PAPER

Testing cognitive function in elderly populations: the PROSPER study

P J Houx, J Shepherd, G-J Blauw, M B Murphy, I Ford, E L Bollen, B Buckley, D J Stott, W Jukema, M Hyland, A Gaw, J Norrie, A M Kamper, I J Perry, P W MacFarlane, A Edo Meinders, B J Sweeney, C J Packard, C Twomey, S M Cobbe, R G Westendorp

J Neurol Neurosurg Psychiatry 2002;**73**:385–389

See end of article for authors' affiliations

Correspondence to:
Dr R G J Westendorp,
Section of Gerontology
and Geriatrics, Leiden
University Medical Center,
C-2-R, PO box 9600,
2300 RC Leiden, The
Netherlands;
R.G.J.Westendorp@LUMC.nl

Received 29 April 2002
In revised form 1 July 2002
Accepted 3 July 2002

Objectives: For large scale follow up studies with non-demented patients in which cognition is an end-point, there is a need for short, inexpensive, sensitive, and reliable neuropsychological tests that are suitable for repeated measurements. The commonly used Mini-Mental-State-Examination fulfils only the first two requirements.

Methods: In the PROspective Study of Pravastatin in the Elderly at Risk (PROSPER), 5804 elderly subjects aged 70 to 82 years were examined using a learning test (memory), a coding test (general speed), and a short version of the Stroop test (attention). Data presented here were collected at dual baseline, before randomisation for active treatment.

Results: The tests proved to be reliable (with test/retest reliabilities ranging from acceptable ($r=0.63$) to high ($r=0.88$) and sensitive to detect small differences in subjects from different age categories. All tests showed significant practice effects: performance increased from the first measurement to the first follow up after two weeks.

Conclusion: Normative data are provided that can be used for one time neuropsychological testing as well as for assessing individual and group change. Methods for analysing cognitive change are proposed.

Now that causes of cognitive impairment are being unravelled, there is an urgent need for sensitive and reliable neuropsychological tests for large scale application. The number of clinical trials evaluating experimental treatments of cognitive decline is increasing rapidly. Areas of cognitive functioning that tend to change with time or in response to medical interventions that bring about subtle cognitive effects include: memory, attention, and general cognitive speed. These are listed among the so called fluid abilities. Many other cognitive domains are far less likely to change, for example, reading, general knowledge, and language abilities. These are called crystallised abilities,¹ and by definition remain relatively constant over the normal adult's life span. For tracking individual changes in cognition, you test for fluid abilities such as memory, general speed, and attention, a decline in which also marks incipient dementia.² As these abilities can be assessed without time consuming or effort consuming procedures they lend themselves well to testing in large scale intervention studies.

Until now, many follow up studies involving cognition in medical settings have been less successful than they might have been. One important reason for this can be found in the tests that were used. In many cases, only one test is used: the Mini-Mental-State-Examination (MMSE).³ It is used as the standard cognitive screening instrument in virtually all studies involving the elderly population and cognition, but for reasons specified below, it is not very suitable in follow up studies. It aims at screening various areas of cognition, including crystallised functions, which, as we have noted, are unlikely to change. Moreover, the MMSE aims at screening cognitive functions in people suffering from or at risk for dementia.

Another drawback of the MMSE and of various other tests is that there is little room for improvement or deterioration over time. Many subjects will perform at a near maximum level on a test that is too easy, leaving no opportunity to identify any detectable improvement—that is, a ceiling effect.

Conversely, a test that is too difficult, especially for the less able population, will elicit the poorest possible performance and is thus insensitive to further deterioration. This phenomenon is called a floor effect. Ceiling and floor effects are often overlooked threats to the sensitivity of a study and are discussed in detail by Rasmussen and colleagues.⁴

A further problem the MMSE shares with many other tests is the possibility of direct learning effects. For example, in the Maastricht Aging Study,⁵ memory recognition for items that were learnt as long as six years before was well above chance. Even elderly or diseased people may remember parts of the content of a test for surprisingly long periods of time. For this reason, parallel forms are needed. For two versions to be parallel they need to meet two important criteria; a high performance on one version should be mirrored by a high score with the other, and the various versions of the test should elicit the same performance when used at baseline. In summary, then, screening instruments for general cognitive dysfunction should not be used for repetitive measurements in follow up studies with non-demented persons.

Finally, a test should be interesting, attractive, and challenging enough to keep the respondent motivated to complete the test to the best of his ability. It should not appear too complicated to people whose cognitive abilities are compromised. It should also be brief: many elderly subjects tire quickly and long testing sessions are costly. For this reason, full scale neuropsychological assessments are often quite impossible.

Here we report on the performance of a neurocognitive test battery that has been applied in a large scale follow up study in an elderly population. To our knowledge, this is the first study reporting on test/retest effects with often used cognition tests and with such large numbers. We attempt to provide data for baseline and clinical measurement together with normative data for practice effects. Moreover, in dealing with test/retest data there is need for a straightforward way of analysing individual change as a function of interventions or

events. For this reason, we also propose a method of analysing longitudinal data.

METHODS

Setting

The data in this manuscript are drawn from the PROSPER (A PROspective Study Of Pravastatin in the Elderly at Risk) cohort of 5804 people aged 70 to 82 years in whom the effects of cholesterol lowering with pravastatin are being tested using a double blind, placebo controlled design.⁶ All subjects have had or are at increased risk of experiencing cardiovascular events. The main risk factors are history of smoking, overweight, hypertension, diabetes, and history of cardiovascular disease. Subjects are randomly assigned in a double blind fashion to placebo or pravastatin. All are followed up for at least three years. Apart from cognition, end points include death and cardiovascular events. Three European coordinating centres—Glasgow (Scotland), Cork (Ireland), and Leiden (the Netherlands)—are collaborating in the project, which is based in primary care (general practice) or in trial centres in close proximity to each of the three coordinating centers.⁶ Data on repeated measurements presented here were assessed at baseline, two weeks apart. As this all occurred before randomisation the treatment effects do not feature in this manuscript.

Subjects

Inclusion and exclusion criteria have been described in detail elsewhere.⁶ The age range was 70 to 82 years at study entry, and subjects were enrolled independent of sex and educational status. Education was recorded as the age at leaving school and not as the number of years of formal education, as the age of entering school differed across the countries and the school systems are very difficult to compare. Three classes were used. Low education was defined as an age of leaving school up to and including 12 years, middle education as an age 13 up to 16 years, and high education as an age of leaving school of 17 years or more.

Tests

The Mini-Mental-State-Examination (MMSE) is widely used to screen for cognitive dysfunctions compatible with dementia.³ In PROSPER the generally used cut off criterion of 24 points was used. Subjects scoring below that level were not enrolled. As MMSE scores were used as exclusion criteria before the trial, and for screening purposes for cognitive disorders suggesting dementia once a year, test/retest effects for the MMSE are not reported here.

To serve as outcome variables, three cognitive tests were used to evaluate cognitive performance. As PROSPER takes place in three countries with two languages, care was taken to select tests or test versions that are not sensitive to language or cultural differences. Therefore, pictures, colours, and digits were used rather than words. This procedure worked well in earlier studies.⁷

The Picture-Word Learning Test (referred to as Learning Test below) was used as a verbal learning test of long term memory. It was derived from the Groningen-Fifteen-Words-Test,^{8,9} originally devised by Rey.¹⁰ Fifteen pictures were successively presented at a rate of one per two seconds. Next, the subject was asked to recall as many pictures as possible. This procedure was carried out three times. After 20 minutes, delayed recall was tested. Pictures instead of words were chosen to circumvent the language problem in the study. Also, the pictures had to be named and were thus encoded verbally as well as visually. Finally, using a picture guarantees that the subject can make a mental image of each stimulus. The main outcome variables are the accumulated number of recalled pictures over the three learning trials and the number of pictures recalled at delayed recall.

The Stroop-Colour-Word-Test (Stroop) has often been used to test selective attention.¹¹ The test involved three parts that displayed 40 stimuli each, which the subject was asked to read or name as quickly as possible: (1) colour names, (2) coloured patches, and (3) colour names printed in incongruously coloured ink; for example, “green” printed in blue letters, where the subject is to say “blue”. Performance on part 3 is determined for a large part by the time needed to discard irrelevant but very salient information (verbal), in favour of a less obvious aspect (colour naming), also known as cognitive interference. Usually, each element of the test contains a hundred stimuli, but for PROSPER, an abbreviated version was used with 40 items per test part that still proved a very reliable estimate of performance on the complete test.¹² The main outcome variables are the times needed for each of the three test parts.

The Letter-Digit Coding Test (Coding Test) is a modification of the procedurally identical Symbol-Digits Modalities Test.¹³ In neuropsychological assessment, this test is often used as a measure of the speed of processing of general information—that is, the test is devised to draw upon several processes simultaneously, such as visual scanning, perception, visual memory, visuoconstruction, and motor functions. The subject is asked to fill in digits near letters, according to a key presented at the top of the test sheet (with nine consonants in random order), and to work as fast as possible. The outcome variable is the total number of correct entries, in 60 seconds. On the test sheet, there are 125 randomly distributed letters from the key at the top. An elderly subject will typically make 15–30 correct entries, which rules out ceiling effects.

The cognitive tests mentioned in this paper are intellectual property of the Maastricht Brain and Behavior Institute of the Maastricht University. Researchers can use the tests for free within a scientific cooperation.

Procedures

In and around Glasgow and Cork subjects were tested in their own general practice surgeries, in which a nurse used a quiet room with an empty desk. In Leiden, all subjects were tested in a study centre with dedicated testing rooms. In all centres, testers were trained study nurses supervised by at least one nurse manager per centre. All nurses were trained by a neuropsychologist and two experienced testers for two days or more before performing the tests. Yearly training sessions and check up visits occurred at all centres. Also, the nurse managers were trained to check regularly on the testing performance of all nurses. There is frequent contact between the nurse managers and the neuropsychologist (PH).

All subjects are to be followed up for an average of 3.5 years (3 to 4 years). Two baseline cognition tests were performed two weeks apart, before randomisation. Thereafter, subjects are being retested at 9, 18, and 30 months, and for some subjects at 42 months. Only the first two measurements at baseline are reported here.

Regarding the Learning and Coding tests, seven parallel forms, using identical procedures, but with different items, were available. This has been done to obviate the learning phenomena discussed above. Different pictures were used for each learning test: every version consisted of 15 different items. For the Coding Test, the key at the top of the form differed for every version. For the Stroop test, parallel versions are not necessary, as the location of the many colour names on the test sheets cannot be remembered. Memorisation of the stimuli is extremely unlikely, as there are $3 \times 40 = 120$ randomly distributed colour names/patches, which are very hard to cluster into meaningful wholes. Also, incidental learning is unlikely, as the test requires performance that is not related to memory in any way, in contrast with both other tests. For the coding test, there are nine letter-digit pairs, making incidental learning far more likely than for the Stroop test. Test versions were assigned by means of a latin square design.

Table 1 Subject characteristics

Age (y)	N	MMSE (mean (SD))	Male (n (%))	Age at leaving school (mean (SD))	Testing unreliable*		
					Learning T (n (%))	Stroop (n (%))	Coding test (n (%))
≤71	1174	28.20 (1.49)	643 (54.8)	15.13 (2.22)	88 (7.5)	110 (9.4)	110 (9.4)
72–73	1198	28.19 (1.47)	626 (52.3)	15.20 (2.21)	86 (7.2)	102 (8.5)	102 (8.5)
74–75	1077	28.08 (1.54)	509 (47.3)	15.17 (2.08)	102 (9.5)	115 (10.7)	115 (10.7)
76–77	966	27.86 (1.62)	449 (46.5)	14.97 (1.99)	121 (12.5)	120 (12.4)	120 (12.4)
78–79	751	27.81 (1.65)	324 (43.1)	15.20 (2.02)	88 (11.7)	101 (13.5)	101 (13.5)
≥80	638	27.66 (1.57)	253 (39.7)	15.24 (2.29)	73 (11.4)	96 (15.1)	96 (15.1)
Total	5804	28.01 (1.56)	2804 (48.3)	15.15 (2.14)	558 (9.6)	644 (11.1)	644 (11.1)

MMSE; Mini Mental Status Examination, Learning T; picture-word learning test, Stroop; Stroop colour-word test, Coding test; letter-digit coding test. *As judged by the study nurse.

After administering the tests the nurse was asked to rate whether the procedure went well and whether the test outcomes were, in her opinion, valid. For this purpose, a field with the following choices was added to the case report form:

- 1 Complete and reliable: data from this test can be used.
- 2 Technical problems: for example, problems with the stopwatch.
- 3 Refusal/insufficient motivation of the subject.
- 4 Physical limitations: for example, insufficient vision/hearing or missing glasses/hearing aids.
- 5 Cognitive limitations: inability to grasp instructions.
- 6 Subject deviated from instructions, and could not be corrected in this: or example, the subject works very neatly and slowly; subject does not do all elements of a test (for example, skips a line in the Stroop test, skips items in the Coding Test).
- 7 Test not administered: for example, forgotten; not enough time.

Whenever a status value did not equal <1>, the nurse wrote a comment on the case report form to explain why this was the case. For subsequent statistical analysis only test data that the nurse assigned a status value of <1> were used.

RESULTS

Table 1 summarises the distribution of age, sex, and educational level within the PROSPER sample. In the total sample, the proportion of women was about 50%. Mean age

was 75.3 years. The average age at which subjects left school was 15.2 years. For Scotland, this means an average of 0.2 years of formal education.

Together, the whole set of tests took 30 to 40 minutes: about 10 minutes for the MMSE, eight minutes for the first three trials of the learning test, two minutes for the delayed recall, eight minutes for the Stroop, and five minutes for the coding test, all including test instructions. In experienced subjects test duration was shorter. Between the first three trials and the delayed recall of the learning test, there was a mandatory delay of 20 minutes. This could be used for the coding test, Stroop and other procedures, as long as they were not memory related (such as the MMSE). If, after completion of Stroop and the coding test, less than 15 minutes had passed after the last learning test trial, the extra time was used for administrative procedures with the subject.

Standardising the administration of the MMSE proved to be by far the most time consuming for the study nurses and the neuropsychologist in PROSPER. For example: one of the 30 points that make up the MMSE score, can be gained by telling the season. In recent years, this has become ambiguous, as there is an astronomical start of the new season (about the 21st of a month) and a meteorological start (the first of the same month). Furthermore, in Ireland, the seasons change one month earlier than in Scotland and the Netherlands. This was solved by using the local standards and allowing for both starting dates. For instance, in Scotland “winter” was correct from 24 November to 28 March. For other questions local differences of interpretation also arose.

Table 2 Average scores on the four test outcomes at baseline and practice effects

	Learning test (number of pictures/words recalled)				Stroop test		Coding test	
	cum recall tr 1–3		delayed recall		seconds for part 3		number of entries in 60 seconds	
	1st meas	pract eff	1st meas	pract eff	1st meas	pract eff	1st meas	pract eff
≤71	27.18 (5.09)	1.71 (4.45)	10.18 (2.35)	0.35 (2.09)	64.40 (24.72)	–4.15 (14.72)	23.12 (7.28)	1.84 (3.72)
72–73	26.81 (5.23)	1.91 (4.57)	10.06 (2.39)	0.38 (2.19)	64.77 (22.84)	–4.45 (15.71)	22.85 (7.12)	1.35 (3.82)
74–75	26.22 (5.36)	1.88 (4.38)	9.84 (2.50)	0.35 (2.15)	69.55 (30.08)	–4.38 (18.32)	21.88 (7.08)	1.44 (3.91)
76–77	25.89 (5.32)	1.85 (4.49)	9.63 (2.52)	0.31 (2.12)	74.68 (30.71)	–5.48 (18.10)	21.08 (7.00)	1.34 (3.73)
78–79	25.14 (5.40)	2.15 (4.40)	9.56 (2.56)	0.36 (2.16)	78.19 (32.64)	–4.97 (19.77)	20.49 (7.14)	1.06 (3.66)
≥80	24.26 (5.39)	2.30 (4.49)	9.04 (2.45)	0.50 (2.21)	80.57 (33.21)	–6.07 (20.65)	19.19 (6.80)	1.10 (3.51)
Total	26.14 (5.36)	1.92 (4.47)	9.80 (2.47)	0.37 (2.15)	70.61 (29.10)	–4.78 (17.53)	21.73 (7.21)	1.40 (3.76)

Cum recall tr 1–3; cumulative recall learning trials 1–3. Columns headed 1st meas denote test performance at the first measurement. Columns headed pract eff give the practice effect—that is, 2nd minus 1st measurement. For the Stroop test, a lower score means improvement. Data are presented as means and standard deviation (SD). All differences in means over age bands at $p < 0.0001$ on one way analysis of variance.

Almost all subjects were able to complete the tests without difficulty. For each test individually, the amount of incomplete data was about 6%. In about 90% of all assessments, the nurses judged the data as complete and reliable for both measurements.

Table 2 shows the mean and standard deviation at the first measurement of the four outcome variables, along with the practice effects observed at the second measurement (second minus first). Error scores are not given: as the average numbers of errors proved to be too low to tabulate meaningfully. For instance, the average number of errors on Stroop part 3 was less than 1.5. Data are given for five age classes. Each test showed gradual and highly statistical significant differences over age (analysis of variance, all $p < 0.001$).

Given the very large number of subjects per cell it is possible to analyse the scores on the various parallel versions of the Learning Test and the Coding Test. There were minor, but nevertheless statistically significant differences between the various versions of the Learning Test, the version of the test explaining less than 2% of the total variance in the study population (analysis of variance, $p < 0.0001$). No such effects were found for the Coding Test.

The second measurement took place after two weeks. Performance on this occasion was compared with the first measurement. For instance, in the column headed Coding Test, first line, the value of 1.84 can be interpreted as an average 8% improvement from the first to the second measurement in that age group. If interpreted as a normative value, the Coding Test score of the average 70 year old should increase from the first to the second measurement by 1.84 correct entries.

High test/re-test correlations are found for the Stroop ($r = 0.80$) and the coding test ($r = 0.88$). The total score for the three trials together and the delayed recall show acceptable reliability coefficients ($r = 0.66$ and $r = 0.63$, respectively). The test/re-test correlations were not affected by age or education (data not shown). There were, however, small differences between the countries, Pearson's correlations varying less than 0.05 for each of the three test sites (not shown).

DISCUSSION

The PROSPER neuropsychological test battery (Picture-Word-Learning-Test, Stroop-Colour-Word-Test and Letter-Digit-Coding-Test) proved to be quite manageable for the vast majority of elderly participants in PROSPER, as can be seen from the high percentages of cases in which testing was complete and reliable. Moreover, the tests were sensitive enough to detect even small age differences. This is in sharp contrast with the test performance of the MMSE. This instrument seemed to have other disadvantages when used as a measure of cognition in repeated measurements: firstly, performance at the MMSE was near maximum in even the oldest subjects (ceiling effect), and secondly, there was very little age related decline in MMSE score from age 70 to over 80 years, suggesting poor sensitivity to factors other than age as well. To this, it might be added that the MMSE is really a short battery of various cognitive tasks, including crystallised functions, as it was devised for overall cognitive screening. The other tests are dedicated to one area of cognition only, and test outcomes are hence more easily interpretable.

Need for standardisation

Administration and scoring of the MMSE proved to be by far the most laborious of all tests used in PROSPER. This suggests that the MMSE is not an instrument as readily usable and as easy to administer as many researchers and clinicians like to think it is. However, the other tests, especially the learning test also required effort before they could be used reliably by the

nurses. This calls for devising a standard way of administration and scoring for all tests that are used (the standards used for PROSPER are available from the authors on request). Furthermore, although the tests used for PROSPER are ubiquitously used in clinical and experimental settings, they are variations on a theme. Until now, there has been little standardisation in test versions or test procedures and instructions. Assuming standard versions for each test would improve comparability of different studies, however, we propose these same versions be used in similar studies.

Reliability

Test/re-test reliability was acceptable for the memory test and high for the Stroop test and Coding test even when applied at such a large scale in various centres within the framework of a clinical trial. The good test/re-test reliability fulfils the first criterion for the different versions of the test to be truly parallel. The second criterion appears also to be fulfilled as the various versions of the tests explained only very little of the total variance.

Practice effects

Performance on the learning test improved by half a word or more from the first to the second measurement. For Stroop and Coding Test the practice effects were roughly five seconds faster and 1.5 more entries correct, respectively. These differences are usually regarded as "relevant" in the clinical assessment of individual patients. It now seems that in clinical settings, practice effects have to be taken into account when following the course of cognitive functions in subjects over time. Because parallel versions are available for the learning test and the coding test and redundant for the Stroop test, the possibility that the subjects had remembered specific material from a test version can be ruled out.

Although direct learning has been ruled out by administering parallel test versions, other forms of learning cannot be circumvented. One of these is procedural learning, improving performance by merely doing something more than once. All tests that are cognitively demanding are subject to this phenomenon. Performance on a test may even get "automatic", when given enough practice,⁴ or participants may develop a conscious strategy for best carrying out a given task.⁵ For instance, in a typical verbal learning task, subjects are not informed that after some time the words are to be recalled. However, confronted again with a parallel version of the test, the subject may rehearse the items during the interval between learning trials and the delayed recall trial. Also, the perceptual and motor processing involved in the Coding Test may improve through repeated administration.

Selectivity of the PROSPER population

Half of the 5804 subjects in PROSPER had a history of cardiovascular disease, whereas the other half of the study population was free of clinical symptoms of cardiovascular disease but were at increased risk for events at the times of testing. Arguably, therefore, they did not perform as well as people might have done without such risk factors. However, a history of cardiovascular diseases or existence of one or more of the risk factors that served as inclusion criteria for PROSPER is very common among the elderly and the same features may be regarded as risk factors for cognitive impairment. Furthermore, subjects have to be interested and sufficiently motivated to stay in the study for three to four years. For this reason, all people who volunteer to participate in research projects should be considered an elite of better performing persons. People were selected on the basis of MMSE outcome, which had to be 24 or higher. So, again a cognitively elite group of subjects was selected. It is noteworthy, however, that although selection may have influenced the absolute values, it is unlikely that it can explain the observed trends in the test/re-test effects and reliability.

Proposal for outcome variables

For reasons of clarity, in many studies it is preferable to minimise the number of outcome variables. We used one memory test and two speed tests. The memory test consisted of three immediate recall trials and one delayed recall trial. Two variables were selected from the many that can be derived from the test:

- (A) the total recall of the first three trials, as a measure of learning capacity;
- (B) delayed recall, ranging from 0 to 15 as a measure for long term retention.

For the speed tests, the two main variables were:

- (C) the time needed for Stroop part III, involving naming incongruous colour in which names were printed, this being an estimate of attention and speed;
- (D) the number of correct responses to the coding test, as a measure of general cognitive and perceptuomotor speed.

From these four, eventually two were selected to serve as principal cognitive outcome variables in PROSPER on the basis of their applicability and reliability:

- (1) memory: the delayed recall of the learning test;
- (2) cognitive speed: the number correct of the coding test.

The other test outcomes will be used for additional, post hoc analyses. Also, the extra tests provided useful cross validation of the main outcome variables: in both areas of cognition (memory and speed), several highly correlated outcome measures are needed for a proper interpretation of the concept. Finally, the first three trials of the learning test are necessary for sufficient imprinting of the stimuli before testing the delayed recall. This procedure cannot be avoided if an estimate of memory is needed.

Proposal for analysis and interpretation of follow up data

The existence of substantial practice effects prohibits direct comparison of repeated cognition measurements. Ideally, test scores of subjects receiving active treatment are compared with those of control subjects going through the same routine. However, as was pointed out, this can only be done in a randomised controlled trial. A straightforward way of analysing and presenting these data is to calculate difference scores, which allow for correction for practice effects. As Collins put it, there is nothing inherently unsound with this approach.¹⁶ We therefore propose that individual difference scores be the primary source for statistical analysis in longitudinal work in medical, pharmacological, and (neuro)psychological studies. In experimental settings, the intervention effect on the difference scores can be compared with the performance change in the control group. A further advantage being that difference scores may be normally distributed, even when the original outcomes are not, fulfilling a major condition for parametric analyses (linear regression, analysis of variance).

Just as the absolute test outcomes can be used for individual test performance, the difference scores may serve as norms for individual change. An example may serve to clarify this. Suppose, a 75 year old subject remembers nine words at the delayed recall. This is $(9-9.84)/2.50 = -0.34$ standard deviation below average for his/her age. At a second measurement, eight words are remembered, amounting to a difference score of $8-9 = -1$. This is $(-1-0.35)/2.15 = -0.63$ standard deviation below average, indicating no severe deterioration.

The standard deviations for these changes will be useful in power calculations whereas the age related trends will help to define effect sizes that would be clinically meaningful to achieve. Inspection of these tables also shows that the differences between the various age bands tend to be small, also relative to the practice effects, but within PROSPER they are highly statistically significant. This means that a clinical trial

such as PROSPER is powered to reveal even small differences in cognitive function between the treatment and the placebo group. In the PROSPER study it is hypothesised that cardiovascular risk reduction using the HMG-CoA-reductase inhibitor pravastatin as a cholesterol lowering drug will reduce cognitive impairment in elderly people, as there is accumulating evidence that cognitive impairment and dementia at old age is caused by atherosclerotic disease.¹⁷ Group differences such as one or two words at the learning test or one or two entries correct at the coding test are equivalent to several years of age related decline and are thus of utmost clinical importance.

ACKNOWLEDGEMENTS

This work was supported by an unrestricted grant from Bristol-Myers Squibb Company. Several of the authors received research support, staff support, consultancy fees, and reimbursement for invited lectures from various companies who manufacture drugs from the statin class.

.....

Authors' affiliations

P J Houx, University Maastricht, Maastricht, The Netherlands
J Shepherd, I Ford, D J Stoff, A Gaw, J Norrie, P W MacFarlane, C J Packard, S M Cobbe, University Glasgow, Glasgow, Scotland
G-J Blauw, E L Bollen, W Jukema, A M Kamper, A Edo Meinders, R G Westendorp, Leiden University Medical Centre, Leiden, The Netherlands
M B Murphy, B Buckley, M Hyland, I J Perry, B J Sweeney, C Twomey, University Cork, Cork, Ireland

REFERENCES

- 1 **Rabbitt P**, Donlan C, Watson P, *et al.* Unique and interactive effects of depression, age, socioeconomic advantage, and gender on cognitive performance of normal healthy older people. *Psychol Aging* 1995;**10**:307-13.
- 2 **Jolles J**. Early diagnosis of dementia: possible contributions from neuropsychology. In: Gispen WH, Traber J, eds. *Aging of the brain*. Berlin: Springer, 1985:84-100.
- 3 **Folstein MF**, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;**12**:189-98.
- 4 **Rasmussen LR**, Larsen K, Houx P, *et al.* The assessment of postoperative cognitive function in research. *Acta Anaesthesiol Scand* 2001;**45**:275-89.
- 5 **Jolles J**, Houx PJ, Bostel MPJv, *et al.*, eds. *The Maastricht Aging Study: Determinants of cognitive aging*. Maastricht: Neuropsych Publishers, 1995.
- 6 **Shepherd J**, Blauw GJ, Murphy MB, *et al.* The design of a prospective study of pravastatin in the elderly at risk (PROSPER). *Am J Cardiol* 1999;**84**:1192-7.
- 7 **Møller JT**, Cluitmans P, Rasmussen L, *et al.* Prolonged postoperative cognitive dysfunction in the elderly. *Lancet* 1998;**351**:857-61.
- 8 **Brand N**, Jolles J. Learning and retrieval rate of words presented auditorily and visually. *J Gen Psychol* 1985;**112**:201-10.
- 9 **Deelman BG**, Brouwer WH, Zomer AHv, *et al.* Functiestoornissen na trauma capitis. In: Jennekens-Schinkel A, Diamant JJ, Diesfeldt HFA, *et al.*, eds. *Neuropsychologie in Nederland*. Deventer: Van Loghum Slaterus, 1980.
- 10 **Rey A**. L'examen psychologique dans les cas d'encéphalopathie traumatique [Psychological assessment in the case of traumatic encephalopathy]. Paris: Presses Universitaires de France, 1964.
- 11 **Houx PJ**, Jolles J, Vreeling FW. Stroop interference: aging effects assessed with the Stroop Color-Word Test. *Exp Aging Res* 1993;**19**:209-24.
- 12 **Klein M**, Ponds RWHM, Houx PJ, *et al.* Effect of test duration on age-related differences in Stroop interference. *J Clin Exp Neuropsychol* 1997;**19**:66-81.
- 13 **Smith A**. The Symbol Digit Modalities Test: a neuropsychologic test for economic screening of learning and other cerebral disorders. *Learning Disorders* 1968;**3**:83-91.
- 14 **MacLeod CM**, Dunbar K. Training and Stroop-like interference: evidence for a continuum of automaticity. *J Exp Psychol Learn Mem Cogn* 1988;**14**:126-35.
- 15 **Lowe C**, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. *Neuropsychologia* 1998;**36**:915-23.
- 16 **Collins LM**. Measurement of change in research on aging: old and new issues from an individual growth perspective. In: Birren JE, Schaie KW, eds. *Handbook of the psychology of aging*. 4th edn. San Diego: Academic Press, 1996:38-56.
- 17 **Anonymous**. Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. Neuropathology Group of the Medical Research Council Cognitive Function and Ageing Study (MRC CFAS). *Lancet* 2001;**357**:169-75.