Bipolar disorder and schizophrenia share a similar deficit in semantic inhibition: A meta-analysis based on Hayling Sentence Completion Test performance

Kui Wang a,1, Li-Ling Song a,1, Eric F.C. Cheung b, Simon S.Y. Lui a,b,c, David H.K. Shum d, Raymond C.K. Chan a,⁎

a Neuropsychology and Applied Cognitive Neuroscience Laboratory, Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
b Castle Peak Hospital, Hong Kong Special Administrative Region, China
c University of Chinese Academy of Sciences, Beijing, China
d Behavioural Basis of Health Program, Griffith Health Institute and School of Applied Psychology, Griffith University, Brisbane, Australia

A B S T R A C T

Bipolar disorder (BD) is associated with deficits in executive function similar to that found in schizophrenia (SZ). However, very few studies have examined whether a specific component of executive function, namely, semantic inhibition, is differentially impaired in BD and SZ. The present study reports the results of a meta-analysis of performance on a theory-driven test of semantic inhibition, namely, the Hayling Sentence Completion Test (HSCT), in patients with BD and SZ, and to examine differential group impairments. The Comprehensive Meta-Analysis Software package was used to calculate the mean effect sizes for group differences on different measures of HSCT. A total of 13 studies were included in the meta-analysis. Effect sizes for six HSCT measures were calculated. These included: Total Latency of Task A, Total Latency of Task B, Suppression Time, Total Error of Task B, Type A Error of Task B, and Type B Error of Task B. When compared with healthy controls, medium-to-large effect sizes were observed in both groups for each HSCT measure. Interestingly, the effect sizes for BD and SZ groups were comparable. These results suggest that patients with SZ and patients with BD are impaired in both task initiation and task inhibition of executive function and these impairments are similar in magnitude for both disorders.

© 2013 Elsevier Inc. All rights reserved.
1. Introduction

Traditionally, while bipolar disorder (BD) has abnormal emotional processes as its key feature, schizophrenia (SZ) is characterized by cognitive deficits. However, recent studies have found that patients with BD also have deficits in various cognitive functions (Arts et al., 2008; Krabbendam et al., 2005; Quraishi and Frangou, 2002). Considering that the two disorders also share many similar symptoms, these recent findings raised the question about whether the observed cognitive deficits in the two patient groups are more related to diagnosis or to pattern of symptoms.

Various definitions of executive functions have been put forward (Burgess et al., 2000; Royall et al., 1993; Stoddart et al., 2007), with the one proposed by Burgess et al. (2000) being most widely accepted. According to them, executive function refers to a wide range of cognitive processes and behavioral abilities such as problem-solving, sequencing, verbal reasoning, the ability to sustain attention, planning, resistance to interference, utilization of feedback, multitasking, cognitive flexibility and the ability to deal with novelty. Among the various cognitive deficits identified, previous studies have highlighted executive dysfunctions in BD and SZ for several reasons. First, executive functions are closely associated with everyday functioning, and are probably most impaired compared to other cognitive processes in individuals with SZ (Frangou, 2010) and BD (Arts et al., 2008). Second, executive dysfunctions are related to symptoms observed in BD and SZ. Executive functions are primarily sub-served by the frontal lobes (Andreasen et al., 1996) and frontal lobe lesions are associated with both the failure to suppress inappropriate responses and the lack of responses (Shallice, 1988). While failure to suppress inappropriate responses is associated with thought disorder and reality distortion (delusion and hallucination) in SZ and mania in BD, lack of responses is associated with psychomotor poverty in SZ and depression in BD (Kraraviti et al., 2005). Third, impaired executive functions may be a common trait marker in BD and SZ (Breton et al., 2011; Frangou et al., 2005a; Morey et al., 2005).

In this study we focused on the inhibition function in individuals with BD and SZ because factorial-analytic studies have repeatedly shown that semantic inhibition is a very important factor in a battery of executive function tests, in both healthy (Chan, 2001) and patient populations (Chan et al., 2004). The Hayling Sentence Completion Test (HSCT) is a test developed to assess inhibition (Burgess et al., 2000) and it has been widely used in clinical practice. In the HSCT, individuals are presented with incomplete sentences with the final word omitted but is strongly suggested by the context. Individuals are asked to complete the sentence in either a logical (Task A, initiation section) or illogical manner (Task B, inhibition section). In Task B any word which is semantically associated with the sentence should be avoided, thus test takers have to inhibit a strongly cued and automatic response. For instance, responding with the word “ship” to the sentence “the captain went down with the sinking _” is correct when undertaking Task A, but incorrect when undertaking Task B (Type A error). Moreover, words such as ‘airplane’, ‘bus’, ‘waterman’ which are semantically associated with the whole sentence context are also scored incorrect for Task B (Type B error). Thus, in completing Task B, test takers are required not only to suppress a pre-potent response but also to plan and manipulate information in working memory. Shorter latency and fewer errors in Task A or Task B indicate better initiation or inhibition function.

Significant correlations between HSCT scores and self-reported measures of attentional impulsivity have been established in remitted BD patients, suggesting that poor response inhibition may be related to impulsivity in these patients (Christodoulou et al., 2006). Impaired HSCT performance has been observed in patients with different symptomatology, including patients with BD (Stoddart et al., 2007). Moreover, Dixon et al. (2004) studied the relationship between inhibition function and symptomatology by recruiting 15 manic, 15 depressed, 15 remitted patients with BD and 30 controls. Even with the modest number of participants in each group, the authors found that each of the three BD subgroups had longer latency, larger error rate in Task A, and decreased use of strategy in Task B (e.g., reporting objects in the testing environment). Similarly, compromised HSCT performance in euthymic BD patients (de Almeida Rocca et al., 2008) and remitted BD patients (Frangou et al., 2005b) have been reported. These findings, therefore, suggest that impaired performance in HSCT may be an enduring feature of BD and not a secondary deficit due to mood symptoms.

Nathaniel-James et al. (1996) first reported that patients with SZ had difficulties in performing the HSCT when compared to healthy controls. Patients with SZ showed longer response latency in Task A and more errors in Task B, indicating clear deficits in response initiation and inhibition. Since then, a large number of studies have repeatedly found that SZ patients exhibit impaired performance on the HSCT, especially in Task B where inhibition is needed (Chan et al., 2012; Chan and Chen, 2004; Chan et al., 2004, 2010; Groom et al., 2008; Joshua et al., 2009; Marczewski et al., 2001; Nathaniel-James et al., 1996, Royer et al., 2005a; Waters et al., 2003). Patients with SZ were found to either commit more errors, or had longer response latency, or both. Significant relationships between HSCT measures and symptoms in individuals with SZ have also been established (Chan et al., 2010; Waters et al., 2003). Results of previous study also suggest that patients with SZ probably showed the most severe impairment on the HSCT task in comparison with other executive function tasks such as verbal fluency and the Modified Wisconsin Card Sorting Test (Nathaniel-James et al., 2004). Average effect sizes for other executive function measures, such as verbal fluency (d = 1.39), the Stroop Color–Word Test (d = 1.22), the Trail Making Test B (d = 1.07) and the Wisconsin Card Sorting Test (d = 0.95) have been reported (Heinrichs and Zakkas, 1998). However, the average effect size for performance on the HSCT, a task in which patients with SZ have potentially the greatest difficulty, is still not known.

A small number of studies directly compared the performance on HSCT in individuals with BD and SZ. Kravariti et al. (2005) compared the performance on HSCT in 30 BD patients and 30 SZ patients. While the performance of BD patients in the manic stage was found to resemble those of patients with SZ with thought disorder and/or reality distortion (delusion and hallucination), the performance of BD patients in the depressive stage was found to be similar to SZ patients with negative symptoms. In another study, Joshua et al. (2009) compared HSCT performance between 39 patients with SZ and 40 patients with BD (as well as a healthy control group) on several measures, including the overall scaled score (which takes into consideration both response latency and error rate), the Task A scaled score, the Task B scaled score, response suppression (subtracting Task A response latency from Task B response latency), the Task B Error scaled score, Type A Error score in Task B and Type B Error score in Task B. Results of the study suggested that the overall scaled score difference observed between the two groups was mainly due to an increase in Type B Error in the SZ group. There was no reliable difference between the two groups on other HSCT measures. Results from Kravariti et al.’s (2005) and Joshua et al.’s (2009) study seem to suggest that there are more similarities than differences between the performance of patients with SZ and BD on the HSCT. However, more evidence is needed before a firm conclusion can be drawn. Apart from further investigation on this topic using large samples, a meta-analysis can be used as a good alternative method to clarify the issue.

Using published studies which compared either/or both patients with BD or SZ with healthy controls, we could compute effect sizes for the HSCT measures, and then obtain a grand effect size for each of these measures for both groups. The meta-analytic method also allows us to directly compare the effect sizes generated by comparing individuals with BD with healthy controls and those generated by comparing individuals with SZ and healthy controls. Therefore, the present meta-analysis had two aims. The first was to obtain a general profile of performance deficits on HSCT measures in patients with SZ and BD separately.
Secondly, we also aimed to compare the effect sizes generated for patients with BD and SZ directly. For any HSCT measures where a grand effect size could not be computed due to small number of available studies, a qualitative review based on published data was provided instead.

2. Methods

2.1. Literature search

The flowchart of data extraction for the meta-analysis of each HSCT measure is shown in Fig. 1. Potential articles used in the meta-analysis were identified from Elsevier, EBSCOHost (PsychINFO, PsychACTICLE) and MedLine databases between January 2000 and June 2012. The keywords used were ‘Hayling Sentence Completion Test’ or ‘HSCT’ or ‘executive functioning’ or ‘inhibition’, and ‘schizophrenia’ or ‘schizophrenic’ or ‘bipolar disorder’. Additional articles were obtained from the reference list of the articles identified from the databases. These search procedures yielded an initial pool of 32 potential articles.

The following inclusion criteria were used to select studies from the initial article pool for quantitative analysis:

(a) Patients met diagnostic criteria for SZ or BD according to various versions of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III, DSM-III-R, DSM-IV) or Research Diagnostic Criteria (RDC);
(b) The HSCT was administered to the patient group(s) (either patients with SZ or BD or both) and healthy controls;
(c) Means and standard deviations (SDs) for at least one of the HSCT measures were available, or exact t-values/p-values/F-values for comparing any HSCT measures between patients and controls were reported (note: one study used median and IQR).
(d) The article was written in English.

The exclusion criteria were:

(a) Studies that included data already reported in a previous study;
(b) Case studies.

Based on these criteria, the final number of articles used in the meta-analysis was 13. Among these, one included a group of BD and a group of SZ patients (Joshua et al., 2009); one included a group of BD patients in the manic phase, a group of BD patients in the depression phase, and a group of BD patients in the remitted phase (Dixon et al., 2004); and one included a group of patients with SZ with clear negative symptoms and a group of patients with SZ with clear thought disorders (Kravariti et al., 2005). For these three articles, each group of patients together with their corresponding healthy controls was regarded as an independent data set. Altogether, 17 valid data sets were used in the meta-analysis. The profile of studies and data sets included in the meta-analysis are presented in Table 1.

2.2. Meta-analysis procedure

The meta-analysis was carried out using the Comprehensive Meta-Analysis software package (Borenstein et al., 2005). Effect sizes for different HSCT measures were calculated as Cohen’s $d$ which is defined as the difference in group means divided by the pooled SDs for the two groups. When means and SDs were not given, the exact t-values, p-values or F-values were used to calculate the effect sizes. The HSCT measures included Total Latency of Task A, Total Latency of Task B, Suppression Time (subtracting Total Latency of Task A from Total Latency of Task B), Total Error of Task B, Type A Error of Task B, Type B Error of Task A, HSCT total scaled score, Task A scaled score, and Task B scaled score. Shorter latency and fewer errors in Task A indicate better initiation and shorter latency and fewer errors in Task B indicate better inhibition. According to Cohen (1992), a small effect size was defined as $d = 0.2$–0.5; a medium effect size as $d = 0.5$–0.8; and a large effect size as $d > 0.8$.

If the number of available studies for computing the effect size of a certain HSCT measure was larger than two, a weighted grand mean was calculated for each set of effect sizes. Random models were used throughout the meta-analysis. Q-statistics was calculated to evaluate the homogeneity of the studies. This statistics is used to test whether the effect sizes of the studies could be assumed to have come from a single population. A significant Q statistics suggests heterogeneity of the individual study effect sizes which implies a greater variation between studies. A fail-safe number was calculated to estimate the number of unpublished studies with nil or minimal effect sizes required to reduce an overall effect size to some specified negligible value (Rosenberg, 2005).
3. Results

Thirteen studies compared the HSCT scores of SZ or BD patients with healthy controls. The total number of patients included in these studies was 292 for SZ and 176 for BD. Mean effect sizes and their 95% CIs and Q-value are presented in Table 2.

### 3.1. Total Latency of Task A

Of the 13 studies, 13 sets of data were used to compare Total Latency of Task A. Among them, six data sets compared patients with BD and healthy controls and seven data sets compared patients with SZ and healthy controls. As seen in Table 2, for the BD group results, a significant effect size of 0.719 (0.231, 1.207) was obtained, Z = 2.888, p < 0.01, indicating that patients with BD performed significantly slower than healthy controls in Task A. However the Q-statistics indicated significant heterogeneity for the studies, Q(5) = 22.044, p < 0.001. The fail-safe number of studies was 48. As illustrated in Fig. 2A, the study by De Almeida Rocca et al. (2008) showed a low-to-medium effect size of 0.355 in which healthy controls had higher latency scores in Task A compared to SZ patients. After excluding this paper from the pool, a large effect size of 0.896 in which SZ patients had higher latency scores in Task A compared to healthy controls was obtained (0.592, 1.201). All the studies in the pool became homogenous (I² = 0.0). As illustrated in Fig. 2B, the study by De Almeida Rocca et al. (2008) showed a low-to-medium effect size of 0.140 in which healthy controls had higher latency scores in Task B compared to BD patients. After excluding this paper from the pool, a large effect size of 0.896 in which SZ patients had higher latency scores in Task B compared to healthy controls was obtained (0.592, 1.201). All the studies in the pool became homogenous (I² = 0.0).

### Table 1

Profile of studies and data sets included in the meta-analysis (numbers in parentheses are standard deviations). Note: a = Total Latency of Task A; b = Total Latency of Task B; c = Suppression time; d = Type A Error of Task A; e = Type B Error of Task A; f = Bipolar disorder; BD = Bipolar disorder; BDD = Bipolar disorder in depression phase; BDM = Bipolar disorder in manic phase; BDR = Bipolar disorder in remission; SZ = Schizophrenia; SZ-TD = Schizophrenia with predominant symptoms of psychomotor poverty; SZ-TD = Schizophrenia patients with predominant symptoms of disorganization.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Patient group</th>
<th>Number of patients</th>
<th>Number of controls</th>
<th>Mean age of patients</th>
<th>Mean age of controls</th>
<th>Mean IQ in patients</th>
<th>Mean IQ in controls</th>
<th>Duration of illness (yrs)</th>
<th>Years of education in controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chan et al., 2010</td>
<td>SZ 6 181</td>
<td>214</td>
<td>15.6</td>
<td>17.0</td>
<td>18.6 (6.6)</td>
<td>23.1 (2.1)</td>
<td>2.0 (0.8)</td>
<td>11.0 (3.0)</td>
<td>22.0 (6.0)</td>
</tr>
<tr>
<td>Dixon et al., 2004</td>
<td>BD 6:9</td>
<td>17:13</td>
<td>33.9 (8.2)</td>
<td>35.2 (9.8)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>63.5 (5.8)</td>
<td>13.9 (2.6)</td>
<td>12.8 (2.0)</td>
</tr>
<tr>
<td>Groom et al., 2008</td>
<td>BD 6:16</td>
<td>22:22</td>
<td>42.1 (11.5)</td>
<td>41.0 (12.0)</td>
<td>107.6 (10.3)</td>
<td>111.6 (11.1)</td>
<td>19.8 (11.6)</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Krishnan et al., 2010</td>
<td>BD 6:16</td>
<td>22:22</td>
<td>42.3 (10.7)</td>
<td>41.0 (12.0)</td>
<td>109.2 (9.4)</td>
<td>111.6 (11.1)</td>
<td>18.7 (9.9)</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Marczewski et al., 2001</td>
<td>BD 15:22</td>
<td>28:37</td>
<td>33.4 (11.0)</td>
<td>35.2 (9.8)</td>
<td>6.1 (2.1)</td>
<td>10.0 (2.2)</td>
<td>11.0 (3.0)</td>
<td>12.8 (2.0)</td>
<td>12.8 (2.0)</td>
</tr>
<tr>
<td>Nathaniel et al., 1996</td>
<td>BD 9:6</td>
<td>15 in total</td>
<td>29.93 (7.2)</td>
<td>30.4 (6.9)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>8.1 (5.8)</td>
<td>10.6 (2.0)</td>
<td>11.0 (2.0)</td>
</tr>
<tr>
<td>Nathaniel et al., 2010</td>
<td>BD 20:24</td>
<td>22:22</td>
<td>34.3 (11.6)</td>
<td>35.2 (9.8)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>10.7 (10.0)</td>
<td>13.1 (2.9)</td>
<td>14.2 (2.8)</td>
</tr>
<tr>
<td>Stoddart et al., 2007</td>
<td>BD 6:16</td>
<td>22:22</td>
<td>48.7 (9.7)</td>
<td>36.6 (14.3)</td>
<td>115.7 (11.6)</td>
<td>119.5 (10.3)</td>
<td>10.5 (7.8)</td>
<td>13.3 (2.8)</td>
<td>12.8 (2.0)</td>
</tr>
<tr>
<td>Waters et al., 2003</td>
<td>BD 35:7</td>
<td>20:4</td>
<td>36.73 (8.41)</td>
<td>34.67 (8.81)</td>
<td>100.2 (9.3)</td>
<td>103.6 (4.8)</td>
<td>13.64</td>
<td>11.0 (2.0)</td>
<td>11.8 (1.9)</td>
</tr>
</tbody>
</table>

### Table 2

Results of meta-analysis of differences in HSCT scores between healthy controls and patients with bipolar disorder and between healthy controls and patients with schizophrenia.

<table>
<thead>
<tr>
<th>HSCT measure</th>
<th>Patients</th>
<th>N</th>
<th>No. of patients</th>
<th>No. of HC</th>
<th>d*</th>
<th>SE</th>
<th>95% CI</th>
<th>Q value</th>
<th>Fail safe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Latency of Task A</td>
<td>BD 6</td>
<td>136</td>
<td>205</td>
<td>205</td>
<td>0.719</td>
<td>0.374</td>
<td>0.431, 0.847</td>
<td>22.044</td>
<td>48 (59)</td>
</tr>
<tr>
<td>SZ 7</td>
<td>184</td>
<td>221</td>
<td>0.195</td>
<td>0.367, 1.132</td>
<td>18.695</td>
<td>74 (78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Latency of Task B</td>
<td>BD 5</td>
<td>92</td>
<td>161</td>
<td>161</td>
<td>0.390</td>
<td>0.460</td>
<td>0.336, 0.514</td>
<td>14.778</td>
<td>52 (44)</td>
</tr>
<tr>
<td>SZ 4</td>
<td>109</td>
<td>136</td>
<td>0.156</td>
<td>0.461, 1.131</td>
<td>8.078</td>
<td>33 (33)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suppression time</td>
<td>BD 8</td>
<td>84</td>
<td>134</td>
<td>0.240</td>
<td>0.313, 0.653</td>
<td>9.1 (1.7)</td>
<td>12.58 (1.7)</td>
<td>11.74 (2.0)</td>
<td></td>
</tr>
<tr>
<td>SZ 5</td>
<td>132</td>
<td>153</td>
<td>0.124</td>
<td>0.065, 0.549</td>
<td>10.038</td>
<td>4 (10)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Error of Task A</td>
<td>BD 5</td>
<td>92</td>
<td>161</td>
<td>161</td>
<td>0.896</td>
<td>1.400</td>
<td>0.402, 1.330</td>
<td>11.68 (50)</td>
<td>45 (40)</td>
</tr>
<tr>
<td>SZ 8</td>
<td>181</td>
<td>266</td>
<td>0.944</td>
<td>0.698, 1.190</td>
<td>10.018</td>
<td>163 (163)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type A Error of Task B</td>
<td>BD 2</td>
<td>62</td>
<td>84</td>
<td>84</td>
<td>0.678</td>
<td>0.221</td>
<td>0.336, 1.021</td>
<td>5.828</td>
<td>7 (1)</td>
</tr>
<tr>
<td>SZ 6</td>
<td>181</td>
<td>214</td>
<td>0.106</td>
<td>0.431, 0.847</td>
<td>0.639</td>
<td>47 (47)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type B Error of Task B</td>
<td>BD 2</td>
<td>62</td>
<td>84</td>
<td>84</td>
<td>0.896</td>
<td>0.221</td>
<td>0.472, 2.211</td>
<td>13.561</td>
<td>1 (1)</td>
</tr>
<tr>
<td>SZ 6</td>
<td>181</td>
<td>214</td>
<td>0.170</td>
<td>0.247, 0.912</td>
<td>12.178</td>
<td>40 (40)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BD: bipolar disorder; CI: confidence interval; d: Cohen's d; HC: healthy control; N: number of data sets; SE: standard error. SZ: schizophrenia.

* p < 0.05.
** p < 0.01.
† Numbers in parentheses indicate the effect size after excluding outliers.
‡ Numbers in parentheses indicate the Fail-safe Number after excluding outliers.
IQ and more years of education than controls (though not statistically significant \( p = 0.07 \)). This could explain their unexpected findings. A positive correlation between IQ and Task A scaled score \((r = 0.56)\) has been reported elsewhere for BD patients (Joshua et al., 2009).

A significant effect size of \(0.749 \, (0.367 - 1.132)\) was obtained from seven data sets with SZ patients, \( Z = 3.840, p < 0.01 \). The fail-safe number was 74, which supported the validity of the results. The studies were heterogeneous, \(Q(6) = 18.695, p < 0.01\). As presented in Fig. 2A, the study by Chan et al. (2010) produced a small negative effect size of \(-0.077\). Among these studies, the study by Chan et al. (2010) was the only study that used Chinese participants. The HSCT is on the one hand language based and on the other hand is related to language ability (Task A specifically). In another study carried out in Hong Kong, the researchers did not report any difference in Task A measures between Chinese first-episode SZ patients and healthy controls (Chan et al., 2012). Chinese is a language that is very different from alphabetic languages in terms of orthographic form, orthographic–phonologic transformation, the way by which children learn the written language, and so on. It is therefore highly possible that there was a cultural effect related to language which affected the results of this study. After excluding the Chan et al. (2010) study from the pool, the effect size changed to \(0.881 \, (0.583, 1.178)\). All the studies in the pool were homogeneous, \(Q(5) = 7.722, p > 0.05\). The fail-safe N value was 78.

A between-group comparison procedure was carried out to compare the effect sizes obtained in the two clinical groups and their controls. The effect sizes obtained when comparing SZ patients with healthy controls were not significantly different from effect sizes obtained when comparing BD patients with healthy controls, \(Q(1) = 7.722, p > 0.05\), indicating a similar extent of impairment in the two patient groups.

### 3.2. Total Latency of Task B

Five data sets compared the Total Latency of Task B between BD patients and healthy controls. As seen in Table 2, a large effect size of \(0.930 \, (0.403, 1.457)\) was obtained, indicating that patients with BD performed Task B significantly slower than healthy controls. The fail-safe number was 52. The Q-statistic suggested heterogeneity, \(Q(4) = 14.778, p < 0.05\). As shown in Fig. 2B, the study by Stoddart et al. (2007) generated an extremely large effect size \((1.842)\). The control group in this study was significantly younger than the patients (a mean difference of 12 years), which might be a potential reason for the unusually large effect size. After excluding the study by Stoddart et al. (2007) from the pool, the rest of the studies became homogeneous and produced an overall effect size of 0.674.

Four data sets compared the Total Latency of Task B between SZ patients and healthy controls and generated a significant effect size of \(0.840 \, (0.566 - 1.113)\), \( Z = 6.014, p < 0.001\). The large effect size indicated that patients with SZ were slower than healthy controls when performing Task B. The effect sizes of the four studies were homogeneous, \(Q(3) = 0.414, p > 0.05\).

The effect sizes obtained between patients with SZ and healthy controls were similar to those obtained between BD patients and controls, \(Q(1) = 0.066, p > 0.05\). As indicated by the between-group comparison procedure, this suggests that the extent of impairment in Task B latency was similar in the two patient groups.

### 3.3. Suppression time

Four data sets compared suppression time between patients with BD and healthy controls, and five data sets compared suppression time between patients with SZ and healthy controls. Although patients with BD had a longer suppression time, the overall effect size of \(0.156 \, (0.313, 0.653)\) was not significant, \( Z = 0.654, p > 0.05\). The studies were heterogeneous, \(Q(3) = 8.078, p < 0.05\). On inspection of Fig. 3A, BD patients in the depressive phase in the study by Dixon et al. (2004) produced an effect size of \(-0.574\). After excluding this study from the pool, a significant effect size of \(0.397\) was obtained, \( Z = 2.517, p = 0.01\). The studies were homogeneous \((Q(2) = 0.754, p > 0.05)\), but the fail-safe number was only 2. These results indicated that patients with BD took a longer time to suppress a pre-potent response, except patients in the depressive phase who needed less time to do so.

A non-significant effect size of \(0.325\) was obtained from the five data sets that compared patients with SZ and healthy controls, \( Z = 1.627, p > 0.05\). The Q-statistic indicated significant heterogeneity of the studies, \(Q(4) = 10.038, p < 0.05\). As shown in Fig. 3A, the study by Chan et al. (2010) generated an effect size of \(-0.398\). When this study was excluded, the other studies became homogeneous, \(Q(3) = 3.327, p > 0.05\), and a low-to-moderate effect size of \(0.460\) was obtained, which was significant, \( Z = 3.20, p = 0.001\). The fail-safe N value was 10. The unique feature of the study by Chan et al. (2010) is that the researchers recruited a group of Chinese participants to undertake the HSCT in Chinese. Overall, the results showed that patients with SZ took longer time to suppress a strong response, except Chinese patients who took a little less time to do so.

The effect sizes obtained between patients with SZ and healthy controls were similar to those obtained between BD patients and controls, \(Q(1) = 0.202, p > 0.05\) as indicated by the between-group comparison procedure. This suggests that the extent of impairment in the two patient groups was similar.

### 3.4. Total Error of Task B

Five data sets compared the Total Error of Task B between BD patients and healthy controls. A significant effect size of \(0.866 \, (0.402,

\[157\]
1.330) was obtained, $Z = 3.656, p < 0.001$. The effect sizes of these studies were not homogeneous ($Q(4) = 11.628, p < 0.01$). As shown in Fig. 3B, the heterogeneity could mainly be attributed to the small effect size in the study by De Almeida Rocca et al. (2008) in which BD patients had more years of education and higher IQ than healthy controls (though not statistically significant; $p = 0.07$). After excluding this study from the pool, the effect sizes of the four remaining studies became homogeneous, $Q(3) = 5.162, p > 0.05$, generating a significant effect size of 1.040 (0.626–1.190), $Z = 4.25, p < 0.001$, with a fail-safe number of 40.

Eight data sets were included to compare the Total Error of Task B between patients with SZ and healthy controls. A large effect size of 0.944 (0.698, 1.190) was obtained, $Z = 7.531, p < 0.01$, indicating that SZ patients made significantly more errors than healthy controls. All the studies were homogeneous, $Q(6) = 10.018, p > 0.05$. The fail-safe number was 163.

The effect sizes obtained between patients with SZ and controls were not significantly different from those obtained between BD patients and controls, $Q(1) = 0.085, p > 0.05$, as indicated by the between-group comparison procedure. Fig. 3B summarizes the studies used for computing effect sizes, the effect sizes and confidence intervals.

### 3.5. Type A Error in Task B

Only two data sets were available for computing the effect size for Type A Error in Task B between patients with BD and healthy controls, producing a significant effect size of 0.678 (0.366, 1.021), $Z = 3.881, p < 0.05$. The two studies were heterogeneous ($Q(1) = 5.828, p < 0.05$). As illustrated in Fig. 4A, the study by Stoddart et al. (2007) generated an unusually large effect size as compared with the other study. The significantly younger control group might be the reason for the heterogeneous result. The fail-safe number could not be computed due to the small number of studies available.

A medium effect size of $0.639 (0.431, 0.847)$ was obtained from the six data sets comparing patients with SZ and healthy controls, $Z = 6.013, p < 0.01$. The studies were homogeneous, $Q(5) = 0.639, p > 0.05$. The fail-safe number was 47. The effect sizes generated by comparing SZ with healthy controls were similar to those generated by comparing BD with healthy controls, $Q(1) = 0.089, p > 0.05$. Thus, patients with both SZ and BD made significantly more Type A error during task inhibition to a similar extent.

### 3.6. Type B Error in Task B

Only two data sets were available for computing the effect size between BD and healthy controls. An effect size of 0.869 was obtained, but it was not significant, $Z = 1.270, p > 0.05$. The two studies were highly heterogeneous, $Q(1) = 13.561, p < 0.001$. The study by Stoddart et al. (2007) generated a significantly larger effect size than the study by Joshua et al. (2009). The fail-safe number could not be calculated because there were only two studies.

Six data sets compared patients with SZ and healthy controls on this measure. A significant effect size of $0.578 (0.247, 0.912)$ was obtained, $Z = 3.415, p = 0.001$. The results were heterogeneous, $Q(5) = 12.178, p < 0.05$. The fail-safe number was 33. The study by Chan et al. (2010) with chronic Chinese SZ patients reported an effect size of $−0.152$, indicating a relatively normal performance in the patient group. After excluding this study from the pool, the overall effect size changed to 0.706, $Z = 5.504, p < 0.001$. The four remaining studies were homogeneous, $Q(4) = 4.938, p > 0.05$, with a fail-safe number of 44. Another study in

---

**Fig. 3.** Effect sizes for the number of Total Error in Task B (3B) of Hayling Sentence Completion Test between patients and healthy controls. BD: bipolar disorder; SZ: schizophrenia.

**Fig. 4.** Effect sizes for the number of Type A Error (4A) and Type B Error (4B) of Task B of Hayling Sentence Completion Test between patients and healthy controls. BD: bipolar disorder; SZ: schizophrenia.
the pool with first episode Chinese SZ patients did report worse performance in patients as compared to healthy controls (Chan et al., 2012). The potential difference in the two studies might indicate that performance on the HSCT task could be improved with the stability of symptoms, at least in Chinese patients.

We also compared the effect sizes generated by comparing the two patient groups with healthy controls and found no significant difference, Q(1) = 0.018, indicating that patients with SZ and patients with BD were comparable in terms of Type B Error in Task B. However there was significant heterogeneity among the small number of studies available for meta-analysis. Fig. 4B shows the effect sizes and confidence intervals generated.

4. Discussion

Consistent with findings from many previous studies, we found in this study, that patients with SZ and BD were both impaired on the HSCT. They were slower in task initiation, as indicated by the medium effect size in Total Latency in Task A (d = 0.881 for SZ, d = 0.896 for BD). Compared with other measures, very few studies have measured errors in Task A in patients with SZ (Chan et al., 2010) and in patients with BD (de Almeida Rocca et al., 2008; Dixon et al., 2004). The limited number of studies on this issue seems to suggest a relatively high correct rate in Task A. Nevertheless, Dixon et al. (2004) did find BD patients in the manic and remitted phases making more errors in Task A. The combined results suggest that both patient groups used a good strategy for Task A, namely they slowed their response to generate a correct answer. This strategy, however, seems insufficient for performing Task B when additional inhibition and planning are needed. They responded more slowly and committed more errors in Task B, as indicated by the large effect sizes in Total Latency in Task B (d = 0.840 for SZ, d = 0.674 for BD) and Total Error of Task B (d = 0.944 for SZ, d = 1.040 for BD). Inhibition appeared to be difficult for them and as a result they made more Type A errors (d = 0.639 for SZ, d = 0.678 for BD) and Type B errors (d = 0.706 for SZ, d = 0.869 for BD). The findings from the present meta-analysis, together with results from other meta-analyses on executive functions in BD and/or SZ (Fioravanti et al., 2005; Heydebrand, 2006; Laws, 1999; Reichenberg, 2010), confirm that executive functions are impaired in both patient groups.

Some researchers have associated impaired inhibition on the HSCT with working memory function (Royer et al., 2009b). To what extent is the observed impaired inhibition in patients contributed to by deficiency in working memory is an interesting topic. This is because deficient working memory capability in both patient groups has been well documented (Arts et al., 2008; Reichenberg, 2010). One study that used Chinese SZ participants may be relevant here. Compared to the same materials in other languages, Chinese sentences had fewer syllables. Therefore, listening to the same sentence in Chinese might be less taxing on phonological working memory than in an alphabetic language such as English. Chan et al. (2010) found that there was no reliable difference in HSCT performance between a group of Chinese patients with chronic SZ and healthy controls, even though the patients performed worse than healthy controls in a Stroop-like task. Based on the findings of Chan et al. (2010), we speculate that due to the difference in language, Chinese patients with chronic SZ may expend less working memory in context building and have more working memory reserve for both initiation and inhibition. Compared with healthy controls, patients with SZ did not utilize strategies which were commonly used by healthy controls, such as providing answers with items around the room or all from one particular semantic category (Joshua et al., 2009). Patients with SZ or BD seem not to have sufficient working memory resource to launch such a strategy for the task. Further studies are needed to clarify the relationship between inhibition and working memory. In a recent longitudinal study on a group of SZ patients in Hong Kong, researchers found that patients performed worse during the first episode, but in the third year after onset, the performance in patients became similar to those of healthy controls (Chan et al., 2012).

Findings from the present meta-analysis suggest that patients with SZ and patients with BD are impaired to a similar extent on the HSCT. It has been suggested that executive dysfunctions might be a potential marker of familial vulnerability to SZ and BD (Zalla et al., 2004). We also notice that previous meta-analyses comparing the executive functions of patients with both disorders found that patients with SZ are moderately more impaired in executive functions than patients with BD (Daban et al., 2006; Krabbendam et al., 2005). This may be related to the different tasks used in different studies. For example, the Trail Making Test (Reitan, 1958) and the Stroop Task (Stroop, 1935) were used to assess executive functions in the meta-analysis by Krabbendam et al. (2005). Even though both tests are widely used to assess executive functions, it is possible that the two tests target different domains of executive functions than the HSCT, which mainly targets inhibition function. While the Trail Making Test focuses on task switching and sequencing (Gray, 2006), the Stroop task examines interference and control of attention (Cohen et al., 1990). As pointed out by Frangou et al. (2005a), while the correct answer in the Stroop task relates simply to the colors presented, the correct answer in the HSCT requires self-generation as well as planning. Therefore it is not surprising that patients with BD are more impaired in the HSCT than in the Stroop task.

In conclusion, patients with SZ and BD both showed impaired performance on the HSCT. They had problems in both task initiation and task inhibition. The extent of the impairment is similar in the two groups of patients on many HSCT measures. The relationship between working memory and disinhibition in both patient groups needs to be further studied. One limitation of the present meta-analysis is the comparatively small number of studies included. Another limitation is that we were unable to compare different subtypes of patients due to the small number of studies.

Acknowledgments

This study was supported partly by the National Key Technologies R&D Program (2012BAI36B01), the Knowledge Innovation Project of the Chinese Academy of Sciences (KSCX2-EW-J-8), the Key Laboratory of Mental Health, Institute of Psychology, and a grant from the National Science Fund China Young Investigator Award (81088001) to Raymond Chan. This study was also supported by National Natural Science Foundation of China (31200775) and the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (Y1XC281005) to Kui Wang, and by an Australian Academy of Science Visit to China grant and a Queensland International Fellowship to David Shum. Finally, it was also supported by a grant from the initiation fund of the CAS/SAFEA International Partnership Programme for Creative Research Team (Y2XC131003).

References

Chan RCK, Chen EYH. Executive dysfunctions and neurological manifestations in schizophrenia. Hong Kong J Psychiatry 2004;14:2–6.