<td>4167.76</td>
</tr>
</tbody>
</table>
| (Continued)
<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>0</th>
<th>df</th>
<th>$G^2$</th>
<th>$p$</th>
<th>AIC</th>
<th>$\Delta$AIC</th>
<th>BIC</th>
<th>$\Delta$BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT task-oriented</td>
<td>DR</td>
<td>6</td>
<td>2</td>
<td>4.93</td>
<td>0.085</td>
<td>1922.02</td>
<td>2.86</td>
<td>1952.30</td>
<td>0.54</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>6</td>
<td>2</td>
<td>4.39</td>
<td>0.111</td>
<td>1921.48</td>
<td>2.33</td>
<td>1951.77</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>DR</td>
<td>7</td>
<td>1</td>
<td>4.00</td>
<td>0.046</td>
<td>1923.08</td>
<td>3.93</td>
<td>1958.42</td>
<td>6.65</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>3.43</td>
<td>0.064</td>
<td>1922.51</td>
<td>3.36</td>
<td>1957.85</td>
<td>6.08</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>7</td>
<td>1</td>
<td>0.07</td>
<td>0.795</td>
<td>1919.15</td>
<td>0.00</td>
<td>1954.49</td>
<td>2.72</td>
</tr>
<tr>
<td>Xu &amp; Bellezza, 2001 ($N = 14,800$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>DR</td>
<td>8</td>
<td>4</td>
<td>131.64</td>
<td>0.000</td>
<td>27812.68</td>
<td>109.26</td>
<td>27891.92</td>
<td>98.65</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>DR</td>
<td>8</td>
<td>4</td>
<td>33.69</td>
<td>0.000</td>
<td>27714.73</td>
<td>11.30</td>
<td>27793.97</td>
<td>0.70</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>10</td>
<td>2</td>
<td>111.50</td>
<td>0.000</td>
<td>27796.54</td>
<td>93.12</td>
<td>27895.59</td>
<td>102.32</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>10</td>
<td>2</td>
<td>18.38</td>
<td>0.000</td>
<td>27703.42</td>
<td>0.00</td>
<td>27802.47</td>
<td>9.20</td>
</tr>
</tbody>
</table>

**Note.** The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column 0 displays the number of functionally independent model parameters given these restrictions.

for the datasets with two levels of old-new response bias. For the datasets for which both multi-memory models were rejected, the best model is unchanged the task-oriented model with only $D$ restricted to be equal, with either two high thresholds (Eldridge et al., 2002, Experiment 1; Postma, 1999) or one high threshold (Hicks & Marsh, 1999, 1s). The 2HT task-oriented model with restrictions on $D$ and $R$ is the optimal candidate model according to the BIC for the remaining datasets to which it can meaningfully be applied. In particular, the 2HT task-oriented model restricted such as not to waste parsimony performed better than the previously BIC-best 1HT multi-memory model and the data-equivalent 1HT task-oriented model with restrictions on both $D$ and $R$, where all three models have the same number of functionally independent models parameters. Finally, for those datasets to which the 2HT task-oriented model cannot meaningfully be applied, the 1HT multi-
6 Evaluation of the Measurement Models

memory model and the data-equivalent 1HT task-oriented model with restrictions on both $D$ and $R$ are BIC-optimal.

Thus, the BIC converges with the other indicators in favouring the 2HT task-oriented model as the most suitable measurement tool among the candidate models under the condition that it is restricted such as not to waste parsimony that can be achieved by virtue of the model structure. At times, this involves restricting $R$ in addition to $D$, whereas for other datasets this additional restriction causes deterioration of the BIC value. Note that for these latter datasets (Eldridge et al., 2002, Experiment 1; Hicks & Marsh, 1999, 1s; Postma, 1999), the hypothesis that 2HT task-oriented model with restrictions on $D$ and $R$ fits the data, assessed by the log-likelihood ratio statistic $G^2$ with $\alpha = .05$, was rejected, underlining the convergence of results of BIC and null-hypothesis testing.

6.2.3 Conclusion

Two preliminary conclusions may be drawn from the systematic evaluation of the MPT models for the remember-know paradigm. First, the methods of null-hypothesis testing, AIC and BIC provide converging evidence that the 2HT task-oriented model is the most promising measurement tool among the evaluated models. However, additional data are necessary to give the 2HT task-oriented model the chance to prove globally valid with an extended database. Second, there exist datasets (Eldridge et al., 2002; Hicks & Marsh, 1999, 1s; Postma, 1999) that are incompatible with the multi-memory systems model as implemented by the multi-memory model with one or two high thresholds, but that can be fit well by the task-oriented model with one or two high thresholds that simply imply that the recognition memory performance is constant.

6.3 Remember-Know-Guess Paradigm

6.3.1 Database

Turning to the remember-know-guess paradigm, a search of the literature yielded a total of four experiments in which the old-new response bias was manipulated (cf. chapter 3). The range of independent variables comprised manipulating information about the (actually constant) percentage of targets in the recognition test (Gardiner et al., 1997); the actual percentage of targets in the recognition test (Xu & Bellezza, 2001); pay-off matrices (Erdfelder, 2000); and the test procedure of having participants report their decision for each item in two successive steps (old-new response, then in case of an old response
remember-know-guess response) or simultaneously in one step with the response option remember-know-guess-new (Eldridge et al., 2002, Experiment 2). In two of these experiments (Erdfelder, 2000; Gardiner et al., 1997), the manipulation of old-new response bias was crossed with a manipulation known to affect recognition memory performance. Accordingly, it is possible to evaluate the effect of manipulating the old-new response bias separately at each level of recognition memory performance. However, in the experiment by Erdfelder (2000), there is only one set of distractors common to both levels of expected memory performance, rendering it meaningless to apply the 2HT task-oriented models (which allow versus preclude guess responses to truly recognised targets) to pairs of response-bias conditions at either level of expected memory performance for the reason detailed above. Hence, the 2HT task-oriented models can meaningfully be applied to four datasets, whereas the 1HT task-oriented models (which allow versus preclude guess responses to truly recognised targets) as well as the multi-memory models with one or two high thresholds can be applied to six datasets.

### 6.3.2 Results and Discussion

**Null-hypothesis testing.** Table 6.3 displays the results of fitting the six MPT models to each dataset with their respective memory parameters restricted to be equal across the levels of the old-new response bias manipulation. As outlined above, a model is a suitable measurement tool if it may be retained with these equality restrictions as assessed by the log-likelihood ratio statistic $G^2$. The significance level for each model test was set to $\alpha = .05$. The power to detect violations of small effect size $w = .10$ (Cohen, 1988) exceeded 98% for all datasets. Both 2HT task-oriented models, regardless of whether guess responses to truly recognised targets are allowed or precluded, were rejected for each of the four datasets to which they could meaningfully be applied. Hence, the 2HT task-oriented models for the remember-know-guess paradigm did not receive any support. The other candidate models were applied to six datasets. Both the 1HT task-oriented model that precludes guess responses to truly recognised targets and the 1HT multi-memory model were rejected for five of six datasets, and the 2HT multi-memory model was rejected for four of six datasets. The significance of these results for global model validity was assessed by testing the hypothesis $H_0: \pi \leq .05$ that the relative frequency of rejecting a candidate model, $\pi$, does not exceed the expected value for incorrectly rejecting a valid model due to committing Type I errors given $\alpha = .05$ for each model test. The probability
Table 6.3
Results of the Evaluation of the MPT Models for the Remember-Know-Guess Paradigm with Parameter Restrictions Required by the Validity Criterion

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>df</th>
<th>$G^2$</th>
<th>$p$</th>
<th>AIC</th>
<th>$\Delta$AIC</th>
<th>BIC</th>
<th>$\Delta$BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eldridge, Sarfatti, &amp; Knowlton, 2002, Experiment 2 ($N = 2,592$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory $rk$</td>
<td>8</td>
<td>4</td>
<td>61.86</td>
<td>0.000</td>
<td>4978.56</td>
<td>54.04</td>
<td>5025.44</td>
<td>38.52</td>
</tr>
<tr>
<td>2HT multi-memory $rkd$</td>
<td>9</td>
<td>3</td>
<td>15.48</td>
<td>0.001</td>
<td>4934.18</td>
<td>9.65</td>
<td>4986.92</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>57.76</td>
<td>0.000</td>
<td>4976.46</td>
<td>51.94</td>
<td>5029.20</td>
<td>42.28</td>
</tr>
<tr>
<td>2HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>19.46</td>
<td>0.000</td>
<td>4938.16</td>
<td>13.64</td>
<td>4990.90</td>
<td>3.98</td>
</tr>
<tr>
<td>1HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>1.82</td>
<td>0.177</td>
<td>4924.52</td>
<td>0.00</td>
<td>4988.99</td>
<td>2.07</td>
</tr>
<tr>
<td>2HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>19.46</td>
<td>0.000</td>
<td>4942.16</td>
<td>17.64</td>
<td>5006.62</td>
<td>19.70</td>
</tr>
<tr>
<td>Erdfelder, 2000, deep processing ($N = 4,616$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory $rk$</td>
<td>8</td>
<td>4</td>
<td>19.69</td>
<td>0.001</td>
<td>10040.47</td>
<td>13.64</td>
<td>10091.97</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT multi-memory $rkd$</td>
<td>9</td>
<td>3</td>
<td>13.36</td>
<td>0.004</td>
<td>10036.15</td>
<td>9.31</td>
<td>10094.08</td>
<td>2.11</td>
</tr>
<tr>
<td>1HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>11.74</td>
<td>0.008</td>
<td>10034.53</td>
<td>7.69</td>
<td>10092.46</td>
<td>0.49</td>
</tr>
<tr>
<td>1HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>0.05</td>
<td>0.824</td>
<td>10026.84</td>
<td>0.00</td>
<td>10097.65</td>
<td>5.68</td>
</tr>
<tr>
<td>Erdfelder, 2000, shallow processing ($N = 4,619$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory $rk$</td>
<td>8</td>
<td>4</td>
<td>6.25</td>
<td>0.181</td>
<td>9866.37</td>
<td>1.72</td>
<td>9917.88</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT multi-memory $rkd$</td>
<td>9</td>
<td>3</td>
<td>2.53</td>
<td>0.469</td>
<td>9864.66</td>
<td>0.00</td>
<td>9922.60</td>
<td>4.72</td>
</tr>
<tr>
<td>1HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>5.93</td>
<td>0.115</td>
<td>9868.05</td>
<td>3.40</td>
<td>9926.00</td>
<td>8.12</td>
</tr>
<tr>
<td>1HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>0.00</td>
<td>0.952</td>
<td>9866.13</td>
<td>1.47</td>
<td>9936.94</td>
<td>19.07</td>
</tr>
<tr>
<td>Gardiner, Richardson-Klavehn, &amp; Ramponi, 1997, low word frequency ($N = 2,304$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory $rk$</td>
<td>8</td>
<td>4</td>
<td>31.09</td>
<td>0.000</td>
<td>4785.33</td>
<td>22.57</td>
<td>4831.27</td>
<td>16.82</td>
</tr>
<tr>
<td>2HT multi-memory $rkd$</td>
<td>9</td>
<td>3</td>
<td>6.53</td>
<td>0.089</td>
<td>4762.77</td>
<td>0.00</td>
<td>4814.45</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>31.05</td>
<td>0.000</td>
<td>4787.29</td>
<td>24.52</td>
<td>4838.97</td>
<td>24.52</td>
</tr>
<tr>
<td>2HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>9.09</td>
<td>0.028</td>
<td>4765.33</td>
<td>2.56</td>
<td>4817.01</td>
<td>2.56</td>
</tr>
<tr>
<td>1HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>4.92</td>
<td>0.026</td>
<td>4765.16</td>
<td>2.40</td>
<td>4828.33</td>
<td>13.88</td>
</tr>
<tr>
<td>2HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>9.09</td>
<td>0.003</td>
<td>4769.33</td>
<td>6.56</td>
<td>4832.50</td>
<td>18.05</td>
</tr>
<tr>
<td>Gardiner, Richardson-Klavehn, &amp; Ramponi, 1997, high word frequency ($N = 2,304$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory $rk$</td>
<td>8</td>
<td>4</td>
<td>22.26</td>
<td>0.000</td>
<td>4934.66</td>
<td>16.07</td>
<td>4980.60</td>
<td>6.35</td>
</tr>
<tr>
<td>2HT multi-memory $rkd$</td>
<td>9</td>
<td>3</td>
<td>8.17</td>
<td>0.043</td>
<td>4922.57</td>
<td>3.98</td>
<td>4974.25</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>22.19</td>
<td>0.000</td>
<td>4936.59</td>
<td>18.00</td>
<td>4988.27</td>
<td>14.02</td>
</tr>
<tr>
<td>2HT task-orient'd $g=0$ $D$</td>
<td>9</td>
<td>3</td>
<td>10.34</td>
<td>0.016</td>
<td>4924.74</td>
<td>6.14</td>
<td>4976.42</td>
<td>2.17</td>
</tr>
<tr>
<td>1HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>0.19</td>
<td>0.660</td>
<td>4918.59</td>
<td>0.00</td>
<td>4981.76</td>
<td>7.51</td>
</tr>
<tr>
<td>2HT task-oriented $D$</td>
<td>11</td>
<td>1</td>
<td>8.67</td>
<td>0.003</td>
<td>4927.07</td>
<td>8.48</td>
<td>4990.24</td>
<td>15.98</td>
</tr>
</tbody>
</table>

(Continued)
### Evaluation of the Measurement Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>θ</th>
<th>df</th>
<th>$G^2$</th>
<th>p</th>
<th>AIC</th>
<th>ΔAIC</th>
<th>BIC</th>
<th>ΔBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT multi-memory rk</td>
<td>11</td>
<td>7</td>
<td>178.60</td>
<td>0.000</td>
<td>33733.60</td>
<td>141.05</td>
<td>33816.61</td>
<td>133.50</td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory rkd</td>
<td>12</td>
<td>6</td>
<td>35.55</td>
<td>0.000</td>
<td>33592.55</td>
<td>0.00</td>
<td>33683.11</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1HT task-orient'd g=0 D</td>
<td>13</td>
<td>5</td>
<td>177.60</td>
<td>0.000</td>
<td>33736.60</td>
<td>144.05</td>
<td>33834.71</td>
<td>151.60</td>
<td></td>
</tr>
<tr>
<td>2HT task-orient'd g=0 D</td>
<td>13</td>
<td>5</td>
<td>167.90</td>
<td>0.000</td>
<td>33726.90</td>
<td>134.35</td>
<td>33825.01</td>
<td>141.90</td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented D</td>
<td>16</td>
<td>2</td>
<td>75.97</td>
<td>0.000</td>
<td>33640.97</td>
<td>48.42</td>
<td>33761.72</td>
<td>78.61</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented D</td>
<td>16</td>
<td>2</td>
<td>167.90</td>
<td>0.000</td>
<td>33732.90</td>
<td>140.35</td>
<td>33853.65</td>
<td>170.54</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column θ displays the number of functionally independent model parameters given these restrictions.

The probability of incorrectly rejecting a candidate model in at least four of six datasets is $p = .0001$, indicating that neither the multi-memory model with either one or two high thresholds nor the 1HT task-oriented model that precludes guess responses to truly recognised targets is globally valid. The last candidate model, the 1HT task-oriented model that allows for guess responses to truly recognised targets, was rejected for two of six datasets. The probability of incorrectly rejecting a candidate model in at least two of six datasets is $p = .0328$. Hence, $H_0: \pi \leq .05$ is to be rejected at the significance level of $\alpha = .05$ and there is reason to doubt the global validity of the 1HT task-oriented model that allows for guess responses to truly recognised targets, too, if not quite as overpowering as for the other candidate models. Thus, according to this reanalysis, none of the candidate models can be trusted to be a globally valid measurement model.

Further insights can be gleaned from considering how the model rejections are distributed across the datasets. The dataset by Xu & Bellezza (2001) is the only one for which all candidate models were rejected. Its number of observations, $N = 14,000$, is by far the largest of the datasets considered, leading to a disproportionate power to detect even minute model violations. However, the memory parameter estimates per condition suggest that the differences may be relevant. For four datasets, only one model could be retained, once the 2HT multi-memory model (Gardiner et al. 1997, low word frequency), and three times the 1HT task-oriented model that allows for guess responses to truly recognised targets (Eldridge et al., 2002, Experiment 2; Erdfelder, 2000, deep processing; Gardiner et al., 1997, high word frequency). In the remaining dataset (Erdfelder, 2000, shallow processing), all models could be retained.
AIC. Further comparisons of the performance of the MPT models are afforded by information criteria. Table 6.3 shows the results. According to the AIC, the 1HT task-oriented model that allows for guess responses to truly recognised targets was ranked best for those datasets for which only this model could be retained as tested by the log-likelihood ratio test (Eldridge et al., 2002, Experiment 2; Erdfelder, 2000, deep processing; Gardiner et al., 1997, high word frequency). For the remaining three datasets, the 2HT multi-memory model was ranked best by the AIC (Erdfelder, 2000, shallow processing; Gardiner et al., 1997, low word frequency; Xu & Bellezza, 2001).

Note that without parameter restrictions, both candidate models have six parameters per experimental condition. Among these is a single memory parameter for the 1HT task-oriented model that allows for guess responses to truly recognised targets, but three memory parameters for the 2HT multi-memory model. Consequently, if the models are restricted according to the validity criterion that memory parameters be equal across levels of old-new response bias, the 2HT multi-memory model is more parsimonious than the 1HT task-oriented model that allows for guess responses to truly recognised targets. However, note that the latter model restricted according to the validity criterion may well waste parsimony that it could in principle achieve by virtue of its structure: Two parameters that capture the retrieval experience for truly recognised targets, $g$ and $R$, are not restricted to be constant across conditions, because this is not required for the model to be valid. However, they may not be affected by the experimental manipulation, in which case two parameters are wasted. With these restrictions added, the 1HT task-oriented model that allows for guess responses to truly recognised targets has the same number of functionally independent model parameters as the 2HT multi-memory model.

Table 6.4 displays the results of the extended model comparison that includes the four task-oriented models with equality restrictions on $D$, $g$ and $R$. The 1HT task-oriented model that allows for guess responses to targets is the AIC-optimal model for five of the six datasets, either with equality restrictions on $D$, $g$ and $R$ (Eldridge et al., 2002, Experiment 2; Erdfelder, 2000, shallow processing; Gardiner et al., 1997, low word frequency; Xu & Bellezza, 2001).

8 For the 1HT and 2HT task-oriented models that preclude guess responses to truly recognised targets by means of the restriction $g = 0$, the only additional equality restriction is imposed on $R$. Note that the 1HT task-oriented model with $g = 0$ and equality restrictions imposed on $R$ is data-equivalent to the 1HT multi-memory model with equality restriction imposed on $r$ and $k$. 
frequency) or with an equality restriction only on \( D \) (Erdfelder, 2000, deep processing; Gardiner et al., 1997, high word frequency). Only for the dataset from Xu & Bellezza (2001), for which no model could be retained, was the 2HT multi-memory model ranked best by the AIC.

In sum, the AIC provides evidence that the 1HT task-oriented model that allows for guess responses to targets is the most suitable measurement tool for the remember-know-guess paradigm. This converges with the finding from null-hypothesis testing that there is

<table>
<thead>
<tr>
<th>Model</th>
<th>( g=0 )</th>
<th>( \theta )</th>
<th>( df )</th>
<th>( G^2 )</th>
<th>( p )</th>
<th>AIC</th>
<th>( \Delta )AIC</th>
<th>BIC</th>
<th>( \Delta )BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT task-orient'd</td>
<td>DR</td>
<td>8</td>
<td>4</td>
<td>61.86</td>
<td>0.000</td>
<td>4978.56</td>
<td>54.26</td>
<td>5025.44</td>
<td>48.40</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>8</td>
<td>4</td>
<td>28.48</td>
<td>0.000</td>
<td>4945.18</td>
<td>20.88</td>
<td>4992.06</td>
<td>15.02</td>
</tr>
<tr>
<td>2HT task-orient'd</td>
<td>DR</td>
<td>9</td>
<td>3</td>
<td>15.48</td>
<td>0.001</td>
<td>4934.18</td>
<td>9.88</td>
<td>4986.92</td>
<td>9.88</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>rkd</td>
<td>9</td>
<td>3</td>
<td>5.60</td>
<td>0.133</td>
<td>4924.30</td>
<td>0.00</td>
<td>4977.04</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>( DgR )</td>
<td>9</td>
<td>3</td>
<td>19.46</td>
<td>0.000</td>
<td>4947.18</td>
<td>22.88</td>
<td>4999.92</td>
<td>22.88</td>
</tr>
<tr>
<td>1HT task-orient'd</td>
<td>( D )</td>
<td>9</td>
<td>3</td>
<td>19.46</td>
<td>0.000</td>
<td>4938.16</td>
<td>13.86</td>
<td>4990.90</td>
<td>13.86</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>( D )</td>
<td>9</td>
<td>3</td>
<td>57.76</td>
<td>0.000</td>
<td>4976.46</td>
<td>52.16</td>
<td>5029.20</td>
<td>52.16</td>
</tr>
<tr>
<td>1HT task-orient'd</td>
<td>( D )</td>
<td>11</td>
<td>1</td>
<td>5.60</td>
<td>0.133</td>
<td>4924.30</td>
<td>0.00</td>
<td>4977.04</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>( D )</td>
<td>11</td>
<td>1</td>
<td>5.60</td>
<td>0.133</td>
<td>4924.30</td>
<td>0.00</td>
<td>4977.04</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>( D )</td>
<td>11</td>
<td>1</td>
<td>57.76</td>
<td>0.000</td>
<td>4976.46</td>
<td>52.16</td>
<td>5029.20</td>
<td>52.16</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>( \theta )</th>
<th>df</th>
<th>( G^2 )</th>
<th>( p )</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>( \Delta BIC )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gardiner, Richardson-Klavehn, &amp; Ramponi, 1997, low word frequency ((N = 2,304))</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT task-orient'd ( g=0 ) ( DR )</td>
<td>1HT multi-memory ( rk )</td>
<td>8</td>
<td>4</td>
<td>31.09</td>
<td>0.000</td>
<td>4785.33</td>
<td>23.76</td>
<td>4831.27</td>
<td>21.94</td>
</tr>
<tr>
<td>2HT task-orient'd ( g=0 ) ( DR )</td>
<td>9</td>
<td>4</td>
<td>9.16</td>
<td>0.057</td>
<td>4763.40</td>
<td>1.83</td>
<td>4809.33</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2HT multi-memory ( rkd )</td>
<td>1HT task-oriented ( DgR )</td>
<td>9</td>
<td>3</td>
<td>6.53</td>
<td>0.089</td>
<td>4762.77</td>
<td>1.20</td>
<td>4814.45</td>
<td>5.11</td>
</tr>
<tr>
<td>1HT task-oriented ( DgR )</td>
<td>9</td>
<td>3</td>
<td>5.33</td>
<td>0.149</td>
<td>4761.57</td>
<td>0.00</td>
<td>4813.25</td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td>1HT task-orient'd ( g=0 ) ( D )</td>
<td>9</td>
<td>3</td>
<td>31.05</td>
<td>0.000</td>
<td>4787.29</td>
<td>25.72</td>
<td>4838.97</td>
<td>29.64</td>
<td></td>
</tr>
<tr>
<td>2HT task-orient'd ( g=0 ) ( D )</td>
<td>9</td>
<td>3</td>
<td>9.09</td>
<td>0.028</td>
<td>4765.33</td>
<td>3.76</td>
<td>4817.01</td>
<td>7.68</td>
<td></td>
</tr>
<tr>
<td>1HT task-oriented ( D )</td>
<td>9</td>
<td>3</td>
<td>6.53</td>
<td>0.043</td>
<td>4762.57</td>
<td>3.98</td>
<td>4814.25</td>
<td>5.57</td>
<td></td>
</tr>
<tr>
<td>2HT task-orient'd ( D )</td>
<td>9</td>
<td>3</td>
<td>5.20</td>
<td>0.158</td>
<td>4761.60</td>
<td>1.00</td>
<td>4813.28</td>
<td>2.59</td>
<td></td>
</tr>
<tr>
<td>2HT task-oriented ( D )</td>
<td>11</td>
<td>1</td>
<td>4.92</td>
<td>0.026</td>
<td>4765.16</td>
<td>3.59</td>
<td>4828.33</td>
<td>18.99</td>
<td></td>
</tr>
</tbody>
</table>

| **Gardiner, Richardson-Klavehn, & Ramponi, 1997, high word frequency \((N = 2,304)\)** |      |               |    |         |      |        |       |       |                 |
| 1HT task-orient'd \( g=0 \) \( DR \)     | 1HT multi-memory \( rk \) | 8    | 4   | 22.26  | 0.000| 4934.66| 16.07 | 4980.60| 11.91           |
| 2HT task-orient'd \( g=0 \) \( DR \)     | 9    | 4             | 10.35| 0.035 | 4922.75| 4.15  | 4968.69| 0.00            |
| 2HT multi-memory \( rkd \)                | 1HT task-oriented \( DgR \) | 9    | 3   | 8.17   | 0.043| 4922.57| 3.98  | 4974.25| 5.57            |
| 1HT task-oriented \( DgR \)              | 9    | 3             | 5.20| 0.158  | 4919.60| 1.00   | 4971.28| 2.59            |
| 1HT task-orient'd \( g=0 \) \( D \)      | 9    | 3             | 22.19| 0.000 | 4936.59| 18.00  | 4988.27| 19.59           |
| 2HT task-orient'd \( g=0 \) \( D \)      | 9    | 3             | 10.34| 0.016 | 4924.75| 6.15  | 4976.43| 7.74            |
| 1HT task-oriented \( D \)                | 9    | 3             | 10.35| 0.016 | 4924.75| 6.15  | 4976.43| 7.74            |
| 2HT task-oriented \( D \)                | 11   | 1             | 0.19| 0.660  | 4918.59| 0.00   | 4981.76| 13.07           |
| 2HT task-oriented \( D \)                | 11   | 1             | 8.67| 0.003  | 4927.07| 8.48  | 4990.24| 21.55           |

| **Xu & Bellezza, 2001 \((N = 14,000)\)** |      |               |    |         |      |        |       |       |                 |
| 1HT task-orient'd \( g=0 \) \( DR \)     | 1HT multi-memory \( rk \) | 11   | 7   | 178.60 | 0.000| 33733.60| 141.05| 33816.61| 133.50          |
| 2HT task-orient'd \( g=0 \) \( DR \)     | 12   | 6             | 169.83| 0.000 | 33724.83| 132.28| 33807.85| 124.73          |
| 2HT multi-memory \( rkd \)                | 1HT task-oriented \( DgR \) | 12   | 6   | 35.55  | 0.000| 33592.55| 0.00  | 33683.11| 0.00            |
| 1HT task-oriented \( DgR \)              | 12   | 6             | 91.46| 0.000  | 33648.45| 55.90 | 33739.02| 55.90           |
| 2HT task-oriented \( DgR \)              | 13   | 5             | 169.83| 0.000 | 33726.83| 134.28| 33817.39| 134.28          |
| 1HT task-orient'd \( g=0 \) \( D \)      | 13   | 5             | 177.60| 0.000 | 33736.60| 144.05| 33834.71| 151.60          |
| 2HT task-orient'd \( g=0 \) \( D \)      | 16   | 2             | 167.90| 0.000  | 33726.90| 134.35| 33825.01| 141.90          |
| 1HT task-oriented \( D \)                | 16   | 2             | 75.97| 0.000  | 33640.97| 48.42 | 33761.72| 78.61            |
| 2HT task-oriented \( D \)                | 12   | 5             | 169.80| 0.000  | 33726.90| 140.35| 33853.65| 170.54           |

*Note.* The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column \( \theta \) displays the number of functionally independent model parameters given these restrictions.
less reason to doubt the validity of the 1HT task-oriented model that allows for guess responses to targets than for any other candidate model.

**BIC.** Applied to the models restricted according to the validity criterion (see Table 6.3), the BIC ranked the multi-memory model with either one or two high thresholds best for each dataset. For two datasets, the BIC converged with the AIC on the 2HT multi-memory model as optimal (Gardiner et al., 1997, low word frequency; Xu & Bellezza, 2001). Else, the BIC preferred a model with fewer functionally independent model parameters than the AIC, reflecting the fact that the BIC penalises the number of model parameters more strictly than the AIC and hence tends to prefer more parsimonious models.

Next, the model comparison was extended to comprise the task-oriented models with those equality restrictions that are not required by the validity criterion, but are required to ascertain that these models do not waste parsimony that they could achieve by virtue of their structure (see Table 6.4). For two datasets, the same model as in the original model set is BIC-optimal (Erdfelder, 2000, shallow processing; Xu & Bellezza, 2001). For four datasets, a different model is optimal given the extended model set. For both datasets from Gardiner et al. (1997), the 2HT task-oriented model with \( g = 0 \) that precludes guess responses to targets with restrictions on \( D \) and \( R \) was ranked optimal instead of the 2HT multi-memory model. Twice, the 1HT task-oriented model that allows for guess responses to targets with restrictions on \( D \), \( g \) and \( R \) was BIC-optimal instead of the multi-memory model with either two high thresholds (Eldridge et al., 2001) or one high threshold (Erdfelder, 2000, deep processing).

In sum, the comparison clearly reflects that the BIC, given the present numbers of several thousand observations, has a strong preference for models with few functionally independent parameters. However, the BIC does not suggest one of the candidate models as the most promising measurement model.

### 6.3.3 Conclusion

What can be concluded from the systematic evaluation of the MPT models for the remember-know-guess paradigm? Null-hypothesis testing and AIC converge in suggesting the 1HT task-oriented model that allows for guess responses to targets as the most promising measurement tool among the evaluated models. The BIC does not converge, but does not suggest any model as most promising either. Hence, the preliminary, somewhat tentative conclusion is to consider the 1HT task-oriented model that allows for guess
responses to targets as the most promising measurement tool for the time being. Additional datasets that extend the database might make the picture clearer.
7 Experiments Manipulating Old-New Response Bias

The evaluation of the measurement models for the remember-know-(guess) paradigm by means of reanalysing existing datasets provided evidence that favours the 2HT task-oriented model for the remember-know paradigm and 1HT task-oriented model that allows for guess responses to truly recognised targets for the remember-know-guess paradigm over the other candidate models for the respective paradigm. However, both models failed the test of global validity. The experiments reported in this chapter served to generate two further datasets that extend the empirical database for model evaluation so that the most promising measurement model for each paradigm gets a second chance to prove globally valid.

Doubts in the measures of retrieval experience traditionally employed in the remember-know paradigm, the rates of remember and know responses to targets, have been raised by the theoretical argument and empirical demonstration\(^9\) that differences in old-new response bias between experimental conditions may contaminate know response rates (Strack & Förster, 1995) and also remember response rates (Eldridge, Sarfatti, & Knowlton, 2002; Erdfelder, 2000; Hicks & Marsh, 1999; Hirshman & Henzler, 1998; Postma, 1999). Consequently, the traditional analysis of remember and know response rates to targets may lead to erroneous conclusions about the effect of experimental manipulations on retrieval experience. The first goal of the experiments reported in this chapter was to replicate empirical demonstrations of this contamination by manipulating old-new response bias.

One suggestion for solving the problem of contamination with old-new response bias has been to give participants the opportunity to report the retrieval experience of guessing, which yields the remember-know-guess paradigm, and to use the observed response rates

---

\(^9\) As discussed in chapter 3, a pure measure of retrieval experience, which is required to be invariant to changes of old-new response bias, can not be required to be invariant to experimental manipulations of old-new response bias, because it is possible that the manipulation affects the true retrieval experience in addition to the old-new response bias. Hence, the cited experimental evidence only demonstrates that remember-know judgments are not suitable as two-in-one measures of retrieval experience and a component of recognition memory performance. However, as remember-know response rates are obtained by partitioning the old response rate, and the old-response rate in recognition memory tests varies as a function of old-new response bias as well as recognition memory performance, there is nevertheless reason to doubt the suitability of remember-know response rates as pure measures of retrieval experience.
as before (Gardiner et al., 1996, 1997). This approach is successful if neither remember nor know response rates, but only guess response rates vary with manipulations of old-new response bias. While Gardiner et al. (1997) demonstrated empirically that this approach may be successful, other experiments showed that one cannot rely on its success (Eldridge et al., 2002; Erdfelder, 2000). However, data are sparse in that only three experiments have examined the effects of manipulations of old-new response bias in the remember-know-guess paradigm.

An alternative approach to solving the problem of contamination is to use measures of retrieval experience that control for old-new response bias stochastically. Such measures have been specified within the measurement models for the remember-know and remember-know-guess paradigms introduced in chapter 5. The second goal of the experiments reported in this chapter was to further examine both approaches to solving the problem of contamination with old-new response bias. Therefore, the manipulation of old-new response bias was crossed with varying whether the response option guess was available for classifying the retrieval experience or not, and the traditional analysis of remember, know, and guess response rates was supplemented with a model-based analysis.

In both experiments, old-new response bias was manipulated by varying the percentage of targets in the test. This is a well-established manipulation of old-new response bias (e.g., Parks, 1966; Ratcliff, Sheu, & Gronlund, 1992). In the only previous application of this manipulation to the remember-know-(guess) paradigm that I know of, Xu and Bellezza (2001) crossed it with manipulating the availability of the response option guess. However, they did not analyse response rates separately for the remember-know and remember-know-guess paradigms, and the model-based reanalysis of their data reported in the previous chapter showed that none of the models was a valid measurement tool for their data. In other investigations of the remember-know-(guess) paradigm, only the information about the percentage of targets in the test given to the participants was varied, while the actual percentage of targets in the test was kept constant (Gardiner et al., 1997; Hirshman & Henzler, 1998; Strack & Förster, 1995). In two of those experiments, participants in one condition received truthful information, whereas participants in the other condition received false information. In the present experiments, this confound was avoided by manipulating the actual percentage of targets in the test, using the levels 25%, 50% and 75% targets in the test.
If participants know about the percentage of targets in the test and behave rationally, their bias to respond old instead of new in the old-new decision should increase with increasing percentage of targets. Thus, in terms of response rates, the rate of old responses to both targets and distractors is expected to increase with increasing percentage of targets. However, in the remember-know paradigm, each old response has to be specified as either remember or know. Accordingly, for items of both types, at least one of the remember and know response rates is expected to increase with increasing percentage of targets in the test, in which case this traditional measure of retrieval experience is contaminated with old-new response bias. Strack and Förster (1995) argued that due to the character of the remember-know instruction, manipulations of old-new response bias should only affect know, but not remember response rates, and found corresponding empirical evidence. However, it has since been demonstrated that frequently not only know, but also remember response rates are affected in the same way (Eldridge et al., 2002; Erdfelder, 2000; Hicks & Marsh, 1999; Hirshman & Henzler, 1998; Postma, 1999). The directional hypotheses that remember and know response rates in the remember-know condition increase with increasing percentage of targets in test were examined with one-tailed tests for linear trends.

In the remember-know-guess condition, the additional response option guess is available for specifying the old responses. Accordingly, the expected increase of the old response rate with increasing percentage of targets must be reflected in the increase of at least one of the remember, know, and guess response rates. The problem of contaminated remember-know response rates is solved if neither remember nor know response rates increase with increasing percentage of targets, but only guess response rates. Conversely, the problem of contamination persists if remember and/or know response rates increase with increasing percentage of targets in the test in spite of the option to explicitly report guessing. Accordingly, the hypothesis that the remember-know-guess paradigm successfully prevents contamination was examined with one-tailed tests for linear trends for the remember, know and guess response rates.

Considering both response option conditions simultaneously, if remember and/or know response rates are contaminated in the remember-know condition, but not in the remember-know-guess condition, the linear trend across percentage of targets is expected to interact with the response options available for classifying retrieval experience. This was tested with two-tailed tests.
Analyses of the linear trends use only one of the two degrees of freedom available in each case. The remaining degree of freedom was used for two-tailed tests of whether there is a (U-shaped or inversely U-shaped) quadratic trend across percentage of targets in the test and whether there is an interaction of the quadratic trend with the response options available for classifying retrieval experience. Note that a quadratic trend across percentage of targets conditions would also indicate a contamination of remember-know response rates, though in another way than expected.

This traditional analysis of remember, know and, where available, guess response rates was supplemented with a model-based analysis. First, all multinomial processing tree (MPT) models introduced in chapter 5 were fit to the data in order to assess their suitability as measurement tools. Second, if one or more measurement models were suitable for a dataset, the optimal model was used to examine the effects of manipulating old-new response bias on the cognitive events underlying the responses in the remember-know-guess paradigm. Finally, the new datasets were integrated into the database of experiments manipulating old-new response bias in order to assess the global validity of the measurement models.

7.1 Experiment 1

7.1.1 Method

Participants and design. One hundred thirty-two students from Jena University were quasi-randomly assigned to one of six experimental conditions of the 3 × 2 between-participants design with the factors percentage of targets in the test (25% vs. 50% vs. 75%) and response options for classifying the retrieval experience (remember-know vs. remember-know-guess). Three participants were excluded from the analysis because they were not native speakers of German. Three further participants were excluded from the analysis because they reported that they did not understand the remember-know distinction. Of the remaining 126 participants, 21 were tested in each condition. All participants received 5€, but students of psychology could opt to receive course credit instead.

Materials and procedure. All experiments were implemented with E-Prime (Schneider, Eschman, & Zuccolotto, 2002). The instructions and materials were presented visually on the monitor of a personal computer, unless stated otherwise. Between one and five participants were tested simultaneously.
For each participant in Experiment 1, a random set of 108 study words was chosen from a pool of 223 highly imageable, concrete German nouns with four to nine letters, which had been selected from two word-norm studies (Baschek, Bredenkamp, Oehrle, & Wippich, 1994; Offe, Anneken, & Kessler, 1994). The study words were presented one at a time for 2s with an interstimulus interval of 1s. Participants were instructed to study a list of words for a later unspecified memory test. Following the study phase, they solved figural reasoning problems taken from the Advanced Progressive Matrices (Raven, 1962) for 15 minutes as a filler task.

Two features of the recognition memory test were varied orthogonally between participants. First, the percentage of targets in the test was either 25%, 50%, or 75%. Second, the response options available for specifying the retrieval experience for an item called old were either remember-know or remember-know-guess. In all conditions, the memory test consisted of 144 test items. Depending on the percentage of targets condition, these were composed of 36 targets and 108 distractors, 72 targets and 72 distractors, or 108 targets and 36 distractors. The targets were selected randomly from the 108 study words previously presented to the participant (unless all study words were required as targets), and the distractors were drawn randomly from the word pool for each participant. As a means of distributing targets and distractors evenly across the test phase in a way that reflects their relative proportions, the 144 test items were divided into 12 sequences of 12 items. Within each sequence, the percentage of targets was equal to the percentage of targets in the whole test, and the order of targets and distractors was determined randomly for each sequence and participant. In the test instructions, participants were explicitly informed about the percentage of targets and distractors in the test, both verbally and by means of a pie chart.

For each test item, participants first had to decide whether it had been presented in the study phase. The response options old and new were displayed underneath the test item, together with a pie chart representing the percentage of targets in red and of distractors in blue. The responses were entered via colour-coded keys of the keyboard, and the selected response option was highlighted by a thin frame coloured red or blue, respectively. In case of an old response, the response options for describing the retrieval experience were added to the display. The response options remember and know were displayed next to each other and could be selected via keys indicated by yellow labels. If available, the response option guess was displayed underneath the response options remember and know, and the
corresponding response key had a green label. Thus, the response option guess was set off both by position and colour. The selected response option was highlighted by a thin frame in the corresponding colour. Before the next trial started, the screen turned blank for 0.6s.

Prior to the test, participants received thorough instructions about the meaning of the response options available for classifying their retrieval experience. To remember a test word was explained to mean that the recognition of the word was accompanied by recollecting its occurrence in the study phase. This was described as the ability to become consciously aware again of details of the episode of studying the word, supplemented by examples of the kinds of aspects that could be brought back to mind by a test item. To know a test word was described as recognising that the word had been in the study phase without recollecting its prior occurrence. This was explained to mean that the test word was familiar from the study phase, but was not accompanied by a specific recollection of the episode of studying the word. If available, the response option guess was introduced by explaining that it could be difficult to decide whether a word had been in the study phase or not. Participants were instructed that if they responded old in this situation, this response should be specified as guess, thereby indicating a lack of certainty that the word had indeed occurred. The verbatim remember-know-guess instructions can be found in Appendix B.1.

7.1.2 Results and Discussion

The number of targets and distractors that participants had to respond to was varied between conditions. However, in statistical analyses of multinomial processing tree models, the power depends on the number of observations in each processing tree. Hence, in order to prevent that conditions contribute differently to the power of the multinomial analyses, they were based on responses to 36 targets and distractors per participant in each condition, the minimal number of targets and distractors presented to each participant in the experiment. For those conditions where participants responded to more targets or distractors, random samples of 36 items were used. Not only the multinomial model-based analyses, but also the traditional analyses were based on responses to 36 targets and distractors per participant. For all analyses, the significance level $\alpha = .05$ was used.

Traditional analysis. Table 7.1 displays the means and standard deviations of remember, know and guess response rates to targets and distractors for each condition. As detailed in the introduction to this chapter, the analysis of the remember, know, and guess
Table 7.1

Means (and Standard Deviations) of Remember, Know, and Guess Response Rates as a Function of Itemtype and Condition (Experiment 1)

<table>
<thead>
<tr>
<th></th>
<th>Remember</th>
<th></th>
<th></th>
<th>Know</th>
<th></th>
<th></th>
<th>Guess</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td>Targets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K</td>
<td>.44</td>
<td>.43</td>
<td>.47</td>
<td>.22</td>
<td>.24</td>
<td>.29</td>
<td>.15</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>(.17)</td>
<td>(.17)</td>
<td>(.20)</td>
<td>(.11)</td>
<td>(.12)</td>
<td>(.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K-G</td>
<td>.29</td>
<td>.46</td>
<td>.34</td>
<td>.20</td>
<td>.21</td>
<td>.30</td>
<td>.15</td>
<td>.11</td>
</tr>
<tr>
<td></td>
<td>(.14)</td>
<td>(.19)</td>
<td>(.19)</td>
<td>(.15)</td>
<td>(.13)</td>
<td>(.15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distractors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K</td>
<td>.03</td>
<td>.05</td>
<td>.06</td>
<td>.07</td>
<td>.10</td>
<td>.17</td>
<td>.09</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.05)</td>
<td>(.10)</td>
<td>(.06)</td>
<td>(.09)</td>
<td>(.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K-G</td>
<td>.02</td>
<td>.04</td>
<td>.04</td>
<td>.02</td>
<td>.04</td>
<td>.09</td>
<td>.09</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.05)</td>
<td>(.07)</td>
<td>(.04)</td>
<td>(.05)</td>
<td>(.08)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The columns 25%, 50%, and 75% denote the percentage of targets in the test. R-K = remember-know paradigm; R-K-G = remember-know-guess paradigm.

response rates as the traditional measures of retrieval experience served to test the hypothesis that remember-know response rates are contaminated with old-new response bias in the remember-know paradigm, but not in the remember-know-guess paradigm. That is, in the remember-know condition, the remember and/or know response rates were expected to increase with increasing percentage of targets in the test, whereas in the remember-know-guess condition, only the guess response rate, but neither the remember nor the know response rate was expected to increase with increasing percentage of targets. Hence, for remember and/or know response rates an interaction was expected of response options available for classifying retrieval experience with the linear trend across percentages of targets. The directional hypotheses of increasing, but not decreasing response rates with increasing percentage of targets were examined with one-tailed tests for linear trends. The supplementary tests for quadratic trends were two-tailed. Results for remember, know and guess responses are reported in turn.

Remember response rates to targets were subjected to a 3 × 2 analysis of variance (ANOVA) with percentage of targets in the test and response options as between-participants factors. Neither the linear trend nor the quadratic trend across percentage of targets was significant, |t|<1 and t(120)=-1.78, p=.077. There was a significant effect of
response options, $F(1, 120)=7.62, p=.007$, indicating a higher remember response rate if the response option guess was available. This main effect was qualified by an interaction of response options with the quadratic trend across percentage of targets, $t(120)=2.52, p=.013$, but there was no interaction of the linear trend, $|t|<1$. According to the a priori hypothesis that the remember response rate is contaminated with old-new response bias in the remember-know paradigm, but not in the remember-know-guess paradigm, the effect of percentage of targets was analysed separately for each response option condition. In the remember-know condition, neither the linear trend nor the quadratic trend was significant, both $|t|<1$. In contrast, in the remember-know-guess condition, the data showed a significant quadratic trend, $t(60)=-3.12, p=.003$, indicating an inversely U-shaped relation. There was no linear trend, $|t|<1$. That is, unexpectedly, the rate of remember responses to targets was contaminated with old-new response bias in the remember-know-guess condition, but not in the remember-know condition.

Corresponding hypotheses were advanced for remember response rates to targets and distractors. However, there was not more than one remember response to distractors for 52% to 86% of the participants in each condition. These low levels of remember responses to distractors prevent a meaningful ANOVA. Instead, simple main effects were analysed with the rank-based Kruskal-Wallis test. There was no significant effect of percentage of targets with response options remember-know, $\chi^2(2)=1.93, p=.191$ (one-tailed), or response options remember-know-guess, $\chi^2(2)=.50, p=.389$ (one-tailed). Thus, the remember response rate to distractors was not contaminated with old-new response bias in either condition.

Turning to know response rates to targets, a $3 \times 2$ ANOVA with percentage of targets in the test and response options as between-participants factors revealed a significant linear increase of the know response rate with increasing percentage of targets in the test, $t(120)=2.77, p=.003$ (one-tailed). The quadratic trend was not significant, $t(120)=1.17, p=.243$ (two-tailed). Neither the effect of response options nor the interaction of the linear or quadratic trend was significant, $F<1$ and both $|t|<1$. According to the a priori hypotheses, the effect of percentage of targets was analysed separately for each response option condition. In the remember-know condition, the know response rate increased linearly with increasing percentage of targets, $t(60)=1.78, p=.040$ (one-tailed). The quadratic trend was not significant, $|t|<1$. Also in the remember-know-guess condition, the know response rate increased linearly with increasing percentage of targets, $t(60)=2.13,$
Corresponding hypotheses were advanced for know response rates to targets and distractors. However, there was not more than one know response to distractors for 19% to 76% of the participants in each condition. The low levels of know responses to distractors prevent a meaningful ANOVA. Kruskal-Wallis tests yielded significant effects of percentage of targets with response options remember-know, χ²(2)=6.48, \( p = .020 \) (one-tailed), and with response options remember-know-guess, χ²(2)=8.59, \( p = .007 \) (one-tailed).

Table 7.1 shows that in both conditions, know response rates to distractors increased with increasing percentage of targets in the test. Unexpectedly, the know response rate was contaminated with old-new response bias not only in the remember-know condition, but also in the remember-know-guess condition.

For guess response rates to targets and distractors, an ANOVA with the between-participants factor percentage of targets was supplemented with a Kruskal-Wallis test to safeguard the interpretability of results. The guess response rate to targets exhibited no linear trend, \(|t|<1\). The quadratic trend missed significance, too, \( t(60)=1.84, \ p = .071 \) (two-tailed). This finding was corroborated by the nonsignificant result of the Kruskal-Wallis test, \( \chi^2(2)=2.39, \ p = .152 \) (one-tailed). Deviating from the targets, the guess response rate to distractors increased linearly with the percentage of targets in the test, \( t(60)=3.93, \ p < .001 \) (one-tailed). The quadratic trend was not significant, \( t(60)=1.21, \ p = .231 \) (two-tailed). The significant effect of percentage of targets was corroborated by the Kruskal-Wallis test, \( \chi^2(2)=11.82, \ p = .002 \) (one-tailed).

To summarise, in the remember-know condition, the experimental manipulation of old-new response bias resulted in a selective increase of the know response rate with percentage of targets in the test. This finding demonstrates empirically that know response rates may be contaminated with old-new response bias and replicates the finding by Strack and Förster (1995). In the remember-know-guess condition, a variety of effects on remember, know and guess response rates was found. Unexpectedly, the remember response rate to targets exhibited an inverse U-shape across percentage of targets in the test. The know response rate to both targets and distractors increased with percentage of targets. The guess response rate to distractors increased with the percentage of targets in the test, too, whereas the guess response rate to targets did not. The effects on remember
and know response rates indicate that providing the response option guess did not have the intended effect of preventing old-new response bias from contaminating remember and know response rates. Such a failure has been observed before (Eldridge et al., 2002; Erdfelder, 2000). The model-based analysis reported next served to elucidate whether the modelling approach to solving the problem of contamination with old-new response bias is more successful.

Models for the remember-know paradigm. The upper panel of Table 7.2 displays the results of fitting the four MPT models for the remember-know paradigm to the data with their respective memory parameters restricted to be equal across percentages of targets in the test. As outlined in the previous chapter, a model is a valid measurement tool if may be retained with these equality restrictions, which was assessed by the log-likelihood ratio statistic $G^2$ with $\alpha = .05$. Only the 1HT task-oriented model was rejected. According to Akaike’s Information Criterion AIC, the 2HT multi-memory model is optimal, whereas the Bayesian Information Criterion BIC ranked the 1HT multi-memory best. The latter model has the fewest functionally independent parameters of all models, given the restrictions required by the validity criterion. Next, the model comparison was extended to comprise the task-oriented models with those equality restrictions that are necessary to ascertain that the task-oriented models do not waste parsimony that they could achieve by virtue of their structure, though the additional restrictions are not required by the validity criterion (lower panel of Table 7.2). AIC and BIC converged on the 2HT task-oriented model with equality restrictions not only on parameter $D$, which captures recognition memory performance, but also on parameter $R$, which captures remember-know judgments to truly recognised targets. The hypothesis that the 2HT task-oriented model with restrictions on $D$ and $R$ fits the data could be retained, too. Hence, while both the 2HT task-oriented model and the multi-memory models qualify as valid measurement tools for this dataset, the 2HT task-oriented model can achieve greater parsimony and is the most suitable measurement tool. Thus, this experiment provides further evidence in line with the conclusion from the reanalysis reported in the previous chapter that the 2HT task-oriented model is the most promising measurement tool for the remember-know paradigm.

The 2HT task-oriented model was used to examine the effects of manipulating the percentage of targets in the test more fully. Figure 7.1 displays the parameter estimates and their 95% confidence intervals. Reiterating a point made before, the model’s validity is
### Table 7.2

**Results of the Evaluation of the Models for the Remember-Know Paradigm (Experiment 1)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>( \theta )</th>
<th>df</th>
<th>( G^2 )</th>
<th>( p )</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ABIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter restrictions required by the validity criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>( rk )</td>
<td>8</td>
<td>4</td>
<td>7.41</td>
<td>0.116</td>
<td>7268.59</td>
<td>3.39</td>
<td>7319.95</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>( rkd )</td>
<td>9</td>
<td>3</td>
<td>2.02</td>
<td>0.568</td>
<td>7265.21</td>
<td>0.00</td>
<td>7322.99</td>
<td>3.03</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>( D )</td>
<td>10</td>
<td>2</td>
<td>7.04</td>
<td>0.030</td>
<td>7272.23</td>
<td>7.02</td>
<td>7336.43</td>
<td>16.47</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>( D )</td>
<td>10</td>
<td>2</td>
<td>2.36</td>
<td>0.308</td>
<td>7267.54</td>
<td>2.34</td>
<td>7331.74</td>
<td>11.79</td>
</tr>
</tbody>
</table>

Extended sets of parameter restrictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>( \theta )</th>
<th>df</th>
<th>( G^2 )</th>
<th>( p )</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ABIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT task-oriented</td>
<td>( DR )</td>
<td>8</td>
<td>4</td>
<td>7.41</td>
<td>0.116</td>
<td>7268.59</td>
<td>4.22</td>
<td>7319.95</td>
<td>4.22</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>( rk )</td>
<td>8</td>
<td>4</td>
<td>3.19</td>
<td>0.527</td>
<td>7264.37</td>
<td>0.00</td>
<td>7315.73</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>( DR )</td>
<td>8</td>
<td>4</td>
<td>2.02</td>
<td>0.568</td>
<td>7265.21</td>
<td>0.83</td>
<td>7322.99</td>
<td>7.25</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>( rkd )</td>
<td>9</td>
<td>3</td>
<td>7.04</td>
<td>0.030</td>
<td>7272.23</td>
<td>7.86</td>
<td>7336.43</td>
<td>20.70</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>( D )</td>
<td>10</td>
<td>2</td>
<td>2.36</td>
<td>0.308</td>
<td>7267.54</td>
<td>3.17</td>
<td>7331.74</td>
<td>16.01</td>
</tr>
</tbody>
</table>

**Note.** The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column \( \theta \) displays the number of functionally independent model parameters given these restrictions.

**Figure 7.1** Parameter estimates and their 95% confidence intervals for the 2HT task-oriented model for the remember-know paradigm (Experiment 1). \( D \) = recognition memory performance; \( b \) = old-new response bias; \( R^* \) = probability of remember responses to items guessed to be old; \( R \) = probability of remember responses to truly recognised targets.
reflected in the fact that the recognition memory performance, captured by parameter $D$, did not differ significantly across percentages of targets in the test, $G^2(2) = 2.36$, $p = .308$. Conversely, the bias to respond old instead of new, captured by parameter $b$, increased significantly with increasing percentage of targets in the test, $G^2(2) = 66.58$, $p < .001$. This shows that the experimental manipulation of old-new response bias was successful. The old-new response bias was conservative with both 25% and 50% targets in the test, $b < .5$, $G^2(1) = 125.78$, $p < .001$ and $G^2(1) = 63.93$, $p < .001$. With 75% targets in the test, there was no systematic preference for either old or new responses, $b = .5$, $G^2(1) = 0.13$, $p = .717$. Turning to retrieval experience, neither the probability of a remember response to an item guessed to be old, captured by parameter $R^*$, nor the probability of a remember response to a truly recognised target, captured by parameter $R$, differed significantly as a function of percentage of targets in the test, $G^2(2) = 1.23$, $p = .541$ and $G^2(2) = 0.91$, $p = .633$.

Thus, the 2HT task-oriented model for the remember-know paradigm separated old-new response bias not only from recognition memory performance, but also from retrieval experience: Varying the percentage of targets in the test selectively affected the old-new response bias.

**Models for the remember-know-guess paradigm.** Table 7.3 displays the results of fitting the six MPT models for the remember-know-guess paradigm to the data with their respective memory parameters restricted to be equal across percentages of targets in the test. For this dataset, each model was rejected, with all $p$s $< .001$. Thus, none of the models considered for the remember-know-guess paradigm is a valid measurement tool for this dataset. The information criteria can nevertheless be applied. AIC and BIC converged in ranking the 1HT task-oriented model that allows guess responses to truly recognised targets as the best model in the set.

Thus, the results of this experiment resemble the findings of the reanalysis reported in the last chapter very much: While there is some limited evidence that favours the 1HT task-oriented model that allows guess responses to truly recognised targets over the other candidate models, modelling the remember-know-guess paradigm is far less successful than modelling the remember-know paradigm. Because none of the models is a valid
**Table 7.3**

*Results of the Evaluation of the Models for the Remember-Know-Guess Paradigm With Parameter Restrictions Required by the Validity Criterion (Experiment 1)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>θ</th>
<th>df</th>
<th>$G^2$</th>
<th>$p$</th>
<th>AIC</th>
<th>ΔAIC</th>
<th>BIC</th>
<th>ΔBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>11</td>
<td>7</td>
<td>198.72</td>
<td>0.000</td>
<td>9334.84</td>
<td>169.51</td>
<td>9405.46</td>
<td>137.42</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>rkd</td>
<td>12</td>
<td>6</td>
<td>59.48</td>
<td>0.000</td>
<td>9197.61</td>
<td>32.28</td>
<td>9274.65</td>
<td>6.60</td>
</tr>
<tr>
<td>1HT task-orient’d g=0</td>
<td>D</td>
<td>13</td>
<td>5</td>
<td>178.93</td>
<td>0.000</td>
<td>9319.06</td>
<td>153.73</td>
<td>9402.51</td>
<td>134.47</td>
</tr>
<tr>
<td>2HT task-orient’d g=0</td>
<td>D</td>
<td>13</td>
<td>5</td>
<td>44.52</td>
<td>0.000</td>
<td>9184.64</td>
<td>19.32</td>
<td>9268.10</td>
<td>0.06</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>D</td>
<td>16</td>
<td>2</td>
<td>19.20</td>
<td>0.000</td>
<td>9165.33</td>
<td><strong>0.00</strong></td>
<td>9268.04</td>
<td><strong>0.00</strong></td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>D</td>
<td>16</td>
<td>2</td>
<td>24.78</td>
<td>0.000</td>
<td>9170.90</td>
<td>5.57</td>
<td>9273.62</td>
<td>5.57</td>
</tr>
</tbody>
</table>

*Note.* The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column θ displays the number of functionally independent model parameters given these restrictions.

measurement tool for this dataset, the effects of manipulating the percentage of targets in the test on cognitive events underlying the observed responses in the remember-know-guess paradigm could not be examined further.

### 7.1.3 Summary

In this experiment, a manipulation of old-new response bias was crossed with varying whether the response option guess was available for classifying the retrieval experience. The traditional analysis of the remember-know condition showed that the know response rate was contaminated with old-new response bias. However, the task-oriented model with two high thresholds proved to be a valid measurement tool for the remember-know condition of this experiment and successfully separated the old-new response bias from retrieval experience as well as from recognition memory performance. In the remember-know-guess condition, the traditional analysis showed that not only the know response rate, but also the remember response rate was contaminated with old-new response bias. Thus, providing the option to classify retrieval experience as guess did not per se solve the problem of contamination with old-new response bias. More severely, none of the probabilistic models for the remember-know-guess paradigm was a valid measurement tool in this experiment. In short, in Experiment 1 the problem of contamination could only be solved by probabilistic modelling of the remember-know condition.
7 Experiments Manipulating Old-New Response Bias

7.2 Experiment 2

In Experiment 2, the same experimental manipulations were used as in Experiment 1. The goal of the second experiment was to strengthen the impact of old-new response bias. To this end, the difficulty of the recognition memory test was increased by reducing the time available for studying the items and by lengthening the retention interval. Furthermore, after completing the recognition memory test, participants did not only report whether they understood the remember-know distinction. In addition, they were asked to write down how they understood the terms remember and know and to report for some self-chosen example test items their reasons for responding remember and know, respectively. These responses were used to assess whether participants had understood the remember-know distinction correctly, which is a necessary condition for obtaining valid data in the remember-know-(guess) paradigm.

7.2.1 Method

Participants and design. One hundred twenty-seven students from Jena University were quasi-randomly assigned to one of six experimental conditions in the $3 \times 2$ between-participants design with the factors percentage of targets in the test (25% vs. 50% vs. 75%) and response options for classifying the retrieval experience (remember-know vs. remember-know-guess). Two participants were excluded from the analysis because they were not native speakers of German. Five further participants were excluded from the analysis because they reported that they did not understand the remember-know distinction. Of the remaining 120 participants, 20 were tested in each condition. Participants received 10€ for taking part in this experiment as well as an unrelated study that served as a filler task during the retention interval.

Materials and procedure. The procedure employed in Experiment 1 was altered with the primary aim of increasing the difficulty of the recognition memory test, thereby strengthening the impact of old-new response bias. The word pool was changed slightly to prevent the possibility of selective reminding during the retention interval. It consisted of 216 highly imageable, concrete German nouns with four to nine letters selected from the same word-norm studies as the original word pool (Baschek et al., 1994; Offe et al., 1994). The numbers of study words, targets and distractors in each condition remained unchanged, as did the procedures for random selection of the item sets and the quasi-random ordering of targets and distractors in the memory test.
The 108 study words were presented one at a time for 1.5s with an interstimulus interval of 0.5s. Participants were instructed to study a list of words for a later unspecified memory test. After the study phase, participants went to another room, where they took part in an unrelated study on the perception and evaluation of German versus Turkish men and young versus old women. When they had completed this filler task, participants were offered a soft drink. After this 1 hour retention interval, participants returned to their original computer and proceeded with the recognition memory test.

The recognition memory test consisted of 144 test items with a percentage of targets of 25%, 50%, or 75%. The participants were explicitly informed about the percentage of targets and distractors in the test both verbally and by means of a pie chart exactly as in Experiment 1. They entered their responses via mouse clicks on response fields shown underneath the test item. First, the response options old in a red frame and new in a blue frame were displayed next to each other above a pie chart representing the percentage of targets in red and of distractors in blue. Upon clicking a response field, it was coloured completely for 0.4s. In case of an old response, the display underneath the test item was replaced by the response options for describing the retrieval experience. The response options remember and know were displayed next to each other in yellow frames. If available, the response option guess was displayed underneath the response options remember and know and in a green frame. Upon clicking a response field, it was coloured completely for 0.4s. Before the next trial started, the screen turned blank for 1s.

The instructions concerning the response options available for classifying the retrieval experience were changed slightly. The explanations what it means to remember a test word were extended by further examples of the kinds of aspects of the study episode that, if brought back to mind by a test item, would justify a remember response. To know a test word was described as being certain of recognising the word from the study phase in the absence of a specific recollection of the episode of studying the word. That is, deviating from Experiment 1, recognition in the know sense was explicitly characterised as involving certainty by adding a single word to the instruction. The response options guess, if available, was explained exactly as in Experiment 1. The verbatim remember-know-guess instructions can be found in Appendix B.2. After completing the recognition memory test, participants answered the yes-no question whether they understood the terms remember and know. Subsequently, they were asked to write down in their own words how they
understood these terms and to report for some self-chosen example words their reasons for responding remember and know, respectively.

### 7.2.2 Results and Discussion

As the power of multinomial analyses depends on the number of observations in each processing tree, the strategy for preventing that the percentage of targets conditions contribute differently to the power of the multinomial analyses was to use responses to 36 targets and distractors per participant in each condition. For those conditions where participants responded to more targets or distractors, random samples were used. However, by mistake one item was contained twice in the word pool. All responses to this item were excluded from the analysis. Therefore, the numbers of observations summed across all participants who reported that they understood the remember-know distinction did not amount to 720 for each processing tree, but varied slightly between 711 and 719, which is negligible with regard to power considerations. Multinomial and traditional analyses were based on the same subset of responses and for all analyses, the significance level $\alpha = .05$ was used.

At the end of this experiment, 120 participants answered affirmatively to the yes-no question whether they understood the terms remember and know, $n = 20$ in each condition. In addition, participants wrote down in their own words how they understood these terms. Eight descriptions (i.e., 6.7%) indicated misunderstandings in participants who reported that they understood the distinction. Excluding these participants from the analysis leads to sample sizes between $n = 18$ and $n = 20$. Due to power considerations for the multinomial analyses and in order to facilitate the specification of the polynomial contrasts used in the traditional analyses, I decided to base the analysis on $n = 18$ participants per condition. Each participant who met the additional criterion of adequately describing the remember-know distinction was included, starting with the lowest participant number, until the limit of $n = 18$ was reached for each condition.

In Experiment 1, only the participants themselves judged whether they understood the remember-know distinction. To allow for a direct comparison of the results of both experiments, the data from Experiment 2 were analysed on the basis of this self-report inclusion criterion. As this analysis included some participants who actually did not understand this crucial distinction, the data were reanalysed on the basis of the additional inclusion criterion of description content as delineated above. Reassuringly, the pattern of
results of both the traditional analysis and the multinomial model-based analysis did not
differ between the inclusion criteria. In order not to inflate the presentation of the results
and to allow for a direct comparison with Experiment 1, the following presentation is
limited to the self-report inclusion criterion.

**Traditional analysis.** Table 7.4 displays the means and standard deviations of
remember, know and guess response rates to targets and distractors for each condition. To
reiterate, the analysis of the remember, know, and guess response rates as the traditional
measures of retrieval experience served to test the hypothesis that remember-know
response rates are contaminated with old-new response bias in the remember-know
paradigm, but not in the remember-know-guess paradigm. That is, in the remember-know
condition, the remember and/or know response rates were expected to increase with
increasing percentage of targets in the test, whereas in the remember-know-guesss
condition, only the guess response rate, but neither the remember nor the know response
rate was expected to increase with increasing percentage of targets. Hence, for remember
and/or know responses an interaction was expected of response options available for
classifying retrieval experience with the linear trend across percentages of targets. The
directional hypotheses of increasing, but not decreasing response rates with increasing
percentage of targets were examined with one-tailed tests for linear trends. The
supplementary tests for quadratic trends were two-tailed.

Remember response rates to targets were subjected to a 3 × 2 ANOVA with percentage
of targets in the test and response options as between-participants factors. The remember
response rate increased linearly with increasing percentage of targets in the test,
t(114)=2.13, p=.018 (one-tailed). The quadratic trend was not significant, t(114)=1.30,
p=.195 (two-tailed). Neither the effect of response options nor the interactions of the linear
and quadratic trends were significant, F(1, 114)=1.38, p=.242 and both |t|<1. According to
the a priori hypotheses, the effect of percentage of targets was analysed separately for each
response option condition. In the remember-know condition, the remember response rate
increased with increasing percentage of targets in the test, t(57)=1.77, p=.041 (one-tailed).
The quadratic trend was not significant, t(57)=1.14, p=.260 (two-tailed). In the remember-
know-guess condition, neither the linear trend nor the quadratic trend were significant,
t(57)=1.26, p=.107 (one-tailed) and |t|<1. That is, as predicted, the rate of remember
responses to targets was contaminated with old-new response bias in the remember-know
Table 7.4

Means (and Standard Deviations) of Remember, Know, and Guess Response Rates as a Function of Itemtype and Condition (Experiment 2)

<table>
<thead>
<tr>
<th></th>
<th>Remember</th>
<th></th>
<th></th>
<th>Know</th>
<th></th>
<th></th>
<th>Guess</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td>Targets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K</td>
<td>.33</td>
<td>.32</td>
<td>.42</td>
<td>.24</td>
<td>.33</td>
<td>.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.16)</td>
<td>(.16)</td>
<td>(.17)</td>
<td>(.13)</td>
<td>(.11)</td>
<td>(.15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K-G</td>
<td>.30</td>
<td>.30</td>
<td>.36</td>
<td>.20</td>
<td>.21</td>
<td>.21</td>
<td>.13</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>(.15)</td>
<td>(.19)</td>
<td>(.15)</td>
<td>(.13)</td>
<td>(.12)</td>
<td>(.11)</td>
<td>(.08)</td>
<td>(.09)</td>
</tr>
<tr>
<td>Distractors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K</td>
<td>.04</td>
<td>.04</td>
<td>.14</td>
<td>.08</td>
<td>.13</td>
<td>.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.06)</td>
<td>(.05)</td>
<td>(.15)</td>
<td>(.08)</td>
<td>(.10)</td>
<td>(.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-K-G</td>
<td>.02</td>
<td>.05</td>
<td>.03</td>
<td>.04</td>
<td>.09</td>
<td>.08</td>
<td>.10</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.07)</td>
<td>(.04)</td>
<td>(.06)</td>
<td>(.09)</td>
<td>(.09)</td>
<td>(.08)</td>
<td>(.09)</td>
</tr>
</tbody>
</table>

Note. The columns 25%, 50%, and 75% denote the percentage of targets in the test. R-K = remember-know paradigm. R-K-G = remember-know-guess paradigm.

to distractors, there was not more than one remember response for 30% to 90% of the participants in each condition. These low levels of remember responses to distractors prevent a meaningful ANOVA. Instead, simple main effects were analysed with the rank-based Kruskal-Wallis test. In the remember-know condition, there was a significant effect of percentage of targets, \( \chi^2(2)=10.01, p=.004 \) (one-tailed). Table 7.4 shows that there was a selective increase of remember responses to distractors with 75% targets in the test. In contrast, in the remember-know-guess condition, there was no significant effect of percentage of targets, \( \chi^2(2)=2.54, p=.140 \) (one-tailed). As predicted, the remember response rate to distractors was contaminated with old-new response bias in the remember-know condition, but not in the remember-know-guess condition.

Turning to know response rates to targets, a 3 × 2 ANOVA with percentage of targets in the test and response options as between-participants factors revealed a significant linear increase of the know response rate with increasing percentage of targets, \( t(114)=1.91, p=.029 \) (one-tailed). The quadratic trend was not significant, |\( t |<1. There was a significant effect of response options, \( F(1,114)=17.70, p<.001 \), indicating a lower rate of know
responses if the response option “guess” is available than if it is not available. No interaction occurred for the linear trend, $t(114)=1.50$, $p=.136$, or the quadratic trend, $|t|<1$. According to the a priori hypotheses, the effect of percentage of targets was analysed separately for each response option condition. In the remember-know condition, the know response rate increased linearly with increasing percentage of targets, $t(57)=2.31$, $p=.012$ (one-tailed). The quadratic trend was not significant, $t(57)=-1.02$, $p=.311$ (two-tailed). In contrast, in the remember-know-guess condition, neither the linear trend nor the quadratic trend was significant, both $|t|s<1$. As predicted, the know response rate to targets was contaminated with old-new response bias in the remember-know condition, but not in the remember-know-guess condition, even though the predicted interaction was not found.

To distractors, there was not more than one know response for 10% to 70% of the participants in each condition. The low levels of know responses to distractors prevent a meaningful ANOVA. Instead, simple main effects were analysed with Kruskal-Wallis tests. In the remember-know condition, there was a significant effect of percentage of targets, $\chi^2(2)=12.96$, $p=.001$ (one-tailed). Table 7.4 shows that the know response rate to distractors increased with increasing percentage of targets in the test. In contrast, in the remember-know-guess condition, there was no significant effect of percentage of targets, $\chi^2(2)=3.37$, $p=.093$ (one-tailed). As predicted, the know response rate to distractors was contaminated with old-new response bias in the remember-know condition, but not in the remember-know-guess condition.

For guess response rates to targets and distractors, an ANOVA with the between-participants factor percentage of targets was supplemented with a Kruskal-Wallis test to safeguard the interpretability of results. For guess response rates to targets, there was no linear or quadratic trend, both $|t|s<1$. This finding was corroborated by the Kruskal-Wallis test, $\chi^2(2)=1.30$, $p=.261$ (one-tailed). Deviating from the targets, the guess response rate to distractors increased linearly with increasing percentage of targets in the test, $t(57)=3.50$, $p<.001$ (one-tailed). The quadratic trend was not significant, $|t|s<1$. The significant effect of percentage of targets was corroborated by the Kruskal-Wallis test, $\chi^2(2)=10.53$, $p=.003$ (one-tailed).

To summarise, in the remember-know condition, the experimental manipulation of old-new response bias led to an increase of both remember and know response rates with increasing percentage of targets in the test. That is, unlike the results of Experiment 1 and Strack and Förster (1995), not only the know response rate, but also the remember
response rate was contaminated with old-new response bias in this experiment that aimed at providing more room for old-new response bias by means of a more difficult task. This is in line with previous evidence that the remember response rate may be contaminated with old-new response bias just as the know response rate (Eldridge et al., 2002; Erdfelder, 2000; Hicks & Marsh, 1999; Hirshman & Henzler, 1998; Postma, 1999). In the remember-know-guess condition, only the guess response rate to distractors increased with the percentage of targets in the test. Thus, in contrast to Experiment 1 and two previous experiments (Eldridge et al., 2002; Erdfelder, 2000), introducing the response option guess prevented old-new response bias from contaminating remember-know response rates, as suggested and successfully demonstrated in one experiment by Gardiner et al. (1997). The model-based analysis reported next served to elucidate whether the modelling approach to solving the problem of contamination is also successful in this experiment.

Models for the remember-know paradigm. The upper panel of Table 7.5 displays the results of fitting the four MPT models for the remember-know paradigm to the data with their respective memory parameters restricted to be equal across percentages of targets in the test. Both the multi-memory model and the task-oriented model with one high threshold were rejected, whereas both models with two high thresholds could be retained. AIC and BIC ranked the 2HT multi-memory model best. The lower panel of Table 7.5, which displays the extended model comparison including the task-oriented models with the additional equality restrictions necessary to ascertain that they do not waste parsimony achievable by virtue of their structure, shows that both AIC and BIC ranked the 2HT task-oriented model with equality restrictions on $D$ and $R$ best. The hypothesis that the 2HT task-oriented model with restrictions on $D$ and $R$ fits the data could be retained, too. Thus, both the 2HT task-oriented model and the 2HT multi-memory model qualify as valid measurement tools for this dataset, but the 2HT task-oriented model can achieve greater parsimony. Hence, this experiment provides further evidence in line with the conclusion from the reanalysis reported in the previous chapter that the 2HT task-oriented model is the most promising measurement tool for the remember-know paradigm.

The 2HT task-oriented model was used to examine the effects of manipulating the percentage of targets in the test more fully. Parameter estimates and their 95% confidence intervals are displayed in Figure 7.2. Reiterating a point made before, the model’s validity is reflected in the fact that the recognition memory performance, captured by parameter $D$,
Table 7.5

Results of the Evaluation of the Models for the Remember-Know Paradigm (Experiment 2)

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>θ</th>
<th>df</th>
<th>$G^2$</th>
<th>p</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ABIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter restrictions required by the validity criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>rk</td>
<td>8</td>
<td>4</td>
<td>14.28</td>
<td>0.006</td>
<td>7262.06</td>
<td>8.57</td>
<td>7312.96</td>
<td>2.21</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>rkd</td>
<td>9</td>
<td>3</td>
<td>3.71</td>
<td>0.294</td>
<td>7253.49</td>
<td>0.00</td>
<td>7310.75</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>$D$</td>
<td>10</td>
<td>2</td>
<td>10.72</td>
<td>0.005</td>
<td>7262.50</td>
<td>9.01</td>
<td>7326.12</td>
<td>15.37</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>$D$</td>
<td>10</td>
<td>2</td>
<td>4.54</td>
<td>0.103</td>
<td>7256.32</td>
<td>2.83</td>
<td>7319.94</td>
<td>9.19</td>
</tr>
</tbody>
</table>

Extended sets of parameter restrictions

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>θ</th>
<th>df</th>
<th>$G^2$</th>
<th>p</th>
<th>AIC</th>
<th>AAIC</th>
<th>BIC</th>
<th>ABIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HT task-oriented</td>
<td>$DR$</td>
<td>8</td>
<td>4</td>
<td>14.28</td>
<td>0.006</td>
<td>7262.06</td>
<td>8.70</td>
<td>7312.96</td>
<td>8.70</td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>$rk$</td>
<td>8</td>
<td>4</td>
<td>5.58</td>
<td>0.233</td>
<td>7253.36</td>
<td>0.00</td>
<td>7304.26</td>
<td>0.00</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>$DR$</td>
<td>8</td>
<td>4</td>
<td>3.71</td>
<td>0.294</td>
<td>7253.49</td>
<td>0.13</td>
<td>7310.75</td>
<td>6.49</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>$rkd$</td>
<td>9</td>
<td>3</td>
<td>4.54</td>
<td>0.103</td>
<td>7256.32</td>
<td>2.96</td>
<td>7319.94</td>
<td>15.68</td>
</tr>
</tbody>
</table>

Note. The column Res. = Restrictions displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column θ displays the number of functionally independent model parameters given these restrictions.

Figure 7.2 Parameter estimates and their 95% confidence intervals for the 2HT task-oriented model for the remember-know paradigm (Experiment 2). $D$ = recognition memory performance; $b$ = old-new response bias; $R^*$ = probability of remember responses to items guessed to be old; $R$ = probability of remember responses to truly recognised targets.
did not differ significantly across percentages of targets in the test, $G^2(2) = 4.54$, $p = .103$. Conversely, the bias to respond old instead of new, captured by parameter $b$, increased significantly with increasing percentage of targets in the test, $G^2(2) = 173.48$, $p < .001$. This shows that the experimental manipulation of old-new response bias was successful. The old-new response bias was conservative with both 25% and 50% targets in the test, $b < .5$, $G^2(1) = 192.61$, $p < .001$ and $G^2(1) = 70.84$, $p < .001$. With 75% targets in the test, the old-new response bias was liberal, $b > .5$, $G^2(1) = 14.32$, $p < .001$. Turning to retrieval experience, there was a significant effect on the probability of a remember response to an item guessed to be old, captured by parameter $R^*$, $G^2(2) = 14.32$, $p < .001$, which indicates a U-shaped relationship. This effect mirrors the high rate of remember responses to distractors only with 75% targets in the test in conjunction with the gradual increase of know responses to distractors.\(^{10}\) The probability of a remember response to a truly recognised target, captured by parameter $R$, did not differ significantly as a function of percentage of targets in the test, $G^2(2) = 2.23$, $p = .328$.

Thus, the 2HT task-oriented model for the remember-know paradigm separated the old-new response bias not only from recognition memory performance, but also from retrieval experience: As in Experiment 1, increasing the percentage of targets in the test did not affect recognition memory performance or retrieval experience for truly recognised targets, but led to an increase of the bias to respond old instead of new. Furthermore, a U-shaped relationship between percentage of targets in the test and judgments of retrieval experience for items guessed to be old was found.

**Models for the remember-know-guess paradigm.** The upper panel of Table 7.6 displays the results of fitting the six MPT models for the remember-know-guess paradigm to the data with their respective memory parameters restricted to be equal across percentages of targets in the test. All models were rejected apart from the 1HT task-oriented model that allows for guess responses to truly recognised targets. The same model was ranked best by AIC, whereas the BIC preferred the 2HT multi-memory model, which has fewer functionally independent model parameters. The lower panel of Table 7.6, which displays

\[ R^* = R_D / (R_D + K_D), \]  
where $R_D$ denotes the rate of remember responses to distractors and $K_D$ denotes the rate of know responses to distractors (cf. Appendix A.1).
### Table 7.6

**Results of the Evaluation of the Models for the Remember-Know-Guess Paradigm (Experiment 2)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Res.</th>
<th>df</th>
<th>$G^2$</th>
<th>$p$</th>
<th>AIC</th>
<th>$\Delta$AIC</th>
<th>BIC</th>
<th>$\Delta$BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter restrictions required by the validity criterion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1HT multi-memory</td>
<td>$rk$</td>
<td>11</td>
<td>82.00</td>
<td>0.000</td>
<td>9309.48</td>
<td>68.69</td>
<td>9379.44</td>
<td>53.45</td>
</tr>
<tr>
<td>2HT multi-memory</td>
<td>$rkd$</td>
<td>12</td>
<td>20.19</td>
<td>0.003</td>
<td>9249.68</td>
<td>8.88</td>
<td>9326.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1HT task-orient’d $g=0$ D</td>
<td>$D$</td>
<td>13</td>
<td>79.06</td>
<td>0.000</td>
<td>9310.54</td>
<td>69.74</td>
<td>9393.22</td>
<td>67.22</td>
</tr>
<tr>
<td>2HT task-orient’d $g=0$ D</td>
<td>$D$</td>
<td>13</td>
<td>18.85</td>
<td>0.002</td>
<td>9250.33</td>
<td>9.53</td>
<td>9333.01</td>
<td>7.01</td>
</tr>
<tr>
<td>1HT task-oriented</td>
<td>$D$</td>
<td>16</td>
<td>3.31</td>
<td>0.191</td>
<td>9240.80</td>
<td>$\textbf{0.00}$</td>
<td>9342.56</td>
<td>16.56</td>
</tr>
<tr>
<td>2HT task-oriented</td>
<td>$D$</td>
<td>16</td>
<td>17.43</td>
<td>0.000</td>
<td>9254.91</td>
<td>14.11</td>
<td>9356.67</td>
<td>30.67</td>
</tr>
</tbody>
</table>

Extended sets of parameter restrictions

| 1HT task-orient’d $g=0$ D $gR$ | $DG$       | 11 | 82.00  | 0.000  | 9309.48  | 68.87       | 9379.44  | 62.51       |
| 1HT multi-memory             | $rk$       | 11 | 22.92  | 0.002  | 9250.40  | 9.79        | 9320.36  | 3.43        |
| 2HT task-orient’d $g=0$ D $gR$| $DG$       | 11 | 22.92  | 0.001  | 9252.40  | 11.79       | 9328.73  | 11.79       |
| 1HT task-oriented            | $Dg$       | 12 | 11.14  | 0.084  | 9240.62  | $\textbf{0.00}$ | 9316.94  | $\textbf{0.00}$ |
| 2HT task-oriented            | $Dg$       | 12 | 22.92  | 0.001  | 9252.40  | 11.79       | 9328.73  | 11.79       |
| 1HT task-orient’d $g=0$ D     | $D$        | 13 | 79.06  | 0.000  | 9310.54  | 69.92       | 9393.22  | 76.28       |
| 2HT task-orient’d $g=0$ D     | $D$        | 13 | 18.85  | 0.002  | 9250.33  | 9.71        | 9333.01  | 16.07       |
| 1HT task-oriented            | $D$        | 16 | 3.31   | 0.191  | 9240.80  | 0.18        | 9342.56  | 25.62       |
| 2HT task-oriented            | $D$        | 16 | 17.43  | 0.000  | 9254.91  | 14.29       | 9356.67  | 39.73       |

*Note.* The column `Res. = Restrictions` displays the model parameters that were restricted to be equal across the levels of the manipulation of old-new response bias. The column `df` displays the number of functionally independent model parameters given these restrictions.

The extended model comparison including the task-oriented models with the additional equality restrictions necessary to ascertain that they do not waste parsimony achievable by virtue of their structure, shows that with the extended equality restrictions on $D$, $g$ and $R$, the 1HT task-oriented model that allows for guess responses to truly recognised targets was ranked best by both AIC and BIC and could be retained. Thus, the 1HT task-oriented model that allows for guess responses to truly recognised targets is a valid tool for measuring the cognitive events that underlie the observed responses in this dataset. The results of this experiment are in line with the tentative conclusion from the reanalysis that for the time being this is the most promising measurement model for the remember-know-guess paradigm.
Figure 7.3  Parameter estimates and their 95% confidence intervals for the 1HT task-oriented model that allows for guess responses to targets for the remember-know-guess paradigm (Experiment 2). \( D = \) recognition memory performance; \( b = \) old-new response bias; \( g^* = \) probability of guess responses to items guessed to be old; \( R^* = \) probability of remember responses to items guessed to be old but not labelled guess; \( g = \) probability of guess responses to truly recognised targets; \( R = \) probability of remember responses to truly recognised targets that are not labelled guess.

Having shown that the 1HT task-oriented model that allows for guess responses to truly recognised targets is a valid measurement tool for this dataset, it was used to examine the effects of manipulating the percentage of targets in the test more fully. Parameter estimates and their 95% confidence intervals are displayed in Figure 7.3. Reiterating a point made before, the model’s validity is reflected in the fact that the recognition memory performance, captured by parameter \( D \), did not differ significantly across percentages of targets in the test, \( G^2(2) = 3.31, \ p = .191 \). Conversely, the bias to respond old instead of new, captured by parameter \( b \), increased significantly with increasing percentage of targets in the test, \( G^2(2) = 60.58, \ p < .001 \). This shows that the experimental manipulation of old-new response bias was successful. The old-new response bias was conservative with each percentage of targets in the test, \( b < .5 \) (25%: \( G^2(1) = 357.80, \ p < .001 \); 50%: \( G^2(1) = 102.19, \ p < .001 \); 75%: \( G^2(1) = 83.46, \ p < .001 \)). That is, even when participants were explicitly informed that there are more targets than distractors in the test, they preferred new responses over old responses, despite the possibility to label responses as guess. With regard to retrieval experience for items guessed to be old, neither the probability of a guess response, captured by parameter \( g^* \), nor the conditional probability
of a remember response, given that the item was not called guess, captured by parameter $R^*$, differed significantly across percentages of targets in the test, $G^2(2) = 5.02, \ p = .081$ and $G^2(2) = 1.73, \ p = .420$. Also with regard to retrieval experience for truly recognised targets, neither the probability of a guess response, captured by parameter $g$, nor the conditional probability of a remember response given that the item was not called guess, captured by parameter $R$, differed significantly across percentages of targets in the test, $G^2(2) = 1.54, \ p = .463$ and $G^2(2) = 1.70, \ p = .428$.

Thus, the 1HT task-oriented model that allows for guess responses to truly recognised targets separated the old-new response bias not only from memory performance, but also from retrieval experience: Varying the percentage of targets in the test selectively affected the old-new response bias.

### 7.2.3 Summary

As in the previous experiment, a manipulation of old-new response bias was crossed with varying whether the response option guess was available for classifying the retrieval experience. The goal of this experiment was to strengthen the impact of old-new response bias, pursued by making the recognition memory test more difficult. Lower levels of recognition memory performance in Experiment 2 than in Experiment 1 indicate that the recognition memory test was indeed more difficult. The traditional analysis of the remember-know condition showed that not only the know, but also the remember response rate was contaminated with old-new response bias. However, the task-oriented model with two high thresholds proved to be a valid measurement tool for the remember-know paradigm and successfully separated old-new response bias from retrieval experience as well as from recognition memory performance. In the remember-know-guess condition, the traditional analysis showed that neither remember nor know response rates were contaminated with old-new response bias. Thus, in this experiment, providing the option to classify retrieval experience as guess solved the problem of contamination. Furthermore, the task-oriented model with one high threshold that allows for guess responses to truly recognised targets proved to be a valid measurement tool for this dataset and successfully

---

11 Recognition memory performance can only be compared for the remember-know condition, because there was no valid measure of recognition memory performance for the remember-know-guess condition in Experiment 1.
separated old-new response bias from retrieval experience as well as from recognition memory performance.

### 7.3 Discussion of Experiments 1 and 2

Two new experiments demonstrated empirically that response rates in the remember-know paradigm may be contaminated with old-new response bias. In Experiment 1, only know response rates were contaminated (cf. Strack & Förster, 1995), whereas in Experiment 2, also remember response rates were contaminated (cf. Eldridge et al., 2002; Erdfelder, 2000; Hicks & Marsh, 1999; Hirshman & Henzler, 1998; Postma, 1999). As discussed before, this experimental evidence can only demonstrate that remember-know response rates are not suitable as two-in-one measures of retrieval experience and a component of recognition memory performance. In contrast, these data are mute with regard to the suitability of remember-know response rates as pure measures of retrieval experience: A pure measure of retrieval experience, which is required to be invariant to changes of old-new response bias, can not be required to be invariant to experimental manipulations of old-new response bias, because it is possible that the manipulation affects the true retrieval experience in addition to the old-new response bias. Importantly, there is nevertheless reason to doubt the suitability of remember-know response rates as pure measures of retrieval experience, because remember-know response rates are obtained by partitioning the old response rate and the old-response rate in recognition memory tests varies as a function of old-new response bias and recognition memory performance. The goal of the present experiments was to examine two approaches to solving the problem of contamination with old-new response bias.

The first approach, which yields the remember-know-guess paradigm, is simply to give participants the option to report that they are guessing and to use the observed response rates as before (Gardiner et al., 1997). In the remember-know-guess condition of Experiment 2, neither remember nor know response rates were contaminated with old-new response bias, whereas both remember and know response rates were contaminated in Experiment 1. Traditional analyses of remember and know response rates are available for a total of five experiments that manipulated old-new response bias in the remember-know-guess paradigm.\(^\text{12}\) In three experiments, at least one of the remember and know response

\(^\text{12}\) In all of these experiments, each response rate was tested at the significance level \(\alpha = .05\).
rates to targets and distractors was contaminated with old-new response bias (Eldridge et al., 2002; Erdfelder, 2000; Experiment 1), whereas in two experiments none of these response rates differed significantly across old-new response bias conditions (Experiment 2; Gardiner et al., 1997). To assess whether this finding has any relevance for the suitability of remember-know response rates obtained in the remember-know-guess paradigm as two-in-one measures of retrieval experience and a component of recognition memory, which represent retrieval from episodic and semantic memory, one can apply the strategy used in the tests of global validity of the measurement models. If, in the traditional analysis, a statistical test is conducted for each of the remember and know response rates to targets and distractors with \( \alpha = .05 \), then, considering these tests as independent, the probability of incorrectly rejecting the null hypothesis in at least one of the four tests per dataset is given by \( \alpha_v = .185 \). Hence, \( H_0: \pi \leq .185 \) represents the hypothesis that remember-know response rates obtained in the remember-know-guess paradigm are valid measures of retrieval from episodic and semantic memory. The probability of incorrectly rejecting at least one of four null hypotheses in at least three of five datasets is \( p = .0474 \). Hence, \( H_0: \pi \leq .185 \) is rejected. This analysis indicates that remember-know response rates obtained in the remember-know-guess paradigm are not suitable as two-in-one measures of retrieval experience and a component of recognition memory, which represent retrieval from episodic and semantic memory. Furthermore, adding the response option guess for specifying old responses is not suitable for remedying the concerns about the suitability of remember-know response rates as pure measures of retrieval experience, because the doubts stem from the fact that the old response rate in recognition memory tests varies not only as a function of old-new response bias, but also as a function of recognition memory performance. The latter aspect is not addressed by providing the response option guess.

The second approach to solving the problem of contamination with old-new response bias is to use probabilistic models for the remember-know-(guess) paradigm that specify measures of retrieval experience that stochastically control for old-new response bias. In the remember-know condition of both Experiment 1 and 2, the 2HT task-oriented model was a valid measurement tool and could achieve the greatest parsimony, as indicated by

\[ 13 \text{ If the tests are not independent, this value constitutes the upper limit of the risk of committing a Type I error. Hence, the analysis based on } \alpha_v = .185 \text{ gives the upper limit of the probability of incorrectly rejecting at least one of four null hypotheses in at least three of five datasets.} \]
the information criteria. Regarding the remember-know-guess paradigm, the 1HT task-oriented model that allows for guess responses to truly recognised targets was the only valid measurement model in Experiment 2, whereas no measurement model was valid in Experiment 1. These experiments were conducted in order to extend the database of experiments manipulating old-new response bias, into which they are formally integrated in a final step.

7.4 Evaluation of the Measurement Models with the Extended Database

In evaluating the measurement models, global model validity was assessed by testing the hypothesis $H_0: \pi \leq 0.05$ that the relative frequency of rejecting a candidate model, $\pi$, does not exceed the expected value for incorrectly rejecting a valid model due to committing Type I errors given $\alpha = 0.05$ for each model test. This strategy is now applied to the extended database.

7.4.1 Remember-Know Paradigm

After adding Experiments 1 and 2 to the database, the 2HT task-oriented model could meaningfully be applied to 9 datasets. It was rejected for 2 of 9 datasets. The probability of incorrectly rejecting a model in at least 2 of 9 datasets is $p = 0.0712$. Hence, the 2HT task-oriented model is a globally valid measurement model. The other three models for the remember-know paradigm could be applied to 13 datasets. The 2HT multi-memory model was rejected for 5 of 13 datasets, and both the 1HT multi-memory model and the 1HT task-oriented model were rejected for 7 of 13 datasets. The probability of incorrectly rejecting a model in at least 5 of 13 datasets is $p = 0.0003$, indicating that these models are no globally valid measurement tools for the remember-know paradigm. Thus, the 2HT task-oriented model is not only the most promising, but the only globally valid measurement model for the remember-know paradigm.

7.4.2 Remember-Know-Guess Paradigm

The extended database shows that both 2HT task-oriented models, regardless of whether guess responses to truly recognised targets are allowed or precluded, were rejected for all six datasets to which they were meaningfully applicable. The other four models could be applied to eight datasets. Both the 1HT task-oriented model that precludes guess responses to truly recognised targets and the 1HT multi-memory model were rejected for
seven of eight datasets, and the 2HT multi-memory model was rejected for six of eight datasets. The probability of incorrectly rejecting a model in at least six of eight datasets is $p < .0001$. Finally, the 1HT task-oriented model that allows for guess responses to truly recognised targets was rejected for three of eight datasets. The probability of incorrectly rejecting a model in at least three of eight datasets is $p = .0058$. Thus, also given the extended database, none of the models considered for the remember-know-guess paradigm is a valid measurement tool. Note that with two new experiments, the 1HT task-oriented model that allows for guess responses to truly recognised targets had the opportunity to prove globally valid: If it could have been retained in Experiments 1 and 2, it would have proven globally valid.

### 7.4.3 Conclusion

The analyses of the measurement models for the remember-know-(guess) paradigm yielded two main results. First, the 2HT task-oriented measurement model for the original remember-know paradigm is a valid measurement model that provides pure measures of retrieval experience. Thus, it constitutes a suitable tool for examining the distinctiveness-fluency framework of retrieval experience suggested by Rajaram (1996). The experiments reported in the next chapter served this purpose.

Second, interestingly, none of the measurement models that implement the multiple-memory systems model (Tulving, 1985b) proved globally valid: The multi-memory models with one and two thresholds for both the remember-know and the remember-know-guess paradigm clearly failed the test of global validity. While the variation of remember-know response rates with manipulations of old-new response bias can easily be attributed to flawed measures of the memory systems that neglect the possibility of remember and know responses due to guessing, this reasoning is not applicable to the multi-memory measurement models. The measurement models explicitly provide for the possibility that remember and know responses may result from guessing, rather than only from retrieval from the episodic and semantic memory systems. Therefore, the consistent failure of all multi-memory measurement models for the remember-know-(guess) paradigm suggests that the multiple-memory systems model can not explain remember-know judgments.
8 Experiments Manipulating Recognition Memory Performance

The experiments reported in this chapter served to re-investigate the effect of classic experimental manipulations of recognition memory performance on retrieval experience, which were the object of some of the earliest experiments conducted with the remember-know paradigm, namely full versus divided attention at study (Gardiner & Parkin, 1990) and semantic versus phonemic orienting tasks (Gardiner, 1988). Both experiments found selective effects on the rate of remember, but not know responses to targets.

Interpreting these findings along the lines of Tulving's (1985b) model of multiple memory systems, Gardiner (1988; Gardiner & Parkin, 1990) suggested that these variables selectively affected the episodic memory system. However, this conclusion is questionable for two reasons. First, remember-know response rates are not suitable as measures of the memory systems, because they are susceptible to contamination with old-new response bias. Second, the multi-memory measurement models, which provide measures of the memory systems that are not contaminated with old-new response bias by virtue of their construction, nevertheless did not meet the validity criterion of invariance to manipulations of old-new response bias, suggesting that Tulving's (1985b) model of multiple memory systems cannot explain remember-know judgments. This evidence suggests that it is not warranted to interpret the selective effects of dividing attention and phonemic as compared to semantic processing on the remember response rate to targets as a selective effect on the episodic memory system.

In an alternative interpretation of the decrease of the remember response rate to targets with dividing attention and phonemic as compared to semantic processing, Rajaram (1996) subsumed these findings under the distinctiveness-fluency framework, which holds that the retrieval experience of remembering depends on the processing of distinctive or salient features at study. In particular, Rajaram (1996) argued that

\[
\text{semantic encoding of stimuli in the levels of processing paradigm (Gardiner, 1988, Experiment 1; Rajaram, 1993, Experiment 1) ... [and] processing information under the undivided-attention condition (Gardiner & Parkin, 1990) ... are factors that facilitate the processing of the distinctive and salient aspects of the material. (p. 374)}
\]
Recall that the distinctiveness-fluency framework is intended to explain retrieval experience per se, rather than recognition memory performance. Consequently, pure measures of retrieval experience are required for examining this framework. In particular, because previous research has shown that distinctiveness manipulations affect memory performance (e.g., Hunt & Elliott, 1980), it is critical that measures of retrieval experience be independent of memory performance in order to permit a test of the notion that distinctiveness manipulations affect retrieval experience per se. Focusing on the experimental manipulations at hand, it is well-established that dividing attention (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996; Parkin, Reid, & Russo, 1990) and phonemic as compared to semantic processing (e.g., Craik & Tulving, 1975; Jacoby & Dallas, 1981) impair recognition memory performance. Because remember-know rates are expected to vary as a function of recognition memory performance and old-new response bias, using them as measures of retrieval experience involves a considerable risk of arriving at erroneous conclusions about the effect of these experimental manipulations on retrieval experience per se. In contrast, pure measures of retrieval experience are provided by the 2HT task-oriented model for the remember-know paradigm, which has been proven to be a valid measurement tool. Accordingly, this measurement model allows for a critical test of whether dividing attention and phonemic as compared to semantic processing do indeed affect the retrieval experience per se, as suggested by the distinctiveness-fluency framework (Rajaram, 1996).

Two experiments were conducted on order to re-investigate the effect of full versus divided attention at study (Experiment 3) and semantic versus phonemic orienting tasks (Experiment 4) on retrieval experience. These experiments had a twofold goal: On the one hand, the aim was to demonstrate empirically that remember-know response rates may lead to erroneous conclusions about the the effect of independent variables on retrieval experience by comparing the results of the traditional analysis and the model-based analysis. On the other hand, the experiments served to examine the distinctiveness-fluency framework by testing the notion that retrieval experience is a function of processing distinctive and salient features at study with valid measures of retrieval experience. In particular, inasmuch as dividing attention at study and phonemic as compared to semantic orienting tasks reduce the processing of distinctive or salient features at study, these manipulations should reduce the conditional probability of remember responses to truly recognised targets, which is captured by parameter $R$ of the 2HT task-oriented model. In
contrast, the distinctiveness-fluency framework is mute with regard to a possible effect of
the processing of distinctive and salient features at study on the conditional probability of
remember responses to items that are not recognised, but guessed to be old, captured by
parameter $R^*$, because the lack of true recognition implies the failure to truly recollect
genuine features of the study episode.

8.1 Experiment 3

Two traditionally analysed experiments converged in showing that divided as compared
to full attention at study selectively affects the rate of remember, but not know responses to
targets (Parkin & Gardiner, 1990; Parkin et al., 1995). In the present experiment,
participants were asked to study a visually presented list of words. While this was the only
task for participants in the full attention condition, participants in the divided attention
condition simultaneously performed a 1-back task with auditorily presented digits.

8.1.1 Method

Participants and design. Forty-seven students from Jena University were quasi-
randomly assigned to either the full attention condition or the divided attention condition.
All participants were native speakers of German. Two participants were excluded from the
analysis because they failed to follow the instructions for the secondary task. Four further
participants were excluded from the analysis because they reported that they did not
understand the remember-know distinction. Of the remaining 41 participants, 21 were
tested in the full attention condition and 20 in the divided attention condition. All
participants received 5€.

Materials and procedure. For each participant, a random set of 60 targets and 60
distractors was chosen from the word pool of 223 highly imageable, concrete German
nouns with four to nine letters used in Experiment 1.

In the first phase of the experiment, participants in both conditions performed the 1-
back task in isolation that was subsequently used as the secondary task in the divided
attention condition. Via headphones, participants heard a random sequence of numerals
between 1 and 5 spoken in a female voice. One digit was presented every two seconds. The
task was to compare each digit with the preceding digit and to press one key if they
differed and another key if they were the same. After a practice block of 20 trials,
participants were reminded of the task and response keys and then worked through a block
of 60 trials. This second block provided a baseline for assessing the performance in the 1-back task when used as secondary task.

In the subsequent study phase, 60 words were presented on the computer monitor one at a time for 2s with an interstimulus interval of 2s. In the full attention condition, the only task was to study these words for a later recognition memory test. In the divided attention condition, the visual presentation of the words was combined with the auditory presentation of a random sequence of numerals between 1 and 5, such that one digit was presented per word and one per interstimulus interval at the previously practiced rate of one digit every two seconds. Participants were instructed to study the words for a later recognition memory test while continuing to solve the 1-back task as before. They were admonished to work through the 1-back task as accurately as in the previous block. Following the study phase, participants solved figural reasoning problems taken from the Advanced Progressive Matrices Test (Raven, 1962) for 15 minutes as a filler task.

The recognition memory test was identical in both conditions. It consisted of 60 targets and 60 distractors. To spread targets and distractors evenly across the test as in the previous experiments, the 120 test items were divided into 10 sequences of 12 items. Each sequence contained six targets and six distractors in a random order determined for each sequence and participant anew. Contrary to the previous experiments, participants were not informed of the percentage of targets and distractors. They entered their responses by pressing colour-coded keys of the keyboard. For each test item, participants first had to decide whether it had been presented in the study phase. The response options old and new were displayed underneath the test item. The corresponding response keys had yellow labels, and the selected response option was highlighted by a yellow frame. In case of an old response, the response options remember and know were added to the display. The corresponding response keys had green labels, and the selected response option was highlighted by a green frame. Before the next trial started, the screen turned blank for 0.6s.

The instructions on the remember-know distinction were the same as in Experiment 1. After completing the recognition memory test, participants reported both whether and then how they understood the terms remember and know. The verbatim remember-know instructions can be found in Appendix B.3.
8.1.2 Results and Discussion

Apart from two participants who failed to follow the instructions for the secondary task, 41 participants reported that they understood the the terms remember and know, \( n = 21 \) in the full attention condition and \( n = 20 \) in the divided attention condition. However, three descriptions (i.e., 7.3%) indicated misunderstandings in participants who reported that they understood the distinction. Excluding these participants leaves 19 participants in each condition. The data were analysed both with the inclusion criterion self-report and with the additional inclusion criterion description content, as it was done for Experiment 2. Reassuringly, the pattern of results of both the traditional analysis and the model-based analysis again did not differ between inclusion criteria. In line with the previous results sections, the following presentation is limited to the self-report inclusion criterion. The significance level was set to \( \alpha = .05 \) for all analyses.

Performance in the 1-back task. In the divided attention condition, performance in the 60-items 1-back task did not differ between training block (average proportion correct .96, \( SD = .05 \)) and study phase (\( M = .95; SD = .06 \)), \(|t| < 1\).

Traditional Analysis. Table 8.1 displays the means and standard deviations of remember and know response rates to targets and distractors for each condition. Regarding responses to targets, dividing attention at study significantly decreased the remember response rate, \( t(39) = 3.89, p < .001 \), but did not affect the know response rate, \(|t| < 1\). Regarding responses to distractors, there was not more than one remember response to distractors for 57\% and 45\% of the participants in the full and divided attention conditions, respectively. Further, 33\% of the participants in the full attention condition did not give more than one know response to distractors. These low levels of remember and know responses to distractors prevent meaningful \( t \)-tests. Instead, responses to distractors were analysed with rank-based Wilcoxon tests. There was no significant effect of attention condition for remember responses, \( z = -1.10, p = .272 \). For know responses, there was a significant effect, \( z = -3.35, p = .001 \). Table 8.1 shows that the know response rate to distractors was higher in the divided attention condition than in the full attention condition.

Thus, focussing on responses to targets, dividing attention at study led to a decrease of the remember response rate, but did not affect the know response rate, replicating prior studies (Gardiner & Parkin, 1990; Parkin et al., 1995). An analysis with the 2HT task-
oriented measurement model served to elucidate whether this result reflects a genuine effect of dividing attention on retrieval experience, which is predicted by the distinctiveness-fluency framework of retrieval experience, or whether it merely reflects an impairment of recognition memory performance that is to be expected in the divided attention condition (e.g., Craik et al., 1996; Parkin et al., 1990). In particular, while the distinctiveness-fluency framework is mute with regard to a possible effect of a distinctiveness manipulation at study on the retrieval experience for items guessed to be old, impeding the processing of distinctive features at study by requiring participants to attend to a secondary task should reduce the conditional probability of remember responses to truly recognised targets.

Model-based analysis. Figure 8.1 displays the parameter estimates and their 95% confidence intervals. As expected, the recognition memory performance, captured by parameter $D$, was worse in the divided attention condition, $G^2(1) = 147.40, p < .001$. The response bias in the old-new decision, captured by parameter $b$, did not differ between attention conditions, $G^2(1) = 0.71, p = .400$. Participants responded conservatively in both conditions as reflected by $b < .5, G^2(1) = 89.07, p < .001$ and $G^2(1) = 147.79, p < .001$ in the full and divided attention condition, respectively. Turning to retrieval experience, the probability of a remember response to an item guessed to be old, captured by parameter $R^*$, was lower after divided than full attention at study, $G^2(1) = 13.08, p < .001$. The retrieval experience for truly recognised items, captured by parameter $R$, was not affected by dividing attention at study, $G^2(1) = 0.49, p = .486$, contrary to the prediction derived from the distinctiveness-fluency framework.
In short, the model-based analysis confirmed that recognition memory performance was worse after divided than full attention at study. Furthermore, the model-based analysis showed that the decrease of the remember response rate to targets with divided as compared to full attention at study resulted from a decrease of recognition memory performance in conjunction with a lower conditional probability of responding remember to an item guessed to be old. Importantly, dividing attention did not affect the retrieval experience for truly recognised targets, contrary to the prediction of the distinctiveness-fluency framework and contrary to the traditional interpretation of the difference in the remember response rates to targets as a selective effect on retrieval experience. Thus, this experiment demonstrates empirically that the use of remember-know response rates as measures of retrieval experience per se may lead to erroneous conclusions and casts doubt on the distinctiveness-fluency framework.

### 8.2 Experiment 4

Experiment 4 served to re-investigate the effect of semantic versus phonemic orienting tasks on retrieval experience. Previous experiments with the remember-know paradigm (Erdfelder, 2000; Gardiner, 1988; Gregg & Gardiner, 1994; Java, Gregg, & Gardiner, 1997; Perfect, Williams, & Anderton Brown, 1995; Rajaram, 1993; Yonelinas, 2001) have
consistently found a reduced rate of remember responses to targets following phonemic as compared to semantic orienting tasks. The results for the know response rates to targets are less consistent: Sometimes no differences were observed (e.g., Gardiner, 1988; Yonelinas, 2001), and sometimes the know response rate to targets was higher following a phonemic orienting task (e.g., Perfect et al., 1995; Rajaram, 1993). In the present experiment, different groups of participants were asked to write down a semantic associate versus a rhyme for each word of a visually presented list.

### 8.2.1 Method

**Participants and design.** Fifty-one students from Jena University were quasi-randomly assigned to either the semantic condition or the phonemic condition. All participants were native speakers of German. One participant was excluded from the analysis because she reported that she did not understand the remember-know distinction. The remaining participants’ explanations of the terms remember and know did not indicate misunderstandings. Of these 50 participants, 25 were tested in each condition. All participants received 5€.

**Materials and procedure.** The word pool of highly imageable, concrete German nouns with four to nine letters used in Experiments 1 and 3 was adapted for the purposes of this experiment as follows. To ascertain that participants would likely find rhymes to the target words, a pretest was conducted. Four students of psychology were asked to generate a rhyme for each of the 223 pool words presented in a random order. A word was excluded from the pool unless at least three of the four pretesters wrote down a valid rhyme. Furthermore, to reduce the risk that participants respond with a word from the pool, it should not contain sets of words which rhyme with each other or which are strongly semantically associated. Care was taken to control the word pool accordingly. By accident, two rhyme pairs remained in the final pool, which consisted of 108 words. For each participant, a random set of 54 targets was chosen from the word pool. The remaining 54 words served as distractors.

In the study phase, 54 targets were presented one at a time on the computer monitor for 5s. The 1s interstimulus interval was filled with an acoustic signal presented via headphones. The participants had been recruited for a study on word processing. Depending on the condition, they were set the orienting task of generating either a semantic associate or a rhyme to each word. The answers had to be written down on lined
response sheets. Participants were informed that they had 5s time for each word. When this time was up, the next word would be announced by a warning tone. They were instructed to attend to the next word when they heard the signal, also in case they had not finished the previous task. Following the study phase, participants solved figural reasoning problems taken from the Advanced Progressive Matrices Test (Raven, 1962) for 15 minutes as a filler task.

The recognition memory test consisted of 54 targets and 54 distractors. To spread targets and distractors evenly across the test as in the previous experiments, the 108 test items were divided into nine sequences of 12 items. Each sequence contained six targets and six distractors in a random order determined for each sequence and participant anew. Participants entered their responses by mouse clicks on response fields. First, the response options old and new were displayed underneath the test item in yellow frames. Upon clicking a response field, it turned yellow completely for 0.6s. In case of an old response, the response options remember and know were added to the display at the bottom of the screen in green frames. Upon clicking a response field, it turned green completely for 0.6s. Before the next trial started, the screen turned blank for 1s. For both the old-new and the remember-know response, it was determined quasi-randomly which response option was presented on the left and right position at each trial.

With regard to the remember-know distinction, the instructions on the response option remember were identical to Experiment 2, whereas the instructions on the alternative know were identical to Experiments 1 and 3. That is, the explanations what it means to remember a test word contained numerous examples of the kinds of aspects of the study episode that, if brought back to mind by a test item, would justify a remember response. To know a test word was described as recognising the word from the study phase in the absence a specific recollection of the episode of studying the word. After completing the recognition memory test, participants reported both whether and how they understood the terms remember and know. The verbatim remember-know instructions can be found in Appendix B.4.

### 8.2.2 Results and Discussion

*Orienting task.* On average, participants in the phonemic condition wrote down a response to 83.2% of the words ($SD = 8.6$), compared to 97.5% in the semantic condition ($SD = 4.1$). Indeed, 44% of the participants in the semantic condition wrote down a
response to every target and 32% left only one blank. A Wilcoxon test confirmed that the semantic orienting task was easier than the rhyme task, $z = -5.47, p < .001$.

It may suggest itself to argue that this difference in the baserate of written semantic or phonemic context features inherently favours remembering in the semantic condition as compared to the phonemic condition. However, importantly, not only recollecting a correct solution of the orienting task, but rather recollecting any aspect of the study episode justifies a remember response. For instance, any detail of how one tried find a solution suffices. Indeed, when asked for examples of remembered items, several participants described how they had unsuccessfully tried to find a solution.

_Traditional analysis._ Table 8.2 displays the means and standard deviations of remember and know response rates to targets and distractors for each condition. Regarding responses to targets, the remember response rate was significantly lower in the phonemic condition than in the semantic condition, $t(48) = 3.87, p < .001$. The know response rate did not differ as a function of orienting task, $|t| < 1$. Regarding responses to distractors, there was not more than one remember response to distractors for 92% and 84% of the participants in the semantic and phonemic conditions, respectively. Further, 80% of the participants in the semantic condition did not give more than one know response to distractors. These low levels of remember and know responses to distractors prevent meaningful $t$-tests. Instead, responses to distractors were analysed with rank-based Wilcoxon tests. There was no significant effect of orienting task for remember responses, $z = -0.93, p = .350$. For know responses there was a significant effect, $z = -4.18, p < .001$. Table 8.2 shows that the know response rate to distractors was higher in the phonemic condition.

Thus, compared to the semantic orienting task, the phonemic orienting task reduced the remember response rate to targets, replicating the finding consistently reported in the

| Table 8.2 |
|------------------|------------------|
| **Means (and Standard Deviations) of Remember and Know Response Rates as a Function of Item Type and Orienting Task (Experiment 4)** |
|                  | Remember         | Know             |
|                  | Semantic   | Phonemic | Semantic   | Phonemic |
| Targets          | .63 (.21)  | .43 (.15) | .33 (.20)  | .33 (.15) |
| Distractors      | .01 (.01)  | .01 (.02) | .02 (.03)  | .07 (.06) |
literature (Erdfelder, 2000; Gardiner, 1988; Gregg & Gardiner, 1994; Java et al. 1997; Perfect et al., 1995; Rajaram, 1993; Yonelinas, 2001). The know response rate to targets was not affected. This has also been found in some previous experiments (e.g., Gardiner, 1988; Yonelinas, 2001), although other experiments found a higher know response rate to targets in the phonemic condition (e.g., Perfect et al., 1995; Rajaram, 1993). An analysis of the present data with the 2HT task-oriented measurement model served to elucidate whether the traditional result reflects a genuine effect of the semantic versus phonemic orienting task on retrieval experience, which is predicted by the distinctiveness-fluency framework of retrieval experience, or whether it merely reflects an impairment of recognition memory performance that is to be expected in the phonemic condition (e.g., Craik & Tulving, 1975; Jacoby & Dallas, 1981). In particular, while the distinctiveness-fluency framework is mute with regard to a possible effect of a distinctiveness manipulation at study on the retrieval experience for items guessed to be old, impeding the processing of salient stimulus attributes at study by means of a phonemic orienting task should reduce the conditional probability of remember responses to truly recognised targets.

**Model-based analysis.** Figure 8.2 displays the parameter estimates and their 95% confidence intervals. As expected, the recognition memory performance, captured by parameter $D$, was worse in the phonemic condition than in the semantic condition, $G^2(1) = 273.25$, $p < .001$. The response bias in the old-new decision, captured by parameter $b$, was more conservative in the phonemic condition than the semantic condition, $G^2(1) = 5.73$, $p = .017$. However, participants in both conditions responded conservatively, as reflected by $b < .5$, $G^2(1) = 5.38$, $p = .020$ in the semantic condition and $G^2(1) = 124.95$, $p < .001$ in the phonemic condition. Turning to retrieval experience, the conditional probability of a remember response to an item guessed to be old, captured by parameter $R^*$, was lower in the phonemic condition than in the semantic condition, $G^2(1) = 4.72$, $p = .030$. The retrieval experience for truly recognised targets, captured by parameter $R$, also differed significantly as a function of orienting task, $G^2(1) = 4.13$, $p = .042$. The conditional probability of a remember response to a truly recognised target was slightly lower in the phonemic condition, $\hat{R} = .62$, than in the semantic condition, $\hat{R} = .66$, in line with the prediction of the distinctiveness-fluency framework.
In short, the model-based analysis confirmed that recognition memory performance was worse after phonemic than semantic processing, replicating the well-known levels-of-processing effect. Furthermore, the old-new response bias, the retrieval experience for items guessed to be old, and the retrieval experience for truly recognised targets differed as a function of orienting task. Importantly, as predicted by the distinctiveness-fluency framework, the conditional probability of a remember response to a truly recognised target was lower after the phonemic than the semantic orienting task. However, the lower remember response rate to targets after the phonemic orienting task, traditionally interpreted as a selective effect on retrieval experience, does not only reflect the small difference of retrieval experience for truly recognised targets, but also the differences in recognition memory performance, old-new response bias and retrieval experience for items guessed to be old.

8.3 Discussion of Experiments 3 and 4

The experiments reported in this chapter re-investigated the effects of classic manipulations of recognition memory performance on retrieval experience with valid measures of retrieval experience, which control for recognition memory performance as well as old-new response bias. The twofold goal was to demonstrate empirically that the
traditional measures may lead to erroneous conclusions and to test the distinctiveness-fluency framework of retrieval experience.

The distinctiveness-fluency framework posits that the retrieval experience of remembering depends on the processing of distinctive and salient features at study (Rajaram, 1996). Traditionally, this notion is tested in terms of the remember-know response rates to targets. In the present experiments, the processing of distinctive and salient features was impeded by requiring participants to perform an attention-demanding secondary task in the study phase (Experiment 3) and by giving participants a phonemic orienting task rather than a semantic orienting task (Experiment 4). Indeed, both manipulations reduced the rate of remember responses to targets, but did not affect the know response rate to targets. Furthermore, both manipulations increased the know responses to distractors, whereas the remember responses to distractors were not affected. That is, both experiments yielded the same pattern of results in terms of remember-know response rates. This pattern is traditionally interpreted as an effect on retrieval experience that is in line with the distinctiveness-fluency framework.

However, because remember-know response rates are expected to vary as a function of recognition memory performance and old-new response bias, this conclusion is questionable. In contrast, pure measures of retrieval experience are provided by the 2HT task-oriented model. Applying the distinctiveness-fluency framework to the parameters of the 2HT task-oriented model leads to the prediction that impeding the processing of distinctive or salient features at study should reduce the conditional probability of remember responses to truly recognised targets. In contrast, the distinctiveness-fluency framework is mute with regard to a possible effect of processing distinctive or salient features at study on the retrieval experience for items guessed to be old, because the lack of true recognition implies the failure to truly recollect genuine features of the study episode.

The model-based analysis of retrieval experience for truly recognised targets showed that the phonemic as compared to the semantic orienting task (Experiment 4) decreased the conditional probability of remember responses given true recognition, as predicted by the distinctiveness-fluency framework. In contrast, dividing attention at study (Experiment 3) did not affect the retrieval experience for truly recognised targets, contrary to the prediction of the distinctiveness-fluency framework. In Experiment 3, the traditional
analysis of remember-know response rates leads to an erroneous conclusion about the effect of dividing attention on retrieval experience.

Given that Experiments 3 and 4 yielded the same pattern of results on the surface of differences between experimental conditions in terms of remember-know response rates, one might wonder whether the critically differing results with respect to retrieval experience for truly recognised targets reflect some idiosyncrasy of the present datasets instead of a genuine systematic difference between the effects of dividing attention and semantic versus phonemic orienting tasks on retrieval experience. Some evidence can be obtained by reanalysing published experiments with the 2HT task-oriented model. Because this measurement model is only meaningfully applicable to pairs of experimental conditions that are expected to vary with regard to recognition memory performance if separate estimates of $D_{new}$ are available for each condition (see chapter 6), only those experiments were reanalysed that manipulated attention at study and orienting tasks between participants.

Table 8.3
Reanalysis of Published Experiments that Manipulated Attention at Study Between Participants With the 2HT Task-Oriented Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Full attention</th>
<th>Divided attention</th>
<th>Statistical test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>CI</td>
<td>Estimate</td>
</tr>
<tr>
<td>$D$</td>
<td>.66</td>
<td>.62, .70</td>
<td>.50</td>
</tr>
<tr>
<td>$b$</td>
<td>.15</td>
<td>.10, .19</td>
<td>.16</td>
</tr>
<tr>
<td>$R^*$</td>
<td>.19</td>
<td>.05, .32</td>
<td>.12</td>
</tr>
<tr>
<td>$R$</td>
<td>.74</td>
<td>.69, .79</td>
<td>.74</td>
</tr>
</tbody>
</table>

Gardiner & Parkin (1990)

| $D$       | .68            | .64, .72          | .43             | .38, .48          | 66.14  | 1  | < .001 |
| $b$       | .44            | .38, .50          | .40             | .36, .44          | 0.84   | 1  | .358   |
| $R^*$     | .29            | .20, .38          | .35             | .28, .42          | 1.12   | 1  | .290   |
| $R$       | .73            | .68, .79          | .70             | .61, .78          | 0.56   | 1  | .454   |

Parkin, Gardiner, & Rosser (1995)

Note. CI = 95% confidence interval; $D$ = recognition memory performance; $b$ = old-new response bias; $R^*$ = probability of remember responses to items guessed to be old; $R$ = probability of remember responses to truly recognised targets.
Table 8.4
Reanalysis of Published Experiments that Manipulated Orienting Tasks Between Participants With the 2HT Task-Oriented Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Semantic</th>
<th></th>
<th></th>
<th>Phonemic</th>
<th></th>
<th></th>
<th>Statistical test</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>CI</td>
<td></td>
<td>Estimate</td>
<td>CI</td>
<td></td>
<td>G²</td>
<td>df</td>
<td>p</td>
</tr>
<tr>
<td>Gardiner (1988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>.81</td>
<td>.78, .84</td>
<td></td>
<td>.59</td>
<td>.55, .63</td>
<td></td>
<td>63.12</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$b$</td>
<td>.11</td>
<td>.05, .17</td>
<td></td>
<td>.08</td>
<td>.04, .11</td>
<td></td>
<td>1.06</td>
<td>1</td>
<td>.303</td>
</tr>
<tr>
<td>$R^*$</td>
<td>.50</td>
<td>.22, .78</td>
<td></td>
<td>.67</td>
<td>.45, .88</td>
<td></td>
<td>0.83</td>
<td>1</td>
<td>.362</td>
</tr>
<tr>
<td>$R$</td>
<td>.80</td>
<td>.76, .84</td>
<td></td>
<td>.75</td>
<td>.70, .80</td>
<td></td>
<td>3.18</td>
<td>1</td>
<td>.075</td>
</tr>
<tr>
<td>Perfect, Williams, &amp; Anderton-Brown, 1995, young adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>.95</td>
<td>.93, .97</td>
<td></td>
<td>.71</td>
<td>.66, .75</td>
<td></td>
<td>91.83</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$b$</td>
<td>.38</td>
<td>.18, .59</td>
<td></td>
<td>.17</td>
<td>.11, .24</td>
<td></td>
<td>4.29</td>
<td>1</td>
<td>.038</td>
</tr>
<tr>
<td>$R^*$</td>
<td>.50</td>
<td>.15, .85</td>
<td></td>
<td>.00</td>
<td>.00, .00</td>
<td></td>
<td>12.47</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$R$</td>
<td>.71</td>
<td>.66, .75</td>
<td></td>
<td>.56</td>
<td>.50, .62</td>
<td></td>
<td>13.23</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Perfect, Williams, &amp; Anderton-Brown, 1995, old adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>.74</td>
<td>.70, .78</td>
<td></td>
<td>.46</td>
<td>.40, .51</td>
<td></td>
<td>63.01</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$b$</td>
<td>.12</td>
<td>.06, .17</td>
<td></td>
<td>.15</td>
<td>.11, .19</td>
<td></td>
<td>0.80</td>
<td>1</td>
<td>.372</td>
</tr>
<tr>
<td>$R^*$</td>
<td>.69</td>
<td>.44, .94</td>
<td></td>
<td>.37</td>
<td>.21, .53</td>
<td></td>
<td>3.98</td>
<td>1</td>
<td>.046</td>
</tr>
<tr>
<td>$R$</td>
<td>.90</td>
<td>.87, .94</td>
<td></td>
<td>.68</td>
<td>.60, .76</td>
<td></td>
<td>22.91</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Note. CI = 95% confidence interval; $D$ = recognition memory performance; $b$ = old-new response bias; $R^*$ = probability of remember responses to items guessed to be old; $R$ = probability of remember responses to truly recognised targets.

With regard to the effects of dividing attention at study, Table 8.3 shows that the experiments by Gardiner and Parkin (1990) and Parkin et al. (1995) converge with the present Experiment 3 in finding that dividing attention does not affect the retrieval experience for truly recognised targets, contrary to the prediction of the distinctiveness-fluency framework. With regard to semantic versus phonemic processing, Table 8.4 shows that the retrieval experience for truly recognised targets differs as a function of orienting task not only in the present Experiment 4, but also in both age groups investigated by Perfect et al. (1995), whereas in the dataset by Gardiner (1988), the difference missed significance ($p=.075$). These findings suggest that phonemic as compared to semantic
processing does indeed influence the retrieval experience for truly recognised targets, as predicted by the distinctiveness-fluency framework.

That is, with regard to the retrieval experience for truly recognised targets, the results of both Experiment 3 and 4 agree with the results of previous experiments, suggesting a genuine difference between the effects of dividing attention and semantic versus phonemic orienting tasks on retrieval experience. In particular, the finding that dividing attention at study does not affect the retrieval experience for truly recognised targets casts doubt on the distinctiveness-fluency framework of retrieval experience.
9 General Discussion

9.1 Overview

The remember-know-(guess) paradigm investigates the subjective experience that accompanies retrieving information from memory in recognition memory tests. In the remember-know instructions, participants are required to respond remember if they are able to become consciously aware again of one or more aspects of the past episode of studying the test item, whereas they are required to respond know if the test item is confidently recognised, but does not bring back to mind any aspect of the prior occurrence in the study phase. The rates of remember and know responses have been used as measures of retrieval experience in two different lines of research, in which the scientific interest is directed at retrieval experience per se (Rajaram, 1996) or at retrieval experience as a tool for isolating different hypothetical memory systems that jointly produce the recognition memory performance (Tulving, 1985b). The present thesis started by analysing the requirements for suitable measures of retrieval experience for either purpose, which turned out to be partially conflicting, and by analysing whether the remember-know response rates meet these requirements.

If the research goal is to explain "the nature of retrieval experience per se" (Rajaram, 1999, p. 262), a pure measure of retrieval experience is required. As remember and know response rates are obtained by partitioning the old response rate, and the old response rate in recognition memory tests varies as a function of both old-new response bias and recognition memory performance, remember-know response rates are expected to vary as a function of these factors as well, instead of being selectively sensitive to changes of retrieval experience.

In contrast, if retrieval experience is used as a tool for isolating different hypothetical memory systems that simultaneously mediate a state of retrieval experience and contribute to recognition memory performance (Tulving, 1985b), it is desirable to use measures that represent these theoretical constructs. As an increase in recognition memory performance can only be achieved by means of greater retrieval success from at least one memory system, measures of retrieval experience that represent memory systems cannot be independent of recognition memory performance, contrary to what is required of pure measures of retrieval experience. Rather, measures that represent retrieval from episodic
and semantic memory are two-in-one measures of retrieval experience and a component of recognition memory performance. Because experimental manipulations of old-new response bias do not influence the ability to discriminate targets from distractors, it is reasonable to assume that they also do not affect the ability to retrieve information from either memory system. Hence, measures that represent a memory system are required to be invariant to manipulations of old-new response bias.

The existing empirical evidence about the effects of manipulations of old-new response bias on remember-know response rates was extended by two new experiments. In both Experiment 1 and 2, the manipulation of percentage of targets in the test was crossed with whether the response option guess was available for classifying the retrieval experience. In the remember-know condition, remember response rates (Experiment 2) and know response rates (Experiment 1 and 2) varied as a function of the old-new response bias manipulation. These results replicate the consistent finding of several published experiments, which employed a range of other manipulations of old-new response bias, in showing that remember-know response rates obtained in the remember-know paradigm do not meet the invariance criterion.

Adding the response option guess for specifying old responses, which yields the remember-know-guess paradigm, has the potential to eliminate the variation of remember-know response rates with old-new response bias, because it abolishes the restriction that remember and know response rates add up to the old response rate. Indeed, in the remember-know-guess condition of Experiment 2 the remember-know response rates met the invariance criterion. However, both remember and know response rates varied as a function of the old-new response bias manipulation in Experiment 1. Combining these results with those of the published experiments shows that remember-know rates obtained in the remember-know-guess paradigm met the invariance criterion in two of five datasets. This indicates that one cannot rely on obtaining uncontaminated remember-know response rates in the remember-know-guess paradigm.

In short, the empirical evidence shows that, irrespective of whether the response option guess is available for classifying retrieval experience or not, the rates of remember and know responses are not suitable for representing the hypothetical memory systems. However, in contrast to measures that represent retrieval from episodic and semantic memory, pure measures of retrieval experience, though required to be invariant to changes
of old-new response bias, can not be required to be invariant to experimental manipulations of old-new response bias, because it is possible that the experimental manipulation affects the true retrieval experience in addition to the old-new response bias. Consequently, this kind of data cannot provide conclusive evidence that remember-know response rates are unsuitable as pure measures of retrieval experience. However, as noted before, doubts in the suitability of remember-know response rates as pure measures of retrieval experience stem from the well-established fact that the old response rate in recognition memory tests varies as a function of both old-new response bias and recognition memory performance. Also, adding the response option guess does not address the potential contamination of remember-know response rates due to differences between experimental conditions with regard recognition memory performance.

Because the remember-know response rates are not suitable as measures of retrieval experience, the present thesis examined alternative measures, which were formulated within multinomial processing tree models. In an attempt to provide pure measures of retrieval experience per se, the task-oriented measurement models start by specifying measures of recognition memory performance and old-new response bias. These are supplemented by two measures of retrieval experience, one of which is specified conditionally on true recognition, whereas the other is specified conditionally on old responses due to guessing. By virtue of their construction, these measures can be expected to be pure measures of retrieval experience. In contrast, the multi-memory measurement models implement Tulving's (1985b) model of multiple memory systems by specifying parameters that reflect retrieval from episodic and semantic memory. Additionally, these measurement models allow for the possibility that remember and know responses may result from guessing, which is captured with additional parameters. These modelling approaches were applied to both the remember-know paradigm and the remember-know-guess paradigm. The measurement models were evaluated by examining whether model parameters that reflect recognition memory performance or components thereof are invariant to experimental manipulations of old-new response bias, based on reanalyses of published experiments and analyses of Experiment 1 and 2.

Interestingly, none of the measurement models that implement the multiple-memory systems model (Tulving, 1985b) proved globally valid. While the variation of remember-know response rates with manipulations of old-new response bias can easily be attributed
to flawed measures of the memory systems, which neglect the possibility of remember and know responses due to guessing, this reasoning is not applicable to the multi-memory measurement models, which explicitly provide for the possibility that remember and know responses may result from guessing rather than only from retrieval from the episodic and semantic memory system. Therefore, the consistent failure of all multi-memory measurement models for the remember-know-(guess) paradigm suggests that the multiple-memory systems model cannot explain remember-know judgments.

Second, the model evaluation showed that the 2HT task-oriented measurement model for the original remember-know paradigm, which was designed to provide pure measures of retrieval experience, is a valid measurement model. Thus, it constitutes a suitable tool for examining the distinctiveness-fluency framework of retrieval experience suggested by Rajaram (1996). One claim of this framework is that the retrieval experience of remembering depends on the processing of distinctive or salient features at study. In terms of the parameters of the 2HT task-oriented model, impeding the processing of distinctive or salient features at study should affect the retrieval experience for truly recognised targets. Experiments 3 and 4 served to test this prediction of the distinctiveness-fluency framework and to demonstrate empirically that the traditional analysis of remember-know response rates may lead to erroneous conclusions.

In Experiment 3, the processing of distinctive features was impeded by requiring participants to perform an attention-demanding secondary task at study as compared to a full attention condition. Dividing attention at study reduced the rate of remember responses to targets, but did not affect the know response rate to targets, which is traditionally interpreted as an effect on retrieval experience. However, the model-based analysis showed that dividing attention did not affect the retrieval experience for truly recognised targets, contrary to the prediction of the distinctiveness-fluency framework. Thus, this experiment demonstrates empirically that the use of remember-know response rates as measures of retrieval experience per se may lead to erroneous conclusions and casts doubt on the distinctiveness-fluency framework.

In Experiment 4, the processing of salient stimulus attributes was impeded by requiring participants to perform a phonemic orienting task as compared to a semantic orienting task. This manipulation reduced the rate of remember responses to targets, but did not affect the know response rate to targets. The model-based analysis showed that the conditional
probability of remember responses to truly recognised targets was lower in the phonemic condition than in the semantic condition, too. That is, the type of orienting task did indeed influence the retrieval experience for truly recognised targets, as predicted by the distinctiveness-fluency framework.

Not only the traditional results of both Experiments 3 and 4 are in line with the results of published experiments that manipulated these variables, but also the model-based results, as indicated by model-based reanalyses. This supports the notion that there is a genuine difference between the effects of dividing attention and semantic versus phonemic orienting tasks on retrieval experience. In particular, the finding that dividing attention at study does not affect the retrieval experience for truly recognised targets casts doubt on the distinctiveness-fluency framework of retrieval experience. In general, the concomitance of differing findings in terms of the parameters of a validated measurement model, but identical findings in terms of remember-know response rates highlights "the importance of models in interpreting remember-know experiments" (Macmillan et al., 2005, p. 607).

The remainder of this chapter is concerned with the relation between remember-know judgments and the retrieval of context features, the role of judgment processes in the remember-know paradigm, and the relation between retrieval experience as captured by remember-know responses and recognition memory performance.

9.2 Remember-Know Judgments and the Retrieval of Context Features

In the remember-know instructions, participants are required to respond remember if they are able to become consciously aware again of one or more aspects of the past episode of studying the test item, whereas they are required to respond know if the test item is confidently recognised, but does not bring back to mind any aspect of the prior occurrence in the study phase. Accordingly, if participants follow the instructions, remember responses should always be accompanied by the retrieval of at least one feature of the study context, whereas know responses should not be accompanied by the retrieval of any context feature. Is this indeed the case? It is a natural consequence of the first-person approach, which hinges on the participants' introspection and classification of their subjective retrieval experience, that the experimenters' ability to check whether participants apply the remember-know instructions properly is limited.
One step in this direction is to ascertain that participants understand the remember-know instruction correctly, which is a necessary condition for applying them properly and thus for obtaining valid data. One possibility is to ask participants to report back the gist of the remember-know distinction. In an alternative approach, participants are asked to describe the experiential basis of their responses in more detail for a subset of items. These strategies are frequently used in the remember-know literature (e.g., Gardiner & Parkin, 1990; Perfect et al., 1995) and were applied in three of the four experiments reported in this thesis. In Experiments 2 and 3, the detailed descriptions indicated misunderstandings in about 7% of the participants who reported to have understood the remember-know distinction, whereas in Experiment 4, no description raised the suspicion of a misunderstanding. Excluding those participants from the analysis whose descriptions gave reason to suspect a misunderstanding did not change the pattern of results of the traditional and the model-based analyses in Experiments 2 and 3. This finding is reassuring in that it suggests that generally, the results of experiments that did not include such precautions, like the present Experiment 1 or many published experiments (e.g., Dewhurst & Hitch, 1999; Rajaram, 1996), are comparable to the results of experiments that strived to ascertain the correct understanding of the remember-know instruction.

Turning to the contents of what participants report when asked to detail their retrieval experience, the responses obtained in the present experiments correspond to those reported in transcript studies (Gardiner, Ramponi, & Richardson-Klavehn, 1998; Java et al., 1997). Among the kinds of context features that participants reported to recollect from the study episode for items called remember were mental images conjured at study, associations formed with other items in the study list, associations to something of personal relevance triggered by the study item, emotional reactions, or perceptual features. In Experiment 4, in which participants had to generate either a semantic associate or a rhyme to each study word, the details reported for items called remember typically referred to aspects of solving or failing to solve the orienting task. Beyond reproducing the prior solution, these responses also referred, for example, to the ease or difficulty of generating a solution, time pressure, or spelling difficulties. Furthermore, they frequently referred to features that are not specifically associated with the orienting task, like emotional reactions or extra-list associations, as they were reported in studies without orienting tasks.
In contrast, when asked to detail the retrieval experience for items classified as know, participants in all studies reported that the item felt familiar or that they just knew it. In line with the know instructions, they did not refer to any recollections. However, a study that asked for self-reports in a different way (Perfect, Mayes, Downes, & Van Eijk, 1996, Experiment 5) provides evidence that these responses may reflect task demands. Participants received a list of kinds of context features and were asked to indicate for each item which categories applied to their retrieval experience. In contrast to the results obtained with the unspecified request to describe details of the retrieval experience, participants reported recollecting some context feature for about one third of the know responses. Thus, while know responses were far less associated with context features than remember responses, recollection of context features was not absent for know responses, contrary to the instruction. Perfect et al. (1996) suggested that these responses do not simply constitute mistakes and argued that their findings "represent evidence of a continuum of contextual knowledge along which subjects are required to draw a distinction between recollection and knowing" (p. 810).

A more objective way of assessing the relationship between remember-know judgments and the retrieval of context features is to combine the remember-know paradigm with a source memory test, which allows to determine the accuracy of memory for context features that characterise the sources. If participants recollect the source of an item, then according to the remember-know instructions, participants should qualify their retrieval experience not as know, but as remember. On the other hand, if participants do not recollect the source, a remember response may nevertheless be justified if participants recollect some other aspect of the study episode that is not diagnostic in the source memory test. In that each source memory test only tests for a subset of the features that can potentially be recollected, source memory tests provide an estimate of the lower bound of the ability to recollect context features of the study episode.

Several studies have shown that remember responses are associated with source discrimination above chance, testing sources defined by features of the item's study context.

---

14 This point is trivial in that reference to a recollection was one reason for excluding participants due to misunderstanding the instruction.
like presentation in the first versus second study list (Perfect et al., 1996), the item's original spatial position (Meiser & Bröder, 2002; Meiser & Sattler, 2006, 2007; Perfect et al., 1996), horizontal versus vertical orientation of associated pairs of pictures (Eldridge, Engel, Zeineh, Bookheimer, & Knowlton, 2005), the colour of an associated cue picture (Dudukovic & Knowlton, 2006; Eldridge et al., 2006) or presentation in a male versus female voice (Meiser & Sattler, 2006). Furthermore, know responses were associated with worse memory for context features than remember responses in these experiments as well as in an experiment that tested the ability to discriminate whether items had been read or generated from anagrams, but which did not test source memory performance against chance (Dewhurst & Hitch, 1999). With regard to the critical issue whether source memory associated with know responses exceeds chance or not, the evidence is mixed. In line with the instruction that know responses be given if the item does not bring back to mind any aspect of its prior occurrence in the study phase, source memory for know responses did not exceed chance (in contrast to remember responses in the same test) when testing for presentation in the first versus second study list (Perfect et al., 1996), the item's original spatial position (Perfect et al., 1996, Experiment 4), the colour of the associated cue picture (Dudukovic & Knowlton, 2006) or presentation in the male versus female voice (Meiser & Sattler, 2006). However, in other experiments, source memory was above chance not only for remember, but also for know responses for the item's original spatial position (Meiser & Bröder, 2002; Meiser & Sattler, 2006, 2007; Perfect et al., 1996, Experiment 3), the orientation of associated picture pairs (Eldridge et al., 2005), or the colour of the associated cue picture (Eldridge et al., 2005). Further, Dudukovic and Knowlton (2006), who found know responses to be associated with source memory at chance level after a retention interval of one week, mention in their discussion that in a pilot study with a 10-minute retention interval, they observed above-chance source memory associated with know responses (which was again worse than that associated with remember responses). Notably, a retention interval in the order of 10 minutes is far more typical of remember-know experiments than a retention interval of one week. In short, know responses may be associated with some memory for context features from the study episode, which is

15 For the context feature voice, the difference just missed statistical significance, $\Delta G^2(1) = 3.63, p = .057$ (Meiser & Sattler, 2006).
consistently lower than the source memory associated with remember responses, but contrary to the remember-know instructions, not reliably absent.

Beyond a quantitative difference, there may exist a difference in the quality of recollected context information that is associated with remember and know judgments, respectively. In particular, it has been suggested that remember responses are particularly associated with recollecting specific features of the study episode (Meiser & Sattler, 2007; see also Perfect et al., 1996), whereas know responses may be accompanied by partial or residual memory for relatively unspecific context features (Hicks, Marsh, & Ritschel, 2002). This assumption, which implies that remember responses should be particularly associated with the ability to make fine-grained source discriminations, can be tested by means of a hierarchical constellation of sources. This rationale was implemented in an experiment in which study words were presented by one of four persons, two men and two women (Meiser & Sattler, 2007). Hence, person-specific information was necessary for identifying the exact source of a test item, whereas less specific information sufficed for distinguishing whether the item had been presented by a man or a woman. Remember responses were associated with better source memory for the individual person than know responses, whereas the partial source memory for gender did not differ between remember and know responses, in line with the assumption that remember responses are particularly associated with recollecting specific context features. Alternatively, if sources are defined by the conjunction of two context features that are varied orthogonally (e.g., one of two spatial positions and one of two font sizes), specific source memory consists of recollecting the configuration of co-occurring context features, as opposed to residual source memory for one of the context features. Accordingly, remember responses, but not know responses should be associated with source memory for the conjunction of both context features. In line with this prediction, source memory for size was found to be better given source memory for spatial position than given no source memory for spatial position only for remember responses, but not for know responses (Meiser & Bröder, 2002; Meiser & Sattler, 2006, 2007; Starns & Hicks, 2005). This differential pattern of stochastic dependence for remember responses, but stochastic independence for know responses was replicated when the context feature of spatial position was crossed with background colour or voice (Meiser & Sattler, 2006, Experiments 2 and 3). Importantly, this qualitative difference in the stochastic relation of source memory for two context features was also
found when the overall source memory performance associated with remember and know responses was equated (Meiser & Sattler, 2006, Experiment 1).

In sum, the review of evidence concerning the relation between retrieval experience and recollection of context features, as indicated by source memory tests, suggests two conclusions. First, remember judgments are reliably associated with recollecting context features, in line with the remember-know instruction. However, know judgments may also be accompanied by recollecting some context feature, contrary to the instruction that know responses be given if the item does not bring back to mind any aspect of its prior occurrence in the study phase. Second, even though the remember-know instructions may be violated in this manner, there nevertheless exist systematic differences between remember and know responses in terms of the quantity and quality of memory for context features. While know responses may be accompanied by partial or residual source memory for relatively unspecific context features (Hicks et al., 2002), remember responses are particularly associated with specific recollections, for example combinations of co-occurring context features (Dudukovic & Knowlton, 2006; Eldridge et al., 2005; Meiser & Bröder, 2002; Meiser & Sattler, 2006, 2007; Starns & Hicks, 2005). Thus, the results of the source memory tests are in line with the idea that participants differentiate between remembering and knowing by setting a response criterion at some point of a "continuum of contextual knowledge" (Perfect et al., 1996, p. 810), which was originally proposed on the basis of introspective reports.

### 9.3 Judgment Processes in the Remember-Know Paradigm

The remember-know paradigm approaches the phenomenon of retrieval experience not by asking participants to describe the subjective experience that accompanies their recognition of a test item as old as detailed and comprehensively as possible, but rather by

---

16 While the remember-know distinction is discussed in terms of the ability to recollect any versus no aspect of the study episode in the overwhelming majority of the literature, the actual remember-know instructions given to participants are not necessarily that clear-cut. For example, in an instruction that numerous researchers referred to as a model, Rajaram (1993) wrote:

"Know" responses should be made when you recognize that the word was in the study list but you cannot consciously recollect anything about its actual occurrence .... In other words, write "K" (for "know") when you are certain of recognizing the words but these words fail to evoke any specific [italics added] conscious recollection from the study list. (p. 102)
asking them to classify their actual retrieval experience as remember versus know. This section discusses evidence\textsuperscript{17} that points to the importance of judgment processes in the remember-know paradigm beyond old-new response bias.

Some evidence can be gathered by returning to the remember-know experiments that aimed at manipulating old-new response bias and by examining whether these manipulations affect the retrieval experience for truly recognised targets, captured by the parameter $R$ of the 2HT task-oriented model that this thesis has shown to be a globally valid measurement model. In these experiments, study phase and retention interval were identical across sets of old-new response bias conditions: The manipulation of old-new response bias was introduced just prior to the test phase, when different groups of participants were informed about the (actual or alleged) percentage of targets in the recognition test (Experiments 1 and 2 of the present thesis; Strack & Förster, 1995; Xu & Bellezza, 2001), instructed to respond conservatively or liberally (Postma, 1998), or introduced to the test procedure, which either consisted of the standard two-step procedure or provided the response options remember, know, and new in one step (Eldridge et al., 2002; Hicks & Marsh, 1999).\textsuperscript{18} Because the study phase is identical across old-new response bias conditions, it is safe to assume that there are no systematic differences between conditions with respect to what participants encode in the study phase. Consequently, if the experimental manipulation affects the retrieval experience for truly recognised targets, this implies that it does not affect what participants are able to recollect from the study episode, but rather how they map their actual retrieval experience onto the response options for classifying the retrieval experience. In other words, it suggests an effect on judgment processes involved in making a remember-know response.

Table 9.1 shows the results for the reanalysed datasets, whereas the corresponding information about Experiments 1 and 2 was reported in chapter 8. Varying the information about the percentage of targets in the test did not affect the retrieval experience for truly recognised targets, regardless of whether the actual percentage of targets was manipulated (Experiments 1 and 2 of the present thesis; Xu & Bellezza, 2001) or not (Strack & Förster, 1995).

\textsuperscript{17} from sources other than source memory experiments

\textsuperscript{18} This enumeration is restricted to those experiments to which the 2HT task-oriented model is meaningfully applicable, cf. chapter 6.
Table 9.1
Reanalysis of Published Datasets with Respect to Effects of Manipulations Aiming at Old-New Response Bias on the Retrieval Experience for Truly Recognised Targets as Captured by Parameter R of the 2HT Task-Oriented Model

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conservative</th>
<th>Liberal</th>
<th>Very liberal</th>
<th>Statistical test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{R}$ CI</td>
<td>$\hat{R}$ CI</td>
<td>$\hat{R}$ CI</td>
<td>$G^2$ df $p$</td>
</tr>
<tr>
<td>Eldridge et al. (2002)</td>
<td>.65 .60, .70</td>
<td>.86 .81, .92</td>
<td></td>
<td>30.94 1 &lt; .001</td>
</tr>
<tr>
<td>Hicks &amp; Marsh (1999)</td>
<td>.70 .67, .73</td>
<td>.75 .71, .79</td>
<td></td>
<td>3.91 1 .048</td>
</tr>
<tr>
<td>4.5s</td>
<td>.72 .68, .77</td>
<td>.94 .88, .99</td>
<td></td>
<td>36.40 1 &lt; .001</td>
</tr>
<tr>
<td>1s</td>
<td>.60 .55, .65</td>
<td>.79 .73, .85</td>
<td></td>
<td>24.89 1 &lt; .001</td>
</tr>
<tr>
<td>Postma (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strack &amp; Förster (1995)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>low word frequency</td>
<td>.69 .58, .79</td>
<td>.84 .71, .97</td>
<td></td>
<td>3.17 1 .075</td>
</tr>
<tr>
<td>high word frequency</td>
<td>.54 .43, .66</td>
<td>.70 .53, .87</td>
<td></td>
<td>2.44 1 .118</td>
</tr>
<tr>
<td>Xu &amp; Bellezza (2001)</td>
<td>.81 .72, .89</td>
<td>.90 .83, .96</td>
<td>.91 .85, .97</td>
<td>4.27 2 118</td>
</tr>
</tbody>
</table>

Note. CI = 95% confidence interval.

In contrast, both the instruction to respond conservatively versus liberally\(^{19}\) (Postma, 1999) and the test procedure with one versus two steps (Eldridge et al., 2002; Hicks & Marsh, 1999) affected the retrieval experience for truly recognised targets, indicating that these experimental manipulations affected how participants mapped their actual retrieval experience for truly recognised targets onto the response options for classifying the retrieval experience. These findings highlight that judgment processes play an important role in the remember-know paradigm beyond old-new response bias.

Further evidence about the importance of judgment processes in the remember-know paradigm comes from an investigation of context effects on remember-know response rates (Bodner & Lindsay, 2003). Participants had to study two lists of words with two of three different orienting tasks, which were designed to induce a shallow, medium or deep level of processing (LOP). All participants studied one list with the medium LOP task, whereas different groups studied the other list with the shallow versus the deep LOP task. For

\(^{19}\) Note that the instruction to respond conservatively versus liberally explicitly referred to the old-new decision.
medium LOP items, the remember response rate was higher and the know response rate was lower when they were mixed with shallow LOP items rather than with deep LOP items at both study and test. The order in which the two LOP tasks were performed at study did not interact with this list-context effect, suggesting that the first LOP task did not influence how participants performed the second LOP task. This justifies the assumption that there were no systematic differences with respect to what participants encoded about medium LOP items, irrespective of whether they were studied in the context of deep or shallow LOP items, and hence that there were no differences with respect to what participants were able to recollect at test. This notion was corroborated by a control experiment, which asked for source judgments instead of remember-know judgments and found that source memory performance for medium LOP items did not differ between the shallow and deep list contexts. Furthermore, there was no list-context effect on remember and know response rates to medium LOP items when the test list consisted exclusively of medium LOP items, instead of a mixture of medium LOP items with shallow or deep LOP items, respectively. This evidence suggests that the list-context effect is due to differences in judgment processes. However, because remember and know response rates may be contaminated with differences in old-new response bias, it is not clear whether these results reflect judgment processes with regard to the old-new decision or with regard to retrieval experience. However, a reanalysis with the 2HT task-oriented model showed that the test-list context affected the retrieval experience for truly recognised targets (see Table 9.2). Hence, the experiments by Bodner and Lindsay (2003) provide further evidence that judgment processes that influence how participants map their actual retrieval experience onto the response options for classifying retrieval experience play an important role in the remember-know paradigm.

The model-based reanalyses reported in this section showed that factors as diverse as list context, test procedure and instructions to respond conservatively or liberally in the old-new decision affect how participants in remember-know experiments map their actual retrieval experience onto the response options for classifying retrieval experience. These findings indicate that judgment processes play an important role in the remember-know paradigm not only with regard to old-new response bias, but also with regard to judging the retrieval experience per se. This argument is based on effects on the parameter $R$ of the 2HT task-oriented model, which captures the probability of a remember response given
### Table 9.2

*Reanalysis of Datasets by Bodner and Lindsay (2003) with the 2HT Task-Oriented Model For Effects of Shallow Versus Deep List Context on the Retrieval Experience for Truly Recognised Targets Captured by Parameter R*

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Medium / shallow</th>
<th>Medium / deep</th>
<th>Statistical test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{R}$</td>
<td>CI</td>
<td>$\hat{R}$</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original FA rates</td>
<td>.79</td>
<td>.74, .84</td>
<td>.58</td>
</tr>
<tr>
<td>Inverted FA rates</td>
<td>.73</td>
<td>.69, .77</td>
<td>0.62</td>
</tr>
<tr>
<td>Experiment 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original FA rates</td>
<td>.80</td>
<td>.74, .85</td>
<td>.59</td>
</tr>
<tr>
<td>Inverted FA rates</td>
<td>.77</td>
<td>.72, .83</td>
<td>.61</td>
</tr>
</tbody>
</table>

*Note.* CI = 95% confidence interval. The columns medium/shallow and medium/deep denote whether the medium LOP items that entered the model-based reanalyses were studied and tested in the context of shallow LOP items or deep LOP items. Note that each test list contained targets studied under different LOP tasks, associated with different levels of recognition memory performance $D$, but only one set of distractors. Hence, when the medium LOP items from each list-context condition are analysed together with the distractors from the same list-context condition, the assumption made by the 2HT task-oriented model that within each list-context condition $D = D_{\text{old}} = D_{\text{new}}$ is dubitable. In the medium/shallow condition, using the false alarm (FA) rates to the common set of distractors underestimates $D_{\text{new}}$ for the medium LOP items, because the shallow LOP task is expected lead to less identification of distractors as new than the medium LOP task. Conversely, in the medium/deep condition, using the FA rates to the common set of distractors overestimates $D_{\text{new}}$ for the medium LOP items, because the deeper LOP task is expected to lead to more identification of distractors as new than the medium LOP task. Note that the parameter $R$, which is the focus of the current analysis, is specified conditionally on $D$. In order to safeguard that the effect on $R$ is not merely an artifact of the model-inherent assumption $D = D_{\text{old}} = D_{\text{new}}$, a second analysis was conducted which inserted the FA rates obtained in the medium/deep condition in the medium/shallow condition and vice versa, thus inverting over- and underestimation of $D_{\text{new}}$, and hence the direction of artifactual influences on $R$, across the list-context conditions. This inversion did not change the results, as can be seen by comparing the row inverted FA rates with the row original FA rates for each experiment.

true recognition of a target. Because differences between conditions with respect to what participants are able to recollect from the study episode were ruled out by means of design (Eldridge et al., 2002; Hicks & Marsh, 1999; Postma, 1999) or subsidiary analyses (Bodner & Lindsay, 2003), these effects can be taken as reflecting judgment processes. However, note that in principle, the parameter $R$ is not only sensitive to how a particular retrieval
experience is mapped onto the response options remember and know, but also to the content of the actual retrieval experience, that is, to what participants are able to recollect from the study episode, which may differ systematically between experimental conditions. In other words, the 2HT task-oriented measurement model does not generally allow one to distinguish whether an experimental manipulation affects the actual retrieval experience in terms of quantity or quality of recollected features or whether it affects judgment processes concerning the classification of the retrieval experience.

In contrast, these two aspects are captured by separate parameters in the sum-difference theory of remembering and knowing (STREAK) proposed by Rotello et al. (2004), which is a two-dimensional signal-detection model. According to this conception, items in a remember-know experiment vary both on a dimension of global strength and on a dimension of specific strength. The latter dimension "measures the specific strength of details associated with the test items" (Rotello et al., 2004, p. 591). Targets and distractors each have a bivariate normal distribution, with a larger standard deviation of the target distribution. Responses are determined by placing two linear decision bounds in the two-dimensional space. The slope of the old-new decision bound is determined by the relative diagnosticity of global and specific strength information, which is in turn determined by the mean global and specific strengths of the target distribution. The remember-know decision bound is orthogonal to the old-new decision bound. With decision bounds at these angles, the old-new decision is based on a weighted sum of global and specific strength (such that a lower value on one dimension can be compensated for by a higher value on the other dimension), whereas the remember-know decision is based on a weighted difference of global and specific strength (such that a remember response is given if the specific strength of a test item is large relative to its global strength). Thus, STREAK contains separate parameters that capture the diagnosticity of specific strength (or the actual retrieval experience in terms of what is recollected) versus the response bias in the remember-know decision. However, note that according to STREAK, the remember-know judgment is based not only on the test item's specific strength, but depends on specific strength relative to global strength. This model might prove useful in future work, which should attend to the judgmental character of remember-know responses in order to achieve a better understanding of the remember-know paradigm.
9.4 Remember-Know Judgments and Recognition Memory Performance

After considering the role of judgment processes in making remember-know responses, the discussion now turns to the relation between recognition memory performance and remember-know responses, starting at the distinctiveness-fluency framework.

The distinctiveness-fluency framework of retrieval experience (Rajaram, 1996) aims to explain "the nature of retrieval experience per se" (Rajaram, 1999, p. 262), but not recognition memory performance. However, before the concepts of distinctiveness and fluency were used to account for retrieval experience per se, they were proposed to account for recognition memory performance. For example, Jacoby and Dallas (1981) concluded from a series of experiments that compared recognition memory and masked perceptual recognition:

We propose that judgments of relative perceptual fluency can be used as a basis for recognition memory. ... The retrieval of study context serves as a second more reliable basis for recognition memory. ... The possibility of retrieving study context depends on the distinctiveness of the original encoding of the item. (p. 333f.)

The 2HT task-oriented measurement model allows to separate the predicted effects of distinctiveness and fluency manipulations on recognition memory performance on the one hand versus on retrieval experience on the other hand by providing separate parameters that capture recognition memory performance, \( D \), and retrieval experience for truly recognised targets, \( R \), respectively. In Experiments 3 and 4, two different manipulations that served to impede the processing of distinctive or salient features at study, dividing attention in Experiment 3 and a phonemic as opposed to a semantic orienting task in Experiment 4, reliably reduced recognition memory performance. The type of orienting task additionally affected the retrieval experience for truly recognised targets, as predicted by the distinctiveness-fluency framework of retrieval experience. In contrast, dividing attention at study did not affect the retrieval experience for truly recognised targets, contrary to the distinctiveness-fluency framework of retrieval experience.

Thus, Experiment 3 provides an empirical example of a manipulation that affected recognition memory performance, but not the retrieval experience for truly recognised
targets. Is it conversely possible to find a manipulation that affects the retrieval experience for truly recognised targets, but not recognition memory performance? Assuming the distinctiveness-fluency accounts of both recognition memory performance and retrieval experience, which respectively hold that recognition memory performance increases both with increasing distinctiveness and with increasing fluency, and that the probability of remember responses to truly recognised targets increases with increasing distinctiveness, but decreases with increasing fluency, one would need to find a pair of encoding conditions in which the effect of decreased distinctiveness on recognition memory performance is offset by the opposite effect of increased fluency, while decreased distinctiveness and increased fluency coact to decrease the conditional probability of remember responses to truly recognised targets. For example, to achieve high fluency paired with little distinctiveness, one could repeatedly present a list of items at a very fast rate of something like 300ms per item in conjunction with an orienting task that keeps participants from attending to potentially distinctive features. To achieve higher distinctiveness, but lower fluency relative to this condition, one might try presenting the list of items once at a somewhat slower presentation rate in conjunction with an orienting task that directs some attention to distinctive features. However, it seems quite a challenge to find an experimental manipulation that increases distinctiveness, but simultaneously decreases fluency.

Also beyond the distinctiveness-fluency accounts of recognition memory performance (Jacoby & Dallas, 1981) and retrieval experience per se (Rajaram, 1996), the relation between recognition memory performance and retrieval experience in terms of the remember-know distinction is a relevant issue, for they are two different aspects of the same recognition memory test. One perspective is to regard them as being mediated by the same underlying processes in such a way that remember-know responses can be used as a tool for isolating two different bases of recognition memory performance. Along these lines, Tulving (1985b) suggested to use the retrieval experience as captured by remember-know judgments for separating the episodic and semantic memory systems. However, as outlined above, the consistent failure of all multi-memory measurement models investigated in the present thesis suggests that the multiple-memory systems model cannot explain remember-know judgments.
Alternatively, Yonelinas (1994) proposed the dual-process signal-detection model of recognition memory, which he summarised thus:

Recollection is assumed to be an all-or-none retrieval process, such that for any item the subject either succeeds or fails at retrieving something about that specific study event. A successful retrieval is expected to lead to a highly confident response. Familiarity, on the other hand, is assumed to be well described by the standard equal-variance signal detection theory described earlier. The two processes are assumed to contribute independently to overall recognition performance. (p. 1343)

Given the match between the functional characterisation of the retrieval processes recollection and familiarity on the one hand and the kinds of retrieval experience distinguished in the remember-know paradigm on the other hand, it is little surprising that Yonelinas (2001, 2002; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Yonelinas & Jacoby, 1995; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998) suggested that the remember-know paradigm can be used as a tool for isolating these retrieval processes. According to the dual-process signal-detection model, the response to any one test item is either based only on recollection, which leads to a remember response, or, if recollection fails, it is based only on familiarity. Note that in this line of research, the aim is not to assess the additional contribution of familiarity to recognition memory performance over and above the contribution of recollection, but instead to determine the retrieval success of familiarity per se, irrespective of whether it governs the response or not.

In different papers, Yonelinas estimated recollection and familiarity from remember-know data in different ways. In the simplest case (Yonelinas, 2001; Yonelinas & Jacoby, 1995), estimators are based only on responses to targets. Recollection is estimated as the

\[ R_T - R_D \]

Other estimators of recollection and familiarity were based the probabilities of a remember response to targets, \( R_T \), and distractors, \( R_D \), as well as the probabilities of a know response to targets, \( K_T \), and distractors, \( K_D \). Yonelinas (2002) estimated recollection as \( R_T - R_D \) and familiarity as \( (K_T/(1-R_T)) - (K_D/(1-R_D)) \). Yonelinas et al. (1996) used the same estimator of recollection, whereas familiarity was estimated in terms of \( d' \), which was derived from \( R_D + K_D \) as estimator of the probability of a familiarity-based false alarm and \( (K_T + R_D)/(1-R_T + R_D) \) as estimator of the probability of a familiarity-based hit. Finally, Yonelinas et al. (1998) estimated recollection as \( (R_T - R_D)/(1-R_D) \) and familiarity in terms of \( d' \)
probability of a remember response, whereas the probability of a know response would underestimate familiarity. Expressed in terms of the model's retrieval processes, participants are required to respond know if a test item is familiar but not recollected, whereas the independence assumption implies that both retrieval processes may be successful for a given item. Consequently, familiarity is estimated as the conditional probability of a know response given that the item is not recollected. This way of estimating recollection and familiarity, which ignores responses to distractors and hence response bias, is obviously overly simplistic and constitutes an inadequate implementation of the dual-process signal-detection model, which explicitly describes familiarity by an equal-variance signal detection model. Hence, an adequate description of this retrieval process requires a parameter that captures the know-new response criterion for familiarity-based responses as well as an accuracy parameter like $d'$ that captures the retrieval success of familiarity independently from the response criterion.

Based on a well-founded implementation, one can test whether the dual-process signal detection model can explain remember-know judgments according to the logic that was applied in the model evaluations reported in the present thesis. Note that according to the dual-process signal-detection model, the retrieval of any feature of the study episode is mediated by the retrieval process of recollection, such that recognition based purely on familiarity is devoid of any retrieved context features. While this fits nicely with how remember and know responses are defined, the evidence reviewed above suggests that it does not accord with how participants actually use the response options when classifying their retrieval experience. Know responses may be associated with some degree of partial source memory, and the conditions and context of testing can influence how the actual retrieval experience is mapped onto the response options remember and know. Hence, there is reason to suspect that the remember-know paradigm may not be suitable for separating the retrieval processes recollection and familiarity.

The dual-process signal-detection model "does not assume that subjects either recollect all of the study information or none of the study information" (Yonelinas et al., 1996, p. 439), but allows for quantitative and qualitative differences with regard to what is derived from $K_D/(1-R_D)$ and $K_T/(1-K_D)$ as estimators of the probability of a familiarity-based false alarm and hit, respectively.
recollected about the study event. However, irrespective of such variation, any retrieved aspect is assumed to lead to a highly confident recognition response purely on the basis of recollection. In contrast, Wixted (2007; Wixted & Stretch, 2004) argued that given such differences in the output of recollection, it is more reasonable to assume that this retrieval process can lead to varying degrees of confidence, just like the familiarity process. Further, he proposed that the response to a test item is not based on the graded output of either recollection or familiarity alone, but on the additive combination of the outputs of both retrieval processes. The resultant strength-of-evidence variable is described by a unidimensional signal detection model, where the strength-of-evidence distribution of the targets has a larger mean and variance than the distractor distribution. In a remember-know experiment, the response is assumed to be determined by placing two criteria on the strength-of-evidence axis, a stricter remember-know criterion and a more lenient know-new criterion, implying that remember-know judgments are equivalent to confidence judgments. Importantly, whereas previously the unidimensional signal-detection model of the remember-know paradigm has been contrasted to dual-process models (e.g., Dunn, 2004; Hirshman & Master, 1997), this dual-process signal-detection model (Wixted, 2007; Wixted & Stretch, 2004) reconciles the dual-process assumption with the unidimensional signal-detection model. This implies that, though one can test whether the unidimensional signal-detection model provides a valid description of remember-know data according to the logic that was applied in the model evaluations reported in the present thesis, it is impossible to conclude from a consistency of remember-know data with a unidimensional signal-detection description (cf. Dunn, 2004) whether recognition memory performance has a single basis or whether it is based on two retrieval processes, recollection and familiarity.

9.5 Conclusion

This thesis is concerned with how participants classify their retrieval experience in remember-know-(guess) experiments. Based on the conclusion that remember-know response rates, which have traditionally been used as measures of retrieval experience, are neither suitable as measures of retrieval experience per se nor as two-in-one measures of retrieval experience and a component of recognition memory performance, alternative measures for either purpose were evaluated. The 2HT task-oriented model, which was designed to provide pure measures of retrieval experience per se, proved to be a valid
measurement model. In two experiments, it was used to test the distinctiveness-fluency framework of retrieval experience (Rajaram, 1996). Though these experiments yielded the same result in terms of remember-know response rates, the model-based analysis showed that contrary to the prediction of the distinctiveness-fluency framework, only one manipulation affected the retrieval experience for truly recognised targets. Besides casting doubt on the distinctiveness-fluency framework, this pattern of results highlights "the importance of models in interpreting remember-know experiments" (Macmillan et al., 2005, p. 607).
References


References


References


Appendix

A. Model Equations and Proofs of Identifiability

To prove that a multinomial processing tree model is globally identifiable, one has to show that \( p(\theta) = p(\theta') \) implies \( \theta = \theta' \) for all parameter vectors \( \theta \) and \( \theta' \). If there are as many model parameters as there are functionally independent response category probabilities, this proof can be accomplished by finding the inverse function of the model equations, that is, by expressing each model parameter as a function of the response category probabilities.

The following notation is used: \( R_T \) denotes the probability of a remember response to a target, \( K_T \) denotes the probability of a know response to a target, \( G_T \) denotes the probability of a guess response to a target, and \( N_T \) denotes the probability of responding new to a target. Analogously, \( R_D, K_D, G_D \) and \( N_D \) denote the probabilities of a response of each type to a distractor.

A.1 Task-Oriented Models for the Remember-Know Paradigm

2HT Task-Oriented Model

Model Equations

\[
\begin{align*}
R_T &= D \cdot R + (1 - D) \cdot b \cdot R^* \\
K_T &= D \cdot (1 - R) + (1 - D) \cdot b \cdot (1 - R^*) \\
N_T &= (1 - D) \cdot (1 - b) \\
R_D &= (1 - D) \cdot b \cdot R^* \\
K_D &= (1 - D) \cdot b \cdot (1 - R^*) \\
N_D &= D + (1 - D) \cdot (1 - b)
\end{align*}
\]

Proof of Identifiability

\[
\begin{align*}
D &= N_D - N_T \\
b &= \frac{1 - N_D}{1 - (N_D - N_T)} \\
R^* &= \frac{R_D}{R_D + K_D} \\
R &= \frac{R_T - R_D}{N_D - N_T}
\end{align*}
\]
1HT Task-Oriented Model

Model Equations

\[ R_T = D \cdot R + (1 - D) \cdot b \cdot R^* \]  (A.11)
\[ K_T = D \cdot (1 - R) + (1 - D) \cdot b \cdot (1 - R^*) \]  (A.12)
\[ N_T = (1 - D) \cdot (1 - b) \]  (A.13)
\[ R_D = b \cdot R^* \]  (A.14)
\[ K_D = b \cdot (1 - R^*) \]  (A.15)
\[ N_D = 1 - b \]  (A.16)

Proof of Identifiability

\[ D = \frac{N_D - N_T}{N_D} \]  (A.17)
\[ b = 1 - N_D \]  (A.18)
\[ R^* = \frac{R_D}{R_D + K_D} \]  (A.19)
\[ R = \frac{N_D \cdot R_T - N_T \cdot R_D}{N_D - N_T} \]  (A.20)

A.2 Task-Oriented Models for the Remember-Know-Guess Paradigm

2HT Task-Oriented Model that Allows Guess Responses to Truly Recognised Targets

Model Equations

\[ R_T = D \cdot (1 - g) \cdot R + (1 - D) \cdot b \cdot (1 - g^*) \cdot R^* \]  (A.21)
\[ K_T = D \cdot (1 - g) \cdot (1 - R) + (1 - D) \cdot b \cdot (1 - g^*) \cdot (1 - R^*) \]  (A.22)
\[ G_T = D \cdot g + (1 - D) \cdot b \cdot g^* \]  (A.23)
\[ N_T = (1 - D) \cdot (1 - b) \]  (A.24)
\[ R_D = (1 - D) \cdot b \cdot (1 - g^*) \cdot R^* \]  (A.25)
\[ K_D = (1 - D) \cdot b \cdot (1 - g^*) \cdot (1 - R^*) \]  (A.26)
\[ G_D = (1 - D) \cdot b \cdot g^* \]  (A.27)
\[ N_D = D + (1 - D) \cdot (1 - b) \]  (A.28)

Proof of Identifiability

\[ D = N_D - N_T \]  (A.29)
\[ b = \frac{1 - N_D}{1 - (N_D - N_T)} \]  (A.30)
\[ g^* = \frac{G_D}{R_D + K_D + G_D} \quad \text{(A.31)} \]
\[ R^* = \frac{R_D}{R_D + K_D} \quad \text{(A.32)} \]
\[ g = \frac{G_T - G_D}{N_D - N_T} \quad \text{(A.33)} \]
\[ R = \frac{R_T - R_D}{(N_D - N_T) - (G_T - G_D)} \quad \text{(A.34)} \]

**1HT Task-Oriented Model that Allows Guess Responses to Truly Recognised Targets**

**Model Equations**

\[ R_T = D \cdot (1 - g) \cdot R + (1 - D) \cdot b \cdot (1 - g^*) \cdot R^* \quad \text{(A.35)} \]
\[ K_T = D \cdot (1 - g) \cdot (1 - R) + (1 - D) \cdot b \cdot (1 - g^*) \cdot (1 - R^*) \quad \text{(A.36)} \]
\[ G_T = D \cdot g + (1 - D) \cdot b \cdot g^* \quad \text{(A.37)} \]
\[ N_T = (1 - D) \cdot (1 - b) \quad \text{(A.38)} \]
\[ R_D = b \cdot (1 - g^*) \cdot R^* \quad \text{(A.39)} \]
\[ K_D = b \cdot (1 - g^*) \cdot (1 - R^*) \quad \text{(A.40)} \]
\[ G_D = b \cdot g^* \quad \text{(A.41)} \]
\[ N_D = (1 - D) \cdot (1 - b) \quad \text{(A.42)} \]

**Proof of Identifiability**

\[ D = \frac{N_D - N_T}{N_D} \quad \text{(A.43)} \]
\[ b = 1 - N_D \quad \text{(A.44)} \]
\[ g^* = \frac{G_D}{R_D + K_D + G_D} \quad \text{(A.45)} \]
\[ R^* = \frac{R_D}{R_D + K_D} \quad \text{(A.46)} \]
\[ g = \frac{N_D \cdot G_T - N_T \cdot G_D}{N_D - N_T} \quad \text{(A.47)} \]
\[ R = \frac{N_D \cdot R_T - N_T \cdot R_D}{(N_D - N_T) - (N_D \cdot G_T - N_T \cdot G_D)} \quad \text{(A.48)} \]
**2HT Task-Oriented Model that Precludes Guess Responses to Truly Recognised Targets**

Model Equations

\[
R_T = D \cdot R + (1-D) \cdot b \cdot (1-g^*) \cdot R^* \quad (A.49)
\]

\[
K_T = D \cdot (1-R) + (1-D) \cdot b \cdot (1-g^*) \cdot (1-R^*) \quad (A.50)
\]

\[
G_T = (1-D) \cdot b \cdot g^* \quad (A.51)
\]

\[
N_T = (1-D) \cdot (1-b) \quad (A.52)
\]

\[
R_D = (1-D) \cdot b \cdot (1-g^*) \cdot R^* \quad (A.53)
\]

\[
K_D = (1-D) \cdot b \cdot (1-g^*) \cdot (1-R^*) \quad (A.54)
\]

\[
G_D = (1-D) \cdot b \cdot g^* \quad (A.55)
\]

\[
N_D = D + (1-D) \cdot (1-b) \quad (A.56)
\]

**Proof of Identifiability**

\[ p(\theta) = p(\theta') \text{ and Equations } A.52 \text{ and } A.56 \text{ lead to } \]

\[
(1-D) \cdot (1-b) = (1-D') \cdot (1-b') \quad (A.57)
\]

and

\[
D + (1-D) \cdot (1-b) = D' + (1-D') \cdot (1-b') \quad (A.58)
\]

Subtracting Equation A.57 from Equation A.58 yields \( D = D' \), and inserting this into Equation A.57 yields \( b = b' \). \( p(\theta) = p(\theta') \) and Equations A.49 and A.53 lead to

\[
D \cdot R + (1-D) \cdot b \cdot (1-g^*) \cdot R^* = D' \cdot R' + (1-D') \cdot b' \cdot (1-g'^*) \cdot R'^* \quad (A.59)
\]

\[
(1-D) \cdot b \cdot (1-g^*) \cdot R^* = (1-D') \cdot b' \cdot (1-g'^*) \cdot R'^* \quad (A.60)
\]

Subtracting Equation A.53 from Equation A.49 yields \( D \cdot R = D' \cdot R' \), and inserting \( D = D' \) yields \( R = R' \). \( p(\theta) = p(\theta') \) and Equation A.51 lead to

\[
(1-D) \cdot b \cdot g^* = (1-D') \cdot b' \cdot g'^* \quad (A.61)
\]

and inserting \( D = D' \) and \( b = b' \) yields \( g^* = g'^* \). Finally, \( p(\theta) = p(\theta') \) and Equation A.53 lead to

\[
(1-D) \cdot b \cdot (1-g^*) \cdot R^* = (1-D') \cdot b' \cdot (1-g'^*) \cdot R'^* \quad (A.62)
\]

and inserting \( D = D', b = b', \text{ and } g^* = g'^* \) yields \( R^* = R'^* \). Thus, \( p(\theta) = p(\theta') \) implies \( \theta = \theta' \) for all parameter vectors \( \theta, \theta' \) and the 2HT task-oriented model that precludes guess responses to truly recognised targets is globally identifiable.
1HT Task-Oriented Model that Precludes Guess Responses to Truly Recognised Targets

Model Equations

\[ R_T = D \cdot R + (1-D) \cdot b \cdot (1 - g^*) \cdot R^* \]  \hspace{1cm} (A.63)
\[ K_T = D \cdot (1-R) + (1-D) \cdot b \cdot (1 - g^*) \cdot (1-R^*) \]  \hspace{1cm} (A.64)
\[ G_T = (1-D) \cdot b \cdot g^* \]  \hspace{1cm} (A.65)
\[ N_T = (1-D) \cdot (1-b) \]  \hspace{1cm} (A.66)
\[ R_D = b \cdot (1 - g^*) \cdot R^* \]  \hspace{1cm} (A.67)
\[ K_D = b \cdot (1 - g^*) \cdot (1 - R^*) \]  \hspace{1cm} (A.68)
\[ G_D = b \cdot g^* \]  \hspace{1cm} (A.69)
\[ N_D = 1 - b \]  \hspace{1cm} (A.70)

Proof of Identifiability

\( p(\theta) = p(\theta') \) and Equation A.70 lead to \( 1 - b = 1 - b' \), which implies \( b = b' \). Next, \( p(\theta) = p(\theta') \) and Equation A.69 lead to \( b \cdot g^* = b' \cdot g^{*'} \) and inserting \( b = b' \) yields \( g^* = g^{*'} \). Further, \( p(\theta) = p(\theta') \) and Equation A.67 lead to

\[ b \cdot (1 - g^*) \cdot R^* = b' \cdot (1 - g^{*'} ) \cdot R^{*'} , \]  \hspace{1cm} (A.71)

and inserting \( b = b' \) and \( g^* = g^{*'} \) yields \( R^* = R^{*'} \). Furthermore, \( p(\theta) = p(\theta') \) and Equation A.66 lead to \( (1-D) \cdot (1-b) = (1-D') \cdot (1-b') \) and inserting \( b = b' \) yields \( 1-D = 1-D' \), which implies \( D = D' \). Finally, \( p(\theta) = p(\theta') \) and Equation A.63 lead to

\[ D \cdot R + (1-D) \cdot b \cdot (1 - g^*) \cdot R^* = D' \cdot R' + (1-D') \cdot b' \cdot (1 - g^{*'} ) \cdot R^{*'} . \]  \hspace{1cm} (A.72)

Inserting \( D = D' \), \( b = b' \), \( g^* = g^{*'} \), and \( R^* = R^{*'} \) yields \( R = R' \). Thus, \( p(\theta) = p(\theta') \) implies \( \theta = \theta' \) for all parameter vectors \( \theta, \theta' \) and the 1HT task-oriented model that precludes guess responses to truly recognised targets is globally identifiable.
B. Remember-Know-(Guess) Instructions

The remember-know-(guess) instructions used in the present experiments were modelled after Rajaram's (1993) instructions. However, they were changed somewhat for each experiment, hence the complete remember-know-(guess) instruction is reproduced for each experiment separately. In Experiments 1 and 2, different groups of participants were given the response options remember-know versus remember-know-guess for specifying their old responses. Omitting the italicised sections from the remember-know-guess instruction reproduced below results in the remember-know instruction.

B.1 Experiment 1

Bitte geben Sie für jedes Wort, das Sie aus der Lernphase wiedererkennen, zusätzlich an,

• ob Sie sich an das frühere Auftreten des Wortes "erinnern" oder
• ob Sie "wissen", dass es in der Lernphase vorkam.

Was mit "Erinnern" und "Wissen" gemeint ist, wird auf den folgenden Seiten erklärt.

"Erinnern" bedeutet, dass das Wiedererkennen des Wortes einhergeht mit einer bewussten – möglicherweise lebendigen – Erinnerung an sein Auftreten in der Lernphase. "Erinnern" ist also die Fähigkeit, sich Einzelheiten der Episode, als Sie sich das Wort einprägten, wieder zu vergegenwärtigen und diese Lernepisode gleichsam "wiederzuerleben". Wenn Ihnen beim Wiedererkennen eines Wortes zum Beispiel eine Assoziation, eine Vorstellung oder eine Empfindung wieder in den Sinn kommt, die Sie bei der früheren Darbietung des Wortes hatten, oder beispielsweise etwas, das währenddessen geschah, dann "erinnern" Sie das Wort.

Im Unterschied dazu "wissen" Sie, dass ein Wort in der Lernphase vorkam, wenn Sie es zwar wiederkennen, aber sein früheres Auftreten nicht bewusst erinnern und nicht im Gedächtnis wiederbeleben können. Anders ausgedrückt, wenn Ihnen ein Wort aus der Lernphase vertraut ist, ohne dass damit eine spezifische Erinnerung an die Episode einhergeht, als Sie sich das Wort einprägten, dann "wissen" Sie schlicht, dass das Wort in der Lernphase vorkam.

Es kann Zweifelsfälle geben, in denen es Ihnen schwer fällt zu entscheiden, ob ein Wort in der Lernphase wohl vorkam oder nicht. Wenn Sie in dieser Situation mit "alt" antworten, dann geben Sie bitte an, dass Sie "raten". Raten, dass ein Testwort wohl alt ist, bedeutet also im Unterschied zu "Erinnern" und "Wissen", dass Sie das Wort nicht wirklich wiederkennen, es aber auch nicht wirklich als neu identifizieren können.
Für die nähere Bestimmung Ihrer "alt"-Antworten als "Erinnern", "Wissen" oder "Raten" dienen die farbig markierten Tasten mit den Aufschriften E, W und R.

• Bitte drücken Sie die gelb markierte Taste E für "Erinnern", wenn das Wiedererkennen des Wortes mit einer Erinnerung an die frühere Lernepisode einhergeht.

• Bitte drücken Sie die gelb markierte Taste W für "Wissen", wenn Sie das Wort schlichtwiedererkennen, ohne dass Sie eine Erinnerung an die frühere Lernepisode haben.

• *Bitte drücken Sie die grün markierte Taste R für "Raten", falls Sie sich im Zweifel darüber, ob das Wort in der Lernphase wohl vorkam oder nicht, für eine "alt"-Antwort entschieden haben.*

B.2 Experiment 2

Bitte geben Sie für jedes Wort, das Sie aus der Lernphase wiedererkennen, zusätzlich an,

• ob Sie sich an das frühere Auftreten des Wortes "erinnern" oder
• ob Sie "wissen", dass es in der Lernphase vorkam.

Was mit "Erinnern" und "Wissen" gemeint ist, wird auf den folgenden Seiten erklärt.

"Erinnern" bedeutet, dass Sie sich beim Wiedererkennen des Wortes sein früheres Auftreten in der Lernphase wieder vergegenwärtigen können. Das heißt, Sie können die frühere Situation, als das Wort präsentiert wurde, gleichsam im Gedächtnis wiedererleben, möglicherweise lebhaft. Sie "erinnern" sich also an ein Wort, wenn Ihnen ein oder mehrere Aspekte der früheren Situation, als das Wort auf dem Bildschirm dargeboten wurde, wieder bewusst werden; Zum Beispiel etwas, das Sie zu der Zeit dachten oder taten oder das um Sie herum geschah, oder etwas über das Aussehen des Wortes, oder was unmittelbar vor oder nach dem Wort kam. Wenn Ihnen beispielsweise eine bestimmte Vorstellung, Empfindung, Assoziation oder etwas Anderes, das Ihnen während der vorherigen Darbietung des Wortes durch den Kopf ging, wieder in den Sinn kommt, oder etwas Äußerliches, dann "erinnern" Sie sich an das Wort.

Im Unterschied dazu "wissen" Sie, dass ein Wort in der Lernphase vorkam, wenn Sie es zwar sicher wiederkennen, aber sein früheres Auftreten nicht bewusst erinnern und nicht im Gedächtnis wiederbeleben können. Anders ausgedrückt, wenn Ihnen ein Wort aus der Lernphase vertraut ist, ohne dass damit irgendeine spezifische Erinnerung an die frühere
Situation einhergeht, als das Wort auf dem Bildschirm dargeboten wurde, dann "wissen" Sie schlicht, dass das Wort in der Lernphase vorkam.

Es kann Zweifelsfälle geben, in denen es Ihnen schwer fällt zu entscheiden, ob ein Wort in der Lernphase wohl vorkam oder nicht. Wenn Sie in dieser Situation mit "alt" antworten, dann geben Sie bitte an, dass Sie "raten". "Raten", dass ein Testwort wohl alt ist, bedeutet also im Unterschied zu "Erinnern" und "Wissen", dass Sie das Wort nicht wirklich wiedererkennen, es aber auch nicht wirklich als neu identifizieren können.

Für die nähere Bestimmung Ihrer "alt"-Antworten als "Erinnern", "Wissen" oder "Raten" erscheinen drei Felder auf dem Bildschirm, sobald Sie für ein Wort "alt" angeklickt haben.

- Bitte klicken Sie das Feld "Erinnern" an, wenn das Wiedererkennen des Wortes mit einer bewussten Erinnerung an die frühere Darbietungssituation einhergeht.
- Bitte klicken Sie das Feld "Wissen" an, wenn Sie das Wort schlicht wiedererkennen, ohne dass Sie eine Erinnerung an die frühere Darbietungssituation haben.
- Bitte klicken Sie das Feld "Raten" an, falls Sie sich im Zweifel darüber, ob das Wort in der Lernphase wohl vorkam oder nicht, für eine "alt"-Antwort entschieden haben.

B.3 Experiment 3

Bitte geben Sie für jedes Wort, das Sie aus der Lernphase wiedererkennen, zusätzlich an,

- ob Sie sich an das frühere Auftreten des Wortes "erinnern" oder
- ob Sie "wissen", dass es in der Lernphase vorkam.

Was mit "Erinnern" und "Wissen" gemeint ist, wird auf den folgenden Seiten erklärt.

"Erinnern" bedeutet, dass das Wiedererkennen des Wortes einhergeht mit einer bewussten – möglicherweise lebendigen – Erinnerung an sein Auftreten in der Lernphase. "Erinnern" ist also die Fähigkeit, sich Einzelheiten der Episode, als Sie sich das Wort einpräghten, wieder zu vergegenwärtigen und diese Lernepisode gleichsam "wiederzuerleben". Wenn Ihnen beim Wiedererkennen eines Wortes zum Beispiel eine Assoziation, eine Vorstellung oder eine Empfindung wieder in den Sinn kommt, die Sie bei der früheren Darbietung des Wortes hatten, oder beispielsweise etwas, das währenddessen geschah, dann "erinnern" Sie das Wort.

Im Unterschied dazu "wissen" Sie, dass ein Wort in der Lernphase vorkam, wenn Sie es zwar wiedererkennen, aber sein früheres Auftreten nicht bewusst erinnern und nicht im
Gedächtnis wiederbeleben können. Anders ausgedrückt, wenn Ihnen ein Wort aus der Lernphase vertraut ist, ohne dass damit eine spezifische Erinnerung an die Episode einhergeht, als Sie sich das Wort einprägten, dann "wissen" Sie schlicht, dass das Wort in der Lernphase vorkam.

Für die nähere Bestimmung Ihrer "alt"-Antworten als "Erinnern" oder "Wissen" dienen die gelb markierten Tasten mit den Aufschriften E und W.

- Bitte drücken Sie die gelb markierte Taste E für "Erinnern", wenn das Wiedererkennen des Wortes mit einer Erinnerung an die frühere Lernepisode einhergeht.
- Bitte drücken Sie die gelb markierte Taste W für "Wissen", wenn Sie das Wort schlicht wiedererkennen, ohne dass Sie eine Erinnerung an die frühere Lernepisode haben.

**B.4 Experiment 4**

Bitte geben Sie für jedes Wort, das Sie aus der Lernphase wiederkennen, zusätzlich an,

- ob Sie sich an das frühere Auftreten des Wortes "erinnern" oder
- ob Sie "wissen", dass es in der Lernphase vorkam.

Was mit "Erinnern" und "Wissen" gemeint ist, wird auf den folgenden Seiten erklärt.

"Erinnern“ bedeutet, dass Sie sich beim Wiedererkennen des Wortes sein früheres Auftreten in der Lernphase wieder vergegenwärtigen können. Das heißt, Sie können die frühere Situation, als das Wort präsentiert wurde, gleichsam im Gedächtnis wiedererleben, möglicherweise lebhaft. Sie „erinnern“ sich also an ein Wort, wenn Ihnen ein oder mehrere Aspekte der früheren Situation, als das Wort auf dem Bildschirm dargeboten wurde, wieder bewusst werden; zum Beispiel etwas, das Sie zu der Zeit dachten oder taten oder das um Sie herum geschah, oder etwas über das Aussehen des Wortes, oder was vor oder nach dem Wort kam. Wenn Ihnen beispielsweise eine bestimmte Vorstellung, Empfindung, Assoziation oder etwas Anderes, das Ihnen während der vorherigen Darbietung des Wortes durch den Kopf ging, wieder in den Sinn kommt, oder etwas Äußerliches, dann „erinnern“ Sie sich an das Wort.

Im Unterschied dazu „wissen“ Sie, dass ein Wort in der Lernphase vorkam, wenn Sie es zwar wiederkennen, aber sein früheres Auftreten nicht bewusst erinnern und nicht im Gedächtnis wiederbeleben können. Anders ausgedrückt, wenn Ihnen ein Wort aus der Lernphase vertraut ist, ohne dass damit irgendeine spezifische Erinnerung an die frühere
Situation einhergeht, als das Wort auf dem Bildschirm dargeboten wurde, dann „wissen“ Sie schlicht, dass das Wort in der Lernphase vorkam.


- Bitte klicken Sie das Feld "Erinnern" an, wenn das Wiedererkennen des Wortes mit einer bewussten Erinnerung an die frühere Darbietungssituation einhergeht.
- Bitte klicken Sie das Feld "Wissen" an, wenn Sie das Wort schlicht wiedererkennen, ohne dass Sie eine Erinnerung an die frühere Darbietungssituation haben.
Summary

The remember-know paradigm investigates the subjective experience that accompanies retrieving information from memory. Participants in a recognition memory test are asked to specify their retrieval experience for each item called old as remember if they are able to recollect some aspect of the encoding situation, or as know if the item is simply familiar. The rates of remember and know responses have been used as measures of retrieval experience in two different lines of research, although the respective research goals imply partially conflicting requirements for suitable measures.

If the goal is to investigate the nature of retrieval experience per se (Rajaram, 1996), a pure measure of retrieval experience is required. As remember and know response rates are obtained by partitioning the old response rate and the old response rate in recognition memory tests varies as a function of both old-new response bias and recognition memory performance, remember-know response rates are expected to vary as a function of these factors as well, instead of being selectively sensitive to changes of retrieval experience.

In contrast, if the retrieval experience is used as tool for isolating different hypothetical memory systems that jointly produce recognition memory performance and that are characterised by distinct states of retrieval experience (Tulving, 1985b), it is desirable to use measures that represent these theoretical constructs. As an increase in recognition memory performance can only be achieved by means of greater retrieval success from at least one memory system, measures of retrieval experience that represent memory systems cannot be independent of recognition memory performance, contrary to what is required of pure measures of retrieval experience. Further, because experimental manipulations of old-new response bias do not influence the ability to discriminate targets from distractors, it is reasonable to assume that they also do not affect the ability to retrieve information from either memory system. Hence, measures that represent a memory system are required to be invariant to manipulations of old-new response bias. A review of published evidence shows that remember-know response rates do not meet this criterion, regardless of whether the additional response option guess is available for classifying retrieval experience or not. Thus, they are not suitable for representing the hypothetical memory systems.

Because the remember-know response rates are not suitable for either purpose, this thesis investigated various alternative measures, which were formulated within multinominal processing tree models for the remember-know and remember-know-guess
paradigms. A reanalysis of published data from experiments that manipulated old-new response bias was supplemented with two new experiments, in which the manipulation of percentage of targets in the test was crossed with whether the response option guess was available for classifying the retrieval experience. Together, the (re-)analyses showed that the task-oriented measurement model with two high thresholds for the original remember-know paradigm provides valid pure measures of retrieval experience. These measures are specified as conditional probabilities of remember responses given true recognition versus guessed old responses, respectively. Furthermore, the measurement model encompasses measures of old-new response bias and recognition memory performance.

This measurement model was applied in two further experiments in order to re-investigate the effect of classic manipulations of recognition memory performance on retrieval experience by comparing the traditional measures with the validated pure measures of retrieval experience. Replicating previous studies, divided attention at study and phonemic as compared to semantic processing strongly reduced the remember response rate to targets, which has previously been interpreted as selective effect on recollective experience. However, the model-based analysis arrives at more complex conclusions. As expected, divided attention and phonemic processing greatly reduced recognition memory performance. Importantly, the retrieval experience for truly recognised targets was not affected by dividing attention at study, whereas phonemic processing slightly reduced the probability of remember responses to truly recognised targets. Both variables affected the reports of retrieval experience for items guessed to be old.

In short, because remember-know response rates are not suitable as pure measures of retrieval experience or for representing hypothetical memory systems, this thesis examined alternative measures for either purpose. Indeed, a measurement model that qualified as valid could be demonstrated to imply different conclusions about the effect of experimental manipulations on retrieval experience than the traditional measures. To achieve a better understanding of the remember-know paradigm, future work should attend to the judgmental character of remember-know responses and investigate factors that influence whether a particular retrieval experience is judged as remember or know.
Zusammenfassung


Wenn das Ziel darin besteht, das Wesen der Abruf erfahrung per se zu untersuchen (Rajaram, 1996), ist ein reines Maß der Abruf erfahrung nötig. Da Remember- und Know-Antwortraten durch Aufteilung der Rate der Alt-Antworten entstehen und die Rate der Alt-Antworten in Wiedererkennenstests in Abhängigkeit von der Antworttendenz bei der Alt-Neu-Entscheidung sowie der Wiedererkennensleistung variiert, ist zu erwarten, dass auch die Remember-Know-Antwortraten in Abhängigkeit von diesen Faktoren variieren, anstatt nur auf Veränderungen der Abruf erfahrung zu reagieren.

nicht, wenn zusätzlich die Antwortoption Raten zur Bestimmung der Abrufe erfahrung zur Verfügung steht. Somit sind sie nicht dazu geeignet, die hypothetischen Gedächtnissysteme zu repräsentieren.


Kurz, da die Remember-Know-Antwortraten weder als reine Maße der Abrufe erfahrung geeignet sind noch um die hypothetischen Gedächtnissystem zu repräsentieren, untersuchte
die vorliegende Arbeit alternative Maße für beide Zwecke. Es konnte gezeigt werden, dass ein Messmodell, dessen Validität nachgewiesen worden war, ein anderes Bild der Effekte experimenteller Manipulationen auf die Abrufperformance zeichnete als die traditionellen Maße. Um ein besseres Verständnis des Remember-Know-Paradigmas zu ermöglichen, sollte zukünftig der Urteilscharakter der Remember-Know-Antworten beachtet und Faktoren untersucht werden, die beeinflussen, ob eine gegebene Abrufperformance als Remember oder Know beurteilt wird.
Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften bekannt ist.


Andreas Böse, Nicole Driesel, Stefanie Ganz, Carina Giesen, Rebecca Kosan, Beatrix Mauder, Claudia Pohl, Sebastian Pohlack, Gerhard Reese und Kerstin Weißer haben als studentische Hilfskräfte bei der Durchführung der Experimente geholfen.


Die Arbeit wurde weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Weder früher noch gegenwärtig habe ich an einer anderen Hochschule eine Dissertation eingereicht.

Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Ort, Datum

Unterschrift