People with schizophrenia have been shown to have a wide range of cognitive deficits, but the longitudinal course of these deficits remains unclear. Some studies have suggested that cognitive function deteriorates over time, whereas others have reported stability or even an improvement in some functions. This issue is important for several reasons. First, it could help to resolve the perennial debate between those who conceptualise schizophrenia as a neurodegenerative, progressive disorder and those who view it as a stable deficit stemming from a developmental defect. Second, cognitive deficits have come to be considered as potential endophenotypes that might facilitate the identification of genetic factors involved in vulnerability to schizophrenia. For a measure to be considered a good endophenotype, it must be stable over time or the non-genetic factors influencing it must be identified and used to obtain a standardised (adjusted) value. Finally, there is growing interest in improving cognitive functions in patients with schizophrenia through pharmacological treatment or cognitive rehabilitation programmes, as a means of improving functional outcome. We therefore need to identify the cognitive deficits that are not fixed and the tests and variables that are most sensitive to change, to facilitate the monitoring of results.

In this study we analysed the available information by meta-analysis of all studies in which the same sample of people with schizophrenia underwent cognitive testing on two separate occasions more than 1 month apart.

**Method**

We based our search for relevant articles on three complementary procedures. We first searched the Medline database, limiting our search to: HUMAN, ADULT and PUBLICATION DATE 1978–2006. We identified relevant articles by searching for (SCHIZO* or PSYCHOTIC) and (COGNIT*) and (RETEST or LONGITUDINAL). The second term (COGNIT*) was replaced with a succession of terms describing cognitive domains (MEMORY, EXECUTIVE, ATTENTION) or the names of frequently used cognitive tests (WECHSLER MEMORY SCALE or WMS, WISCONSIN or WCST, TRAIL MAKING TEST or TMT, VERBAL FLUENCY, CONTINUOUS or CPT, STROOP, CALIFORNIA or CVLT, DIGIT SPAN). We also obtained additional references from the two practitioner’s guides to evaluating change edited by McCaffrey et al. We then screened the reference lists of these articles to find other studies on the subject. Finally, we manually searched recent issues (from January 2003 to March 2006) of seven major psychiatric journals considered relevant to this study: Schizophrenia Research, Archives of General Psychiatry, American Journal of Psychiatry, British Journal of Psychiatry, Psychiatry Research, Schizophrenia Bulletin and Journal of Nervous and Mental Disease.

**Selection of articles for inclusion in the meta-analysis**

We selected articles satisfying the following criteria:

(a) diagnosis of schizophrenia in patients (according to Research Diagnostic Criteria, DSM–III, DSM–III–R or DSM–IV, ICD–9 or ICD–10);

(b) participants aged over 18 years;

(c) results reported separately for each task (i.e. not only as composite scores) on both occasions;

(d) sufficient data available to calculate the effect (i.e. mean and standard deviation, $F$ or $t$ statistics) in at least three studies for each cognitive test;

(e) data independent of the other data included in our meta-analysis (if several articles dealt with the same population, we selected only the article with the largest sample);

(f) at least 1 month between test and retest and no specific training to improve performance in the test used (to minimise practice effects).
Recorded variables
From each study we extracted the following variables, if available:
(a) name of the first author and year of publication;
(b) the cognitive tests used;
(c) diagnosis of participants (i.e. only schizophrenia, or schizophrenia and other psychotic disorders);
(d) number of participants included at each evaluation;
(e) time (months) between test and retest;
(f) type of treatment and dosage (in chlorpromazine equivalents) at the time of the first and second evaluations. Type of treatment was recorded as a dichotomous variable: typical antipsychotic drugs (neuroleptics) v. atypical (novel) antipsychotic drugs;
(g) percentage change in positive and negative scores between test and retest, measured using the Scale for the Assessment of Negative Symptoms (SANS), the Scale for the Assessment of Positive Symptoms (SAPS) and the Positive and Negative Syndrome Scale (PANSS);
(h) results of tests (mean; s.d.) from the two evaluations or statistics (F, t) and direction of change if the mean and s.d. were unavailable; if data were provided for more than two evaluations we used only data from the first two, to minimise practice effects.

The variables (c) to (g) were recorded because they were considered potential moderators of the difference between the two measures of cognitive performance.

Meta-analytical procedure
We analysed each measure from each cognitive test separately, without grouping variables from several tests into composite scores for large cognitive domains. We felt that this strategy would be the most effective way to identify longitudinal changes in specific cognitive processes and the most useful variables for genetic research (those providing stable measures) or for assessing treatment impact (those changing over time).

Using the data reported in each study, we estimated the effect size by calculating Hedges’ unbiased g9 with positive values reflecting better performances on retesting (second evaluation) than at first evaluation. We tested the homogeneity of effect sizes using the I² statistic, as described by Higgins & Thompson.9 As suggested by these authors, values exceeding 30% were considered to indicate significant heterogeneity.

For homogenous data, we calculated the global effect size, using a fixed effect model as described by Hedges & Olkin.5 In the absence of significant heterogeneity, the use of a fixed effect model is legitimate and may provide greater statistical power than the random effect model. For heterogeneous studies we used the sample-adjusted meta-analytic deviancy (SAMD) and scree plots10 to identify studies with extreme values (outliers). The SAMD statistic compares the value of each study with the mean sample-weighted value calculated with that study excluded from the analysis, adjusting for the sample size of the study. A study is considered to be an outlier if its SAMD value is greater than 3.0 or if it produces a drastic break in the SAMD scree plot. When these procedures identified clear outliers (small numbers of studies isolated by a drastic break in the scree plot and with SAMD values exceeding 3), the data from these studies were removed and analyses were carried out as previously described. Data for outliers were analysed to identify, when possible, the origin of the heterogeneity. When the SAMD and scree plot procedures failed to identify outliers clearly we used a random effect model11 to calculate global effect size.

For comparison purposes we used the same analytical procedures to calculate the effects in control samples. A comprehensive evaluation of the differences between test and retest results in healthy controls was beyond the scope of this study. We therefore restricted our analysis to control samples from the studies included in our meta-analysis, together with data from studies cited by McCaffrey et al.6,7 For each variable, separate analyses were conducted for samples derived from studies included in our meta-analysis and for all samples (from our meta-analysis and from McCaffrey et al.).6,7

For the tests for which more than ten samples were available, we assessed the influence of potential moderator variables with a one-factor fixed-effect model.12 The potential moderator variables tested were time between the two evaluations; diagnosis of the participants (schizophrenia only, or schizophrenia and other psychotic disorders); change of treatment type v. same type of antipsychotic treatment (typical or atypical); difference in dosage of antipsychotic medication (chlorpromazine equivalents) between test and retest; percentage change in negative symptoms; and percentage change in positive symptoms.

Results
We identified 131 potentially relevant studies, 53 of which met the inclusion criteria. The selection process is illustrated in Fig. 1. The tests used for each study are listed in a data supplement to the online version of this paper. In total, 2476 people with schizophrenia participated in the 53 studies selected. The mean time between test and retest was 12 months (median 4 months) and the mean age of these participants was 37 years. Only 12 of the studies included in our analysis had a healthy control group (total number of control participants 324). We analysed the data for 11 tests (see online data supplement) and 31 variables available from more than three studies.

Analysis of effect sizes in the schizophrenia group samples
We analysed each test independently; however, for clarity the results are presented according to the categories of memory, executive function, attention and ‘other tests’.

Memory
The selected articles contained data for nine variables, from six tests exploring various aspects of memory. Two of these tests assessed visual memory: the Rey Complex Figure test and Visual Reproduction from the Wechsler Memory Scale (WMS). The remaining four assessed verbal memory: the California Verbal Learning Test (CVLT), Logical Memory from the WMS, the Hopkins Verbal Learning Test (HVLT) and the Rey Auditory Verbal Learning Test (RAVLT). The results are summarised in Table 1.

For three variables (RAVLT immediate recall, HVLT immediate recall and HVLT delayed recall), the data were heterogeneous. The SAMD and scree plot analyses failed to identify clear outliers for the two variables from the HVLT. Thus, for these variables the global effect was calculated by a random effect procedure. For the RAVLT, one study13 was identified as an outlier (SAMD 3.72, and a clear break in the scree plot). Following the exclusion of this study, the other data were homogeneous and the global effect was therefore calculated using a fixed effect model. Stip et al.13 obtained the largest improvement in the
Szoke et al.

but the intervals between assessments were shorter for participants, the third assessment. Other studies used a similar approach for the last observation carried forward, which was, for most studies, the Hopkins Verbal Learning Test (HVLT) and for others, the California Verbal Learning Test (CVLT). They reported results for cognitive assessment at baseline and after a second assessment.

In our meta-analysis, we identified six variables that showed significant improvement over the test–retest interval: immediate recall in the Rey Complex Figure test, significant improvements were observed in all tests except HVLT delayed recall. The characteristics of the tests (visual or verbal, recall or recognition, immediate or delayed) had no clear influence on the magnitude of improvement at retest (Table 1).

**Executive functions**

We were able to calculate a global estimate of the difference between test and retest for ten variables from five tests of executive functions (Table 2): (lexical and semantic Verbal Fluency, the Stroop test, the Wisconsin Card Sorting Test (WCST) and the Trail Making Test part B (TMT–B)).

Data were homogeneous for all but two variables: percentage perseverative error in the WCST and number of coloured words in the interference task of the Stroop test. For the percentage of perseverative errors in the WCST, the SAMD identified the study by Penades et al. as an outlier (SAMD 3.73, and break in the scree plot). The remaining data were homogeneous once the data from this study had been excluded. Penades et al. studied patients who underwent cognitive rehabilitation therapy between test and retest, potentially accounting for the greater improvement between test and retest observed in this study (estimated effect 1.35, 95% CI 0.72–1.98).

The results obtained for the number of coloured words in the interference task of the Stroop test were highly heterogeneous, with the effects in the various studies ranging from −0.32 to 0.99 (P=0.004). The use of different formats of this classic test in the various studies is the most likely explanation for this difference. Indeed, a 60 s format was used in one study, the 90 s format was used in two studies, a 100 s task in another study and two more studies referred to the original article by Stroop published in 1935 – which, however, did not use the number of coloured words as a response variable. Two studies did not describe the format of the Stroop task used. No clear outlier was identified through the SAMD or scree plot procedures, so global effect was calculated using the random effect procedure.

The estimated effects for the executive functions tasks ranged from 0.02 to 0.28, with slight improvements observed for certain tasks, and no significant difference for others.

**Attention**

In the studies included in our analysis, we identified six variables from four tests generally considered to measure attentional processes. The results for these variables are summarised below (Table 3).

The data for the Digit Span Distractibility Task were heterogeneous. We therefore carried out analyses to identify outliers.

---

**Table 1** Change between test and retest in variables derived from the memory tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Visual/verbal</th>
<th>Recall/recognition</th>
<th>Immediate/delayed</th>
<th>Studies n</th>
<th>Samples n</th>
<th>Participants n</th>
<th>Estimated effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rey Complex Figure</td>
<td>Visual</td>
<td>Recall</td>
<td>Immediate</td>
<td>5</td>
<td>9</td>
<td>161</td>
<td>0.53 (0.31 to 0.76)</td>
</tr>
<tr>
<td>WMS Visual Reproduction</td>
<td>Visual</td>
<td>Recall</td>
<td>Delayed</td>
<td>10</td>
<td>14</td>
<td>450</td>
<td>0.30 (0.16 to 0.44)</td>
</tr>
<tr>
<td>RAVLT</td>
<td>Verbal</td>
<td>Recall</td>
<td>Immediate</td>
<td>11</td>
<td>15</td>
<td>400</td>
<td>0.20 (0.06 to 0.34)</td>
</tr>
<tr>
<td>HVLT</td>
<td>Verbal</td>
<td>Recognition</td>
<td>Immediate</td>
<td>5</td>
<td>10</td>
<td>252</td>
<td>0.33 (0.10 to 0.55)</td>
</tr>
<tr>
<td>HVLT</td>
<td>Verbal</td>
<td>Recall</td>
<td>Delayed</td>
<td>3</td>
<td>6</td>
<td>167</td>
<td>0.33 (−0.05 to 0.70)</td>
</tr>
<tr>
<td>CVLT</td>
<td>Verbal</td>
<td>Recall</td>
<td>Immediate</td>
<td>4</td>
<td>4</td>
<td>159</td>
<td>0.32 (0.10 to 0.54)</td>
</tr>
<tr>
<td>WMS Logical Memory</td>
<td>Verbal</td>
<td>Recall</td>
<td>Immediate</td>
<td>10</td>
<td>13</td>
<td>380</td>
<td>0.30 (0.15 to 0.44)</td>
</tr>
<tr>
<td>WMS Logical Memory</td>
<td>Verbal</td>
<td>Recall</td>
<td>Delayed</td>
<td>9</td>
<td>12</td>
<td>360</td>
<td>0.30 (0.15 to 0.45)</td>
</tr>
</tbody>
</table>

CVLT, California Verbal Learning Test; HVLT, Hopkins Verbal Learning Test; RAVLT, Rey Auditory Verbal Learning Test; WMS, Wechsler Memory Scale/Wechsler Memory Scale Revised.

a Data from one study excluded (outlier).

b Calculated using a random effect model.
Cognition in schizophrenia

The SAMD for the study by Green et al\textsuperscript{17} was 4.07, and this study was therefore excluded from the analysis for this task. Green's study reported the greatest improvement of any of the studies using this test (estimated effect 1.01, 95% CI 0.48–1.56). No potential moderator variable that could account for the heterogeneity of the data was identified.

The estimated effects of attentional tasks ranged from 0.08 to 0.27, with only that for the Trail Making Test part A (time) being significant.

Other tests

Six other tasks that could not be grouped into any meaningful category were analysed (Table 4). Four were from the Wechsler Adult Intelligence Scale (WAIS): two verbal tasks (Similarities and Vocabulary), one task assessing psychomotor performance (Digit Symbol Substitution) and one assessing visuospatial conceptualisation (Block Design). The effects obtained for these tasks were small but (with the exception of Vocabulary) significant, showing a slight improvement between test and retest. The other two tests analysed, the Boston Naming Test and Rey Complex Figure (copy), showed a small, non-significant change between test and retest.

The estimated effects for all the tasks analysed are presented in graphical form to facilitate comparison (Fig. 2).

Potential moderator factors

We analysed the impact of six potential moderator factors (time between evaluations; participants' diagnosis; change in treatment; difference in chlorpromazine equivalents; percentage change in negative symptoms; percentage change in positive symptoms) on the test–retest difference for variables for which data from at least ten samples were available – i.e. 17 variables: Visual Reproduction (immediate and delayed recall); Logical Memory (immediate and delayed recall); Verbal Fluency (Letters and Categories); WCST (number of perseverative errors, percentage perseverative errors, number of categories completed); TMT–A

<p>| Table 2 Change between test and retest in variables derived from the executive tests |
|--------------------------------------------------|-----------------|----------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Test Variable</th>
<th>Studies n</th>
<th>Samples n</th>
<th>Participants n</th>
<th>Estimated effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Fluency (Letters) Total words</td>
<td>25</td>
<td>38</td>
<td>1186</td>
<td>0.20 (0.12 to 0.28)</td>
</tr>
<tr>
<td>Verbal Fluency (Categories) Total words</td>
<td>7</td>
<td>13</td>
<td>601</td>
<td>0.02 (-0.10 to 0.14)</td>
</tr>
<tr>
<td>Stroop Coloured words</td>
<td>8</td>
<td>11</td>
<td>358</td>
<td>0.28 (0.02 to 0.53)</td>
</tr>
<tr>
<td>TMT–B Errors</td>
<td>3</td>
<td>4</td>
<td>71</td>
<td>0.28 (-0.06 to 0.61)</td>
</tr>
<tr>
<td>TMT–B Time</td>
<td>27</td>
<td>36</td>
<td>1191</td>
<td>0.23 (0.14 to 0.31)</td>
</tr>
<tr>
<td>WCST Percentage of perseverative errors</td>
<td>8</td>
<td>11</td>
<td>297</td>
<td>0.21 (0.04 to 0.37)</td>
</tr>
<tr>
<td>WCST Total number of errors</td>
<td>5</td>
<td>8</td>
<td>287</td>
<td>0.17 (0.01 to 0.34)</td>
</tr>
<tr>
<td>WCST Number of perseverative errors</td>
<td>14</td>
<td>23</td>
<td>637</td>
<td>0.16 (0.05 to 0.28)</td>
</tr>
<tr>
<td>WCST Number of perseverative responses</td>
<td>5</td>
<td>7</td>
<td>144</td>
<td>0.13 (-0.11 to 0.36)</td>
</tr>
<tr>
<td>WCST Number of categories completed</td>
<td>22</td>
<td>29</td>
<td>1013</td>
<td>0.12 (0.03 to 0.21)</td>
</tr>
</tbody>
</table>

TMT–B, Trail Making Test part B; WCST, Wisconsin Card Sorting Test.  
\( a \) Calculated using a random effect model.  
\( b \) Data from one study excluded (outlier).

<p>| Table 3 Change between test and retest in variables derived from the attention tests |
|--------------------------------------------------|-----------------|----------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Test Variable</th>
<th>Studies n</th>
<th>Samples n</th>
<th>Participants n</th>
<th>Estimated effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Span Number of digits</td>
<td>11</td>
<td>15</td>
<td>330</td>
<td>0.13 (-0.03 to 0.28)</td>
</tr>
<tr>
<td>DSDT Correct responses</td>
<td>5</td>
<td>8</td>
<td>251</td>
<td>0.11 (-0.06 to 0.28)</td>
</tr>
<tr>
<td>Stroop Coloured dots</td>
<td>3</td>
<td>4</td>
<td>153</td>
<td>0.15 (-0.08 to 0.37)</td>
</tr>
<tr>
<td>Stroop Words</td>
<td>3</td>
<td>4</td>
<td>153</td>
<td>0.13 (-0.10 to 0.35)</td>
</tr>
<tr>
<td>TMT–A Time</td>
<td>17</td>
<td>23</td>
<td>840</td>
<td>0.27 (0.18 to 0.37)</td>
</tr>
<tr>
<td>TMT–A Errors</td>
<td>3</td>
<td>4</td>
<td>73</td>
<td>0.08 (-0.29 to 0.41)</td>
</tr>
</tbody>
</table>

\( \text{DSDT, Digit Span Distractibility Task; TMT–A, Trail Making Test part A.} \)

\( a \) Data from one study excluded (outlier).

<p>| Table 4 Change between test and retest in variables derived from the ‘other’ (not classified) tests |
|--------------------------------------------------|-----------------|----------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Test Cognitive process</th>
<th>Studies n</th>
<th>Samples n</th>
<th>Participants n</th>
<th>Estimated effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarities Verbal concept formation</td>
<td>4</td>
<td>8</td>
<td>152</td>
<td>0.36 (0.13 to 0.59)</td>
</tr>
<tr>
<td>Digit Symbol Psychomotor performance, sustained attention</td>
<td>7</td>
<td>11</td>
<td>215</td>
<td>0.28 (0.10 to 0.48)</td>
</tr>
<tr>
<td>Block Design Visuospatial conceptualisation</td>
<td>9</td>
<td>14</td>
<td>368</td>
<td>0.22 (0.08 to 0.37)</td>
</tr>
<tr>
<td>Vocabulary Vocabulary skills</td>
<td>4</td>
<td>5</td>
<td>73</td>
<td>0.10 (-0.23 to 0.42)</td>
</tr>
<tr>
<td>Rey Complex Figure test (copy) Perceptual organisation</td>
<td>5</td>
<td>9</td>
<td>157</td>
<td>0.09 (-0.14 to 0.31)</td>
</tr>
<tr>
<td>Boston Naming Test Verbal naming</td>
<td>4</td>
<td>6</td>
<td>437</td>
<td>-0.03 (-0.17 to 0.11)</td>
</tr>
</tbody>
</table>

251
and TMT–B (time); RAVLT (immediate recall); HVLT (immediate recall); Digit Symbol Substitution; Block Design; Digit Span; and Similarities. The interval between the two evaluations and the participants’ diagnoses were reported in all cases. Data for the other potential moderators were reported less frequently, but it was none the less possible – except for dosage of antipsychotic drugs – to analyse their influence on all 17 variables. We were able to analyse the influence of antipsychotic drug dosage on only 7 variables: Verbal Fluency (Letters), WCST (number of perseverative errors, percentage of perseverative errors, number of categories completed); time for TMT–A and B; and Block Design. The results of most of these analyses were not significant. The significant results are summarised below.

For TMT–B (time), patients who changed treatment from neuroleptics to novel antipsychotic drugs showed a significantly larger improvement in performance (estimated effect 0.33, 95% CI 0.19 to 0.47) than patients who remained on the same type of treatment (either neuroleptics or atypical antipsychotic drugs) for both test and retest (estimated effect 0.12, 95% CI 0.01 to 0.25). Similar results were obtained for Visual Reproduction (delayed recall). Patients who changed treatment between the two assessments performed better on retest (estimated effect 0.45, 95% CI 0.18 to 0.72) than those who remained on the same treatment (estimated effect 0.07, 95% CI –0.42 to 0.28). For the Block Design task, patients remaining on the same dosage of antipsychotic medication (in chlorpromazine equivalents) performed better (estimated effect 0.23, 95% CI –0.11 to 0.57) than those whose dosage of antipsychotic medication was decreased (estimated effect –0.06; 95% CI –0.36 to 0.25). The time between the two trials significantly affected performance in the Logical Memory (delayed recall) and Visual Reproduction (delayed recall) tests. As expected, the improvement in performance

**Fig. 2** Change (estimated effects and confidence intervals) in the cognitive variables analysed in participants with schizophrenia.

for these variables was inversely related to the time between test
and retest.

**Test–retest results for control groups**

The studies included in our analysis provided sufficient data for
estimation of the differences between test and retest in ‘internal’
controls for only 9 variables. With the addition of supplementary
data (samples of controls from other studies: ‘external’ controls),
we had sufficient data to estimate the effects for 19 of the 31
variables. Table 5 summarises our findings for the controls. We
present data for controls from studies included in our meta-
analyse and for these data combined with data from the studies
cited by McCaffrey et al3–5 – ‘all controls’. Data for the same
variables in participants with schizophrenia are provided for
comparison.

The data suggested a definite practice effect (improvement
statistically different from 0) in 10 of the 19 variables, a possible
practice effect (improvement between test and retest, but with a
confidence interval including 0) in 6 variables and no improve-
ment in the remaining 3 variables (with no variable significantly
deteriorating). Comparisons of those data with data from the
schizophrenia group may be summarised as follows. For variables
showing a significant practice effect, no significant difference was
observed for 6 variables and controls improved significantly more
than the participants with schizophrenia for the other 4 variables.
For the variables with a possible practice effect, improvement was
significantly greater for controls on one measure, with no signifi-
cant difference for the other five. Participants with schizophrenia
showed no significant difference from controls for 2 of the 3 vari-
bles for which no improvement was observed in controls, but a

<table>
<thead>
<tr>
<th>Test Variable</th>
<th>Studies Participants</th>
<th>Estimated effect (95% CI)</th>
<th>n</th>
<th>Estimated effect (95% CI)</th>
<th>Estimated effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logical Memory Immediate</td>
<td>4 93</td>
<td>0.54* (–0.31 to 1.38)</td>
<td>9 805</td>
<td>0.26* (0.01 to 0.52)</td>
<td>0.29 (0.14 to 0.44)</td>
</tr>
<tr>
<td>Logical Memory Delayed</td>
<td>Not enough data</td>
<td>0.46* (0.20 to 0.72)</td>
<td>9 704</td>
<td>0.30 (0.16 to 0.45)</td>
<td></td>
</tr>
<tr>
<td>Visual Reproduction Immediate</td>
<td>5 126</td>
<td>–0.28* (–0.87 to 0.32)</td>
<td>9 717</td>
<td>0.19* (–0.07 to 0.45)</td>
<td>0.15 (0.01 to 0.29)</td>
</tr>
<tr>
<td>Visual Reproduction Delayed</td>
<td>5 126</td>
<td>0.32* (–0.43 to 1.16)</td>
<td>7 518</td>
<td>0.37* (0.25 to 0.49)</td>
<td>0.30 (0.16 to 0.43)</td>
</tr>
<tr>
<td>WCST Categories completed</td>
<td>3 89</td>
<td>0.24* (–0.06 to 0.54)</td>
<td>5 129</td>
<td>0.19* (–0.05 to 0.44)</td>
<td>0.08 (–0.03 to 0.18)</td>
</tr>
<tr>
<td>WCST Errors</td>
<td>Not enough data</td>
<td>0.31* (–0.10 to 0.72)</td>
<td>3 46</td>
<td>0.17 (0.01 to 0.34)</td>
<td></td>
</tr>
<tr>
<td>WCST Perseverative errors</td>
<td>4 61</td>
<td>0.43 (0.05 to 0.82)</td>
<td>6 101</td>
<td>0.38* (0.10 to 0.66)</td>
<td>0.18 (0.04 to 0.31)</td>
</tr>
<tr>
<td>TMT–A Time</td>
<td>4 64</td>
<td>0.44* (0.13 to 0.76)</td>
<td>12 361</td>
<td>0.34* (0.19 to 0.48)</td>
<td>0.26 (0.15 to 0.36)</td>
</tr>
<tr>
<td>TMT–B Time</td>
<td>5 126</td>
<td>0.15* (–0.12 to 0.42)</td>
<td>15 541</td>
<td>0.36* (0.18 to 0.42)</td>
<td>0.18 (0.10 to 0.27)</td>
</tr>
<tr>
<td>Verbal Fluency Letters</td>
<td>3 75</td>
<td>0.28* (–0.04 to 0.61)</td>
<td>11 582</td>
<td>0.28* (0.17 to 0.40)</td>
<td>0.18 (0.09 to 0.26)</td>
</tr>
<tr>
<td>Verbal Fluency Categories</td>
<td>Not enough data</td>
<td>–0.10* (–0.18 to 0.19)</td>
<td>5 198</td>
<td>0.02 (–0.10 to 0.14)</td>
<td></td>
</tr>
<tr>
<td>Boston Naming Test Immediate</td>
<td>Not enough data</td>
<td>0.06 (–0.11 to 0.24)</td>
<td>5 245</td>
<td>0.02 (–0.17 to 0.11)</td>
<td></td>
</tr>
<tr>
<td>Stroop words</td>
<td>Not enough data</td>
<td>0.17* (–0.11 to 0.45)</td>
<td>4 98</td>
<td>0.13 (–0.10 to 0.35)</td>
<td></td>
</tr>
<tr>
<td>Stroop coloured dots</td>
<td>Not enough data</td>
<td>–0.06 (–0.37 to 0.25)</td>
<td>3 80</td>
<td>0.15 (–0.08 to 0.37)</td>
<td></td>
</tr>
<tr>
<td>Stroop coloured words</td>
<td>Not enough data</td>
<td>–0.03* (–0.27 to 0.21)</td>
<td>6 134</td>
<td>0.28* (0.02 to 0.53)</td>
<td></td>
</tr>
<tr>
<td>Similarities</td>
<td>Not enough data</td>
<td>0.15 (0.03 to 0.27)</td>
<td>5 513</td>
<td>0.31 (0.10 to 0.52)</td>
<td></td>
</tr>
<tr>
<td>Block Design</td>
<td>Not enough data</td>
<td>0.16* (0.06 to 0.27)</td>
<td>10 680</td>
<td>0.22 (0.10 to 0.35)</td>
<td></td>
</tr>
<tr>
<td>Digit Span</td>
<td>Not enough data</td>
<td>0.09 (–0.04 to 0.22)</td>
<td>10 472</td>
<td>0.13 (–0.03 to 0.28)</td>
<td></td>
</tr>
<tr>
<td>Digit Symbol Substitution</td>
<td>5 136</td>
<td>0.38 (0.13 to 0.63)</td>
<td>18 1115</td>
<td>0.21* (0.13 to 0.30)</td>
<td>0.29 (0.20 to 0.39)</td>
</tr>
</tbody>
</table>

Table 5: Changes in the cognitive variables in the control samples compared with changes in the schizophrenia samples

Several reviews14,18 and meta-analyses19–21 have already been published on the topic of cognitive change in schizophrenia. Overall, these analyses suggest that there is no progressive impairment for most cognitive variables and that, for some cognitive domains, improvement is possible after the onset of the disorder. These previ-
ous studies were, in our view, subject to several limitations. First,
with the exception of Rund’s review, they aimed to examine the
effects of medication on cognitive functions in schizophrenia. This
resulted in the exclusion of several primary studies and over-
representation of studies with short test–retest intervals. Another

**Discussion**

Several controls from the studies included in the meta-analysis combined with samples from other studies.

b. Calculated using a random effect model.

c. Significantly older than schizophrenia groups included in our meta-analysis.

d. Interval between the two evaluations significantly longer than in the schizophrenia groups included in our meta-analysis.

5 for difference compared with participants with schizophrenia.
The observed improvements in the performances of participants with schizophrenia may result from real improvements in cognition, a practice (learning) effect or a combination of the two. In samples of adult controls, improvements in cognitive performance assessed on two separate occasions are mostly due to practice effects. In older control group participants, this effect may be combined with a slight deterioration of performance (especially in memory and timed tests).

We were able to estimate the test–retest effect in controls for only 19 of the 31 variables. Test variables showing a possible or definite practice effect among patients with schizophrenia fell into two categories: variables for which the schizophrenia and control groups showed similar improvement, and variables for which the improvement was smaller than expected in the schizophrenic group. In other words, for these tests, the improvement in the schizophrenia group never exceeded the practice effect. As the control groups were older than the schizophrenia groups for most of the variables (and the test–retest interval was also longer for two of the variables), differences between control and schizophrenia groups might have been underestimated. For only one variable (number of words in the interference task of the Stroop test), which showed no practice effect, improvement was greater in the schizophrenic group than in controls. However, this result should be interpreted with caution because the total control group was small and the data for the schizophrenia group were heterogeneous. These data suggest that, for most variables, the practice effect alone might account for the improvement observed in
people with schizophrenia and might mask an actual deterioration in some cognitive processes.

Given the size and extensive impact of the practice effect, this effect should be taken into account in the design of future studies. The use of a control group is therefore of paramount importance, to ensure that results can correctly be interpreted as indicating improvement or deterioration in the cognitive abilities of people with schizophrenia. For example, in our meta-analysis, patients with schizophrenia showed similar improvements for delayed recall in the Logical Memory task and in the number of words in the interference condition of the Stroop test; however, when these results were compared with those for controls, conflicting interpretations were obtained (deterioration in the memory task but improvement in the Stroop test). It may also be important to match the two groups – schizophrenia and controls – not only in terms of demographic characteristics (age, gender, etc.), but also for familiarity with the tests used and, more generally, familiarity with testing situations. This is likely to be true for longitudinal studies, but is probably even more important for studies comparing the performances of different populations with a single evaluation.

### Semantic verbal fluency as a potential endophenotype

Our results suggest that semantic Verbal Fluency (Categories), for which stable results were obtained in patients (estimated effect 0.02) and a slight (statistically non-significant) decrease over time was observed in controls (estimated effect −0.10), may represent the most promising potential endophenotype. The slight decrease observed in controls might have resulted from the inclusion of a large number of older people in the group (mean age 73.4 years). There are also other arguments to support the use of semantic verbal fluency as a potential endophenotype. The Categories Verbal Fluency test is one of the measures showing the highest degree of impairment in patients with schizophrenia and controls – not only in first-degree relatives of such patients. Furthermore, in our analysis Verbal Fluency (Letters) was observed in controls might have resulted from the inclusion of a large number of older people in the group (mean age 73.4 years). There are also other arguments to support the use of semantic verbal fluency as a potential endophenotype. The Categories Verbal Fluency test is one of the measures showing the highest degree of impairment in patients with schizophrenia and controls – not only in first-degree relatives of such patients.

Lexical verbal fluency, which is similar to semantic verbal fluency in the test format and, to some extent, in the cognitive processes involved (e.g. general retrieval), does not share these qualities. Lexical verbal fluency is less impaired than semantic verbal fluency in people with schizophrenia and in their relatives. Furthermore, in our analysis Verbal Fluency (Letters) scores improved significantly in both participants with schizophrenia and controls.

Keefe et al. and Woodward et al. identified verbal fluency as one of the cognitive domains showing significant improvement in patients treated with novel antipsychotic drugs. Heinrichs & Zakzanis found a strong trend for patients taking high dosages of medication (chlorpromazine equivalents) to show lower levels of verbal fluency impairment. However, all these studies used a composite score based on data from both the Letters and Categories Verbal Fluency measures, and they therefore do not contradict our conclusion that semantic verbal fluency is stable.

The only other test showing similar, stable results in control and schizophrenia groups was the Boston Naming Test, which is also sensitive to the integrity of the semantic store.

### Limitations

The results of our analysis must be interpreted bearing its limitations in mind. Most of these limitations result from the small number of primary studies available, and from heterogeneity in the tests used and in data collection and reporting. More than half the theoretically relevant studies were excluded for various reasons.

Some studies were excluded because of major methodological differences (e.g. inclusion of patients with diagnoses other than psychotic disorders), but 39 studies were excluded simply because data for individual tests were not provided. This clearly represents a major loss of information, although it is not clear what effect this information would have had on our results. Differences in the variables reported limited our ability to detect a significant effect of moderators. Finally, the lack of a healthy control group in most studies limited the interpretation of the results. We tried to palliate this problem by including controls from other studies, but this resulted in large differences in demographic (e.g. age) and study (e.g. time between test and retest) characteristics. Furthermore, we did not carry out a systematic review of the data for controls and we included no recent data (subsequent to the publication of two books by McCaffrey et al.1,6,7)

**References**


18 Henry JD, Crawford JR. A meta-analytic review of verbal fluency deficits in schizophrenia relative to other neurocognitive deficits. Cogn Neuropsychiatry 2005; 10: 1–33.


William Battle’s Treatise on Madness (1758) and John Monro’s Remarks on Dr Battle’s Treatise (1758)

William Battle, resident physician and driving force behind the foundation of St Luke’s asylum, published what was probably the first English medical monograph devoted to madness in 1758. A proponent of Enlightenment pedagogy, Battle advocated an optimistic view of the treatability of insanity, by management – rather than the ineffectual and brutal purges, vomits and blood-letting then regularly practiced at Bethlem.

‘Madness is frequently taken for one species of disorder, nevertheless, when thoroughly examined, it discovers as much variety with respect to its causes and circumstances as any distemper whatever: Madness, therefore, like most other morbid cases, rejects all general methods, e.g. bleeding, blisters, caustics, rough catharticks, the gurms and faecid anti-hysterics, opium, mineral waters, cold bathing and vomits.’

John Monro, physician of Bethlem like his father before him, read Battle’s Treatise as a barely veiled attack on the elder Monro’s running of the ancient hospital. With filial piety he rushed to publish his Remarks in defence of his father’s regime.

‘Notwithstanding we are told in this treatise, that madness rejects all general methods, I will venture to say, that the most adequate and constant cure of it is by evacuation; which can alone be determined by the constitution of the patient and the judgment of the physician. The evacuation by vomitting is infinitely preferable to any other, if repeated experience is to be depended on . . .’

Dynastic stagnation fumigated against Enlightenment progress and humanity.

Battle’s Treatise was a turning point in the medical approach to mental illness. His division of madness into ‘original’ and ‘consequential’ illnesses were forerunners to the ‘organic’ and ‘functional’ terms used to this day, and his promotion of therapeutic optimism through engagement with the patient, rather than restraint and other physical affronts, prefigured the ‘moral therapy’ of the Tukes at the York Retreat later in the 18th century.

Then again, William Battle was an ambitious career physician. His attack on Monro’s practice was as much an attempt to break that family’s near monopoly on mad-doctoring in London – both John and his father James Monro operated the Bethlem as a private fiefdom, with no medical students allowed in to learn their trade (although sightseers could gape at the inmates for a penny a time). Despite Bethlem’s cruel reputation, it had a waiting list – there was clearly a substantial demand for madhouse places. In parallel with his position at St Luke’s (only a stone’s throw near monopoly on mad-doctoring in London – both John and his father James Monro operated the Bethlem as a private fiefdom, with no medical students allowed in to learn their trade (although sightseers could gape at the inmates for a penny a time). Despite Bethlem’s cruel reputation, it had a waiting list – there was clearly a substantial demand for madhouse places. In parallel with his position at St Luke’s (only a stone’s throw from Bethlem and operating in direct competition), Battle ran his own private madhouses. Whatever else Battle’s ‘management’ might have entailed, confinement (at a fee) was prerequisite – and he died a very, very rich man.

Andrew Morris, Scientific Editor, British Journal of Psychiatry
## Data supplement

### Studies included in the meta-analysis

<table>
<thead>
<tr>
<th>Study</th>
<th>Rey Figure</th>
<th>WMS Visual Reproduction</th>
<th>BAVLT</th>
<th>HVLT</th>
<th>CVLT</th>
<th>WMS Logical Memory</th>
<th>Verbal Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albus et al (2002)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Bilder et al (2002)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Bryan &amp; Tameirine (2000)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buchanan et al (1994)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chua et al (2004)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cuesta et al (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de Reus et al (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deutsch et al (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feldt et al (1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fujii et al (1993)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>GARD et al (1996)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goldberg &amp; Leonardi (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green et al (1997)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green et al (2002)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hagger et al (1993)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harris et al (1997)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvey et al (1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvey et al (2005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hof et al (2001)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoff et al (1996)</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hughes et al (2002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Khan et al (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pfister et al (2004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puskar et al (1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purdon et al (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resnick et al (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resnick et al (1997)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell et al (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARG (1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seidman et al (1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seidman et al (1999)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shmueli et al (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smith et al (2002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stip et al (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stirling et al (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swanson et al (1991)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Towse et al (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van der Gaag et al (2002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verfaille et al (2002)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widdershans et al (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verdoux et al (2006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CVLT, California Verbal Learning Test; DSOFT, Digit Span Distractibility Test; HVLT, Hopkins Verbal Learning Test; INHCT, Rey Auditory Verbal Learning Test; TMT, Trail Making Test; WAIS, Wechsler Adult Intelligence Scale; WCST, Wisconsin Card Sorting Test; WMS, Wechsler Memory Scale.

Longitudinal studies of cognition in schizophrenia: meta-analysis
Andrei Szőke, Anca Trândafir, Marie-Estelle Dupont, Alexandre Méary, Franck Schürhoff and Marion Leboyer
Access the most recent version at DOI: 10.1192/bjp.bp.106.029009

Supplementary Material
Supplementary material can be found at: http://bjp.rcpsych.org/content/suppl/2008/04/01/192.4.248.DC1

References
This article cites 69 articles, 2 of which you can access for free at: http://bjp.rcpsych.org/content/192/4/248#BIBL

Reprints/permissions
To obtain reprints or permission to reproduce material from this paper, please write to permissions@rcpsych.ac.uk

You can respond to this article at /letters/submit/bjprcpsych;192/4/248

Downloaded from http://bjp.rcpsych.org/ on September 8, 2017
Published by The Royal College of Psychiatrists