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,\(^1,2\) whereas others have reported stability or even an improvement in some functions.\(^3,4\) 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.\(^5\) 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 (REFETEST 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 cognitive remediation to account for most of the improvements observed. Semantic verbal fluency may be the best candidate cognitive endophenotype.

### Results

Participants with schizophrenia showed a significant improvement in most cognitive tasks. The available data for controls showed, with one exception (the Stroop test), a similar or greater improvement. Performance in semantic verbal fluency remained stable in both individuals with schizophrenia and controls.

### Conclusions

Participants with schizophrenia displayed improvement in most cognitive tasks, but practice was more likely than cognitive remediation to account for most of the improvements observed. Semantic verbal fluency may be the best candidate cognitive endophenotype.

### Declaration of interest

None.

### 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 schizo-
    phrenia 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) anti-
    psychotic 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 statis-
    tics (F, t) and direction of change if the mean and s.d. were
    unavailable; if data were provided for more than two evalua-
    tions we used only data from the first two, to minimise
    practice effects.

The variables (c) to (g) were recorded because they were con-
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 g*, 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.1 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
identified 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 hetero-
genesis. 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 pro-
cedures to calculate the effects in control samples. A comprehen-
sive 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.).8,9

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 on-
line 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 hetero-
genous. 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 pro-
cedure. 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 homogenous and
the global effect was therefore calculated using a fixed effect
model. Stip et al13 obtained the largest improvement in the...
Improvements were observed in all tests except HVLT delayed immediate recall in the Rey Complex Figure test. Significant immediate recall in the Visual Reproduction test to 0.53 for potentially accounting for the observed results.

In Stip’s study (4 weeks), increasing the effect of practice and design,14,15 but the intervals between assessments were shorter participants, the third assessment. Other studies used a similar and for the last observation carried forward, which was, for most studies, n=53

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 0.43)</td>
</tr>
<tr>
<td>WMS Visual Reproduction</td>
<td>Visual</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>RAVLT*</td>
<td>Verbal</td>
<td>Recall</td>
<td>Immediate</td>
<td>5</td>
<td>11</td>
<td>377</td>
<td>0.31 (0.16 to 0.45)</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 (0.15 to 0.55)*</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 (0.16 to 0.45)</td>
</tr>
</tbody>
</table>

RAVLT. They reported results for cognitive assessment at baseline and for the last observation carried forward, which was, for most participants, the third assessment. Other studies used a similar design,14,15 but the intervals between assessments were shorter in Stip’s study (4 weeks), increasing the effect of practice and potentially accounting for the observed results.

The estimated effects for memory tests ranged from 0.20 for immediate recall in the Visual Reproduction test to 0.53 for 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 al16 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 al16 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.

Fig. 1 Selection process for the articles included in the meta-analysis.
The SAMD for the study by Green et al 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).

### Table 2 Change between test and retest in variables derived from the executive tests

<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>
</tbody>
</table>

TMT-B, Trail Making Test part B; WCST, Wisconsin Card Sorting Test.

### Table 3 Change between test and retest in variables derived from the attention tests

<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>DSST 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 (-0.25 to 0.41)</td>
</tr>
</tbody>
</table>

DSST, Digit Span Distractibility Task; TMT-A, Trail Making Test part A.

### Table 4 Change between test and retest in variables derived from the ‘other’ (not classified) tests

<table>
<thead>
<tr>
<th>Test Variable</th>
<th>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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>Boston Naming Test Verbal naming</td>
<td>4</td>
<td>6</td>
<td>437</td>
<td>-0.02 (-0.17 to 0.11)</td>
<td></td>
</tr>
</tbody>
</table>

https://doi.org/10.1192/bjp.bp.106.029009 Published online by Cambridge University Press
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
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-analysis and for these data combined with data from the studies cited by McCaffrey et al.4,19 – ‘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 improvement 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 significant difference for the other five. Participants with schizophrenia showed no significant difference from controls for 2 of the 3 variables for which no improvement was observed in controls, but a significant improvement was observed in the number of words in the interference task of the Stroop test. Comparisons of data from ‘internal controls’ and participants with schizophrenia gave similar results, with no significant difference in improvement for 4 variables and significantly higher levels of improvement in controls for the other 5 variables.

We evaluated the differences between studies assessing controls or participants with schizophrenia by comparing the testing (interval between test and retest) and demographic (age of the participants) characteristics available for all samples. The significant differences are indicated in Table 5. Similar improvements in performance were observed in controls and in participants with schizophrenia whose medication type changed, for the TMT–B (0.03 to 0.44) and for Visual Reproduction (delayed recall) (0.37 to 0.49), whereas for the same variables, controls showed a significantly larger improvement in performance than patients remaining on the same medication.

**Discussion**

Several reviews4,19 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 previous 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
limitation is that the results of these studies were presented as composite scores for broad cognitive functions (e.g., executive functions, memory). This is a convenient way to summarise findings, but it combines data from tests exploring different cognitive processes (e.g., memory tests assessing immediate or delayed recall, learning or recognition), generating results of questionable theoretical relevance. Furthermore, the results obtained do not help researchers to select the most useful tests or variables based on stability or sensitivity to change. Another important methodological problem is that changes in the samples of people with schizophrenia were not compared with changes in control groups. Thus, when improvement is observed it is unclear how much is due to learning or practice effects and how much is due to a genuine improvement in cognitive abilities. We tried to overcome these limitations by including in our meta-analysis all the available studies in which the same group of people with schizophrenia underwent cognitive testing on two occasions more than 1 month apart. We analysed each variable separately and, when possible, we also calculated changes over time in the cognitive performances of normal controls.

Given the methodological differences cited above, it is interesting to compare our results with those of previous studies, especially that of Woodward et al., which is the most recent of these studies and used similar statistical methods. Our estimated effect sizes were smaller than those in Woodward’s study and the estimated effects showed a broader distribution (from 0.02 to 0.53 vs. 0.17 to 0.46 in Woodward’s study). Our estimated effect sizes might be lower because we tried to limit the effect of practice by including only studies in which the test–retest interval was greater than 1 month (Woodward et al. also included studies with a test–retest interval of between 1 week and 1 month), and by using only the results of the first and second evaluations from studies reporting several successive evaluations (whereas Woodward et al. used the first and last evaluations). The broader distribution of the estimated effects probably results from the effects being reported separately for each variable. However, there are also some similarities in the results reported by these two meta-analyses. In both Woodward’s review and our own analysis, the greatest differences in effect sizes were observed in the memory tests (Learning and Delayed Recall domains, 0.46 and 0.43 respectively), with tests of ‘cognitive flexibility and abstraction’ (0.38) and ‘vigilance and attention’ (0.35) showing milder improvement.

**Influence of moderator factors**

When enough data were available we assessed the effect of potential moderator factors on cognitive changes between test and retest. Participants with schizophrenia showed significantly greater improvements in performance if the total antipsychotic dosage was maintained (for Block Design), if the test–retest interval was shorter (delayed recall for Logical Memory and Visual Reproduction) and if treatment was changed from conventional to novel antipsychotic drugs (TMT–B time and Visual Reproduction delayed recall tests).

The two tests showing a greater improvement with shorter test–retest intervals were both memory tests. Memory measures are among the most susceptible to the effects of practice and the test–retest interval has a substantial influence on the magnitude of the practice effect. Thus, the pattern of results for these two tests can be accounted for by the effect of practice. For two other variables, TMT–B time and Visual Reproduction delayed recall, a significantly greater improvement was observed if the patient’s medication had been switched from a conventional to a novel antipsychotic drug. These results may suggest that atypical antipsychotic drugs have more beneficial effects on cognition than typical neuroleptics. However, this difference was relatively small and limited to a few cognitive variables (significant improvements in only 2 of the 17 cognitive variables tested).

Furthermore, there are at least two potential sources of bias that might lead to these conclusions: publication bias (particularly for studies sponsored by pharmaceutical companies), and the fact that changes between the two antipsychotic drug categories were always in the same direction (conventional to novel antipsychotic drugs). Patients are generally included in such studies because of the inefficacy of their current treatment and/or the presence of adverse effects. Thus, the cognitive improvement might result from the withdrawal of an ineffective treatment or the removal of an adverse effect, rather than from a specific positive action of a new antipsychotic drug. If treatment change is itself the factor associated with improvement, then changes in medication for the same reasons (inefficacy and/or adverse effects) in the other direction (i.e. from atypical to typical antipsychotic drugs) should also result in cognitive improvement. To our knowledge, this hypothesis has not been tested. In addition, some of the observed differences may not be due to the specific action of the two classes of antipsychotic drug, and may instead be due to atypical antipsychotic drugs having fewer extrapyramidal adverse effects and, in some cases, normothymic effects. Thus, patients taking such medication require fewer prescriptions of anticholinergic and/or normothymic drugs, both of which are known to have mildly deleterious effects on cognition.

Overall, our results concerning the role of potential moderators must be regarded as exploratory and interpreted with caution for two reasons. First, not all the studies provided data, limiting our ability to assess the influence of some of these moderators. Second, the large number of statistical tests might have led to spurious findings due to type I errors.

**Comparison with performances in controls: role of the practice effect**

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 schizophrenia 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

---

254

https://doi.org/10.1192/bjp.bp.106.029009 Published online by Cambridge University Press
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 schizophrenia22 and in first-degree relatives of such patients.23,24

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 schizophrenia22 and in their relatives.23,24 Furthermore, in our analysis Verbal Fluency (Letters) scores improved significantly in both participants with schizophrenia and controls.

Keefe et al.9 and Woodward et al.38 identified verbal fluency as one of the cognitive domains showing significant improvement in patients treated with novel antipsychotic drugs. Heinrichs & Zakzanis27 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 potentially relevant studies were excluded for various reasons. Some studies were excluded because of major methodological issues (e.g. inclusion of patients with diagnoses other than psychiatric 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).6,7

Acknowledgements

This research was supported by grants from Institut National de la Santé et de la Recherche Médicale (INSERM) and the Délégation à la Recherche Clinique de l’Assistance Publique – Hôpitaux de Paris (PHRC AOM 98152). 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.

References

1 Goldberg TE, Weinberger DR, Berman KF, Plikas NH, Podd MH. Further evidence for dementia of the prefrontal type in schizophrenia? A controlled study of teaching the Wisconsin Card Sorting Test. Arch Gen Psychiatry 1987; 44: 1008–14.


Henry JD, Crawford JR. A meta-analytic review of verbal fluency deficits in schizophrenia. Schizophr Bull 2006; 32: 179–94.


William Battie ran his own private madhouses. Whatever else Battie’s ‘management’ might have 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), Battie ran his own private madhouses. Whatever else Battie’s ‘management’ might have

Dynastic stagnation fulminated against Enlightenment progress and humanity.

Battie’s ‘Treatise’ was a turning point in the medical approach to mental illness. His division of madness into ‘original’ and ‘consequential’ illnesses are 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 Battie was an ambitious career physician. His attack on Monro’s practice was as much an attempt to break that family’s

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 vomiting is infinitely preferable to any other, if repeated experience is to be depended on . . .’

John Monro, physician of Bethlem like his father before him, read Battie’s Treatise as a barely veiled attack on the elder Monro’s running of the

Then again, William Battie was an ambitious career physician. His attack on Monro’s practice was as much an attempt to break that family’s

Battie’s ‘Treatise’ was a turning point in the medical approach to mental illness. His division of madness into ‘original’ and ‘consequential’ illnesses are 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 Battie 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 from Bethlem and operating in direct competition), Battie ran his own private madhouses. Whatever else Battie’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

250 years ago

William Battie, 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, Battie advocated an optimistic view of the treatability of insanity, by management – rather than the ineffective 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 cathartics, the gums and faetid anti-hysterics, opium, mineral waters, cold bathing and vomits.’

John Monro, physician of Bethlem like his father before him, read Battie’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 vomiting is infinitely preferable to any other, if repeated experience is to be depended on . . .’

Dynastic stagnation fulminated against Enlightenment progress and humanity. Battie’s ‘Treatise’ was a turning point in the medical approach to mental illness. His division of madness into ‘original’ and ‘consequential’ illnesses are 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 Battie 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 from Bethlem and operating in direct competition), Battie ran his own private madhouses. Whatever else Battie’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

250 years ago

William Battie, 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, Battie advocated an optimistic view of the treatability of insanity, by management – rather than the ineffective 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 cathartics, the gums and faetid anti-hysterics, opium, mineral waters, cold bathing and vomits.’

John Monro, physician of Bethlem like his father before him, read Battie’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 vomiting is infinitely preferable to any other, if repeated experience is to be depended on . . .’

Dynastic stagnation fulminated against Enlightenment progress and humanity. Battie’s ‘Treatise’ was a turning point in the medical approach to mental illness. His division of madness into ‘original’ and ‘consequential’ illnesses are 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 Battie 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 from Bethlem and operating in direct competition), Battie ran his own private madhouses. Whatever else Battie’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