Health Professions Education

Volume 4 | Issue 4 Article 4

2018-10-18

A Recognition Study Testing the Psychological Validity and Development of Illness Scripts

Eugène J.F.M. Custers
University Medical Center Utrecht, School of Medical Sciences, Center for Research and Development of Education, HB Building 4.05 P. O. Box 85500, 3508 GA Utrecht, The Netherlands, ecusters@umcutrecht.nl

Recommended Citation

Custers, Eugène J.F.M. (2018) "A Recognition Study Testing the Psychological Validity and Development of Illness Scripts," *Health Professions Education*: Vol. 4: Iss. 4, Article 4.
DOI: 10.1016/j.hpe.2017.12.001
Available at: https://hpe.researchcommons.org/journal/vol4/iss4/4

Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 4 (2018) 259–277

www.elsevier.com/locate/hpe

A Recognition Study Testing the Psychological Validity and Development of Illness Scripts

Eugène J.F.M. Custers

University Medical Center Utrecht, School of Medical Sciences, Center for Research and Development of Education, HB Building 4.05 P. O. Box 85500, 3508 GA Utrecht, The Netherlands

Received 1 December 2017; accepted 2 December 2017
Available online 6 December 2017

Abstract

Purpose: This study investigates whether the recognition memory phenomena previously found for script-based stories also apply to illness scripts, the hypothesized mental structures expert physicians apply in medical diagnosis. In addition, the development of these scripts is investigated.
Method: Second and sixth year students and experienced family physicians participated; the influence of typicality of information (prototypical versus atypical statements), textual presence (verbatim or implicit), and delay (15 min or 1 week) on recognition memory discrimination was investigated in a 3×2×2 ANOVA design and on recognition reaction times (RTs) in a 3×2×2×2 ANOVA design.
Results: The expected developmental differences could not be replicated; all participants appear to dispose of illness script structures, which explains poorer memory discrimination for prototypical than atypical information. The results also show that at a longer delay, medical students and physicians are more inclined to infer unstated, but script-typical information. With regard to the RTs, the interaction between typicality and textual presence on RTs could be replicated: RTs for prototypical unstated items were longer than for any of the other types of information. Apart from this, RTs for different statements did not show a consistent pattern.
Discussion: The superior memory discrimination for script atypical, compared with script prototypical, information, and at immediate retention, compared to delayed retention supports theoretical notions as well as previous research on illness scripts as general event representations with actual case information "tagged" to these stored representations. This tagged information decays over time. In terms of script development, all participants appear to have their knowledge structured in illness scripts, even students who have little experience with the diseases included in the study.

Keywords: Medical expertise; Illness scripts; Diagnosis; Recognition memory; Memory discrimination

1. Introduction

For a long time, it has been known that humans do not literally remember real-world knowledge, such as stories, but recollect them by their gist[1]. The concept of a 'schema' has been used to represent the basic units people use to remember these gists[2–5]. A more specific type of schema, a *script*, has been proposed as the knowledge structure that represents generalized events as a unit[6,7]. In the years following the publication of the Schank and Abelson work[7], several researchers have investigated the representational characteristics and behavioral aspects of

E-mail address: ecusters@umcutrecht.nl

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2017.12.001

scripts[8–15]. The following seven aspects are a common denominator that distinguishes scripts from other knowledge representation formats: They are [1] high-level, precompiled, conceptual knowledge structures, which are [2] stored in long-term memory, which [3] represent general (stereotyped) event sequences, in which [4] the individual events are interconnected by temporal and possibly also causal or hierarchical relationships, that [5] can be activated as integral wholes in appropriate contexts, that [6] contain variables and slots that can be filled with information present in the actual situation, retrieved from memory, or inferred from the context, and that [7] develop as a consequence of routinely performed activities or viewing such activities being performed; in other words, through direct or vicarious experience[6,7,16,17]. Scripts serve some important functions: [1] they provide structured knowledge necessary for understanding behavioral sequences, [2] they enable subjects to smoothly integrate new incoming information with existing knowledge, [3] they guide memory retrieval and enable predictions about future events, [4] they guide actual behavior, and [5] they contain knowledge that can be used to explain – at least superficially – why a specific action or sequence of actions has occurred or might occur.

Scripts are actived in short-term memory when the individual is in the appropriate context or is remembered of such a context. The classic example is the "restaurant script"[11]. When a script is activated, its central aspects will become available in a more or less fixed manner, while less central aspects will have the form of variables or "slots" that might be filled in using actual information in the context (e.g., that an appetizer is served) or by default (e.g., that food can be ordered from a menu). Slot values that can be inferred by default or by actually present information provide the necessary flexibility to scripts. The process of assigning values to variables and filling slots with information from the context, information retrieved from memory, or by default, is called *script instantiation*. Basically, an instantiated script is a composite memory representation, which consists of both generic script knowledge and situation-specific information. This situation-specific information is "tagged" to the representation of the generic script[5,7,9,10,12,18] which enables storage and retrieve of individual events as instantiated scripts in long term memory. However, whereas the generic script will be a stable memory representation, tagged knowledge will decay over time; hence, specific memories of instantiated scripts will increasingly be dominated by knowledge of the generic script. This is why we usually forget detailed knowledge of events that took place a long time ago, unless this knowledge was very salient or disrupted the script (e.g., memory of the occasion when we were forced to leave a restaurant because the room filled with smoke).

1.1. The psychological validity of scripts

A large number of studies has provided evidence for the psychological validity of the script concept[19–24]. Script theory specifically predicts memory performance for [a] different types of information (i.e., typical versus atypical); [b] differential relevance of information (i.e., important versus unimportant); [c] different retrieval tasks (i.e., recall versus recognition)[3,5,10,13,18,21,22]; and [d] different retention intervals (i.e., immediate versus delayed memory test)[13,25,26]. For example, recall studies have shown that memory for very typical information is disturbed by *recall intrusions*, that is, such information is easily "recalled" even if it was not present in the original context[3,5]. The recognition memory equivalent of recall intrusions is *memory discrimination*: the ability to tell apart presented from unpresented information, when tested for recognition memory. Memory discrimination is assessed by two parameters: false alarms (false "recognition" of something that was not there) and misses (failure to recognize what in fact was there). It has been demonstrated that memory discrimination is poorer for script typical information than atypical information, at immediate as well as delayed memory testing[3,9–15,18,21,25,27,28]. In other words, script typical information is both more often falsely recognized and missed than script atypical information on a recognition test.

Apart from memory discrimination scores, reaction times (RTs) or response latencies have also been used to assess recognition memory performance. The assumption is that recognition reaction times can be used to distinguish information processing stages involved in determining recognition responses for particular test items[11,14,15,21]. Examples of such stages are searching the memory trace for a tag and judging the likelihood of presentation by judging script relatedness of the information presented on the test. The RT is seen as a cumulative measure of the time it takes for information to be processed and judged for presence by passing through these stages. For example, if recognition memory is probed by script atypical information that was not present, the individual's response will usually be a correct rejection and RT will be short, because there is neither a memory trace nor an expectation that such trace will be present. However, if the probe consists of script typical, but unstated, information, the individual will experience a conflict in deciding whether this information was actually presented, or merely inferred. Resolving this conflict will take (some) time, and the probability of an incorrect response will increase. Findings by Nakamura and Graesser[27] and Yekovich and Walker[15] confirm these hypotheses. For example, Yekovich and Walker[15] found hit rate RTs of 889 and 1093 msec for presented peripheral (atypical) and central (typical) script probes, respectively, while correct rejection times for unmentioned central script information were about twice as long (2032 msec). However, the finding of short RTs for false alarms to unstated typical probes (1124 msec) suggests that such errors are a consequence of a quick, but incorrect, decision that a memory tag is in fact present.

1.2. Illness scripts

The script concept has been applied to the medical domain, in particular to clinical diagnosis. The analogy between “real life” scripts and “illness” scripts is not hard to see: Most diseases can easily be viewed as a generic sequence of events occurring in a patient. Any individual patient is the equivalent of an instantiated illness script, with both prototypical (central) or atypical (peripheral) features, which appear in a certain order and are interconnected in a doctor's memory representation of a disease[29].

The term “illness script” was coined by William Clancey[30] only a few years after the general script concept emerged in the psychological literature. The illness script concept appeared to nicely resolve the tension between two existing approaches to medical diagnosis: one which emphasizes that diagnosis is basically a reasoning process (i.e., the use of biomedical knowledge to explain complaints, symptoms, and other findings in a patient), and one which conceives of diagnosis as a quick categorization process by which patterns of complaints and symptoms are directly mapped to diagnostic categories. Feltovich and Barrows[31] further elaborated the illness script concept, by distinguishing between Enabling Conditions (factors that influence the probability that someone gets a disease, such as age, sex, occupation, and risk behavior), the Fault (the underlying pathophysiological process), and the Consequences (the complaints, signs, and symptoms the Fault gives rise to). With repeated experience, a practitioner's disease knowledge will rather quickly develop into illness scripts, i.e., precompiled knowledge structures that can be activated without explicit reasoning about the underlying Fault[32]. In fact, studies outside and in the medical domain suggest illness scripts can even be learned on an experience-only basis, without the underlying explanatory knowledge[33–35].

There is already some evidence that the memory phenomena discussed earlier in the context of everyday scripts also apply to illness scripts. For example, Arkes and Harkness[36] found that symptoms consistent with the diagnosis, but not explicitly mentioned, are often falsely recognized by diagnosticians. Hassebrock and Prietula[37] observed that experienced pediatricians who were probed for memory of patients they had seen many years before, mostly relied on general inferences from (remembered) pathology, or case features that deviated from the script but had been critical for diagnosis or treatment at the time. Thus, these physicians remembered tagged atypical knowledge, but used the general script knowledge to fill in the memory gaps. Third, Custers and colleagues[38] found a consistent relationship between typicality of a case and average processing (reading) time of case statements: Information that was prototypical for a particular disease was processed faster than atypical information, which supports the assumption that information that easily fits into the script slots can be processed faster than atypical information.

1.3. Illness scripts and the development of medical expertise

Expertise in many real world domains can be accounted for largely in terms of the development of a large repertory of schemas or scripts[39–42]. In the field of medical diagnosis, this amounts to experts having a large repertory of illness scripts. Though novices – medical students – may basically possess the relevant knowledge, it is not yet properly structured and not tuned toward use in practical situations. Experts are supposed to benefit from the integration and coherence script structures provide, in accessing knowledge as well as in processing and recalling it[38,43–46]. In contrast, disturbing the underlying organizational structure, for example, by presenting information in a scrambled order, affects experts much more than novices[44,47,48]. Similarly, though illness scripts enable experts to quickly make likely inferences of typical information on the basis of what they see or read, this comes with a cost: After a delay, experts are expected to show poorer memory discrimination for typical information than non-experts[49].

1.4. The present study

The present study has two broad aims: First, to further corroborate the psychological validity of the illness script concept by comparing memory performances on an illness script recognition test with corresponding performances found in previous studies on classic scripts; and second, to investigate performance differences between medical students and experienced physicians. Both issues are studied by a recognition memory performance task, similar to the ones used by Smith and Graesser[13] and Walker and Yekovich[14]. A recognition memory study enables us to study memory discrimination as well as RTs; in addition, recognition scores are considered more sensitive memory parameters than recall scores[22]. The experimental paradigm consists of a learning phase, in which short cases are presented in the form of statements about a patient with a particular disease, and a recognition memory phase, in which participants' memory is probed for recognition of the information in the cases. The nature of the information (prototypical, atypical, inconsistent) and its presence in the cases (stated or unstated) are manipulated. As the task is presented in a computerized form, RTs as well as responses can be recorded. Like in a standard Signal Detection Theory experiment, four types of responses can be distinguished: hits, misses, correct rejections, and false alarms. As informing the participants in advance about the aim of the study might disturb a memory discrimination effect, only incidental memory of the presented materials will be tested.

To investigate the effect of experience, participants at three levels of medical expertise are included in the study: second year students, sixth year students, and experienced family physicians. Experience was also investigated at a more individual level, by asking sixth year students and family physicians to fill in a form in which they indicated, for each of the diseases that were used in the study, how many patients they had seen with this disease.

The study includes an immediate as well as a delayed test. We predict that the difference in memory discrimination between typical and atypical actions and concepts decreases with increasing time interval[27], because the scripted knowledge increasingly dominates recognition memory performance after a delay, to the detriment of memory for actually presented information.

To make the study as sensitive as possible to developmental differences, participants are probed for *verbatim* recognition memory. A verbatim memory task can be used to independently study the influence of the meaning and of the form in which the information is presented[50]. The assumption is that experienced participants, due to more automatic processing, cannot completely "shut out" the meaning of a statement, even if they are asked to judge its "verbatimness" and disregard its meaning. Novices, on the other hand, will not be distracted by meaning and hence have better memory for surface features (i.c., the verbatim form). In addition, a verbatim memory task enables us to distinguish between two types of inferences: paraphrases and script-based inferences. A *paraphrase* is a statement identical in meaning to a statement in the case, but expressed in different wording (e.g., "high serum cholesterol level" versus "hypercholesterolemia"). A *script-based inference* is a statement that is prototypical for the disease in question, but that is not mentioned in the case description. For example, if the diagnosis is "bacterial pneumonia," it might be inferred by default that the patient has fever, even though no information about this is provided in the case description. If a participant falsely recognizes a paraphrase, this suggests that the meaning of a statement has been stored, but not its verbatim appearance: the memory of the exact wording is lost. If a participant falsely recognizes a script-based inference, apparently inferred knowledge is "recognized" as if it had been actually presented. Experienced participants are expected to show high false alarm rates for paraphrases, in particular. False alarm rates for both paraphrases and script based inferences are expected to be ralatively low at short intervals and to increase over time, because exact memory representation of patient information will fade with the passage of time and the general illness script will increasinlgy dominate this representation.

In sum, the following predictions were tested in the experiment:

With respect to recognition memory discrimination:

[1] In general, memory discrimination will be better for atypical items than for prototypical items. Thus, a main effect of typicality will be found; [2] Memory discrimination will not only be affected by typicality, but by expertise level as well. More specifically, an interaction between typicality and expertise level will be predicted: experts will show relatively poor memory discrimination for typical information compared to atypical information, while this difference will be less pronounced or absent in novices. This is a consequence of relative novices lacking fully developed illness scripts; [3] Immediate vs Delayed testing: After a delay of 1 week, memory for surface aspects of the cases will have faded, and processing will rely more on the general script[26]. In all likelihood, memory discrimination for prototypical information will have decreased more than for atypical information. More specifically, paraphrases and script-based inferences will show relatively high false alarm rates.

With respect to recognition memory RTs:

[4] A positive relationship between expertise level and response speed will be expected: more experienced subjects will generally take less time to respond than less experienced subjects. These results would corroborate the

results of a study by Custers and colleagues[38], who found that expert physicians process case information faster than non-expert physicians or advanced students; [5] We expect to find support for the results of Nakamura and Graesser[27] and Yekovich and Walker[15]: In general, RTs for atypical items will be shorter than those for prototypical items; [6] An interaction will be found between textual presence and typicality: Especially unstated prototypical items will show long reaction times, due to large response latencies for correct rejections. Inconsistent information, which is always unstated in the case, will show very short reaction times: subjects will be able to reject these statements immediately and correctly, because no memory trace for these items has been formed and they will not be inferred by default either; [7] For reaction times, the predicted interaction between typicality and textual presence will also be moderated by expertise level: due to their more cohesive illness scripts, it will be more pronounced for experienced participants than for inexperienced participants. Thus, a three-way interaction between expertise level, typicality, and textual presence will be expected, with a stronger interaction effect for the experts than for less experienced subjects; [8] Specific reaction time results of the present study will be compared with the corresponding ones of Yekovich and Walker[15]. These authors found that atypical hits were associated with fast reaction times, whereas correct rejections of typical items showed the largest response latencies. Reaction times for prototypical hits and false alarms fell somewhere in-between; [9] Immediate versus delayed testing: RTs for prototypical information will be script based and faster than for atypical information, in particular at delayed testing. As the atypical information in this study was not highly salient, RTs are predicted to increase over time.

2. Method

2.1. Participants

Participants were second-year students, sixth-year students, and family physicians. The second-year students, who were in the last months of their second year, had virtually no clinical experience. The sixth-year students had almost completed their medical course, including their clerkships. The experienced physicians were recruited from teachers who were employed at the Institute for the Education of Family Physicians at UMC Utrecht.

2.2. Material

From the set of 24 diseases used in the study by Custers et al.[51], nine were selected to be included in the present study. Diseases were selected to cover as broadly as possible a range of seriousness, prevalence, and body region involved, with sufficient variation in Enabling Conditions (patient background factors) and Consequences (complaints, signs, and symptoms). The final selection consisted of the following diseases: aneurysm of the aortic artery, herpes zoster, nervous abdominal pain, dermatitis peri-oralis, pre-infarct syndrome, vaginal candidiosis, epidural hematoma, kidney stones colic, and carcinoma of the head of the pancreas. For each disease, a short case description of 15 to 24 statements was constructed. The statements provided information about the patient's context and background, the setting (e.g., office hour, emergency telephone call, house call), the main complaints, and some symptoms. Though most of the information in each case was of a typical, "textbook" nature, also some more atypical patient features and symptoms were included. Appendix A shows an example of a case description.

For each case, a set of twelve test statements was constructed. Five of these statements were "stated," i.e., verbatim copies of statements that appeared in the case description. The remaining seven statements were "unstated," that is, they differed, *at least in wording*, substantially from any statement that appeared in the case. Their functional role was that of distractors or foils. Orthogonal to this "textual presence" dimension, test statements also varied on a "typicality" dimension: they could refer to either prototypical or atypical case information. Prototypical and atypical case information was selected from protocols gathered in a previous study[51]. In general, atypical information concerned feature less frequently mentioned by physicians when describing a prototypical patient with a particular disease. The unstated prototypical items were further divided in a paraphrase (a feature that was present in the case but expressed in a different wording), and a script-based inference (a feature that was not mentioned in the case but typical for the disease). Finally, two test items contained information that was inconsistent with the case; these items were always unstated. Inconsistent statements were contrasted with atypical items, which were less likely or less prominent for the disease in question, but not excluded or contradictory. Table 1 shows the distribution of the test items over the different categories. Appendix B shows the twelve test statements for the case example provided in Appendix A.

The first of the nine cases, "rupturing aortic aneurysm," was used for practice purposes only. For the remaining eight experimental cases, two different sequences were constructed to control for possible order effects. These sequences were presented alternately to successive

Table 1
Example of a classic script story (left hand panel) and an illness script story (right hand panel).

Jack decided to take his girlfriend, Chris, out for a nice dinner. He called a friend of his who recommended a good restaurant. He then went out to his car and drove over to Chris' house to pick her up. They drove to the restaurant and he stopped out in front. They got out of the car and let the valet park it. Then they walked inside and Jack confirmed their reservations with the hostess. They sat for a few minutes in the waiting area, and then the hostess escorted them to their table. After the waitress had introduced herself, they ordered cocktails. They browsed the evening paper for a while, and then looked at their menus. After they decided what they wanted to eat, they ordered dinner: Jack opted for a steak with pepper sauce, while Chris chose a lobster dish. As their food was being served, they placed their napkins in their laps. They enjoyed their meal, and had another drink. After dessert, Jack wanted to smoke a cigar, but the waitress warned him that they were sitting in the non-smoking area. When they were finished, Jack paid the bill, and left a generous tip. Then they walked out of the restaurant, got their car, and drove home. (from Graesser, Gordon, and Sawyer, 1979, p. 330–331)	You are just about to have dinner, when you receive a telephone call from Mrs. Jones. She wants you to come immediately, because her husband "is having it again": he is rolling across the room because of the pain, and has vomited several times. Mr. Jones is a 47-year old store-keeper, who is married and has three teenage children. At age 30, he was treated for bronchitis. Six years ago, he had his leg broken as a consequence of a car accident. Four years ago, he was treated with medicaments for kidney stones. His older brother is known with coronary disease, while his father died at age 58 from a CVA. His 52-year old sister has diabetes mellitus. When you arrive at the Jones' home, Mr.Jones is sitting on the sofa, smoking a cigarette and recovering a bit; the pain has just subsided. He complains about having had a convulsive abdominal pain at the left side, just abreast of the navel; the pain extended to his groins. The pain emerged all of a sudden, and then gradually subsided; during the attack, he almost couldn't stand it. Earlier that day, he had remarked that his urine showed a red hue, but he had not paid much attention to it, because he had no pain at that time. His wife, who has measured his temperature, says he has a 38.2° Centigrade fever.

participants. All cases consisted of a number of text lines that were presented successively on a computer screen; the whole sequence of nine cases was presented in one block. The nine sets of twelve test statements together also were presented as a block. The whole task was written in the program “Reaction Time” and implemented on a Macintosh laptop computer. The ‘z’ and ‘?’ keys on the Qwerty-keyboard were programmed to register responses (keypresses and reaction times).

2.3. Procedure

Subjects were tested individually, the students at the UMC Utrecht department, the family physicians in a quiet room at the General Practioners' Education department. The experiment comprised two sessions; the first session consisted of a study phase, an interim task and a test phase; the second session, which took place after a delay of 1 week, consisted of only the test phase.

2.3.1. Study phase

After a short general introduction, the study task was started. The nine case descriptions were presented successively. Before each case was started, the name of the disease of that case was displayed on the screen, in order to activate the appropriate illness script. Next, the statements successively appeared on the screen. Presentation duration of every statement was determined by the formula: [$t = 1500$ ms + 35 additional msecs for every text character in the statement]. This time was based on reading times results in a previous study[38]. This display time was sufficient for all participants to read and comprehend the content of the statement, but not long enough to memorize it thoroughly.

Participants were instructed to read each case as attentively as possible and to try to assimilate as much of the presented information as they could. Though they were informed a test would be administered after the presentation of the cases, the nature of this test was not revealed in advance. All nine cases (one practice case and 8 experimental cases) were presented successively as a block, with a short pause in-between cases. Participants were not allowed to make notes.

2.3.2. Interim task

After finishing the study phase, a short (2–3 min) interim task was administered. Sixth year students and family physicians filled in a “patient frequency form” (see Appendix D) on which they had to indicate, for each disease, how many patients they had seen with this disease in their career. Second year students were asked to tell a couple of minutes about medical journals they were familiar with. Apart from this, the primary purpose of the interim task was to clear participants' short-term memory for the study task.

2.3.3. Test phase

Next, the test task was administered. For each case, subjects were shown the twelve test statements, one by one. The set of test statements associated with each individual case was always presented as a block; prior to each block of test statements, the name of the corresponding disease appeared on the screen, in order to enable subjects to re-instantiate the appropriate illness script. The blocks of test statements appeared in the same order as the cases in the study phase. The order of the individual test statements within each block was randomly determined in advance, but remained fixed across all presentations.

Participants were instructed to decide as accurately and as quickly as possible for each individual test statement whether or not it had been presented *verbatim* in the original case presentation. If "yes," they had to press the corresponding key on the keyboard, if "no," they had to press the alternative key. The instruction emphasized that test statements had been either been presented verbatim or differed *considerably* from any statement they had seen in the study phase, *at least in wording*, to avoid participants' focussing on minor surface changes. However, it was also stressed that a particular test statement could be very similar in *meaning* to an item in the associated case, but that they should ignore this, as they only had to judge the literal (verbatim) presence of the test statements.

After each block of twelve test statements corresponding to one case, participants could take a short break if they wished. Unlike the study phase, the speed of presentation in the test phase was participant-paced; test items remained visible until a response was given. Every time a key was pressed, this was recorded, along with the reaction time in milliseconds. After they had finished this task, participants were debriefed.

2.3.4. Delayed test

After a delay of 1 week (6–8 days), participants returned to complete the test task for the second time. The procedure was exactly the same as during the first session. Finally, participants received a token reward for their participation in the study.

2.4. Analysis

Data were collapsed over all eight experimental cases for each participant and experimental condition. Type of test statement was the unit of analysis; for each participant, this procedure yielded 12 recognition scores (12 types of test statements) on a 9-point scale (as there were 8 cases), ranging from 0, i.e., eight "no"-responses to the corresponding test statements, to 8, i.e., eight "yes"-responses. "Yes" responses could be either hits or false alarms, and "no" responses could be either misses or correct rejections. As the 12 test statements for each case were distributed over 5 different types, five recognition scores for every participant were calculated and expressed as proportions "yes" answers for each of the 5 types. This procedure was performed twice, once for the first session (immediate test), and once for the second session (delayed test).

Table 2
Organization of the twelve test statements for each case.

	Prototypical	Atypical	Inconsistent
stated	1 Enabling Condition 2 Consequences	1 Enabling Condition 1 Consequence	
unstated	1 Enabling Condition	1 Enabling Condition	1 Enabling Condition
	2 Consequences (1 paraphrase 1 script-based inference)	1 Consequence	1 Consequence

As the recognition responses can be divided into four categories, i.e., "hits," "false alarms," "misses," and "correct rejections," the data fit the requirements of a signal detection theory (SDT) analysis of recognition memory performance[52]. Thus, it would be possible to calculate d' values as a measure of memory discrimination – participants' ability to discriminate between stated and unstated items of a particular type. However, d' cannot be calculated if either the hit rate for a particular type of item equals 1 or the false alarm rate equals 0. As this situation occurred in the present experiment, we had to revert to A', a nonparametric analog of d'.[53–55] The formulas for calculating A' can be found in Appendix C.

The second dependent variable in this study are reaction times (RTs) or response latencies. Mean RTs of each participant for every statement type were computed. As a previous study[56] revealed no relationship between the length of test statements and corresponding RTs ($r = -.10$), reaction times did not need to be corrected for statement length. For every participant, five mean reaction times were computed for each type of statement and for each of the two sessions.

The A' values were analyzed in a 3×2×2 ANOVA with expertise level as between subjects factor and typicality of statement and time of testing as within subjects factors. As the inconsistent statements were always unmentioned (the empty cell in Table 2), these statements were compared in a separate analysis with the other two types of unstated items (prototypical and atypical); hence, an additional 3×2×3 ANOVA was performed on the false alarm rates, with expertise level as between subjects factor and typicality (three levels: prototypical, atypical and inconsistent) and time of testing as within subjects factors.

The RTs were analyzed in a 3×2×2×2 ANOVA, with expertise level as between subjects factor and time of test, typicality, and textual presence as within subjects factors. Finally, to contrast the effects of the two types of unstated prototypical items, i.c., paraphrases and script-based inferences, two 3×2×2 ANOVAs with expertise level as between subjects factor and paraphrase versus script-based inference and time of testing as within subjects factors were computed, one for false alarm rates and one for RTs.

As the average RTs for a particular item type are a composite of positive recognitions ("yes"-answers) and failures to recognize an item ("no"-answers), they reflect quite different underlying processes. Consequently, additional analyses were performed, which focused on the relationship between RT and type of response. The aim of this analysis was to investigate whether our findings were consistent with the findings of Yekovich and Walker[15]. The RT data of the four relevant item-response combinations (hits, false alarms, and correct rejections for prototypical items, and hits for atypical items) were analyzed in a 3×4×2 ANOVA with expertise level as between subjects factor and response type and time of testing as within subjects factors.

As not all of our data satisfied the distributional criteria for analysis of variance, preliminary analyses were performed after transforming the data into their normal logarithmic values. As analyses of these transformed data yielded essentially the same outcomes as ANOVAs of the original data, we will report only the latter results because the actual values are easier to interpret.

3. Results

Participants were 24 s-year students, 23 sixth-year students, and 8 family physicians. The second-year students, who participated in the last months of their second year, had virtually no clinical experience. The sixth-year students had almost completed their medical course, including their clinical duties. The experienced physicians were recruited from teachers who were employed at the Institute for the Education of Family Physicians at UMC Utrecht. One of these physicians had 1.5 years experience, the other seven had experience in the range from 11 to 25 years (exact quantitative estimates are hard to provide because some of them had worked on a part-time basis for at least some of the time).

Due to an unknown equipment failure, the program failed to present the test items for 4 of the 8 cases for three participants in one of the sessions; hence, the results for these individuals in the corresponding sessions are based on 4 rather than 8 test items of each type. In addition, RTs < 500 ms and $> 10{,}000$ ms were considered outliers and removed from the analysis. This added up to 26 recordings < 500 ms and 45 recordings $> 10{,}000$ ms being removed (0.7% of 10,224 entries).

In addition, the data of one second-year student in the delayed testing condition were also discarded, because this participant showed, for atypical and script-inconsistent information, a hit rate of 0.0 and a false alarm rate of 1.0 (the most likely explanation would be that this participant consistently confused the 'yes' and 'no' buttons during delayed testing).

3.1. Recognition memory scores

3.1.1. Effects of expertise level and typicality on recognition memory discrimination

Table 3 shows the memory discrimination scores (A' values) for prototypical and atypcial test statemens, at immediate and delayed testing, for the three expertise levels. The 3×2×2 ANOVA showed no significant main effect of expertise level; however, as predicted, a significant main effect of typicality, $F(1, 48) = 17.241$, $p < .001$, $MS_e = .065$, and a significant main effect of time of testing $F(1, 48) = 19.594$, $p < .001$, $MS_e = 0.324$ were found. None of the interactions was significant. From Table 3, it can be read that memory discrimination was consistently better for atypical than for prototypical items, and at immediate testing, compared with delayed testing. The effect sizes for immediate versus delayed testing, expressed as Cohen's d, were 1.15 (prototypical statements) and 1.53 (atypical statements), respectively, and for prototypical versus atypical statements 0.54 (immediate testing) and 0.52 (delayed testing), respectively.

The memory discrimination data of the sixth year students were separately analyzed on basis of their actual experience with the cases included in this study. To achieve this, two groups of cases were constructed: cases for which the student in question had no or virtually no experience (at maximum one patient seen with the disease, "low experience cases"), and cases for which he/she had seen "many" patients, i.e., five or more ("high experience cases"). The assumption was that one case would be too few to have formed an illness script on basis of practical experience, whereas 5 or more cases would be sufficient. In total, the sixth year students had seen zero or one patient for 68 (immediate test) to 71 (delayed test) case-by-participant combinations and five or more patients for 47 case-by-participant combinations (immediate as well as delayed test). The results of this analysis of specific experience are almost identical, and show the same pattern as the overall results: better memory discrimination for atypical than for prototypical cases and for immediate, as opposed to delayed, testing (Table 3).

3.1.2. *The effects of expertise level and typicality on false alarm scores for inconsistent information*

Atypcial symptoms or characteristics of a disease may occur relatively infrequently; yet, they must at least be more credible than inconsistent information. To investigate this, false alarm rates for atypical information were contrasted with those for inconsistent information. From Table 4, it can be read that false alarm rates were low for both types of items at immediate test (in the range of 6%-12%), but higher at delayed test (in the range of 19%-28%). Analysis of variance indeed showed a main effect of time of test, $F(1, 48) = 55.972$, $p < .001$, $MS_e =$ 0.809, but no other significant main effect, nor any significant interaction. Thus, atypical and inconsistent items show nonzero, approximately equal false recognition rates, which is both surprising and in contrast with expectations.

3.1.3. *The effect of expertise level on recognition scores for paraphrases and script-based inferences*

Subsequently, recognition data (false alarms) for two types of unstated prototypical items, i.c., paraphrases and script-based inferences, were compared. Fig. 1 shows the results. A 3×2×2 analysis of variance with expertise level as between subjects factor and time of test (immediate or delayed) and type of item (paraphrase or script-based inference) as within subjects factors revealed significant main effects of expertise level, $F(2, 48) = 3.840$, $p < .05$, $MS_e = 0.307$ and of time of test $F(1, 48) = 52.213$, $p < .001$, $MS_e =$ 2.020 and a significant interaction between type of item and time of test $F(1, 48) = 16.090$, $p < .001$, $MS_e =$ 0.369. No significant main effect of type of item was found, and none of the other interactions were significant. Fig. 1 shows that the expertise effect can be accounted for by the results of the second year students, who falsely recognized both types of items more often than either the sixth year students or family physicians. As expected, recognition memory performance was considerably lower at delayed than at immediate testing. The significant interaction between time of test and type of item means participants are more inclined to falsely recognize paraphrases at immediate testing, but script-based inferences at delayed testing. As the third-order interaction between expertise level, time of test, and type of item was not significant, the increase in false recognition responses to script-based inferences over time appears to generalize to all three expertise levels (see Fig. 1).

3.1.4. *The effect of delay on hit rates for prototypical and atypical information*

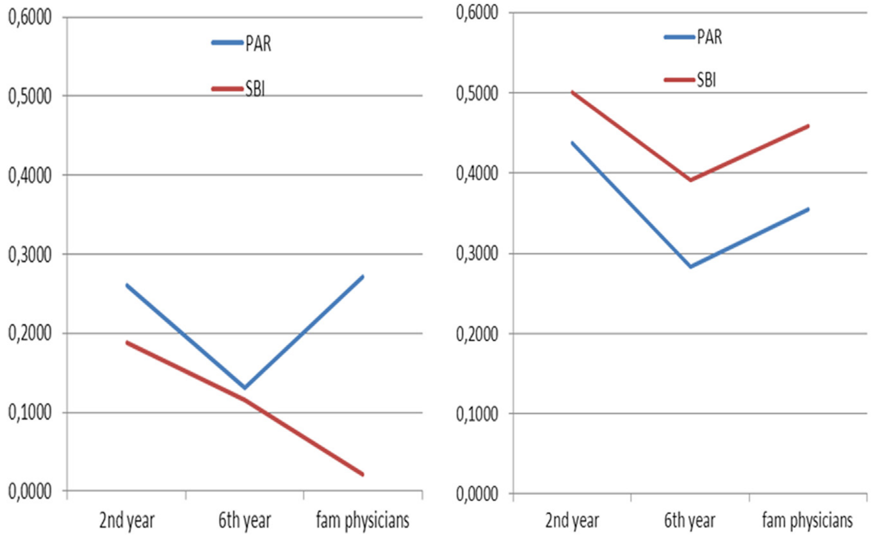
As memory for information in the cases is predicted to decrease over time, hit rates will decrease as well, even more for atypical information, which will be easily forgotten, than for prototypical information, the memory of which will be supported by the script-based inference.

Table 3
Memory discrimination (A', ± SD) as a function of expertise level, textual presence, and time of testing (PT = protoyipical, AT = atypical, IMM = immediate test, DEL = delayed test).

level of expertise:	PT, IMM	PT, DEL	AT, IMM	AT, DEL
2nd-year students (N=23)	.8765 (±0.07)	.7572 (±0.09)	.9055 (±0.06)	.8077 (±0.09)
6th-year students (N=23)	.8835 (±0.06)	.7841 (±0.15)	.9210 (±0.06)	.8253 (±0.16)
family physicians (N=8)	.8808 (±0.04)	.7736 (±0.11)	.9055 (±0.03)	.8488 (±0.06)
mean	.8801 (±0.06)	.7713 (±0.12)	.9125 (±0.06)	.8205 (±0.12)

Table 4
Memory discrimination (A', ± SD) as a function of experience with specific diseases, 6th-year students (PT = protoyipical, AT = atypical, IMM = immediate test, DEL = delayed test).

6th-year students	PT, IMM	PT, DEL	AT, IMM	AT, DEL
low experience cases	.8820 (±0.07)	.7851 (±0.13)	.9264 (±0.08)	.8449 (±0.12)
high experience cases	.8818 (±0.08)	.8084 (±0.09)	.9335 (±0.09)	.8543 (±0.13)



Proportion false alarms at immediate test (PAR=paraphrase, SBI=script-based inference)

Proportion false alarms at delayed test (PAR=paraphrase, SBI=script-based inference)

Fig. 1. Proportion false alarms on paraphrases and script-based inferences, at immediate test (left hand panel) and delayed test (right hand panel).

A 3×2×2 analysis of variance with expertise level as between subjects factor and time of testing and typicality as within subjects factors revealed no significant main effect of expertise level, but significant main effects of time of testing $F(1, 48) = 4.807$, $p < .05$, $MS_e = 0.049$, and of typicality $F(1, 48) = 10.756$, $p < .01$, $MS_e = 0.174$. Fig. 2 shows the results, collapsed over all participants. Though the hit rate for atypical statements appears to decrease more than for prototypical statemens, the interaction between time of testing and typicality was not significant.

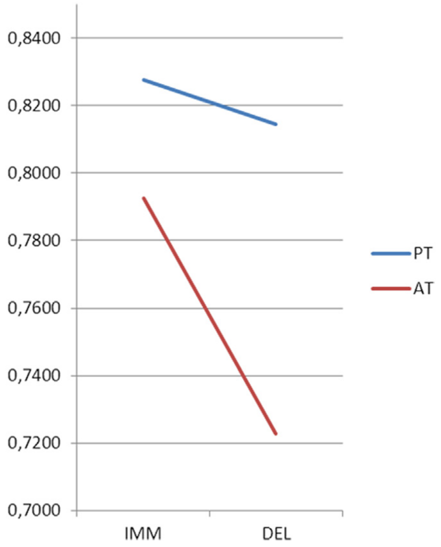


Fig. 2. Hit rates for prototypical (PT) and atypical (AT) information, at immediate (IMM) and delayed (DEL) testing, over all participants (N=51).

3.2. Recognition reaction times (RTs)

3.2.1. Effect of expertise level, typicality, and textual presence on recognition reaction times

A 3×2×2×2 analysis of variance of the RTs for the experimental test statements with expertise level as between subjects factor and time of test, typicality, and textual presence as within subjects factors failed to yield a significant main effect of expertise, $F(2, 53) = 1.696$, $p < .15$, $MS_e = 20{,}328{,}739.70$. Yet, the family physicians in our sample consistently showed longer RTs than sixth year students or second year students, on the average 2875 (±573) ms, 2432 (±589) ms, and 2398 (±638) ms, respectively. There were no significant main effects of time of test, typicality, or textual presence. Significant second-order interactions were found between time of test and typicality, $F(1, 48) = 14.827$, $p < .001$, $MS_e = 768{,}756.564$, and between typicality and textual presence, $F(1, 48) = 23.851$, $p < .001$, $MS_e = 1{,}457{,}253.232$. The interaction between expertise level and textual presence was marginally significant, $F(2, 48) = 3.169$, $p < .051$, $MS_e = 338{,}911.072$. The three-way interaction of expertise level, time of test, and typicality was also significant, $F(2, 46) = 3.626$, $p < .05$, $MS_e = 187{,}986.727$. As this interaction was not predicted and is hard to interpret, we will not discuss it further.

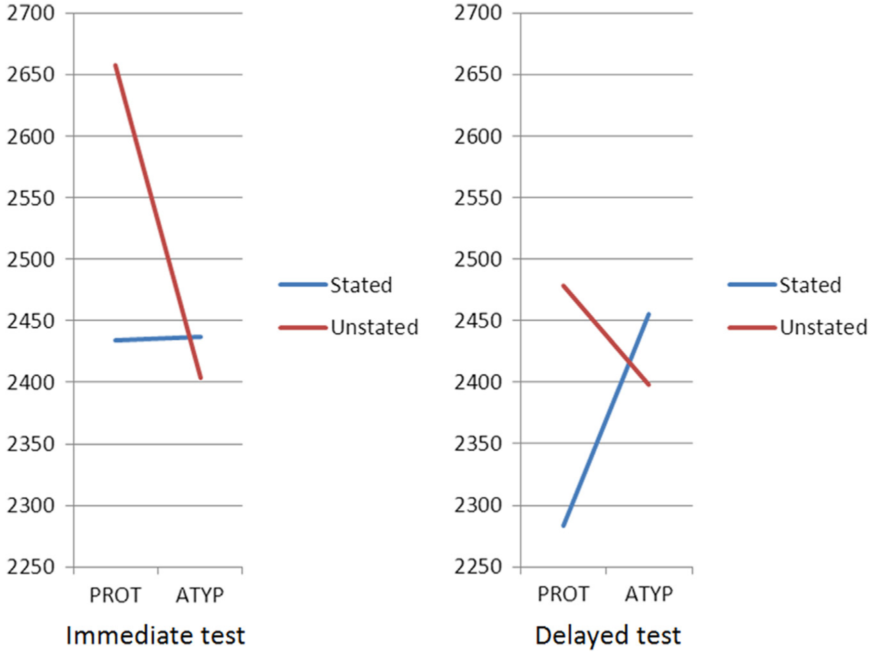


Fig. 3. The influence of interaction of typicality and textual presence on RTs, at immediate (left hand panel) and delayed (right hand panel) test, over all levels of expertise.

Thus, the results do not support the hypothesis that experienced participants would show shorter RTs than less experienced participants. If anything, experienced physicians were slower than students. In addition, given the large SDs of approximately 600 ms, the difference in average RT between 2nd year students and 6th year students (34 ms) can be considered negligible.

Of interest is the significant two-way interaction between typicality and textual presence $F(1, 50) = 27.515$, $p < .001$, $MS_e = 1654{,}684.547$. Over all participants and both times of testing, RTs for prototypical unstated items (2568 ± 97 msec) were longer than for prototypical stated items (2358 ± 89 ms), atypical stated items (2446 ± 100 ms) and atypical unstated items (2401 ± 87 ms). Fig. 3 shows the results, in separate graphs for immediate and delayed testing. At immediate testing, there was a main effect of typicality, $F(1, 52) = 14.775$, $p < .001$, $MS_e = 68{,}833.001$ and a significant typicality by textual presence interaction, $F(1, 52) = 13.539$, $p < .005$, $MS_e = 65{,}175.540$, but no significant effect of textual presence. The results show that RTs for prototypical unstated items (2658 ± 101 ms) were outlyers compared with any of the other three types (RTs in the range of 2283–2477 ms). At delayed testing, there was a main effect of textual presence, $F(1, 52) = 4.729$, $p < .05$, $MS_e = 52{,}137.500$ and a significant typicality by textual presence interaction, $F(1, 52) = 15.192$, $p < .001$, $MS_e = 69{,}359.974$, but no significant effect of typicality. At this time of testing, RTs for prototypical stated items (2288 ± 88 msec) differed from any of the other three types (RTs in the range of 2392–2495 ms).

3.2.2. The effects of expertise level on RTs for unstated and inconsistent information

A 3×2×3 ANOVA with expertise level as between subjects factor and time of test and type of unstated information (three levels: prototypical, atypical, and inconsistent information) as within subjects factor, revealed neither a main effect of expertise level, nor any interaction of expertise level with the other factors. A significant interaction between time of test and typicality was found, however, $F(1, 48) = 28.160$, $p < .001$, $MS_e = 1{,}886{,}532.692$. In line with predictions, at immediate testing, prototypical unstated items showed the longest RT (2793 ± 119 ms). At delayed testing, however, inconsistent items were processed slowest (2845 ± 132 ms). Average RTs for the other types of items at either delay were all within a relatively narrow range, i.e., from 2483–2588 ms. While the finding of long RTs for prototypical items is in line with predictions, the long RTs for inconsistent items were unexpected: participants were predicted to be able to quickly reject these items because of their being inconsistent with the activated script.

3.2.3. *The effect of expertise level on reaction times for paraphrases and script-based inferences*

A 3×2×2 ANOVA with expertise level as between subjects factor and time of testing and paraphrase/script-based inference (two different forms of unstated prototypical information) as within subjects factor revealed no significant differences. The data showed only a tendency for RTs to be shorter at delayed than at immediate testing (2565 ± 149 ms and 2758 ± 127 ms, respectively). Table 5 shows the results.

3.2.4. *Reaction times for different types of recognition responses: comparison of the present results with the Yekovich and Walker[15] data*

According to Yekovich and Walker[15], RTs for hits, correct rejections, and false alarms reflect different underlying memory processes. Thus, we analyzed RTs separately for these responses. Table 6 shows the results.

As the Yekovich and Walker study[15] did not include a delayed test, only the results in the upper panel of Table 6 are relevant for our comparison. They found that correct rejections of prototypical information, in particular, showed much longer RTs than any of the other three types (hits for prototypical and atypical information, and false alarms for prototypical information).

A 3×4 ANOVA of RTs with expertise level as between subjects factor and response type as within subjects factor revealed neither a significant effect of expertise level, nor a significant interaction, but only a main effect of response type, $F(1, 46) = 7.046, p < .05$, $MS_e = 2{,}189{,}819.069$. Analyses of separate contrasts (paired *t*-tests) revealed significant differences between RTs for prototypical hits and atypical hits on the one hand, and RTs for correct rejections on the other. Table 6 shows that, in line with the results of Yekovich and Walker[15], correct rejection responses for (unstated) prototypical statements took significantly more time than

Table 5
False alarm scores for atypcial (AT) and case-inconsistent (INC) items, at immediate (IMM) and delayed (DEL) test.

level of expertise:	AT, IMM	INC, IMM	AT, DEL	INC, DEL
2nd-year students (N=22)	.1193 (±0.10)	.0682 (±0.08)	.2813 (±0.12)	.2330 (±0.12)
6th-year students (N=23)	.0761 (±0.07)	.0815 (±0.09)	.2123 (±0.15)	.1848 (±0.11)
family physicians (N=8)	.0628 (±0.04)	.0627 (±0.08)	.1876 (±0.12)	.2813 (±0.19)
mean	.0932 (±0.08)	.0735 (±0.09)	.2391 (±0.06)	.2169 (±0.13)

Table 6
Average reaction times (expressed in msec) for four types of test item-response combinations, at immediate (upper panel) and delayed (lower panel) testing (HIT, PT = hit, prototypical item; HIT, AT = hit, atypical item, CORR REJ, PT: correct rejection, prototypical item, FA, PT: false alarm, prototypical item).

Immediate test:				
Level of expertise:	HIT, PT	HIT, AT	CORR REJ, PT	FALSE AL, PT
2nd-year students (N=23)	2270 (±654)	2248 (±703)	2771 (±773)	2510 (±743)
6th-year students (N=23)	2342 (±674)	2387 (±685)	2453 (±582)	2707 (±1103)
family physicians (N=7)	2693 (±491)	2729 (±752)	3061 (±746)	2861 (±1060)[a]
mean (N=53)	2357 (±654)	2372 (±711)	2673 (±724)	2655 (±954)
Yekovich and Walker (1986)	1093	889	2032	1124
Delayed test:				
Level of expertise:	HIT, PT	HIT, AT	CORR REJ, PT	FALSE AL, PT
2nd-year students (N=22)	2015 (±617)	2267 (±812)	2402 (±871)	2439 (±942)
6th-year students (N=23)	2220 (±570)	2361 (±634)	2508 (±775)	2542 (±713)
family physicians (N=7)	2815 (±545)	2747 (±740)	2740 (±800)	2977 (±792)
mean (N=52)	2209 (±648)	2371 (±685)	2492 (±829)	2555 (±843)
Yekovich and Walker (1986)	na[b]	na[b]	na[b]	na[b]

[a] Data of one participant were removed because this participant had only one FA,PT entry with a very high value (9800 msec)
[b] Yekovich and Walker[15] did not include a delayed testing condition in their study

hits for either prototypical (t = 4.801, df = 51, p < 0.001) or atypical (t = 4.923, df = 51, p < 0.001) statements. In contrast with the findings by Yekovich and Walker[15], however, in our study RTs for false alarms to prototypical statements were approximately equal to those for correct rejections of these statements. In short, we found that hit-responses (to prototypical as well as atypical statements) are made quicker than either correct rejections or false alarms to prototypical, unstated statements, and that there are no differences in RTs between these latter two types of responses.

4. Discussion

In this study, we investigated the influence of expertise level, typicality, and actual presence of information in an illness script based context on memory performance, which we operationalized as recognition memory discrimination and recognition reaction speed. The two main aims of the study were: First, to find out whether the illness script, as a specific variant of the "classic" script, is a psychologically valid construct, and second, to test some aspects of script development. We will first discuss the consequences of our results for the validity of the illness script concept, and next the developmental implications.

4.1. Implications for validity of the illness script concept

In line with predictions, we found consistently better memory discrimination for script atypical than script prototypical information, and at immediate retention than at delayed retention. This supports theoretical notions about illness scripts as general event representations with actual case information "tagged" to these stored representations. As the memory of a case fades, tagged information is supposed to decay over time; hence, the finding of decreased memory discrimination after one week. Though other memory theories may be able to explain this result, the finding of a typicality effect on memory discrimination is characteristic for (illness) script theory. That is, prototypical and atypical statements only differ in their representativeness for a particular disease; otherwise, they are similar. Yet, participants showed poorer memory performance for the former, both at immediate testing (particularly as a consequence of high false alarm rates for prototypical statements), and at delayed testing (particularly as a consequence of lower hit rates for atypical statements). This is exactly what illness script theory would predict: the underlying script knowledge produces relatively many spurious recognition responses for prototypical, but not for atypical, information, whereas after a delay, the decay of tagged atypical information is responsible for participants' failure to recognize presented information. Finally, the hit rates and false alarm rates for prototypical statements at immediate test in our study (0.83 and 0.22, respectively) are comparable to the results of Bower and Clark-Meyers[19], who reported a 0.84 hit rate for presented script-typical items and 0.31 false alarm rate for script related foils, respectively, after a delay of 20 min.

At immediate testing, memory discrimination for atypical information is high, with A' scores over 0.90. False alarm rates for atypical information at immediate testing are low, approximately 0.10, which is in the same order of magnitude as the 0.16 false alarm rate for atypical items reported in a classical script study by Graesser, Gordon, and Sawyer[11]. The finding of lower memory discrimination at delayed testing, irrespective of the typicality of the information, is in line with the idea that at longer delays, general knowledge, rather than retrieved memory traces of cases, dominates memory representations[57]. Surprisingly, false alarm rates for atypical and inconsistent information showed a very similar pattern: low at immediate testing (in the range of 0.06–0.12), but higher at delayed testing (in the range of 0.19–0.28). The most likely explanation is that though inconsistent information directly contradicted information in the case (e.g., whereas the case read, "the patient's spouse phoned the doctor for a house call", the inconsistent statement said "The doctor saw the patient in the consultation room"), it was not entirely precluded by the disease (i.c., renal stone colic). The finding is also in line with an earlier study, in which we found that students as well as physicians assigned nonzero probabilities (sometimes up to 0.10) to case descriptions that were in fact incompatible with the announced diseases[38]. Physicians appear to be reluctant to declare "probability: zero" to a diagnostic hypothesis, no matter how unlikely it is. Relatively high false alarm rates can also be caused by a bias to give "positive" answers, at the expense of "negative" (i.c., correct rejections and misses), even if the costs of making these mistakes are equal[58].

Overall, at immediate testing, participants make more false alarms to paraphrases than to script-based inferences, whereas the reverse is the case at delayed testing. This suggests that at short delays, memory of the surface structure of the presented information has already decayed, but participants are not strongly inclined to infer unstated typical information on basis of the activated illness script. Family physicians, in particular, make few script-based inferences at short intervals, though they have difficulties rejecting paraphrases of typical information. Apparently, they remember the meaning of presented information, but not its surface structure. At a longer delay, however, all participants show high false alarm rates to script-based inferences (in the range of 0.40–0.50) which suggests that the illness script dominates the memory representation and facilitates false recognition of case-typical, but unpresented, information. Though in everyday life this may be adaptive – inferring unstated typical information will enable fluent conversation and quick comprehension – in a professional practice that emphasizes memory accuracy it may be a drawback. For example, if a physician has generated a working diagnosis that still needs to be confirmed, relying on the script may make this working hypothesis appear more plausible than it actually is, i.e., the incorrectly inferred prototypical information provides additional "evidence" for the working diagnosis.

With regard to the RTs, the most interesting finding is the interaction between typicality and textual presence on RTs. RTs for prototypical unstated information were consistently longer than those for any of the other three types (prototypical stated information and atypical stated and unstated information). This effect appears to be most pronounced at immediate testing (though the three-way interaction between typicality, textual presence, and testing time was not significant). It is in line with the prediction that knowledge activated by the illness script interferes with judgments of the actual presence or absence of a memory trace, and it was found in the earlier study as well[56]. In other words, participants cannot entirely suppress the meaning of incoming information even if asked to focus exclusively on the verbatim expression. This can be considered the script-equivalent of the Stroop-effect[59], the phenomenon that people cannot suppress reading responses in naming on the color of the letters in which a word is printed, if there is a conflict between the two (e.g., the word "green" printed in blue letters). Similarly, it takes time for participants to resolve the conflict between script-based activation of knowledge and the lack of a concrete memory trace for this knowledge. The effect is less pronounced, however, than in the study by Yekovich & Walker[15]. In general, the RTs in our study are across the board much longer than those in the Yekovich and Walker[15] study (Table 5), which indicates that it is harder to judge the presence of illness script information, compared with everyday script information, and that some form of additional processing might be required to produce recognition responses[60]. Unlike Yekovich & Walker[15], we found no RT differences between prototypical and atypical stated items, which suggests there is no effect of implicit activation on judgments of the actual presence of statements, at least at immediate testing. To summarize, it just seems to have been somewhat harder for our participants to come up with a response to unpresented prototypical statements, irrespective of whether this response was a correct rejection or a false alarm, than to respond to presented statements of either typicality. Non-negligible false alarm rates, even at immediate testing, support this conclusion.

The fact that memory discrimination in our study was better at immediate than at delayed testing suggests the influence of a verbatim memory trace at this point in time. These findings are in line with Sulin and Dooling[25], who found that after a delay of 5 min, even thematically related foils (comparable to our unstated prototypical statements) were correctly rejected, showing that participants have a discernible memory trace of the actually presented information.

4.2. Development of illness scripts

In general, we found only very limited evidence of developmental differences between the different expertise levels included in the study. One explanation could be that we did not result in probing such differences by the current set-up. This would be in line with a study by Bishop and colleagues[34], which suggests that even laypeople have disease prototypes ("prototype" being very similar to our script concept). Novices, such as our second-year students, may have sufficient (theoretical) clinical knowledge to show similar performance as experienced family physicians on our recognition task.

With respect to reaction times (RTs), a strinking finding in our study was that the experienced physicians were consistently *slower* than the student participants. This is in contrast with earlier studies, in which experts were faster than nonexperts. Though we had relatively few expert participants, it is extremely unlikely that we would have obtained the reverse result by increasing their number. In our 1995 study[56], we found average RTs of 3103 msec, 2692 msec, and 2467 msec for fourth year students, sixth year students, and family physicians, respectively, for similar statements as the ones used in the present study, at immediate testing. In the present study, the corresponding values are 2482 msec, 2431 msec, and 2891 msec. In other words, the fourth year students (comparable to the second year students in our study) and sixth year students in our

1995 study[56] were much slower, whereas the family physicians were considerably faster, than their current colleagues. However, given the large inter-individual variation in average RTs (SDs in the range of ± 500–1000 ms), it seems safe to conclude there were no expertise effects worth speaking of on RTs in our current study.

In general, our study shows that memory discrimination decreases considerably between immediate and delayed testing, irrespective of typicality of the information. Whereas the proportion of misses is usually fairly low, high false alarm rates for presented prototypical information after a delay can be expected[13,25]. We found false alarm rates (proportions) in the range of 0.40–0.50, for prototypical information. The proportion of script based inferences, in particular, is high after a delay of one week (Fig. 1, right-hand panel). In medicine, illness-script based inferences are a double-edged sword. On the one hand they may be functional, providing suggestions for further diagnostic actions, whereas on the other hand, they may distort representation of the actual case, in particular after a delay – the representation becoming increasingly more that of a prototypcial patient, whereas noting the absence of an important prototypical feature may suggest the presence of an unusual disease or an unusual presentation of a more common disease[61]. Previous studies suggested that participants high in domain knowledge were less susceptible to this effect than participants relatively low in domain knowledge[62,63], but we could not confirm this. This is an issue for further investigation, for there may be a trade-off between the accuracy of memory and the ability to infer unstated knowledge.

5. Conclusions

Our study failed to show any consistent expertise effects with respect to the development of illness scripts, neither when expertise was conceived in academic terms, nor when it was expressed as number of patients seen with a particular disease. This suggests that possible expertise differences may be a matter of activating appropriate scripts in a diagnostic context, rather than the "richness" of the knowledge structures per se. In combination with the results of Hobus et al.[64,65], the ability to activate appropriate illness scripts appears to be the major expertise feature that discriminates between experts and relative novices.

The finding that participants were likely to falsely recognize unpresented, but script-typical, information, particularly after a delay, is consistent with their knowledge being organized into illness scripts. Not only does participants' knowledge of the surface structure of the information presented decay quickly, information consistent with the case (diagnosis) appears to intrude their memory representation.

Acknowledgment

The present study is a replication and extension of a study performed as part of a PhD-thesis by the author, which appeared in 1995. The author is grateful to Remy PJM Rikers, PhD, of Erasmus University Rotterdam, The Netherlands, who did the programming work for this study, and to Nori Elshof, who recruited and tested the second and sixth year students. The author specially wants to thank Dr Ron Pieters of the Institute for the Education of Family Physicians of UMC Utrecht, the Netherlands, for his participation and efforts to encourage the family phiysicians affiliated to the Institute to participate. Finally, I want to thank all the students and family physcians who participated in the study.

Disclosures

At the time the study was performed, the Ethical Review Board of NVMO (Dutch Association of Medical Education) was not yet in operation. Participants voluntarily signed up for the study and provided informed consent; no sensitive information was collected, participating or refusing did not have any consequences for their position as students or professionals/teachers, and only numerical data were stored which could not be retraced to individual participants.

Funding

None.

Conflicts of interest

There are no other conflicts of interest.

Appendix A

Example of a case description.
case: kidney stones colic

1. Man, aged 47
2. He is married and has three teenage children
3. His occupation is store-keeper
4. At age 30, he was treated for bronchitis
5. Six years ago, he had his leg broken as a consequence of a car accident
6. Four years ago, he was treated with medicaments for kidney stones
7. Some of his relatives are known with coronary diseases and diabetes mellitus
8. His wife rings up, asks the physician for an immediate visit: it's happening again
9. Her husband is vomiting almost continuously
10. He is rolling across the room because of the pain
11. At the moment the physician arrives, the pain has just subsided
12. The patient is sitting on the sofa, recovering a little
13. He complains about having had a convulsive abdominal pain at the left side, abreast of the navel
14. The pain extends to his groins
15. The pain emerged all of a sudden, and subsequently gradually subsided
16. During an attack, he almost can't stand it
17. Earlier that day, he had already seen some blood in his urine
18. But he had no pain at that time
19. His wife says she has measured his temperature: 38.2° Centigrade

Appendix B

See Appendix Table B1.

Table B1
Test items for the kidney stones colic case described in Appendix A (the actual order of the items in the test was randomly determined).

Typ[a]	Pres[b]	Script[c]	Item text
P	S	EC	Man, aged 47
P	S	Con1	He is rolling across the room because of the pain
P	S	Con2	The patient is sitting on the sofa, recovering a bit
P	U	EC	Four years ago, he had a kidney stone colic
P	U	Con1[d]	The pain radiated
P	U	Con2[e]	In-between the attacks, he doesn't look very ill
A	S	EC	Six years ago, he had his leg broken as a consequence of a car accident
A	S	Con	Earlier that day, he had already seen some blood in his urine
A	U	EC	He is slightly overweight
A	U	Con	He has a mild fever
I	U	EC	The patient appears at the consulting hour
I	U	Con	The pain gradually increases in severity, but then suddenly disappears

[a] Typ= item typicality (P=prototypical, A=atypical, I=inconsistent)
[b] Pres= textual presence (S=stated, U=unstated)
[c] Script= script component (EC=Enabling Condition, Con=Consequence)
[d] paraphrase (cf. Appendix A, statement 14)
[e] script-based inference (is typical for kidney stones colic, but not mentioned in the case described in Appendix A)

Appendix C

A', the measure of memory discrimination and the nonparametric analog of d', can be calculated by the following formulas, where HR= hit rate and FAR= false alarm rate (cf. Snodgrass et al., p. 451).[55]

$$\begin{aligned}&If\ \ HR > FAR,\\ &A' = 0.5 + \left[(HR–FAR)\ (1 + HR–FAR)/4HR(1–FAR)\right];\end{aligned} \tag{C.1}$$

$$If\ \ HR = FAR, * A' = 0.5 \tag{C.2}$$

$$\begin{aligned}&If\ \ HR < FAR, *\\ &A' = 0.5–\left[(FAR–HR)(1 + FAR–HR)/4FAR(1–HR)\right]\end{aligned} \tag{C.3}$$

* N.B. HR = FAR in a binary decision task equals chance performance, and HR < FAR can indicate worse than chance performance. Alternatively, HR < FAR might point to a task execution problem (e.g., confusion of “yes” and “no” keys).

Appendix D

The patient frequency form, which all sixth year students and family physicians were asked to fill in (to tick the box corresponding to the estimated number of patients they had seen):

Herpes zoster	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Nervous abdominal pain	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Meningitis caused by mumps	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Pre-infarct syndrome	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Vaginal candidiosis	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Epidural hematoma	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Kidney stones colic	0	1	2	3	4	5	6-10	11-15	16-20	20^+
Carcinoma of the head of the pancreas	0	1	2	3	4	5	6-10	11-15	16-20	20^+

References

1. Bartlett FC. *Remembering. A Study in Experimental and Social Psychology*. Cambridge, UK: The Syndics of the Cambridge University Press; 1932/1954.
2. Bobrow DG, Norman DA. Some principles of memory schemata. In: Bobrow DG, Collins A, editors. *Representation and Understanding*. New York: Academic Press; 1975. p. 72–92.
3. Brewer WF, Treyens JC. Role of schemata in memory for places. *Cogn Psychol*. 1981;13:207–230.
4. Mandler JM. *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers; 1984.
5. Schmidt DF, Sherman RC. Memory for persuasive messages: a test of a schema-copy-plus-tag model. *J Pers Soc Psychol*. 1984;47:17–25.
6. Abelson RP. Concepts for representing mundane reality in plans. In: Bobrow DG, Collins A, editors. *Representation and Understanding. Studies in Cognitive Science*. NY: Academic Press; 1975.
7. Schank RC, Abelson RP. *Scripts, Plans, Goals and Understanding. An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers; 1977.
8. Abbott V, Black JB, Smith EE. The representation of scripts in memory. *J Mem Lang*. 1985;24:179–199.
9. Bellezza FS, Bower GH. The representational and processing characteristics of scripts. *Bull Psychon Soc*. 1981;18:1–4.
10. Davidson D. Recognition and recall of irrelevant and interruptive atypical actions in script-based stories. *J Mem Lang*. 1994;33: 757–775.
11. Graesser AC, Gordon SE, Sawyer JD. Recognition memory for typical and atypical actions in scripted activities: tests of a script pointer + tag hypothesis. *J Verbal Learn Verbal Behav*. 1979;18: 319–332.
12. Maki R. Memory for script actions: effects of relevance and detail expectancy. *Mem Cogn*. 1990;18:5–14.
13. Smith DA, Graesser AC. Memory for actions in scripted activities as a function of typicality, retention interval, and retrieval task. *Mem Cogn*. 1981;9:550–559.
14. Walker CH, Yekovich FR. Script-based inferences: effects of text and knowledge variables on recognition memory. *J Verbal Learn Verbal Behav*. 1984;23:357–370.
15. Yekovich FR, Walker CH. Retrieval of scripted concepts. *J Mem Lang*. 1986;25:627–644.
16. Haberlandt K, Bingham G. The effect of input direction on the processing of script statements. *J Verbal Learn Verbal Behav*. 1984;23:162–177.
17. Pryor JB, Merluzzi TV. The role of expertise in processing social interaction scripts. *J Exp Soc Psychol*. 1985;21:362–379.

18. Graesser AC, Woll SB, Kowalski DJ, Smith DA. Memory for typical and atypical actions in scripted activities. *J Exp Psychol Hum Learn Mem.* 1980;6:503–515.
19. Bower GH, Clark-Meyers G. Memory for scripts with organized vs. randomized presentations. *Br J Psychol.* 1980;71:369–377.
20. Rosch EH. Principles of categorization. In: Rosch EH, Lloyd BB, editors. *Cognition and Categorization*. Hillsdale NJ: Lawrence Erlbaum Associates, Publishers; 1978. p. 27–48.
21. Bower GH, Black JB, Turner TJ. Scripts in memory for text. *Cogn Psychol.* 1979;11:177–220.
22. Graesser AC. *Prose Comprehension Beyond the Word*. NY: Springer Verlag; 1981.
23. Barsalou LW, Sewell DL. Contrasting the representation of scripts and categories. *J Mem Lang.* 1985;24:646–665.
24. Haberlandt K, Bingham G. Verbs contribute to the coherence of brief narratives: reading related and unrelated sentence triples. *J Verbal Learn Verbal Behav.* 1978;17:419–425.
25. Sulin RA, Dooling DJ. Intrusion of a thematic idea in retention of prose. *J Exp Psychol* 1974;103:255–262.
26. Tzeng OJL. Sentence memory: recognition and inferences. *J Exp Psychol Hum Learn Mem.* 1975;1:720–726.
27. Nakamura GV, Graesser AC. Memory for script-typical and script-atypical actions: a reaction time study. *Bull Psychon Soc.* 1985;23:384–386.
28. Nakamura GV, Graesser AC, Zimmerman JA, Riha J. Script processing in a natural situation. *Mem Cogn.* 1985;13:140–144.
29. Custers EJFM. Thirty years of illness scripts: theoretical origins and practical applications. *Med Teach.* 2015;37:457–462.
30. Clancey WJ. The epistemology of a rule-based expert system – a framework for explanation. *Artif Intell.* 1983;20:215–251.
31. Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In: Schmidt HG, De Volder ML, editors. *Tutorials in Problem-based Learning. New Directions in Training for the Health Professions*. Assen/Maastricht, The Netherlands: Van Gorcum; 1984. p. 128–142.
32. Kuipers B, Kassirer JP. Causal reasoning in medicine: analysis of a protocol. *Cogn Sci.* 1984;8:363–385.
33. Ahn W, Brewer WF, Mooney RJ. Schema acquisition from a single example. *J Exp Psychol Learn Mem Cogn.* 1992;18:391–412.
34. Bishop GD, Briede C, Cavazos L, Grotzinger R, McMahon S. Processing illness information: the role of disease prototypes. *Basic Appl Soc Psychol.* 1987;8:21–43.
35. Bishop GD, Converse SA. Illness representations: a prototype approach. *Health Psychol.* 1986;5:95–114.
36. Arkes HR, Harkness AR. Effect of making a diagnosis on subsequent recognition of symptoms. *J Exp Psychol Hum Learn Mem.* 1980;6:568–575.
37. Hassebrock F, Prietula MJ Autobiographical memory in medical problem solving. *Paper presented at the Annual meeting of the American Educational Research Association*. Boston, MA: 1990.
38. Custers EJFM, Boshuizen HPA, Schmidt HG. The influence of medical expertise, case typicality and illness script component on case processing and disease probability estimates. *Mem Cogn.* 1996;24:384–399.
39. Gilhooly KJ. Cognitive psychology and medical diagnosis. *Appl Cogn Psychol.* 1990;4:261–272.
40. Glaser R. On the nature of expertise. In: Klix F, Hagendorf H, editors. *Human Memory and Cognitive Capabilities: Mechanisms and Performances*, *2*. Amsterdam, The Netherlands: Elsevier Science Publishers; 1986. p. 915–928.
41. Rumelhart DE, Norman DA. Accretion, tuning, and restructuring: three modes of learning. In: Klatzky R, Cotton JW, editors. *Semantic Factors in Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers; 1978. p. 37–60.
42. VanLehn K. Problem solving and cognitive skill acquisition. In: Posner MJ, editor. *Foundations of Cognitive Science*. Cambridge MA: The MIT Press; 1989. p. 527–579.
43. Chiesi H, Spilich G, Voss JF. Acquisition of domain-related information in relation to high and low domain knowledge. *J Verbal Learn Verbal Behav.* 1979;18:257–273.
44. Coughlin LD, Patel VL. Processing of critical information by doctors and medical students. *J Med Educ.* 1987;62:818–828.
45. Hassebrock F, Johnson PE, Bullemer P, Fox PW, Moller JH. When less is more: representation and selective memory in expert problem solving. *Am J Psychol.* 1993;106:155–189.
46. Spilich GJ, Vesonder GT, Chiesi HL, Voss JF. Text processing of domain-related information for individuals with high and low domain knowledge. *J Verbal Learn Verbal Behav.* 1979;18:275–290.
47. McKeithen KB, Reitman JS, Rueter HH, Hirtle SC. Knowledge organization and skill differences in computer programmers. *Cogn Psychol.* 1981;13:307–325.
48. Norman GR, Brooks LR, Allen SW. Recall by expert medical practitioners and novices as a record of processing attention. *J Exp Psychol Learn Mem Cogn.* 1989;15:1166–1174.
49. Arkes HR, Freedman MR. A demonstration of the costs and benefits of expertise in recognition memory. *Mem Cogn.* 1984;12:84–89.
50. Smith EE, Adams N, Schorr D. Fact retrieval and the paradox of interference. *Cogn Psychol.* 1978;10:438–464.
51. Custers EJFM, Boshuizen HPA, Schmidt HG. The role of illness scripts in the development of medical diagnostic expertise: results from an interview study. *Cogn Instr.* 1998;16:367–398.
52. Parks TE. Signal detectability theory of recognition memory performance. *Psychol Rev.* 1966;73:44–58.
53. Pollack I. A nonparametric procedure for evaluation of true and false positives. *Behav Res Meth Instrum.* 1970;2:155–156.
54. Pollack I, Norman DA. A non-parametric analysis of recognition experiments. *Psychon Sci.* 1964;1:125–126.
55. Snodgrass JG, Levy-Berger G, Haydon M. *Human Experimental Psychology*. Oxford, U.K: Oxford University Press; 1985.
56. Custers EJFM. *The Development and Function Of Illness Scripts*. Maastricht, Netherlands: Datawyse / Universitaire Pers; 1995.
57. Reder LM. Plausibility judgments versus fact retrieval: alternative strategies for sentence verification. *Psychol Rev.* 1982;89:250–280.
58. Kareev Y, Trope Y. Correct acceptance weighs more than correct rejection: a decision bias induced by question framing. *Psychon Bull Rev.* 2011;18:103–109.
59. Stroop JR. Studies of interference in serial verbal reactions. *J Exp Psychol.* 1935;18:643–662.
60. Erdfelder E, Bredenkamp J. Recognition of script-typical versus script-atypical information: effects of cognitive elaboration. *Mem Cogn.* 1998;26:922–938.

61. Sanders L. *Every Patient Tells a Story. Medical Mysteries and the Art of Diagnosis*. New York: Broadway Books; 209–213.
62. Sanbonmatsu DM, Kardes FR, Herr PM. The role of prior knowledge and missing information in multiattribute evaluation. *Organ Behav Hum Decis Process*. 1992;51:76–91.
63. Sanbonmatsu DM, Kardes FR, Sansone C. Remembering less and inferring more: effects of time of judgment on inferences about unknown attributes. *J Pers Soc Psychol*. 1992;61:546–554.
64. Hobus PPM, Boshuizen HPA, Schmidt HG Mental representation of prototypical patients: expert-novice differences. *Paper presented at the First European Congress of Psychology*. Amsterdam: The Netherlands, 1989.
65. Hobus PPM, Schmidt HG, Boshuizen HPA, Patel VL. Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. *Med Educ*. 1987;21:471–476.