Long-term memory disorders: measurement and modeling
Meeter, M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
In this chapter, tests for retrograde amnesia are reviewed. Three types of tests have been developed: public events tests, autobiographical memory tests, and neologism tests. All three can only be used to measure retrograde amnesia indirectly, by comparing the remote memory of the patient with what is normal. Nevertheless, testing for retrograde amnesia is not the same as testing remote memory. From this follow several constraints on retrograde amnesia tests: items must measure memories that are universal and dateable. Subtests measuring different periods must be equivalent in some way. Test scores must be stable, but internal consistency in normal controls need not be high. Existing tests only conform partially to these constraints. This may suffice for research purposes, but whether it is enough for diagnostic purposes remains to be seen.

The overwhelming majority of studies into retrograde amnesia are executed in the Anglophone world. None of the studies on retrograde amnesia listed in chapter 2 were done in the Netherlands, and as far as I am aware, only one Dutch study has ever reported on a gradient in retrograde amnesia (Deelman & Maring, 1985). One reason for this is surely the scarcity of retrograde amnesia tests in anything but the English language. In the Dutch language area, for example, few tests have ever been developed (Deelman & Maring, 1985; Wimbergmuhle, 1996), and these few were all obsolete by the end of the nineties. For this reason, we started the development of a battery of tests for retrograde amnesia for use in the Netherlands and Flemish Belgium. Although the primary motivation at the outset was to enable research on retrograde amnesia, enthusiasm from clinical quarters has broadened the goal to also include diagnostic use. Prerequisites for diagnostic use are the same as those for research, but it is less forgiving: mismeasurement in individual cases may reduce the power of a scientific study, but this is a small cost compared to misdiagnosing a patient.

In this chapter, I will briefly present the developed battery. The main thrust will be a review of the difficulties facing tests of retrograde amnesia, and how existing tests deal with them. Section 3.1 introduces the constraints to test developers by retrograde amnesia. The sections thereafter deal with the kinds of tests for retrograde amnesia that have evolved, and with three of the tests developed here (a fourth test is described in chapter 4). How tests deal with the constraints listed in 3.1 is the topic of section 3.4. The chapter ends with a summary and conclusions.

3.1 MEASURING RETROGRADE AMNESIA

Retrograde amnesia is defined as a loss of memories that were present before the onset of amnesia. For example, if a patient had a hemorrhage in January 1, 2000, he or she
would suffer from retrograde amnesia if memories predating that day were lost after the insult.

The best assessment of retrograde amnesia would be a comparison of the patient's memory before and after the insult, but this is rarely possible: clinicians do not, with few exceptions, have the opportunity to test patients before they present with an amnesic syndrome. The only workable strategy for a clinician is to compare the memory of patients with what can be considered normal for their age and background. A pathological loss of memories for periods predating the onset is then interpreted as in indication for retrograde amnesia.

The fact that retrograde amnesia can only be measured indirectly, means that a developer of retrograde amnesia tests is confronted with a number of major challenges. One of the most important follows from the fact that retrograde amnesia is not the loss of a general ability, but it is tied to a specific content: memories from the past. The content of one's memory is the product of an idiosyncratic life, which implies that there is no gold standard for a normal memory of the past with which all patients can be compared. Nevertheless, a shared culture may instill some memories within most of its members. Some news events, for example, have reached virtually all members of western society (e.g., the murder of Kennedy). Some aspects of private life are similar for all members of western cultures (e.g., going to primary school). If items in retrograde amnesia tests tap only these virtually universal memories, patients can be assumed to have had these memories before the onset of amnesia, and retrograde amnesia can be detected. Thus, the first demand to items in retrograde amnesia tests is universality.

Then there are, of course, the more common concerns of test reliability and validity. All tests must be valid measurements, but with regard to reliability the situation is more complicated. As I will argue, some forms of reliability are desirable, others are not.

Another demand follows from the necessity to distinguish retrograde amnesia from anterograde amnesia, and from the desirability of specifying the period for which retrograde amnesia exists. Items must test memories that are dateable. It should be clear in which period the memories are formed that are tested by the item. At the least, it should be clear that the tested memories stem from the period before the onset of amnesia. If they were formed after this onset, or if in normal controls the memory is only present because of active maintenance in the period after the onset of amnesia in the patient, then the item does not measure retrograde amnesia, but instead anterograde amnesia. To be able to specify periods for which retrograde amnesia exists, more precise dateability is of course required.

If items are dateable enough, it becomes possible to measure a gradient in retrograde amnesia. However, if retrograde amnesia for two or more periods is to be compared, subtests measuring memory for these periods must be equivalent in some way. Items can be, for example, on average equally difficult in each subtest, but there are also other ways in which subtests can be equivalent.

These five requirements of a good retrograde amnesia test will be further discussed in a later section. First, three categories of retrograde amnesia tests will be introduced, which differ in the memory material used for the test. Some test memories with an autobiographical content, others use new vocabulary as material. Most test knowledge for past public events, however.
3.2 EXISTING TESTS

Knowledge of public events

Public knowledge is knowledge that every member of a culture can be expected to share. It is difficult to know what belongs to the public domain in this way, but some news events have such impact that knowledge of them is virtually universal. Examples are the murder of Kennedy, the start of the gulf war, or, within the Dutch context, the marriage of the crown prince to a young Argentinean lady.

Tests of public knowledge can consist of questions about news events (e.g., Deelman & Maring, 1985; Gade & Morteson, 1990; Kopelman, 1989; Leplow & Dierks, 1997; Mayes et al., 1994). Some target a specific type of public events, such as horse races (Squire & Slater, 1975), the Oscar awards (Brandt & Benedict, 1993), or television shows (Squire & Slater, 1975). Knowledge of public events can also be tested indirectly, for example through testing the patient’s identification of people that had a prominent place in the news of a certain period. This is the idea behind famous faces tests, the first retrograde amnesia tests to be developed (Sanders & Warrington, 1971). Many such batteries have later been created (e.g., Albert et al., 1981; Hodges & Ward, 1989). In famous faces tests, participants are presented with portraits of people. They either have to name the person in question or provide identifying information, distinguish the famous faces from foils (faces of not-famous people), or state which famous people are alive and which dead, and how the latter died. Some tests have also focused on famous names: subjects had to select names of famous people from foils, and subsequently answer questions about the famous person (Hodges & Graham, 1998; Schmidtke & Vollmer, 1997; Vollmer-Schmolck, Garbaletto, & Schmidtke, in press).

Autobiographical memories

A second type of test focuses not on the public, but on the personal. Autobiographical memory may suffer disproportionately in retrograde amnesia, and it is thus fortunate that several tests have been developed for remote autobiographical memory. One such test was created to investigate the distribution of memories throughout the lifespan. By asking subjects to generate memories to a set of cues and then date them, the distribution of memories across the lifespan was to be measured (Crovitz & Schiffman, 1974). Neuropsychologists have subsequently used modified versions of this technique (dubbed Galton-Crovitz) to investigate amnesia (e.g., Graham & Hodges, 1997).

The most widely used test for remote autobiographical memory is a more traditional questionnaire. The Autobiographical Memory Interview (AMI) consists of fairly specific questions about situations and events that a typical member of a western society experiences (Kopelman et al., 1990; Kopelman, Wilson, & Baddeley, 1989). For example, most members of western societies have attended primary school, and can therefore generate names of three primary school classmates or teachers (one of the questions for the “childhood” period of the AMI). It also contains less specific questions, in which subjects must generate an anecdote in response to cues.

Another questionnaire is a mixture between the unstructured Galton-Crovitz technique and the AMI. In this test, subjects are asked questions that do not target specific memories, but are nevertheless relatively specific in time (Schmidtke & Vollmer, 1997). As an example, subjects must generate five cities within their country that they visited in the last five years.
CHAPTER 3

All tests of autobiographical memories do not, in the first place, score the correctness of the answers provided, because without help a tester cannot know what is true of the life of this particular subject, and what not. Instead, the quantity of memories is used as dependent variable, relying on the fact that in amnesia omissions are more typical than errors (Kopelman et al., 1990). Sometimes the family of tested patients is asked to verify answers generated by the patient.

Neologisms

Nearly all tests of semantic memory are tests of remote memory in the sense that they measure old memories. This does not make them retrograde amnesia tests, as this name is usually restricted to tests that give information about the temporal extend of memory loss, but semantic memories are often defined as those devoid of spatiotemporal content. In some definitions of semantic memory, however, all tests of remote memory contain semantic items. If semantic memory is “the component of long term memory which represents our knowledge of objects, facts and concepts, as well as words and their meanings” (Garrard, Perry, & Hodges, 1997), then all tests of public knowledge are tests of semantic memory (as they query facts about the news), and many tests of autobiographical memory contain semantic items (in the AMI, some questions are also labeled “personal-semantic”). Nevertheless, some semantic memories are more semantic than others. Semantic memories are “usually overlearned and not temporally specific” (Garrard et al., 1997). Knowledge of news events need not be overlearned, and is to some extent temporally specific. Though some “personal-semantic” questions in the AMI lack this specificity (such as items in which subjects must generate addresses on which they lived), others are specific about time (such as the many questions about a specific wedding in the “young adult” section of the AMI). The prototype of semantic memory is our lexical knowledge. Tests that use word meanings as its material are therefore unequivocally semantic. In the English language domain, one such test of retrograde amnesia has been constructed (Verfaellie et al., 1995). For every five-year period from 1955 to 1990, ten to twelve neologisms were chosen that entered the language after the start of the period. For example, “pro-life” was introduced in American English between 1970 and 1975, and was therefore included in the period ’70-’75. If patients could define remote words but not recent words, this was taken as an indication for retrograde amnesia for recent periods.

Recall vs. recognition

Although retrograde amnesia has been found with tests in many different formats, the retrieval demands posed by the test can have a large influence on performance (Squire et al., 1989). In many patients, amnesia seems larger when retrieval is heavily tasked, such as in open questionnaires, in tests where photos must be named, or in test where definitions have to be provided. Deficits are then smaller in tests in multiple choice format or tests in which portraits or names of famous people must be discriminated from lures. Many tests are therefore available both in a demanding recall or naming format, and in a less demanding recognition format (e.g., Deelman & Maring, 1985; Gade & Morteson, 1990; Leplow & Dierks, 1997; Squire & Slater, 1975; Verfaellie et al., 1995).
3.3 THE NEUROMOD BATTERY OF RETROGRADE AMNESIA

As already stated in the introduction, until recently no test for retrograde amnesia was available in the Dutch language area. In the past years, students, colleagues from the NeuroMod group and I developed a whole battery of such tests. Tests of all types discussed above were developed: two public events questionnaires, one autobiographical test, and a vocabulary test, of which most have both an open questions and multiple choice form. At the moment of writing (fall 2002), many of the tests were undergoing validation studies in populations with retrograde amnesia. Three tests will be shortly presented here; the fourth, a news events test presented over internet, is the topic of the next chapter.

1) The Amsterdamse Media-Vragenlijst or AMV (Amsterdam Media Questionnaire) is a public events questionnaire consisting of 42 questions about public events in the ’70s, ’80s and ’90s (Meeter, Klomps, & Borsboom, 2001). It can be taken orally and in pencil- and-paper form. Questions are in open format, but a multiple choice format is also provided, in which the correct answer must be chosen from among four alternatives (4-AFC format). Some sample questions are given in Table 1.

Questions in the AMV were formulated with help of so-called review-of-the-year books and the media. The resulting list of 120 questions was taken through a two step selection process (Klomps, 2001). In the first, two pilot groups were tested: a group of 33 adults, and 45 adolescents. When items were not answered correctly by at least 50% of adults old enough to have experienced the event, it was either discarded or reformulated. The adolescent group was included to investigate specificity of the items for the period in which the event occurred: questions from the ’70s and ’80s were eliminated if more than 30% of adolescents answered them correctly, as the adolescents could not have consciously experienced the news event which was at the base of the question. Questions surviving the first step were presented to a stratified sample of 76 Dutch adults. Of these, 70% had to provide the correct answer to a question in its open format for it to make the test. From the remaining 55 questions, the 14 from each decade with highest p-values (proportion of subjects that answered it correctly) were included in the test.

2) The second test had an autobiographic content. After some consideration, we decided that the easiest trajectory to a valid test of remote autobiographical memory was to adapt the English-language AMI (Kopelman et al., 1990; Kopelman et al., 1989) to the Netherlands. This resulted in the Autobiografisch Geheugeninterview or AGI, an interview in which the participant is asked open questions about his or her personal life (Meeter & Murre, in prep.). As in the AMI, three time periods are covered: childhood (0-18 years), young adulthood (18-32) and recent time periods (last 5 years). Questions are divided into ‘personal semantic’ questions (precise questions about factual information), and 9 ‘incident’ questions, in which an anecdote must be generated (see Table 1 for sample questions). To develop the AGI, questions in the AMI were first translated to Dutch by me, then translated back into English by a co-author [J.M.] without knowledge of the original wording, to control for translation errors. Some questions were dropped—as they were not applicable to the Dutch situation—and replaced by others. This kept the total number of questions per period the same as in the AMI. The finished AGI was presented to a diverse group of 27 adults (as administering the interview was labor-intensive and privacy-sensitive, testing a larger and stratified group was impossible).

3) The third test was a neologisms test, akin to the one developed in America (Verfaellie et al., 1995). The Neologismen- en Vocabulairetest or NVT consists of 44 words that patients have to define (Meeter & de Wilde, 2001). In the multiple choice form, the
correct definition has to be chosen from among four alternatives (the three lures were constructed in such way to have plausible links to the target word). Eleven are neologisms that, according to etymological dictionaries, entered Dutch in the ’70s, 11 are neologisms from the ’80s, and 11 from the ’90s. The remaining 11 words are baseline words that were included to check for a general decline in vocabulary knowledge, unspecific for recent periods (see Table 1 for sample items). The word is counted as correct when the participant has shown a faint hint of knowing the word. Providing detailed definitions tasks verbal ability, to the extend that it is part of all standard intelligence tests (Brody, 1992; Wechsler & Matarazzo, 1972).

The NVT was developed in a way similar to the AMV. First, items were constructed using, in this case, dictionaries of neologisms. Then items were presented to a pilot group of 8 adults for a first sifting. Remaining words were then sent to a stratified sample of 70 adult. Items answered correctly in their open format by more than 90% of participants were included in the test. Of the remaining 44, the 33 with highest p-value were included in the test.

Table 1: Samples of questions from the tests produced by our group, with English translations. In each test, questions are subdivided into three periods - decades in the case of the AMV and the NVT, life periods in the case of the AGI.

<table>
<thead>
<tr>
<th>AMV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>’70s</td>
<td>Waardoor komt het Drentse plaatsje Beilen in december 1975 in het nieuws? [What brought Beilen in Drenthe in the news in December ’75?]</td>
</tr>
<tr>
<td>’80s</td>
<td>Welke sport beoefent Ivan Lendl? [What was the sport of Ivan Lendl?]</td>
</tr>
<tr>
<td>’90s</td>
<td>Waardoor werd de heer Montignac beroemd? [What made mr. Montignac famous?]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AGI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>childhood</td>
<td>Waar ging u naar de lagere school? [What primary school did you visit?]</td>
</tr>
<tr>
<td>early adulthood</td>
<td>Kunt u een voorval vertellen uit de tijd van uw studie / eerste baan? [Can you recount an incident from the time of your studies/first job?]</td>
</tr>
<tr>
<td>recent</td>
<td>Waar heeft u het afgelopen jaar kerst gevierd? [Where did you celebrate Christmas last year?] (with instruction to ask about other holiday for those who do not celebrate Christmas)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NVT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>’70s</td>
<td>Wat betekent ’intercity’? [What does ‘Intercity’ mean?]</td>
</tr>
<tr>
<td>’80s</td>
<td>Wat betekent ’muismat’? [What does ‘mouse placemat’ mean?]</td>
</tr>
<tr>
<td>’90s</td>
<td>Wat betekent ’viagra’? [What does ‘Viagra’ mean?]</td>
</tr>
<tr>
<td>reference words</td>
<td>Wat betekent ’roddelen’? [What does ‘gossiping’ mean?]</td>
</tr>
</tbody>
</table>

Note: correct answers AMV: hijacking of a train; tennis en/of golf; his diet. Correct answer first NVT-word: “Fast train for domestic connections between large cities; only stops at large stations in the bigger cities, or any answer containing ’train’.”
3.4 CONSTRAINTS TO A TEST OF RETROGRADE AMNESIA

Validity

In evaluating the tests reviewed above, the central question is whether they are valid measures of retrograde amnesia. To answer this, some clarification on which validity concept is used is in order. One definition is that for a test to be a valid measurement of an attribute, variations in the attribute must cause differences in test score (Borsboom, Mellenbergh, & van Heerden, subm.). A slightly different definition will be used here: a test is valid to the extent that differences in test score are caused by variation in the to-be-measured attribute¹. Tests are invalid to the extent that scores reflect noise, or attributes other than the one intended.

Borsboom et al. (subm.) suggest that the most difficult issue to resolve with respect to validity often is what exactly is measured. For tests of retrograde amnesia this is indeed a little ambiguous, as one tests remote memory to diagnose memory loss. Is a test of remote memory loss just a test of remote memory? If it was, normal variation in remote memory (in the number and quality of remote memories) should be measured by the test. However, memory loss in amnesia is not the same as normal variation in remote memory; these are two fundamentally different constructs. Retrograde amnesia tests are tests of pathological loss of memory, not of remote memory per se. From this follows that normal variation in remote memory is one of the attributes which lowers the validity of the test, instead of increasing it.

One way to directly investigate the validity of a test for diagnostic use is to look at the discriminability on the test of patients with an affliction and normal controls (this is a form of criterion validity). Is the test a sensitive indicator for the affliction? Is it specific, in that normal controls are rightly classified as not suffering the affliction? In a similar way, a test of retrograde amnesia must be sensitive to retrograde amnesia, and must also be specific in identifying it. To date and to my knowledge, no such study has ever been performed. Of many other tests, mean scores for patient groups have been compared to those of normal controls (reviewed in chapter 2). Though such studies typically show clear differences between the two populations, this does not show that distributions in score do not overlap. Up to now, it is thus unclear whether tests have good enough sensitivity and specificity.

Given the paucity of discriminability studies, the only way to ascertain that a test of retrograde amnesia is valid is to investigate threats to validity, and show how these cannot explain low scores on the test. One threat to validity is that other deficits of a patient may explain his or her low performance. For example, an inability to concentrate may cause bad performance. Malingering may also be a threat to validity, especially in cases of functional or psychogenic amnesia (Kopelman, 1994). Such alternative explanations can probably only be excluded in individual cases. Several of the constraints to retrograde amnesia tests mentioned in the introduction, however, can also be seen as incorporating threats to validity. If memories are not dateable, low performance may reflect anterograde amnesia instead of retrograde amnesia. If the

¹ The original definition has as consequences that every test and most experiments in psychology are valid tests of intelligence. Moreover, it implies that a thermometer is a valid measurement not only of temperature, but also of season, altitude, distance of the earth to the sun, El Nino, CO₂ concentration in the atmosphere, and via the latter even of the outcome of American elections.
memories that form the basis of a test are not universal, low performance on the test may reflect a lack of knowledge, instead of a pathological memory loss. Sampling universal memories is a way of excluding normal variation in remote memory from having an influence on test scores. Universality is therefore not so much an end in itself, as it is a means to the end of a valid test for pathological memory loss. This also holds for the other constraints for retrograde amnesia tests.

**Universality and intelligence**

If tests of retrograde amnesia would only task universal memories, all normal controls would have a score of perhaps 99% correct (with 1% loss to inattention and confusion). Not a single retrograde amnesia test has ever yielded such control scores, and most tests do not approach it. On some public events tests, normal controls perform below 50% correct (Leplow & Diers, 1997), but for most around 70% correct is typical (Albert et al., 1981; Klomps, 2001; Kopelman et al., 1989; Kritchevsky & Squire, 1989). Tests of autobiographical memory typically do better: performance of normal controls on the AMI is seldom lower than 80-90% (Kopelman et al., 1990; Meeter & Murre, in prep.). Control performance on the NVT, the Dutch vocabulary test, was very high for controls: close to the ceiling (see Table 1). On the English vocabulary test it was lower, at around 75% correct in the recall version, probably because more stringent scoring norms were used (Verfaellie et al., 1995).

A low average performance would not be very problematic if scores were clustered around the mean. This is not the case in many tests, however. Especially news tests can show a large variance in the scores (e.g., Kopelman et al., 1990), and some individuals can have a very low score (Klomps, 2001; see Figure 2). The reason is that it has proven difficult to find memories shared by everyone: there is a striking variance in interest, general knowledge and learning capacity in the population. These differences influence the score of normal controls on retrograde amnesia tests. This does not work to the advantage of the tests: a low score of a CVA-patient on a retrograde amnesia test can be

---

**Figure 2:** Mean and range of scores on the three tests. AMV: stratified sample of 74 normal controls (Klomps, 2001); AGI: 27 interviewed normal controls (Meeter & Murre, in prep.); stratified sample of 70 normal controls (de Wilde, 2001).
caused by the brain insult, but may also reflect a level of general knowledge that was already low before the insult. In the first case it would be correct to conclude that the person suffers from retrograde amnesia, in the second not. Especially extremely low scores in normal controls are thus problematic: they directly limit the ability of the test to discriminate between low-scoring normal controls and high-scoring patients with retrograde amnesia.

Not surprisingly, part of the variance in retrograde amnesia test scores is related to general intelligence (Kapur, Thompson, Karsounis, & Abbott, 1998). Accumulated memories—in the guise of general knowledge and semantic memory—are the domain of several subtests of standard intelligence tests (i.e., the WAIS contains a Vocabulary subtest, and a general knowledge subtest; (Wechsler & Matarazzo, 1972). On top of that, personal interest has a clear influence. For example, media consumption correlates even more strongly with the public knowledge form of retrograde amnesia tests than does intelligence (Kapur et al., 1998). It can be expected that the influence of personal interest is even bigger in more specialized tests, such as the one relying on the Oscars: a film buff will know more about Oscar winners than an occasional movie watcher. In some cases, part of the variance can also be ascribed to societal differences. For example, up to 1990, east and west Germans were exposed to a very different news diet. This led to substantial differences in performance on a public events questionnaire based on west-German news (Leplow, Dierks, Merten, & Hansgen, 1997).

Universality of memories minimizes the effects of intelligence, interests and history on the scores: where there is no variance to explain, none will be explained by intelligence. The NVT offers an example of such effects. Initially, two ways of scoring were envisioned: a strict one in which items were awarded more points if the word was defined with some precision, and a loose one in which any hint of knowledge about word meaning was enough for maximum points. This led to normal controls scoring on average 82% and 93% of maximum respectively. It also led to the correlation with an IQ subtest (WAIS vocabulary) dropping from 0.64 for strict scoring to 0.52 for loose scoring (de Wilde, 2001).

Another way to minimize unwanted effects of interest and personal history is to sample items from as diverse as possible a domain. Instead of asking only about political events or cinema trivia—and run the risk that a healthy person not interested in politics or films might end up with the test score of a patient with retrograde amnesia—they may sample questions from many domains, from politics, sports, culture and many other domains in order to create a set of items that do not correlate highly with one another. This will lead to less variance in the scores, because each respondent will be able to answer those items that lie within a personal interest field. Indeed, such deliberately item wide sampling has been undertaken by many authors of public events tests (Klomps, 2001; Mayes et al., 1994; Schmidtké & Vollmer, 1997; Vollmer-Schmolck et al., in press). Effects of personal interest are smaller with autobiographical tests, as most people are interested in their own life.

**Reliability**

The quest for universal memories and low variance in the scores of normal controls has peculiar consequences for test reliability. One first needs to distinguish two reliability concepts: stability and internal consistency. Stability refers to whether the score of a respondent remains relatively invariant over time; it is usually measured with a test-retest correlation. Internal consistency refers to whether different items within the test measure the same variable; if they do, the item scores will be highly correlated (internal
consistency differs from validity in that this variable does not have to be the construct that the test was intended to measure). Internal consistency is most often assessed using Cronbach's alpha.

A few test constructors have estimated internal consistency by computing Cronbach's alpha for patient populations (Eijackers, 2002: .92 for the AMV, .84 for the AGI; Mayes et al., 1994: 0.86 and 0.91 for items from different time periods). Most alphas were computed from normal controls data. Moderately high alphas were found in the population of normal controls (Eijackers, 2002: .89 for the AMV, .62 for the AGI; Leplow & Dierks, 1997: 0.71; Mayes et al., 1994: 0.71 to 0.63; Vollmer-Schmolck et al., in press: 0.88 to 0.67). These values are, according to the authors, at a satisfactory level, implying that for them a high internal consistency in the population of normal controls is an asset for a test of retrograde amnesia. Indeed, a high reliability is one of the hallmarks of a good test. Tests of retrograde amnesia, however, may be an exception to this rule: in the case of these tests, a high internal consistency in the normal population may be unnecessary, and even undesirable.

This may be argued in two ways. A first line of argumentation starts from the formal definition of reliability. Reliability is defined as the proportion of the variance in the test score that is variance in the true score (T) of the respondents on the test. In analyses of reliability, the test scores are often divided into two parts, the true score (T) of the subject on the one hand, and an unsystematic residual error term (E) with a zero mean on the other. Reliability can thus be written as:

\[
\text{Rel.} = \frac{\text{Var}(T)}{\text{Var}(X)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)}
\]

(Mellenbergh, 1996). This shows that reliability is high if the variance in the true scores within the population is high, and low if there is a large error variance on top of the true score. Reliability is a parameter not only of the test, but also of the population that is tested. Its magnitude depends on the variance in true scores that is present within the population; when that variance is large (i.e., \(\text{Var}(T)\) in equation 1 is large), it is easier to obtain a high reliability estimate. Reliability estimates obtained in one population (e.g., normal controls) thus do not automatically generalize to other populations (e.g., amnesic patients)\(^2\).

Tests of retrograde amnesia are constructed with the intention to measure a pathological loss of remote memories. This is a variable on which patients suffering from amnesia can be assumed to show a considerable variance. Though the logic of striving for a high reliability in amnesic populations is thus clear, this cannot be said for the normal control population. By definition, control populations do not exhibit pathological loss of memory. There is thus no variability on this latent variable to measure within the normal control population. Maximizing reliability in the normal control population, however, implies constructing the test in such way that the variance in the true score is large compared to the variance of the error term. The variables underlying this true score variance cannot be the one that the test purports to measure, since the intended variable

\(^2\) Another consequence of this is that reliability estimates obtained for groups of amnesic patients will depend heavily on the composition of the group of patients, and that estimates for combined groups of normal controls and patients will be highly sensitive to the proportions of patients vs. normal controls in the sample.
(pathological memory loss) has a zero variance in the population of normal controls. Anything that is nonetheless measured in the normal population, must be something else than pathological memory loss (e.g., general knowledge). On a pure test of retrograde amnesia, the population of normal controls will have a true score variance of zero. A pure test of pathological memory loss will therefore have a minimal reliability in the normal population, not a maximal one (if we fill in the value of 0 for $Var(T)$ in equation 1, reliability will be 0 for all nonzero values of the error variance).

While this argument is relevant for internal consistency, it is not for stability. Stability estimates of reliability are in the case of retrograde amnesia tests direct measures of error. Retrograde amnesia is assessed through testing knowledge that is considered to reside in every normal person's memory. This knowledge can only be assumed to be stable from one testing occasion to next. Respondents should not, for example, have a different score depending on the time of day they are tested, or on who rates their answers. From this it follows that any instability in the scores of respondents can only be noise, and noise is as always to be minimized.

A second argument for a low internal consistency is that a lower the internal consistency goes together with a lower the variance in the scores of normal controls. This is desirable, since a high variance in the scores of normal controls means that the distribution of normal control scores will overlap more with that of patient scores, which will lower test sensitivity, test specificity, or both. To minimize the variance in the scores of normal controls, one would have to, next to minimizing the variance in item scores, minimize interitem correlations; and therefore minimize the internal consistency of the test in the normal control population.

These arguments are not specific for tests of retrograde amnesia, but also valid for all other tests that purport to measure pathology in a patient group and nothing in the normal population, whether it is a test for psychopathology, neurological disorders, or a physical disease. To give an example, suppose that a rare new disease, bylia, is discovered with three symptoms: a runny nose, a sore throat, and stomach ache. None of these symptoms invariably occur in patients with the illness, but the presence of each symptom independently heightens with the same amount the likelihood that someone has bylia. Now suppose that doctors were so overburdened that they had only time to question a patient about two symptoms, and diagnose bylia if the patient had both. If doctors chose to ask about a runny nose and a sore throat, they would have put together a 'test' that had a high internal consistency in the non-bylia stricken population: many people without bylia would have both symptoms together for other reasons such as a common cold, resulting in a high interitem correlation, and thus in a high internal consistency. This would, however, lead to many of those tested to be falsely diagnosed with bylia (e.g., patients with a cold). The high reliability of this 'test' of two symptoms in the normal, non-bylia stricken population would thus make the test less valid. Doctors would do better if they composed a less reliable 'test' by asking patients about stomach ache and one of the two other symptoms. This would lead to less false positives, and not at the cost of more false negatives since each symptom adds in the same amount to the likelihood that a patient has bylia.

---

1 This follows from a familiar relation: $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$, i.e. the variance of a sum of random variables is larger when the random variables covary. A high internal consistency implies a high correlation, and thus a high covariance, between item scores. Since the test score is the sum of the item scores, its variance will be higher the more the item scores are correlated.
Bylia is here described as an all-or-none phenomenon, while retrograde amnesia is probably best seen as a dimensional construct that varies from none in normal controls, via mild to severe. This is not relevant for the arguments, which concentrate on those that do not have retrograde amnesia. One situation in which the arguments do not apply is when the to-be measured abnormality is conceived as residing in the tail of a continuous distribution (as in the case of retardation).

Returning to retrograde amnesia, it is unlikely that it is possible to construct a retrograde amnesia test on which normal controls do not show any variance. On the contrary, it is highly likely that normal controls will show genuine variance on all measures of remote memory on which retrograde amnesia patients might be discriminated from normal controls. Considering all tests of retrograde amnesia useless if in populations of normal controls the tests showed variance and had an internal consistency higher than zero would be foolish. The relatively high Cronbach's alpha obtained by several test authors for normal controls indicate that retrograde amnesia tests do measure some variance in that population. As discussed above, people vary in their memory functions, in the amount of information they pick up, in their interests, and in their general knowledge. A high internal consistency in populations of normal controls is thus more a reason for worry than an asset.

Dateability

To be able to assign questions in the test to different periods of the life of a person (and to the period before the onset of amnesia, making it a test of retrograde amnesia), one needs to be able to tie the memory that the question queries to that period. The content of the question must refer to some dateable event. If this is a public event, the question can be dated with help of official history. Personal events can be dated with the help of the personal history of the person undergoing the test, for instance by letting family members date events generated by patients. A more practical way is to query for events that are known to occur at a certain moment in the life of the person. The shift from primary to secondary school, for example, usually occurs at age 12 in most western countries. A question about this shift can thus be assigned to the period around the patient's twelfth birthday.

The ideal memory underlying an item of a retrograde amnesia test is laid down on a specific, identifiable day, and is never afterwards revisited. In reality, dateability of remote memories is probably never that perfect. With few exceptions, events are not so forceful as to instantaneously put down a lasting memory that is detailed enough to allow perfect dating. Possible exceptions are traumatic memories (Williams, 1994) and flashbulb memories (Brown & Kulik, 1977), but the former cannot be part of retrograde amnesia tests while the latter are few in number (and possibly unreliable, see Neisser & Harsh, 1993). Important autobiographical memories are relived, important public memories retold, and that is what saves them from the fate of most memories: forgetting.

Dateability is thus usually a matter of degree. If the content of the memory is dependent on an event that occurred at some moment, this at least provides a lower bound for when the memory was formed (e.g., memories of Kennedy's murder were not formed before November 22, 1963). If one then assumes that with discrete events, most reliving and retelling occurs in the immediate time after the event, then the memory can be dated as coming approximately from the time that the event took place. The assumption just made cannot be rigorously tested, but some findings support it. Anderson and Schooler (1991) found that in both the media and in personal email, mention of events trails off
with a power law after the event, while Meeter, Murre and Janssen (subm.) could not find evidence for extensive relearning of news events after the event took place.

Dateability also does not need to be perfect, as most tests group questions into wide periods. Many news events tests use decades as their periods (Albert et al., 1979; Kopelman, 1989; Leplow & Dierks, 1997; Mayes et al., 1994; Meeter et al., 2001; Sanders & Warrington, 1971; Schmidtk & Vollmer, 1997; Vollmer-Schmolck et al., in press). If the memories forming the basis of the items can thus be dated to the decade in question, this is good enough. Within such bounds, the dateability of public events tests can be tested to some extent by putting items to subjects too young to have formed a memory of the original events (Klomps, 2001; Leplow & Dierks, 1997; Sanders & Warrington, 1971). If, for example, a young adult who was 5 at the time of the Falkland war cannot answer a question about that event, this can be taken as an indication that older subjects who know the correct answer did not learn it through the media in recent years: if they did, there would be no reason why the younger subject would not also have picked up the answer in those years.

Other types of test have even less precision than public events tests. The Autobiographical memory interview uses larger intervals. The neologisms tests use decades (Meeter & de Wilde, 2001) or five-year periods (Verfaellie et al., 1995) as time intervals, but these are only indicative of when memories for the word could have started to be formed: words enter a language at a certain moment, but this is almost certainly not the date that the participant first experience it. Moreover, words usually do not leave the language, so that memories for the word can continuously have formed since the moment the word entered the language to the onset of amnesia.

Dateability of items in retrograde amnesia tests is thus never perfect. Moreover, it is worse for tests that score good on universality (neologism tests, autobiographical tests) than it is for tests that do not score as well on universality (news events questionnaires).

**Equivalence of items**

If items are not only dateable, but also equivalent across periods, then gradients in amnesia can be investigated. This equivalence can be on two variables. The first is acquisition strength, the second retrieval probability. In both cases "item difficulty" is equalized, but at different points in time: either at acquisition, or at recall. If periods have equivalent acquisition strength, this means that the items in each period were on average laid down with the same strength, had the same initial item difficulty. For example, Squire and Slater (1975) constructed their public events questionnaire with help of one-season TV series. As there is no reason to assume that TV series laid down stronger memories in 1962 than they did in 1972, items on the test can be assumed to test memories that were acquired with on average equal strength. As another example, Leplow et al. (1997) constructed items on their public events questionnaire from newspaper headings according to a tight script. In this way, equivalence of questions was at least made plausible.

It is not easy to construct a script that is applicable to all tested periods. The greatest drawback to equivalence of acquisition strength, however, is the cost in terms of item universality. Inflexible scripts often yields more difficult items than they would be, had the constructor had the freedom to select the easiest items. Moreover, equal acquisition strength implies that normal control performance will show a forgetting curve, as older items are forgotten more often than recent ones. On Squire and Slater’s (1975) test of TV series, performance of normal controls drops from 75% to 30% correct. On Leplow et
al.’s (1997) questionnaire, it drops from around 55% to less than 10%, making the test almost useless for measuring retrograde amnesia for remote periods.

Perhaps for this reason, most test constructors have attempted to keep retrieval probability equal across periods, not acquisition strength (e.g., Kopelman et al., 1990; Mayes et al., 1994; Verfaellie et al., 1995). This can be tested in a straightforward manner: performance of normal controls on the test should be more or less a straight line. Indeed, such performance for normal controls is found on many tests of retrograde amnesia (see chapter 2, figure 1). Moreover, performance is typically higher than for tests in which acquisition strength is equated. However, it comes with the price of a loss of interpretability: there are now qualitative differences in the kinds of items making up different periods. Thus, if patients show a different performance for one period than for another, it may be due not to differential memory loss, but to a differential loss of one type of items that happened to be more numerous in one part of the test than in another (Mayes et al., 1994). As is the case with dateability, there is thus something of a trade-off between better equivalence of subperiods (with equivalence of acquisition strength), and better universality (with equivalence of retrieval probability).

### 3.5 DISCUSSION

Tests of retrograde amnesia have tended to fall into one of three categories. Public events tests rely on news as material for items; these can either take the form of questions about events, or of famous faces that have to be named or recognized. Autobiographical tests rely on cues that elicit memories on most people, or on aspects of life that are similar in most members of the surrounding culture. Neologism tests rely on shared vocabulary that has entered the language in a certain period. Of all three types, one test has been constructed by our group.

Several constraints to such tests have been discussed in this chapter. Tests should sample universal memories that are dateable. The first assures that variation in general knowledge, history or interests do not contaminate the test, while the latter assures that one is measuring retrograde, not anterograde amnesia. With dateability, equivalence of subtests measuring different periods enables one to investigate gradients in retrograde amnesia. Test scores must be stable, and must have high internal consistency in populations of patients with amnesia. However, because tests of retrograde amnesia are not tests of remote memory but tests of pathological memory, a high Cronbach’s alpha in normal controls is a reason for worry.

None of the tests of retrograde amnesia reviewed here fulfills these constraints perfectly. If universal memories would suffice to answer all questions normal controls would have perfect performance, but this is not the case. Especially on public events tests, performance is sometimes low. More worrying than a low average is the high variance: some normal controls have a very low score, threatening discriminability of patients with retrograde amnesia. Public events tests do measure something other than pathological memory loss in the normal population, as is also shown by high internal consistency estimates. Performance on these tests correlates with characteristics such as intelligence, interests and personal history, suggesting that these subject variables pollute some retrograde amnesia tests. However, on autobiographical tests and a neologism test (the NVT) normal controls on average had high scores, while on most public events tests mean performance was still safely above 50%. Many tests thus do rely on memories that to some degree are universal.
Dateability was also found to be a matter of degree. On this criterion, tests with higher normal control performance—the AMI and especially neologism tests such as the NVT—are worse than the public events tests that score worse on universality. It is, in a sense, not surprising that tests scoring well on universality perform lower on dateability, and vice versa. Items that are repeated often form durable memories, but repetition already implies that the memories become less dateable. Also between equivalence of test periods and universality there is something of a tradeoff. Of the two ways to approach equivalence, the one that leads to higher scores entails more interpretational difficulties.

Nevertheless, universality is worth emphasizing. The greatest threat to the validity of retrograde amnesia tests is probably the already mentioned effects of intelligence, interests and personal history on scores (also see Kapur et al., 1998). These influences become smaller the more the test relies on universal memories. This was shown with the NVT, where strict scoring lead to lower scores, and a higher correlation of scores with intelligence (de Wilde, 2001). Another way to ensure less influence of interest and personal history is broad sampling of items, as is already done by many constructors of retrograde amnesia tests.

All in all, the perfect test of retrograde amnesia has not yet been created. Present tests conform somewhat to the demand that universal and dateable memories be at their base, and different periods have been made equivalent to some degree. This may be all that is necessary for research on retrograde amnesia. Indeed, populations in which retrograde amnesia is commonly assumed score much lower on tests of retrograde amnesia than do normal controls, suggesting that these tests are in fact to some extend valid measurements of retrograde amnesia. Whether this is enough for clinical use of such tests in individual cases remains to be seen, however, especially as no studies into the sensitivity and specificity of retrograde amnesia tests have been done.