Acta Psychiatrica Scandinavica

Acta Psychiatr Scand 2020: 141: 110–130

DOI: 10.1111/acps.13121

ACTA PSYCHIATRICA SCANDINAVICA

Systematic Review or Meta-analysis

Executive functions and memory in bipolar disorders I and II: new insights from meta-analytic results

Cotrena C, Damiani Branco L, Ponsoni A, Samamé C, Milman Shansis F, Paz Fonseca R. Executive functions and memory in bipolar disorders I and II: new insights from meta-analytic results.

Objective: To perform a systematic review and meta-analysis of executive functions (EF) and episodic memory in bipolar disorder (BD).
Methods: A literature search was conducted on three electronic databases. Results were combined using random-effects meta-analysis.
Results: A total of 126 studies (6424 patients with BDI, 702 with BDII, and 8276 controls) were included. BDI was associated with moderate to large impairments across all cognitive functions and BDII with small-to-medium impairments. Small significant differences were identified between BDI and BDII on all cognitive functions except inhibition. The Trail Making Test (TMT) (g = 0.74, 95% CI: 0.67–0.80), Hayling Test (g = 0.58, 95% CI: 0.34–0.81), Digit Span Total (g = 0.79, 95% CI: 0.57–1.01), and Category Fluency (g = 0.59, 95% CI: 0.45–0.72) tasks were most sensitive to cognitive impairment in BDI. The TMT (g = 0.65, 95% CI: 0.50–0.80) and Category Fluency (g = 0.56, 95% CI: 0.37–0.75) were also sensitive to cognitive alterations in patients with BDII.
Conclusion: BD type I was associated with more severe and widespread impairments than BDII, which showed smaller impairments on all functions except inhibition, where impairments were larger. Education and (hypo)manic symptoms should be further investigated in future studies due to their possible influence on the neuropsychological profile of BD. The instruments identified in this review should be considered for inclusion in cognitive assessment batteries in BD.

C. Cotrena[1], L. Damiani Branco[1], A. Ponsoni[1], C. Samamé[2], F. Milman Shansis[3], R. Paz Fonseca[1]

[1]Department of Health Science, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil, [2]School of Psychology, University of Buenos Aires, Buenos Aires, Argentina and [3]Faculdade de Medicina da Universidade do Vale dos Sinos (UNISINOS) São Leopoldo, Programa de Pos-Graduação em Saude Coletiva, São Leopoldo, Rio Grande do Sul, Brazil

Key words: bipolar disorder; executive functions; memory; inhibition; cognition

Laura Damiani Branco, Av. Ipiranga, 6681, bld 11, rm 932, Porto Alegre, Rio Grande do Sul, Brazil 90619-900. E-mail: lauradbranco@gmail.com

Accepted for publication October 21, 2019

Summations

- Patients with BD type I performed worse than control subjects with moderate to large effect sizes, while patients with BD type II showed impairments with small-to-medium effect sizes. Small significant differences were identified between BD types I and II on all cognitive functions except inhibition.
- Education, (hypo)mania symptom scores, and lithium use moderated cognitive impairments in BD.
- The TMT B, Hayling Test B, Digit Span Total, and Category Fluency were most sensitive to cognitive impairments and to differences between cognitive performance in individuals with BD types I and II.

Limitations

- Small number of studies involving BD type II.
- Large heterogeneity in effect sizes.
- Inconsistent reporting of potential moderator variables (mood symptoms, medication use, comorbidities).

Introduction

Cognitive impairment is currently considered a central feature of bipolar disorder (BD), with alterations in executive functions (EF) and verbal memory affecting a significant proportion of patients with this condition (1). These impairments occur during mood episodes but also during euthymia, and may be present even before illness onset (2–4). Impairments in EF and verbal memory, especially, have been identified as putative cognitive endophenotypes for BD (5). However, though the presence of cognitive impairment in BD is all but established, the *nature* of this impairment continues to be a source of contention.

Questions remain, for instance, regarding possible differences between the neurocognitive functioning of patients with BD type I (BDI) and BD type II (BDII). Studies which have compared these subtypes of the disorder produced discrepant findings, with some reporting no differences between patients with BDI and BDII (6) and others suggesting the former might show greater impairments than the latter (7). Reviews and meta-analyses of cognition in BD have also addressed this issue, but have not been able to reach a definitive conclusion. Most such studies have found that BD subtypes do not significantly differ in terms of their cognitive function (8, 9), or differ only in very specific abilities (10). Yet the small number of studies involving BDII limits the strength of this conclusion: Several reviews explicitly mention their inability to compare BDI and BDII due to the absence of eligible studies in the literature (11, 12). Others retrieve as few as one single study involving this particular disorder (1, 13). As such, further investigation is needed to shed light on cognitive differences between BD subtypes.

Another issue which has not been fully elucidated in the literature is the influence of characteristics which may act as protective or risk factors for cognitive impairment in BD. These include the use of psychoactive medication (14, 15), duration of illness (16, 17), age of onset (17, 18), history of psychotic symptoms (19, 20), current mood state (21, 22), and number of mood episodes (21). Though previous studies of these variables have produced interesting findings, clinical characteristics have not been able to fully explain the variations in cognitive profiles observed in BD. As a result, researchers have begun to look for risk and protective factors among lifestyle or demographic features which may differ between participants. These features include premorbid intelligence quotient (23), sleep disturbances (24), and more recently, variables related to cognitive reserve, such as education levels (25–27). A quantitative synthesis of these studies using methods such as meta-regression may provide important information for future studies. However, none of the recent meta-analyses of cognitive performance in BDI and BDII provided this information (8, 9).

The present study aimed to make a novel contribution to the literature by addressing some of the limitations of previous reviews and meta-analyses. These include the investigation of differences between BDI and BDII, their levels of impairment relative to control participants, the possible influence of protective and risk factors on cognition, the analysis of individual task scores, and the provision of a more comprehensive picture of the existing literature. The recent meta-analysis conducted by Dickinson et al. (8), for instance, was limited to randomized clinical trials, which resulted in a relatively small sample size, especially of studies involving BDII. Randomized trials have also been criticized for their low external validity (28, 29). The present study will include observational studies in addition to randomized trials, which may make for a larger sample, and also a more representative one, whose findings may be generalizable to different patient populations. Another recent meta-analysis, conducted by Bora et al. (9), compared patients with BDI/BDII to each other, but not to healthy control participants. Therefore, though the findings allow for important conclusions regarding the neurocognitive profiles of different subtypes of BD, they do not show whether these profiles would result in cognitive impairments relative to the general population. Another important contribution of the present study is the examination of individual task scores. Existing reviews and meta-analysis provide effect sizes for general cognitive domains (12, 30) or assessment instruments (8, 9), with separate metrics calculated only for a few widely used tests such as the Wisconsin Card Sorting Test (WCST). Given the large variability in tasks used to evaluate cognition, especially the EF (31), it is likely that a summarized effect size would obscure important differences between neuropsychological tasks and scores. The analysis of individual tasks and scores may contribute to the construction of standardized test batteries and help identify the measures which reveal the greatest differences between cognitive profiles in BDI and BDII.

In light of these observations, the aim of the present study was to perform a comprehensive systematic review and meta-analysis of EF and

episodic memory in BD. These particular cognitive functions were selected due to their relevance to the neurocognitive profile of BD. Other cognitive functions, such as attention and processing speed, may also be impaired in patients with BD types I and II (32, 33). However, a recent review of the literature has found that impairments inattention and processing speed are less consistently associated with BD, whereas memory and EF remain among the most promising neurocognitive endophenotypes for this condition (34). In addition to describing, comparing, and quantifying the severity of executive and memory impairments in BD, we aimed to identify possible moderators of the association between BD and cognitive impairment. We also sought to identify whether individual tasks and scores differed in their sensitivity to impairments in EF and episodic memory in patients with BDI and BDII relative to control subjects. We believe our findings will have important implications for research, clinical practice, and treatment of individuals with BD.

Material and methods

Procedures

The present study was conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis) (35) and MOOSE (Meta-analysis of Observational Studies in Epidemiology) (36) guidelines. This review protocol was registered on PROSPERO, under protocol number CRD42018095257 and can be accessed at (http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018095257).

Search strategy and article selection

A literature search was conducted on the PubMed, PsycINFO, and CENTRAL databases in March 2018, using the following keywords: ("bipolar disorder") AND ("executive function" OR "executive functions" OR "working memory" OR "inhibition" OR "cognitive flexibility" OR "shifting" OR "switching" OR "planning" OR "updating" OR "verbal fluency" OR "memory"). The descriptors for EF were selected based on previous reviews of the topic (37), as well as the most widely accepted models of EF in the current literature (38). The search was limited to publications in English, Spanish, French, or Portuguese, and studies with adult samples published between 2008 and 2018.

The titles and abstracts of the articles retrieved were then screened for initial eligibility. Full-text versions of the articles identified as potentially eligible were retrieved and examined for the following inclusion criteria: i) at least one measure of EF or episodic memory; ii) separate scores for patients with BDI and/or BDII; iii) control group of participants with no mood disorders; and iv) providing sufficient information for calculation of effect sizes. When an article met all inclusion criteria save for the provision of means and standard deviations for calculating effect sizes, authors were contacted in order to request the missing data.

The reference lists of included articles, and other publications citing included articles, were also screened for any studies which may have been eligible for inclusion but were not identified in the database search. A flowchart of the search and selection process is shown in Fig. 1.

Data extraction and coding

After the final set of articles was selected for the review, predictors and outcome data were extracted from each publication. Data regarding sample size, age, education, premorbid, and current IQ were extracted for both clinical and control groups. The following data were also extracted for the groups of BD patients: subtype of the disorder, duration of illness, age of illness onset, number of depressive episodes, number of manic/hypomanic episodes, number of hospitalizations, current scores on mania/depression scales, medication use, and number of previous suicide attempts. Studies which did not evaluate these variables were included in the overall analysis, but not in the relevant meta-regression models.

Data for both predictor and outcome variables (i.e. neuropsychological test results) were extracted in the form of means and standard deviations. When a study reported multiple measures of the same construct or several scores from a single EF or episodic memory task, all relevant scores were extracted in order to prevent selective reporting bias.

Data were extracted separately for patients with BDI and BDII. In other words, studies which compared a control group to two sets of patients, each with one type of BD, provided data on two separate comparisons: control patients vs. BDI and control patients vs. BDII. If a study involved more than one group of patients with the same diagnosis (e.g. BDI with history of smoking, BDI with no history of smoking), different procedures were followed depending on whether the study also provided a separate control group for each clinical comparison. If a single control group was

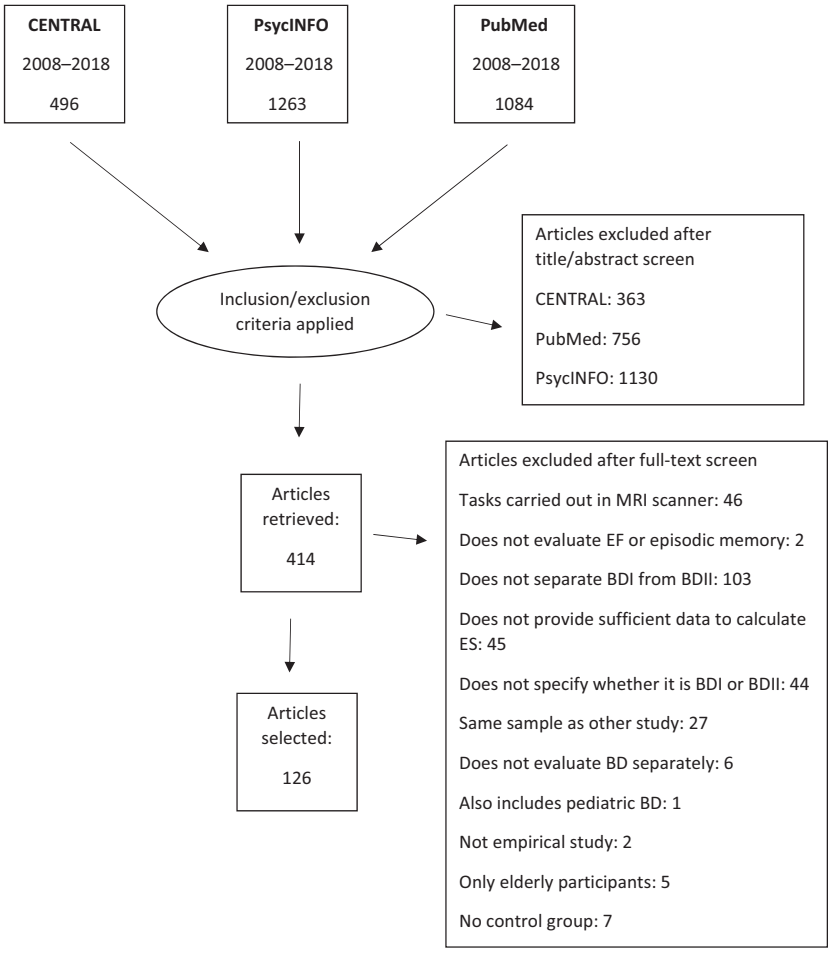

Fig. 1. Article selection flow diagram.

compared with multiple sets of patients with the same condition, but varying on any other clinical or demographic characteristic, the data from all clinical groups were combined using weighted means and standard deviations. This method was chosen rather than entering each comparison separately in order to avoid introducing statistical dependence. Some studies, however, involved multiple control groups and patient groups (e.g. smokers with BDI vs. smoking control subjects; non-smokers with BDI vs. non-smoking control subjects). In these cases, each comparison was considered separately, given that no statistical dependence would be generated by this procedure.

In order to calculate composite effect sizes and organize the analyses, tasks were coded according to the main construct they are/believed to measure. Instruments used to evaluate EF were coded according to the criteria proposed by Snyder (37) in a meta-analysis of EF in major depressive disorder. This led to a classification of instruments into the following categories: inhibition, cognitive flexibility, verbal working memory, visuospatial working memory, verbal fluency, and planning. Episodic memory tasks were coded as described by Scott et al. (39) into the following categories: immediate verbal memory, delayed verbal memory, immediate visual memory, and delayed visual memory. No distinction was made between recall and recognition tasks for the purpose of composite effect sizes, since few studies explicitly stated which of these two methods were used to evaluate episodic memory. As such, measures of episodic memory were categorized only based on the length of time between encoding and recall (immediate vs. delayed) and type of stimulus used (verbal vs. visual). The coding of scores and tasks into the relevant cognitive constructs was performed by the first two authors of the present study, both of whom are neuropsychologists with significant experience in cognitive assessment.

Data analysis

Effect sizes were calculated for each measure using Hedge's adjusted *g,* consisting of a similar formula to Cohen's *d,* using the pooled standard deviation of both participant groups as the denominator,

and a correction for small sample sizes (40). For ease of interpretation, effect sizes were calculated so that positive values indicate inferior performance by patients with BD while negative values suggest the contrary. Outliers were identified based on the standard residuals method (41), adjusted for the use of random-effects models (42). These analyses were performed for each statistical model, so that studies classified as outliers were only removed from the relevant calculations, remaining in all other models in the study.

Data were analyzed separately for patients with BDI and BDII. As such, comparisons were provided for each of the following scenarios: patients with BDI vs. control subjects, BDII vs. control subjects and BDI vs. BDII. These comparisons included a 'summary' effect size for each cognitive ability evaluated in the present study, as well as separate effect sizes for individual tasks and scores reported by at least five studies. The five-study cutoff was selected based on the current literature, which holds that the statistical power associated with meta-analytic procedures is only reliably larger than that observed in single trials when the number of studies analyzed is least five (43).

Summary effect sizes were calculated for each of the following constructs: inhibition, cognitive flexibility, verbal working memory, visuospatial working memory, verbal fluency, planning; immediate verbal memory, delayed verbal memory, immediate visual memory, and delayed visual memory. These analyses included all comparisons of scores or tasks coded into each of these constructs. In some cases, studies may have provided more than one score for a given construct. This was the case, for instance, for studies which provided the number of categories completed and perseverative errors on the Wisconsin Card Sorting Task (WCST). Both scores were coded as measures of cognitive flexibility. In cases such as this, when multiple effect sizes derived from the same participants were relevant to a given analysis, effect sizes were aggregated using the BHHR method (44). This procedure allowed for the calculation of a single effect size for each study, controlling for any potential biases.

The calculation of summary effect sizes for constructs, scores, or tasks was conducted by pooling the effect sizes derived from each study using an inverse variance heterogeneity method, which has been recently introduced as a superior alternative to random-effects models (45). Variability in effect sizes between studies was analyzed using the Q and I^2 statistics. Publication bias was examined through visual inspection of the funnel and doi plots, and quantitatively determined using funnel plot regression tests (46), the LFK asymmetry index (47), and trim-and-fill-adjusted effect sizes (48).

The influence of other variables on cognitive performance was evaluated through meta-regression. According to the literature, moderator analyses require a sample of at least 20 effect sizes in order to provide stable estimates with adequate statistical power. As such, after all relevant data on outcomes and predictors were extracted from the studies included in this review, predictors which met the aforementioned criterion for sample size were entered into meta-regression models for each of the variables analyzed. Data analyses were conducted using MetaXL version 3.0 (49) for Microsoft Excel, as well as packages *metafor* (50), *Mad* (51), and *altmeta* (52)in the R environment (53).

Results

In total, the 126 studies included in the meta-analysis included 15 402 participants: 6424 patients with BDI, 702 patients with BDII, and 8276 control participants. The mean sample size per study was 66.21 participants per control group, with sizes ranging from 14 to 495 participants, and 49.84 patients for clinical groups, which ranged in size from 11 to 375 participants.

Given that the majority of studies compared control participants to patients with BDI (n = 107; 84.9%) while 19 (15.1%) compared control participants to individuals with BDII, the results of comparative analyses will be discussed in this order. The comparison of patients with BDI vs. BDII, analyzed in 13 studies, will be discussed at the end of the comparative analysis section. The results of moderator analyses will be presented after all comparative findings are shown.

BDI vs. controls

A total of 107 studies provided comparisons of patients with BDI and control participants. Four of these studies analyzed two sets of patients with BDI and two separate sets of control participants, all of which included more than 10 individuals each, and these studies were therefore divided into separate samples (54–57).

Control participants in these studies were on average 36.44 years of age, with 13.56 years of education. Patients, on the other hand, were on average 37.33 years old, with 13.02 years of formal study. Clinical data reported by the studies suggested that the majority of samples were comprised entirely of euthymic patients (44%). In 5.65% of

samples, patients were evaluated during a (hypo)-manic episode, while in 0.8% of cases, they experienced a depressive episode during testing. In 19.4% of cases, samples were heterogeneous with regard to mood state and included both euthymic patients and individuals with at least some mood symptoms. The remaining studies did not provide specific information on patients' mood state at the time of testing.

The number of hospitalizations, length of illness, age of onset, and mood episodes were only reported by some of the studies examined. For the studies which did provide this information, the mean number of hospitalizations per patient was 2.96, while the mean length of illness was 12.85 years and the mean age of onset was 25.1. Patients had a mean of 5.28 depressive episodes and 5.76 (hypo)manic episodes over their lifetimes. The results of comparative analyses between these patients and corresponding control groups are presented below.

Inhibition

Inhibition was investigated by 55 studies of BDI. In addition to the instruments mentioned in Table 1, the most frequently used measures of inhibition were the Stop Signal Task ($k = 4$) and the Frontal Assessment Battery (FAB) ($k = 3$). Other measures of inhibition, used by one study each, were the antisaccade task, the number of impulsive errors on a delayed gratification test, and the number of commission errors on a sustained attention to response task, a rapid visual processing task or a tonic alertness test.

The largest difference between control participants and individuals with BDI was observed in the number of errors on the Hayling Sentence Completion Test (HSCT) B ($g = 0.58$; 95% CI: 0.34–0.81). However, this effect size was also associated with moderate heterogeneity ($I^2 = 42.19\%$). The only measure of inhibitory control which did not display a significant effect size, as evidenced by a confidence interval which included the value of 0, was the number of commission errors on the Continuous Performance Test (CPT). This measure also had the greatest heterogeneity of all tests used to evaluate inhibitory control for which it was possible to calculate effect sizes ($I^2 = 58.62\%$). All other measures yielded medium effect sizes between patients and control subjects, except for accuracy on the Stroop Color-Word test, whose effect size was slightly smaller. A forest plot of composite effect sizes for inhibitory control, which displays effect sizes between control subjects and participants with BDI, can be seen in Fig. S1. Importantly, the heterogeneity of the composite effect size for inhibition was classified as low ($I^2 = 24.85\%$).

Risk of bias analyses suggested only minor asymmetries in the reporting of inhibition outcomes, as evidenced by the LFK index. The bias in these cases was in favor of smaller effect sizes between control participants and patients with BDI.

Working memory

Working memory was investigated by 64 studies of BDI, most of which analyzed verbal rather than visuospatial working memory. Variations of the Digit Span Task were the most commonly used instruments in working memory assessment. While some studies provided separate scores for the forward and backward span portions of the Digit Span Test, others provided only a total sum score. As such, effect sizes were calculated separately for these instances. As shown in Table 1, the largest differences in verbal working memory between control subjects and patients with BDI were observed in the Digit Span Sum score ($g = 0.79$; 95% CI: 0.57–1.01), followed by the Digit Span Backward ($g = 0.63$; 95% CI: 0.49–0.77). However, both outcomes were associated with moderate heterogeneity.

Three additional measures of verbal working memory were used in some studies, but could not be separately analyzed since they were present in less than five investigations. These included the Arithmetic subtest from the Wechsler Adult Intelligence Scales III (WAIS-III) ($k = 4$), n-back tests ($k = 3$), a Sentence-Word Span test ($k = 1$), and the verbal working memory score from the Computerized Neurocognitive Battery (58). Nevertheless, scores on all of these measures were included in the verbal working memory composite and Overall Working Memory composite scores. The I^2 statistic revealed moderate to substantial levels of heterogeneity in effect sizes for verbal working memory.

The Spatial Span Task was the most widely used measure of visuospatial working memory. However, as in the Digit Span Task, there was significant variation in the way scores were reported. In this case, since no score was reported by at least five different studies, a single effect size was calculated based on all variables derived from the Spatial Span Task (Forward Span, Backward Span, Total Score). In addition to this instrument, studies also evaluated visuospatial working memory using n-back tests ($k = 1$), the Cambridge

Table 1. Weighted mean effect size analyses between control participants and patients with bipolar disorder type I

					95% CI		Heterogeneity					Risk of bias	
	N_c	N_{BD}	K	*g*	LL	UL	Q	*df*	*P*	I^2	Trim-and-fill-adjusted *g*	Egger's *P*	LFK index
Inhibitory control													
CPT commission errors[29]	195	147	5	0.33	−0.02	0.68	9.67	4	0.05	58.62	0.33	0.70	1.43
CWI set-shifting	762	371	6	0.52	0.39	0.65	1.75	5	0.82	0.00	0.52	0.73	−1.99
Stroop interference speed[15]	1210	1023	19	0.54	0.42	0.66	28.20	18	0.06	36.16	0.57	0.16	−1.86
Stroop interference accuracy	767	757	16	0.46	0.34	0.59	21.31	15	0.13	29.60	0.46	0.54	0.56
Go/No-Go (Accuracy)	106	105	5	0.55	0.21	0.89	5.93	4	0.20	5.98	0.73	0.02	0.13
HSCT B (Errors)	295	250	6	0.58	0.34	0.81	8.65	5	0.12	42.19	0.62	0.34	−1.25
Inhibition composite[16]	2743	2551	55	0.56	0.50	0.62	71.86	54	0.05	24.85	0.59	0.04	−1.51
Working memory													
Digit Span Fwd[5]	1432	976	18	0.35	0.24	0.46	24.78	17	0.10	31.41	0.30	0.05	2.078
Digit Span Bwd[6]	1295	1054	19	0.63	0.49	0.77	42.53	18	0.0009	57.68	0.63	0.80	0.01
Digit Span Fwd + Bwd	400	508	11	0.79	0.57	1.01	22.45	10	0.0130	55.46	0.88	0.002	−2.40
Letter-number sequencing[7]	600	362	8	0.54	0.34	0.73	11.23	7	0.13	37.65	0.44	0.79	0.73
N-back[30]	673	374	7	0.52	0.30	0.73	12.35	6	0.05	51.42	0.52	0.50	1.94
Verbal WM composite[8]	2963	2048	42	0.58	0.47	0.70	145.99	41	<0.0001	71.92	0.63	0.83	0.34
Spatial Span	272	181	7	0.45	0.19	0.71	11.28	6	0.0800	46.82	0.36	0.13	2.21
Visuospatial WM composite	547	437	15	0.52	0.34	0.70	30.31	14	0.0069	53.81	0.52	0.63	0.95
Overall WM composite[9]	2268	2026	64	0.55	0.46	0.64	176.72	63	<0.001	64.35	0.55	0.61	0.71
Flexibility													
TMT A (time)[1,10]	2298	2026	37	0.60	0.54	0.66	34.62	36	0.45	1.78	0.60	0.49	−0.50
TMT B (time)[31]	2027	2104	37	0.74	0.67	0.80	28.81	36	0.80	0.00	0.74	0.59	−0.61
TMT B-A (time)	500	286	10	0.42	0.21	0.63	15.31	9	0.08	41.23	0.33	0.27	0.11
WCST Categories[11]	2515	2182	37	0.63	0.46	0.81	135.11	36	<.0001	73.35	0.80	0.21	−2.31
WCST Correct[25]	638	525	6	0.23	−0.01	0.48	13.27	5	0.02	62.33	0.23	0.71	−0.50
WCST Errors total	1180	981	12	0.70	0.50	0.90	25.43	11	0.0079	56.74	0.78	0.53	−1.05
WCST FMS	288	357	9	0.32	0.12	0.52	11.35	8	0.46	29.53	0.32	0.70	1.24
WCST Non-perseverative errors	832	810	8	0.43	0.33	0.53	5.04	7	0.66	0.00	0.44	0.75	−0.37
WCST Perseverative errors[12]	2656	2265	35	0.53	0.36	0.70	142.18	34	<0.0001	76.09	0.63	0.24	−1.60
WCST Perseverative responses[13]	404	476	9	0.46	0.32	0.60	5.56	8	0.70	0.00	0.42	0.10	0.94
WCST Conceptual level responses	136	200	5	0.64	0.35	0.93	5.80	4	0.21	30.99	0.57	0.23	1.74
Flexibility composite[14]	4825	3979	76	0.52	0.38	0.66	429.30	75	<0.001	82.53	0.52	0.94	−0.29
Fluency													
Letter fluency[2]	2042	1646	31	0.47	0.37	0.58	57.05	30	0.0021	47.42	0.32	0.10	0.98
Category fluency[3]	2155	1299	23	0.59	0.45	0.72	58.00	22	<0.001	62.07	0.62	0.57	−0.99
Fluency composite[4]	2709	1963	38	0.54	0.45	0.64	76.26	37	<0.001	51.48	0.37	0.41	0.42
Episodic memory													
CVLT Trial [1]	254	258	7	0.47	0.29	0.64	5.32	6	0.50	0.00	0.47	0.72	1.15
CVLT 1–5 Learning[17]	869	760	17	0.62	0.50	0.74	20.38	16	0.20	21.49	0.57	0.36	0.42
CVLT Short delay cued recall	227	225	6	0.59	0.40	0.78	3.71	5	0.59	0.00	0.59	0.70	1.33
CVLT Short delay free recall	361	389	10	0.59	0.41	0.77	11.98	9	0.21	24.90	0.59	0.71	−0.72
WMS Auditory memory (immediate)	225	233	5	0.71	0.50	0.92	4.59	4	0.33	12.95	0.63	0.66	0.38
Immediate verbal memory composite[21]	2878	2059	46	0.51	0.42	0.61	108.42	45	<0.0001	58.50	0.49	0.30	−0.45
CVLT Long delay cued recall[18]	227	225	6	0.50	0.24	0.76	8.57	5	0.13	41.63	0.50	0.93	0.42
CVLT Long delay free recall[19]	1011	725	16	0.59	0.48	0.70	15.88	15	0.39	5.55	0.59	0.69	−1.18
CVLT Recognition	364	303	8	0.57	0.41	0.74	3.10	7	0.87	0.00	0.57	0.86	0.61
RAVLT Recognition	144	164	5	0.46	0.22	0.69	3.29	4	0.51	0.00	0.46	0.09	1.04
WMS Auditory (delayed)	225	233	5	0.41	−0.10	0.91	23.17	4	0.0001	82.73	0.41	0.71	
WMS Auditory recognition delayed	225	233	5	0.29	−0.19	0.77	20.99	4	0.0003	80.94	0.29	0.70	
Delayed verbal memory composite	2841	2067	45	0.51	0.40	0.62	139.65	44	<0.0001	68.38	0.40	0.10	0.71

Table 1. (Continued)

					95% CI		Heterogeneity					Risk of bias	
	N_c	N_{BD}	K	*g*	LL	UL	Q	*df*	*P*	I^2	Trim-and-fill-adjusted *g*	Egger's *P*	LFK index
Verbal episodic memory composite[32]	3668	2830	60	0.55	0.46	0.64	187.70	59	<0.0001	68.57	0.45	0.11	0.74
Rey Complex Figure Immediate Recall[20]	277	357	6	0.59	0.42	0.75	1.90	5	0.86	0.00	0.59	0.95	−0.48
Immediate visual memory composite[22]	1140	1128	22	0.48	0.36	0.60	31.94	21	0.06	34.25	0.43	0.34	−0.05
Rey Complex Figure Delayed Recall	340	333	7	0.52	0.33	0.71	7.79	6	0.25	22.99	0.37	0.32	1.63
Delayed visual memory composite	1055	826	17	0.40	0.24	0.56	35.88	16	0.003	55.40	0.32	0.45	0.57
Visual episodic memory composite[27]	1776	1456	30	0.45	0.36	0.54	44.73	29	0.03	35.18	0.37	0.21	−0.01
Episodic memory composite[28]	4600	3331	67	0.59	0.50	0.68	246.12	66	<0.0001	73.18	0.55	0.21	0.93
Planning													
Tower of Hanoi	1149	1013	8	0.64	0.45	0.84	23.98	7	0.0011	70.81	0.60	0.01	5.79
Planning composite[24]	1365	1235	14	0.61	0.44	0.78	33.78	13	0.0013	61.51	0.61	0.65	1.43

CPT, Continuous Performance Test; CVLT, California Verbal Learning Test; CWI, Color-word Interference; Digit Span Fwd, Digit Span Forward; Digit Span Bwd, Digit Span Backward; HSCT, Hayling Sentence Completion Test; RAVLT, Rey Auditory Verbal Learning Test; TMT, Trail Making Test; WCST, Wisconsin Card Sorting Test; WM, Working memory; WMS, Wechsler Memory Scales.

[1]Scores reported for reference only, and not included in composite effect size calculations.

[2]One outlier removed. The effect including the outlier was calculated at g = 0.51 (95% CI: 0.36–0.65).

[3]Two outliers removed. The effect including the outliers was calculated at g = 0.68 (95% CI: 0.43–0.94).

[4]Two outliers removed. The effect including the outliers was calculated at g = 0.65 (95% CI: 0.45–0.85).

[5]Two outliers removed. The effect including the outliers was calculated at g = 0.48 (95% CI: 0.14–0.81).

[6]One outlier removed. The effect including the outlier was calculated at g = 0.68 (95% CI: 0.41–0.95).

[7]One outlier removed. The effect including the outlier was calculated at g = 0.69 (95% CI: 0.12–1.25).

[8]One outlier removed. The effect including the outlier was calculated at g = 0.62 (95% CI: 0.45–0.79).

[9]Two outliers removed. The effect including the outliers was calculated at g = 0.61 (95% CI: 0.48–0.74).

[10]Two outliers removed. The effect including the outliers was calculated at g = 0.58 (95% CI: 0.46–0.70).

[11]One outlier removed. The effect including the outlier was calculated at g = 0.66 (95% CI: 0.47–0.85).

[12]One outlier removed. The effect including the outlier was calculated at g = 0.54 (95% CI: 0.34–0.74).

[13]One outlier removed. The effect including the outlier was calculated at g = 0.58 (95% CI: 0.25–0.91).

[14]One outlier removed. The effect including the outlier was calculated at g = 0.50 (95% CI: 0.34–0.65).

[15]Two outliers removed. The effect including the outliers was calculated at g = 0.64 (95% CI: 0.34–0.94).

[16]Three outliers removed. The effect including the outliers was calculated at g = 0.58 (95% CI: 0.43–0.72).

[17]One outlier removed. The effect including the outlier was calculated at g = 0.65 (95% CI: 0.50–0.81).

[18]One outlier removed. The effect including the outlier was calculated at g = 0.63 (95% CI: 0.21–1.05).

[19]One outlier removed. The effect including the outlier was calculated at g = 0.62 (95% CI: 0.48–0.76).

[20]One outlier removed. The effect including the outlier was calculated at g = 0.73 (95% CI: 0.27–1.19).

[21]Three outliers removed. The effect including the outliers was calculated at g = 0.59 (95% CI: 0.42–0.75).

[22]Two outliers removed. The effect including the outliers was calculated at g = 0.59 (95% CI: 0.39–0.79).

[23]One outlier removed. The effect including the outlier was calculated at g = 0.55 (95% CI: 0.28–0.81).

[24]One outlier removed. The effect including the outlier was calculated at g = 0.30 (95% CI: −0.05 to 0.66).

[25]Two outliers removed. The effect including the outliers was calculated at g = 0.61 (95% CI: 0.47–0.75).

[26]Two outliers removed. The effect including the outliers was calculated at g = 0.52 (95% CI: 0.37–0.67).

[27]Two outliers removed. The effect including the outliers was calculated at g = 0.62 (95% CI: 0.50–0.74).

[28]One outlier removed. The effect including the outlier was calculated at g = 0.47 (95% CI: −0.35 to 1.30).

[29]One outlier removed. The effect including the outlier was calculated at g = 0.72 (95% CI: 0.38–1.06).

[30]Three outliers removed. The effect including the outliers was calculated at g = 0.70 (95% CI: 0.59–0.81).

[31]Two outliers removed. The effect including the outliers was calculated at g = 0.59 (95% CI: 0.46–0.72).

Neuropsychological Test Automated Battery (CANTAB) (k = 3) or CNS Vital Signs (CNS-VS) batteries (k = 1), and delayed matching to sample tasks (k = 2). Effect sizes for the Spatial Span Task and Visuospatial Working Memory as a whole were associated with moderate heterogeneity.

The reason an Overall Working Memory Composite was calculated in addition to the Verbal and Visuospatial composites is that several studies used cross-modal measures of working memory, which tapped into both verbal and non-verbal aspects of this construct. Examples include the working memory subtests from the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Consensus Battery (k = 4), Wechsler Memory Scales (WMS) (k = 4), and CANTAB (k = 3) batteries. The studies in question did not provide separate scores for the verbal and visual portions of these instruments, and it would therefore be impossible to include these in either composite. At the same time, excluding these studies altogether would result in a significant loss of data. As such, their effect sizes were combined with those of all remaining variables in the Overall Working Memory Composite, whose aim is precisely to provide a general estimate of working memory impairments in BD, across all tasks and sensory modalities. The comparison of overall working memory composite effect sizes between control participants and individuals with BDI is illustrated by a forest plot in Fig. S2.

The LFK index did not identify any publication bias in the calculation of composite effect sizes for working memory. However, major asymmetries in the doi plot were detected in effect sizes corresponding to the Spatial Span Task, the Digit Span Forward, and the Digit Span Sum score. In the latter, the bias was toward a smaller effect size between control participants and patients, while in the remaining two tests, the bias was toward larger group differences.

Shifting/Flexibility

Flexibility was investigated by 74 studies of BDI. In addition to the Trail Making Test (TMT) and WCST, whose results are presented in greater detail in Table 1, the instruments used to evaluate cognitive flexibility included the intradimensional/extradimensional (ID/ED) shift task (k = 2), the Computerized Neurocognitive Battery (k = 1), the CNS-VS (k = 1), the Cogtest (k = 1), the FAB (k = 1), and the Penn Conditional Exclusion Test (k = 1). The measures which identified the greatest differences between control participants and subjects with BDI were part B of the TMT (g = 0.74; 95% CI: 0.67–0.80) and the total number of errors on the WCST (g = 0.70; 95% CI: 0.50–0.90). Though the number of errors on the WCST was associated with moderate heterogeneity, the effect size for the TMT B was not associated with any significant heterogeneity across studies.

The heterogeneity in effect sizes for WCST variables may be related to the fact that studies used widely different versions of the instrument. In addition to the traditional 128 card version of the WCST (k = 17), some studies used 64 (k = 8), 48 (k = 6), or 36-card (k = 2) variations. Some studies explicitly mentioned the use of a computerized rather than a physical version of the WCST (k = 3). This may also explain the variations in the way scores for this task were reported. While some studies reported perseverative and non-perseverative errors separately, others provided simply an overall error rate. This is why each of these cases was considered separately when calculating effect sizes in Table 1. Studies also varied in their use of percentage vs. raw scores for the aforementioned variables. However, a major finding of this review was that several studies simply did not provide any details on the version of the WCST used in their investigation or the scoring methods used (raw scores, percentages, etc.). Since these data were not universally available, scores were not separated according to the version of the instrument used, or by percentage vs. raw scores. As such, the effect sizes provided in Table 1 may be thought to provide a comprehensive picture of the results associated with the WCST, regardless of the version used or the method employed to calculate error scores.

The effect sizes between control subjects and participants with BDI on measures of shifting or cognitive flexibility can be seen in Fig. S3. As evidenced by the I^2 statistic, composite scores for cognitive flexibility were associated with considerable heterogeneity.

The LFK index did not reveal any significant publication bias for the majority of effect sizes pertaining to cognitive flexibility, including the composite effect size for all measures of this construct. Major asymmetry in the doi plot was only observed for the number of categories completed in the WCST, where a bias was observed toward smaller differences between patients and control subjects. Minor asymmetries were also observed for other variables from the WCST.

Fluency

Letter fluency tasks were used by 38 studies in BDI. The majority of investigations (k = 20)

used letter fluency tasks from the Controlled Oral Word Association Test (COWAT)/FAS or the Delis-Kaplan Executive Function System (D-KEFS) battery. Category fluency was investigated by 22 studies, and measured using the COWAT or D-KEFS in eight of these investigations. Though control participants outperformed those with BDI on both types of task, the largest effect size was observed for category fluency ($g = 0.59$; 95% CI: 0.45–0.72). The effect sizes for fluency tasks were associated with moderate heterogeneity.

Design fluency was only evaluated by three studies ($k = 3$) in samples with BDI (59–61). Each study used one of the following measures of design fluency: the D-KEFS Design Fluency Task, the 5-Point Test, and the Design Fluency subtest form the Calibrated Ideational Fluency Assessment. Given the small sample available to analyze these findings, and the variability in assessment methods, design fluency was not individually analyzed, though it was included in the composite fluency score. A visual representation of these scores can be seen in Fig. S4. Importantly, effect sizes for fluency tasks were not associated with any level of publication bias.

Episodic memory

Episodic memory was investigated by 67 studies of BDI. Twenty-six studies used the California Verbal Learning Test (CVLT) ($k = 26$), seventeen used the WMS ($k = 17$), nine used the Rey Complex Figure Test ($k = 9$), seven used the Rey Auditory Verbal Learning Test (RAVLT) ($k = 7$), six used the CANTAB ($k = 6$), four used the MATRICS ($k = 4$), three used the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) ($k = 3$), and two each used the Brief Visuospatial Memory Test (BVMT) ($k = 2$) and CNS-VS battery ($k = 2$). Other measures of memory, used by one study each, include the Babcock Story Recall Test, the COGNISTAT, the Computerized Neurocognitive Battery, the Delayed Match to Sample Task, the Hopkins Verbal Learning Test, the Rivermead Behavioral Memory Test, and the Semantic Memory test with Associative Increments. The measures which identified the greatest differences between control participants and subjects with BDI were the WMS Immediate Auditory ($g = 0.71$; 95% CI: 0.50–0.92) and the CVLT 1–5 ($g = 0.61$; 95% CI: 0.50–0.74). Importantly, both tests were associated with relatively low heterogeneity. The I^2 statistics suggested that heterogeneity levels were higher for measures of immediate verbal memory than delayed verbal memory.

Tasks were categorized according to the type of stimulus (verbal vs. visual) and length of time between encoding and recall (immediate vs. delayed). However, some studies did not provide sufficient information on the procedures used to evaluate episodic memory and could only be classified based on the type of stimulus (verbal vs. visual). This was the case for seven studies of visual memory in BDI and 12 studies of verbal memory in BDI. These tasks, in addition to those categorized into immediate and delayed recall categories, were therefore included in summary scores for visual and verbal memory. Heterogeneity was higher for measures of verbal memory than visual memory.

In 14 studies, rather than providing separate scores for verbal and visual measures of episodic memory, these were combined into a single composite effect size for this cognitive ability. These studies therefore could not be classified into any of the categories outlined in the Methods section. To avoid excluding these data altogether, an overall episodic memory score was calculated, including every single measure of episodic memory used to compare patients with BDI to control participants. A forest plot of composite effect sizes can be seen in Fig. S5. The results of all remaining analyses are presented in Table 1.

Risk of bias analyses did not reveal major asymmetries in the doi plot of any effect sizes corresponding to episodic memory. Minor asymmetries were noted, however, for some scores from the CVLT, RAVLT, WMS, and Rey Complex Figure test. All biases were toward larger effect sizes between patients and control participants, except for the CVLT long delay free recall, where the bias was in the opposite direction.

Planning

Fourteen studies evaluated planning in patients with BDI. The Tower of Hanoi (TOH) was the most widely used measure of planning, as can be seen in Table 1. There was significant variation in the way scores were reported, with some studies providing the number of moves required to solve each problem, others calculating error scores, and still others providing measures of planning and executive time. No single measure was used by more than 4 studies, and as such, separate effect sizes for each of these could not be calculated. As a result, the effect size for the TOH includes measures of both time and accuracy.

In addition to the TOH, the other instruments used to evaluate planning were the Stockings of Cambridge task (k = 4), the Tower of London (k = 2), and the D-KEFS Tower test (k = 1). Given the low number of studies with each of these measures, no separate effect sizes were calculated for them, though they were included in the composite planning score, shown in Fig. S6. Effect sizes for planning were associated with moderate to substantial heterogeneity, as well as major risk of bias toward larger effect sizes between patients and control participants.

BDII vs. control subjects

Control participants in these studies were on average 35.50 years of age, with 14.12 years of education. Patients, on the other hand, were on average 37.05 years old, with 14.4 years of formal study. The majority of samples were euthymic (47%), while in 5.26% of samples, patients were evaluated during a depressive episode. In 21.05% of cases, samples were heterogeneous with regard to mood state. The remaining studies did not provide a description of participant mood state at the time of testing.

As was the case with studies of BDI, few investigations of patients with BDII provided data on the number of hospitalizations, length of illness, age of onset, and mood episodes experienced by participants. For the studies which did provide this information, the mean number of hospitalizations per patient was 0.52, while the mean length of illness was 12.28 years and the mean age of onset was 24.51. Patients had a mean of 8.52 depressive episodes and 3.99 hypomanic episodes over their lifetimes. The results of comparative analyses between patients with BDII and corresponding control groups are presented below.

Inhibition

As can be seen in Table 2, only nine studies evaluated inhibition in patients with BDII. The tasks used to evaluate inhibitory control in these samples included the Stroop Test (k = 2), the Go/No-Go test (k = 2), the D-KEFS Color-Word Interference Set-shifting Score (k = 2), the HSCT (k = 1), the INECO Frontal Screening battery (k = 1), and the CPT (k = 1). Since no test was used more than twice, no test-specific effect sizes could be calculated. Instead, all measures were combined into a composite measure, whose effect sizes per study are shown in Fig. S7a. The overall effect size for inhibition was not associated with significant heterogeneity or publication bias.

Working memory

Only 14 studies evaluated working memory in patients with BDII. As was the case with BDI, the most commonly used measures of working

Table 2. Weighted mean effect size analyses between control participants and patients with bipolar disorder type II

					95% CI		Heterogeneity					Risk of bias	
	N_c	N_{BD}	k	g	LL	UL	Q	df	P	I^2	Trim-and-fill-adjusted g	Egger's P	LFK *index*
Inhibition													
Inhibition composite	556	270	9	0.60	0.45	0.74	8.24	8	0.41	2.91	0.60	0.33	0.21
Working memory													
Digit Span Fwd	602	325	7	0.27	0.09	0.45	8.22	6	0.22	27.01	0.27	0.73	0.82
Digit Span Bwd	653	333	7	0.45	0.30	0.60	7.03	6	0.32	14.62	0.33	0.34	1.78
Verbal WM composite	812	472	10	0.44	0.29	0.58	15.75	9	0.11	36.53	0.36	0.15	3.28
Overall WM composite	883	577	14	0.42	0.27	0.57	23.60	13	0.04	44.92	0.37	0.42	1.00
Flexibility													
TMT A time	554	423	9	0.53	0.40	0.67	8.10	8	0.42	1.27	0.59	0.15	−2.32
TMT B time	685	475	11	0.65	0.50	0.80	13.45	10	0.20	25.66	0.70	0.23	−2.31
WCST Composite	429	273	5	0.58	0.45	0.71	1.46	4	0.92	0.00	0.58	0.97	−1.10
Flexibility composite	855	564	15	0.59	0.50	0.69	13.74	14	0.47	0.00	0.61	0.66	−0.99
Fluency													
Letter fluency	498	270	7	0.42	0.23	0.62	10.41	6	0.17	32.78	0.42	0.85	0.28
Category fluency	647	339	7	0.56	0.37	0.75	10.02	6	0.12	40.15	0.63	0.78	−0.34
Fluency composite	785	462	11	0.48	0.35	0.61	12.16	10	0.27	17.77	0.57	0.21	−1.29
Episodic memory													
Immediate verbal memory composite	486	352	11	0.31	0.16	0.46	13.20	10	0.21	24.32	0.31	0.97	−0.30
Delayed verbal memory composite	472	335	10	0.22	0.07	0.36	11.58	9	0.24	22.31	0.14	0.35	0.28
Immediate visual memory composite	376	296	6	0.31	−0.02	0.63	13.44	5	0.02	62.80	0.47	0.01	−3.19
Delayed visual memory composite	236	203	6	0.32	0.14	0.50	4.01	5	0.55	0.00	0.35	0.47	−1.51
Episodic memory composite	737	529	13	0.33	0.19	0.46	21.10	12	0.05	43.14	0.39	0.31	−0.39

memory were variations of the Digit Span Task. As can be seen in Table 2, Digit Span Backward scores had a much higher effect size (g = 0.45; 95% CI: 0.30–0.60) than that observed for the Digit Span Forward. Importantly, the effect size for the Digit Span Backward did not show significant heterogeneity.

In addition to these instruments, measures of working memory reported by the studies included the WM score from the Wechsler Memory Scales (k = 4), the letter-number sequencing task (k = 2), the Paced Auditory Serial Addition Test (PASAT) (k = 1), a Sentence-Word Span Task (k = 1), the INECO Frontal Screening battery (k = 1), and the n-back (k = 1). Measures of verbal working memory used by fewer than five studies were included in the Composite verbal working memory score. Though the composite effect size for verbal working memory showed only moderate heterogeneity, it did evidence a publication bias toward larger effect sizes between control participants and patients with BDII.

Only two studies provided measures of spatial working memory for patients with BDII, and as such, a separate score for visuospatial working memory could not be calculated for these individuals. However, the effect sizes for spatial measures of working memory were included in the Overall Composite, whose effect sizes are shown in Fig. S7b. This score showed moderate heterogeneity and minor publication bias toward larger effect sizes between control participants and patients with BDII.

Shifting/Flexibility

As can be seen in Table 2, only 15 studies evaluated cognitive flexibility in individuals with BDII. The vast majority of these studies used the TMT or the WCST, with only a single study using the IFS to evaluate cognitive flexibility. The instrument with the largest effect size for cognitive flexibility was the TMT B (g = 0.65; 95% CI: 0.50–0.80). Effect sizes for flexibility showed very low heterogeneity, which speaks to the consistency of this finding.

Of the five studies which used the WCST, four employed the 128-card version of the instrument, while one used the modified, 48-card format, developed by Nelson (62). Given the small number of investigations which used this assessment instrument, and the fact that no single score was provided by all five studies, a composite WCST score was calculated by obtaining the average effect size for all WCST variables in each study. The results of composite effect sizes for cognitive flexibility can be seen in Fig. S7c.

The composite effect size for flexibility was not associated with significant publication bias. However, the LFK index revealed a major publication bias toward smaller effect sizes on the TMT A and B.

Fluency

Fluency tasks were used by 11 studies in BDII. The majority of investigations (k = 5) used fluency tasks from the COWAT/FAS or the D-KEFS battery. As was the case for BDI, the largest differences between control participants and those with BDII were observed in category fluency (g = 0.56; 95% CI: 0.37–0.75) rather than letter fluency. However, the effect sizes for category and letter fluency were both associated with moderate heterogeneity.

Only a single study evaluated design fluency in patients with BDII (59). The study in question did so using the D-KEFS Design Fluency task. Forest plots of composite effect sizes for fluency tasks, including letter, category, and design fluency scores, can be seen in Fig. S7d.

The composite score for fluency tasks was not associated with significant heterogeneity. However, the LFK index revealed minor publication bias toward smaller effect sizes between control subjects and patients with BDII.

Episodic memory

As can be seen in Table 2, 13 studies evaluated episodic memory in patients with BDII. Six studies used the WMS (k = 6), three studies used the CVLT (k = 3), two used the Rey Complex Figure Test (k = 2), and one each used the RAVLT (k = 1) and Claeson-Dahl Verbal Learning Test (k = 1). No single variable was reported by at least five studies, and the only test used by more than five investigations—the WMS—was not reported in a similar way by every article. Some studies reported separate scores for every task in the WMS, while others only provided summary scores. As such, in order to obtain a satisfactory sample size for comparative analyses, measures of episodic memory for patients with BDII were only evaluated at a construct level, where effect sizes from multiple tasks could be pooled. Similarly, since many studies did not provide separate episodic memory scores for visual and verbal or immediate and delayed tests, an overall 'episodic memory score' was calculated including every instrument

used to evaluate this construct across all 13 studies.

As can be seen in Table 2, the largest differences between control participants and patients with BDII were observed in immediate verbal memory and delayed visual memory, followed by delayed verbal memory. Effect sizes for these cognitive functions were not associated with significant levels of heterogeneity. Differences between control participants and patients with BDII in immediate visual memory were not significant. A forest plot of the overall composite effect size for episodic memory can be seen in Fig. S7e.

The LFK index did not reveal any significant bias in effect sizes pertaining to verbal memory or the episodic memory composite. However, both immediate and delayed visual memory scores were associated with publication bias toward smaller effect sizes between control participants and patients with BDII.

Planning

Only three studies evaluated planning in patients with BDII. Two of these studies used the TOH, while the other used the D-KEFS Tower test. A composite Planning score could not be calculated given the small sample size.

BDI vs. BDII

As previously mentioned, 13 studies compared patients with BDI to individuals with BDII. Patients with BDII were on average 36.81 years of age with 14.23 years of education, as compared to patients with BDI, who had a mean of 36.98 years of age and 13.67 years of education. The majority of samples were euthymic (46%), while in 7.69% of samples, patients were evaluated during a depressive episode. In 15.38% of cases, samples were heterogeneous with regard to mood state. The remaining studies did not provide a description of participant mood state at the time of testing.

The mean number of hospitalizations per patient was 2.22 for BDI and 0.426 for BDII. Mean values for length of illness and age of onset were 11.24 and 25.14 years for BDI, and 11.46 and 25 years for BDII respectively. Patients with BDI had a mean of 5.37 depressive episodes and 3.02 (hypo)-manic episodes, while those with BDII had a mean of 9.93 depressive episodes and 3.37 hypomanic episodes over their lifetimes. The results of comparative analyses between patients with BDI and BDII are presented in Table 3.

Cognitive flexibility was evaluated by 10 studies, seven of which provided a value for the TMT B time to completion. The remaining three studies used the WCST. As such, comparisons between patients with BDI and II were conducted for both the TMT B separately, and all measures of cognitive flexibility in the form of a summary score. Inhibition was evaluated by six studies, two of which used the D-KEFS Color-Word Interference task ($k = 2$), while the HSCT, Stroop task, Go/No-Go, and CPT commission errors were used by one study each. Eight studies evaluated working memory in patients with BDI and BDII. The tasks used to measure this construct included Digit Span Forward ($k = 5$), Digit Span Backward ($k = 4$), WMS Working Memory score ($k = 2$), n-back ($k = 1$), Sentence-Word Span ($k = 1$), letter-number sequencing ($k = 1$), and a Spatial Span task ($k = 1$). Seven studies compared the performance

Table 3. Weighted mean effect size analyses between patients with bipolar disorder type II and those with bipolar disorder type I

					95% CI		Heterogeneity					Risk of bias	
	N_{BDII}	N_{BDI}	*k*	*g*	LL	UL	*Q*	*df*	*P*	I^2	Trim-and-fill-adjusted *g*	Egger's *P*	LFK *index*
TMT A Time*	395	312	6	0.23	0.04	0.41	6.49	5	0.26	22.94	0.23	0.87	0.01
TMT B Time	351	433	7	0.28	0.03	0.52	14.39	6	0.03	58.31	0.28	0.82	−1.05
FC	424	531	10	0.22	0.03	0.41	18.66	9	0.03	51.77	0.22	0.92	−0.18
IC†	160	282	5	0.09	−0.09	0.27	3.66	4	0.45	0.00	0.09	0.85	−0.57
Digit Span Forward	269	365	5	0.11	−0.06	0.27	3.01	4	0.56	0.00	0.02	0.24	2.94
WMC	382	475	8	0.22	0.08	0.35	8.01	7	0.33	12.64	0.18	0.21	1.21
Category fluency	296	407	6	0.22	0.06	0.37	3.26	5	0.66	0.00	0.22	0.51	−2.18
VFC	312	432	7	0.18	0.04	0.31	4.27	6	0.64	0.00	0.18	0.74	−0.18
IVEM	228	292	7	0.32	0.10	0.54	10.88	6	0.09	44.86	0.13	0.31	1.78
DVEM	211	275	6	0.30	0.02	0.57	14.74	5	0.01	66.09	0.13	0.15	1.22
VEM	266	271	5	0.31	0.05	0.57	9.70	4	0.05	58.77	0.50	0.37	−0.53
EMC	359	384	8	0.35	0.15	0.55	17.64	7	0.01	60.31	0.35	0.66	

DVEM, delayed verbal episodic memory; EMC, episodic memory composite; FC, flexibility composite; IC, inhibition composite; IVEM, immediate verbal episodic memory; VEM, visual episodic memory; VFC, verbal fluency composite; WMC, working memory composite.

*Score reported for reference purposes and not included in composite effect size calculations.

†One outlier removed. The effect including the outlier was calculated at $g = 0.22$ (95% CI: −0.50 to 0.94).

of patients with BDI and BDII on fluency tasks. Six of these used category fluency tests ($k = 6$), while four used letter fluency tasks ($k = 4$) and one of these used a design fluency test ($k = 1$). Summary effect sizes for category fluency separately and fluency tasks as a whole are shown in Table 3.

Eight studies evaluated episodic memory, using predominantly the WMS ($k = 4$). Two studies also used the CVLT ($k = 2$), while the RBANS, Signoret Memory Battery, Claeson-Dahl, and Rey Complex Figure Test were used by one study each ($k = 1$). The coding procedures resulted in sample sizes which were sufficiently large for the calculation of separate effect sizes for immediate and delayed verbal episodic memory. However, visual memory scores could not be separated according to length of time between encoding and recall, since the resulting sample size would be too small for a separate analysis. As such, a summary score was provided for visual episodic memory including measures of both delayed and immediate recall.

Composite effect sizes for inhibition, working memory, and verbal fluency were not associated with significant heterogeneity. However, according to the I^2 statistics, cognitive flexibility and episodic memory were associated with moderate to substantial heterogeneity respectively.

The LFK index revealed evidence of publication bias for the comparison of some cognitive functions between BDI and BDII. A bias toward smaller differences was seen for the TMT B and Category fluency, while biases toward larger group differences were noted for working memory and verbal episodic memory.

Forest plots corresponding to composite effect sizes between patients with BDI and BDII can be seen in Fig. S8.

Moderator analyses

Given the high level of heterogeneity in the results of comparative analyses and the large number of variables known to affect cognition in patients with mood disorders, a series of meta-regressions were conducted in order to investigate which factors may influence the difference in cognitive performance between control subjects and patients with BD. Age, education, employment status, IQ, and reading level were extracted for patients and control subjects as potential moderators of effect size. For patients, specifically, the following data were also examined: medication use; scores on measures of depression and (hypo)mania; number of previous suicide attempts; number of previous hospitalizations; duration of illness; age of onset; mean number of past depressive episodes; mean number of (hypo)manic episodes; and current mood (euthymic/not euthymic).

The majority of variables were extracted and analyzed exactly as reported by the original studies, with three exceptions: depression ratings, age of onset/duration of illness, and medication use. Since depression ratings were obtained using different scales in some investigations, all scores were converted to Hamilton Depression Rating Scale (HDRS) equivalents, since this was the most prevalent assessment instrument in the studies retrieved. The conversion was conducted based on published data regarding the Montgomery-Asberg Depression Rating Scale (MADRS) and Inventory of Depressive Symptomatology (IDS)(63, 64).

Age of onset and duration of illness were only submitted to additional processing when one of these two variables was unavailable. In order to extract as much data as possible from the included studies, we calculated the mean age of onset and duration of illness for every study which provided only one of these variables as well as participants' age at the time of the study. In these cases, the value of the missing variable (either age of onset or duration of illness) was calculated by subtracting the available value (i.e. age of onset or duration of illness) from participants' age at the time of testing. This was carried out for 21 studies which provided only current age and duration of illness, and 18 which provided current age and the age of illness onset. This procedure resulted in a large enough sample size for each of these variables to be included in the meta-regression model for all cognitive components evaluated.

The data pertaining to medication use also required additional processing given the large variability in the format in which these data were provided. While 34 studies provided no quantitative data on medication use by participants, the remaining investigations varied widely in their reporting. In order to identify which format was most prevalent, and therefore, more likely to provide an adequate sample size to be used in moderator analyses, all medication data were extracted from every study, and frequency analyses were used to determine which medication variables were most common across studies. These analyses revealed that the most common method used to describe medication use in patient samples involved reporting the frequency of patients using lithium, anticonvulsants, and antipsychotic agents. These variables were available for 61 samples.

In order to ensure a large enough sample size for each analysis, rather than using specific tasks or scores as dependent variables, meta-regression analyses were conducted using summary measures

Table 4. Simultaneous moderator regression analyses

		95% CI										
	β	LL	UL	SE	*z*	*P*	N_c	N_{BD}	*k*	Q_w (*df*)	Qb (*df*)	*P*
Inhibition												
Patient age	−0.018	−0.112	0.075	0.048	−0.389	0.698	1192	1002	20	7.992 (6)	28.735 (13)	0.007
Age difference	0.002	−0.101	0.104	0.052	0.030	0.976						
Patient education	0.015	−0.120	0.150	0.069	0.223	0.823						
Education difference	−0.032	−0.196	0.131	0.083	−0.386	0.699						
HDRS	−0.091	−0.264	0.083	0.089	−1.027	0.305						
YMRS	−0.022	−0.059	0.014	0.019	−1.199	0.231						
Diagnosis	−0.063	−0.729	0.603	0.340	−0.185	0.853						
Length of illness	0.007	−0.604	0.079	0.036	0.197	0.844						
Age of onset	−0.084	−0.169	0.002	0.044	−1.922	0.055						
Euthymia	−1.286	−2.072	−0.502	0.400	−3.214	0.001						
% lithium	0.007	−0.002	0.017	0.005	1.479	0.139						
% anticonvulsants	0.009	−0.003	0.021	0.006	1.464	0.143						
% antipsychotics	−0.007	−0.020	0.005	0.006	−1.207	0.228						
Cognitive flexibility												
Patient age	−0.282	−0.574	0.010	0.149	−1.892	0.058	1105	1133	21	17.124 (7)	24.395 (13)	0.028
Age difference	−0.004	−0.069	0.061	0.033	−0.113	0.910						
Patient education	−0.123	−0.293	0.047	0.087	−1.420	0.156						
Education difference	−0.216	−0.486	0.053	0.138	−1.571	0.116						
HDRS	−0.009	−0.052	0.034	0.022	−0.414	0.679						
YMRS	0.263	−0.050	0.576	0.160	1.645	0.100						
Diagnosis	0.119	−0.487	0.726	0.309	0.385	0.700						
Length of illness	0.252	−0.041	0.546	0.150	1.688	0.091						
Age of onset	0.313	0.022	0.603	0.148	2.109	0.035						
Euthymia	−0.373	−1.344	0.597	0.495	−0.754	0.451						
% lithium	0.013	0.004	0.023	0.005	2.705	0.007						
% anticonvulsants	0.014	0.003	0.026	0.006	2.442	0.015						
% antipsychotics	−0.001	−0.016	0.015	0.008	−0.069	0.945						
Working memory												
Patient age	0.002	−0.209	0.213	0.108	0.021	0.983	1946	1648	29	13.063 (15)	60.724 (13)	<0.001
Age difference	−0.004	−0.036	0.027	0.016	−0.254	0.799						
Patient education	0.022	−0.067	0.111	0.045	0.486	0.627						
Education difference	0.147	0.023	0.271	0.063	2.324	0.020						
HDRS	−0.006	−0.030	0.018	0.012	−0.457	0.648						
YMRS	−0.189	−0.353	−0.026	0.084	−2.268	0.023						
Diagnosis	−0.081	−0.350	0.188	0.137	−0.591	0.554						
Length of illness	0.026	−0.172	0.223	0.101	0.253	0.800						
Age of onset	0.004	−0.192	0.201	0.100	0.044	0.965						
Euthymia	−0.129	−0.505	0.248	0.192	−0.670	0.503						
% lithium	0.003	−0.003	0.009	0.003	0.955	0.340						
% anticonvulsants	0.000	−0.009	0.009	0.005	0.067	0.946						
% antipsychotics	0.008	0.000	0.017	0.005	1.861	0.063						
Verbal fluency												
Patient age	−0.046	−0.270	0.179	0.115	−0.398	0.691	1709	1259	22	6.448 (8)	41.917 (13)	<0.001
Age difference	0.024	−0.033	0.080	0.029	0.824	0.410						
Patient education	−0.116	−0.210	−0.022	0.048	−2.428	0.015						
Education difference	0.013	−0.135	0.161	0.076	0.166	0.868						
HDRS	−0.005	−0.033	0.022	0.014	−0.391	0.696						
YMRS	0.006	−0.232	0.243	0.121	0.046	0.964						
Diagnosis	0.460	0.146	0.774	0.160	2.874	0.004						
Length of illness	0.018	−0.208	0.245	0.116	0.158	0.875						
Age of onset	0.056	−0.155	0.268	0.108	0.521	0.603						
Euthymia	−0.221	−0.664	0.221	0.226	−0.980	0.327						
% lithium	−0.001	−0.010	0.008	0.005	−0.257	0.798						
% anticonvulsants	−0.003	−0.013	0.008	0.005	−0.514	0.607						
% antipsychotics	−0.007	−0.018	0.005	0.006	−1.151	0.250						
Episodic memory												
Patient age	0.447	0.060	0.834	0.198	2.262	0.024	1774	1429	25	42.413 (11)	73.136 (13)	<0.001
Age difference	−0.228	−0.299	−0.156	0.036	−6.254	0.000						
Patient education	−0.238	−0.510	0.035	0.139	−1.709	0.087						
Education difference	−0.233	−0.499	0.034	0.136	−1.712	0.087						
HDRS	0.027	−0.026	0.080	0.027	0.985	0.324						
YMRS	−0.357	−0.673	−0.042	0.161	−2.218	0.027						
Diagnosis	0.186	−0.449	0.820	0.324	0.573	0.567						
Length of illness	−0.496	−0.851	−0.141	0.181	−2.741	0.006						
Age of onset	−0.396	−0.830	0.038	0.221	−1.787	0.074						
Euthymia	−0.397	−1.058	0.263	0.337	−1.179	0.238						
% lithium	0.019	0.007	0.031	0.006	3.021	0.003						
% anticonvulsants	−0.003	−0.022	0.016	0.010	−0.277	0.782						
% antipsychotics	0.008	−0.013	0.030	0.011	0.757	0.449						

of the effect size for each construct. Effect sizes were also combined across BDI and BDII, with diagnosis entered as a covariate. The need to ensure a large enough sample size and therefore obtain robust statistical power eliminated several potential predictors, which were reported for fewer than 20 samples. This was the case for employment status, reading scores, number of medications taken, percentage of patients on lithium monotherapy, medication load, mean lithium dosage, and number of suicide attempts. From the remaining variables, we sought to identify the most comprehensive model—with the largest number of predictors—which could be developed with a minimum sample size of 20 studies.

This procedure led to the identification of ten variables which were available for at least twenty studies of every major construct evaluated in the present study (inhibition, cognitive flexibility, working memory, verbal fluency, episodic memory). These variables were the following: age, education, depression scores, (hypo)mania scores, diagnosis (BDI or BDII), length of illness, age of onset, current mood (euthymic/not euthymic) and percentage of patients treated with lithium, antipsychotics, and anticonvulsants. The results of the simultaneous moderator analyses are shown in Table 4.

To control for any differences in demographic variables between the samples in each study, differences in age and education between control participants and clinical samples were also calculated and entered as predictors. The variables 'age difference' and 'education difference' shown in Table 4 were both calculated as follows: $age_{controlgroup}$—$age_{patients}$; $education_{controlgroup}$—$education_{patients}$. Positive values suggest that control participants are older and more educated, respectively, while negative values suggest that patients are older or more educated.

As the values in Table 4 show, all models were able to account for a significant proportion of the variability between studies. Inhibition was the only cognitive domain where effect sizes were significantly affected by mood status (euthymic/not euthymic), with euthymic samples showing smaller differences from control participants. There was also a trend toward significance for the effects of age of onset ($P = 0.055$); a younger age of onset was associated with larger differences between patients and control subjects for inhibitory control. Effect sizes for cognitive flexibility increase when larger percentages of patients are medicated with lithium or anticonvulsants, and have a younger age of onset. Group differences for working memory increase when patients are less educated, and decrease with scores on the YMRS. Effect sizes for verbal fluency increase when patients are less educated and have BDI rather than BDII. Lastly, group differences in episodic memory are lower when patients have higher YMRS scores, a longer disease duration and are more similar in age to the control group. Effect sizes for episodic memory increase when participants are older, and when a higher percentage of patients is taking lithium.

Discussion

The aim of the present study was to investigate EF and episodic memory impairments in BDI and BDII. We also sought to identify clinical, demographic, and cognitive reserve variables which may moderate the association between BD and cognitive impairment. The analysis of 126 studies with a total sample of 15402 participants ($n = 6424$ with BDI; $n = 702$ with BDII; $n = 8276$ control) revealed widespread cognitive impairments in BDI, with most effect sizes classified as moderate to large. Patients with BDII did show significant impairments relative to control participants, but these differences were mostly in the small-to-medium range. Small but significant differences were also identified between patients with BDI and BDII on all cognitive functions except inhibitory control. Moderator analyses suggested that each cognitive function may be affected by different variables, with only two factors accounting for significant amounts of variability in more than one regression model: Young Mania Rating Scale (YMRS) scores and lithium use. Each of these findings will be discussed in greater detail in the following paragraphs.

Our findings regarding the cognitive impact of BDI are in agreement with previous meta-analyses (10). The composite effect size for planning was numerically greater than all other composite effect sizes between patients and control subjects. However, these findings must be interpreted with caution, given the high heterogeneity associated with composite effect sizes for all cognitive functions except inhibition. The large effect size for planning must also be interpreted in light of findings from the risk of bias analysis, which identified major bias toward larger effect sizes between control participants and patients with BDI. All remaining effect sizes were classified as medium. The analysis of composite scores within each domain revealed that effect sizes were numerically larger for verbal than visual measures of both working memory and episodic memory, corroborating previous findings in the literature (11, 13). However, heterogeneity

was also larger for verbal memory instruments than those measuring visual memory. The numerical comparison of effect sizes for individual tests showed that the largest differences between control participants and patients with BDI were observed on the following variables: HSCT B (errors), Digit Span Total, Category Fluency, TMT B (time), WMS immediate auditory memory, and the CVLT long delay free recall. Importantly, the latter three were not associated with significant levels of heterogeneity, adding strength to this finding. These results corroborate those of previous meta-analyses, which identified medium to large effect sizes on measures including the TMT B, HSCT B, Digit Span Total, and Category fluency (8).

The comparison between control subjects and those with BDII revealed that only inhibition and cognitive flexibility could differentiate between these participants with at least a medium effect size. This finding is strengthened by the low heterogeneity of effect sizes associated with these two cognitive functions. The presence of differences in cognitive flexibility and inhibition between these participant groups has also been reported in previous meta-analyses (8, 10). These observations support the idea that executive dysfunction is perhaps the most prevalent and severe cognitive impairment in BDII (65). Effect sizes for working memory, verbal fluency, and episodic memory between patients with BDII and control participants were numerically smaller but still significant. The effect sizes calculated in the present study were similar to those reported in a previous meta-analysis of cognition in BDII, which also found that verbal memory showed the smallest differences between these participant groups (10). These findings highlight the fact that patients with BDII do show cognitive impairments relative to healthy control participants, especially in the EF. Though these impairments may be numerically smaller than those associated with BDI (9), they may still be clinically significant and lead to poorer functioning relative to healthy adults. Another important issue pertaining to the nature of cognitive impairment in BDII has to do with publication bias. According to the LFK indices calculated in the present study, the literature appears biased toward publishing smaller effect sizes between control participants and patients with BDII on measures of cognitive flexibility and episodic memory. This may lead to an artificial underestimation of cognitive impairments in BDII, which should be addressed in future studies.

Comparisons between patients with BDI and BDII revealed small but significant differences between these participants on episodic memory, cognitive flexibility, and working memory. Though effect sizes for episodic memory and cognitive flexibility were associated with moderate to substantial heterogeneity, this was not the case for working memory, which did not display significant heterogeneity across studies. In all three cognitive functions, patients with BDII outperformed those with BDI. A similar pattern of findings, where group differences favor patients with BDII on most measures of EF, has been obtained in previous meta-analyses (8). The largest numerical difference between these participant groups was observed in the verbal memory score. This corroborates previous meta-analyses where effect sizes for verbal memory between BDI and BDII were the highest among all cognitive functions examined (9, 10).

The meta-regression models also revealed significant sources of variability in effect sizes between studies. Though corresponding R^2 values must be interpreted with caution, given their low reliability in samples of fewer than 40 studies (66), our findings suggest that these cognitive functions and moderators may be especially promising targets for future studies. Scores on the YMRS and lithium use were significant predictors of effect sizes in two cognitive domains each. Scores on the YMRS decreased effect sizes for inhibition and WM, while samples with a larger proportion of lithium users showed larger effect sizes for cognitive flexibility and episodic memory. Findings regarding the effects of lithium corroborate previous studies suggesting that measures of cognitive flexibility and episodic memory, such as the TMT B and RAVLT, respectively, are especially sensitive to the effects of this drug (67). The positive effects of YMRS scores on effect sizes may have to do with the effects of subclinical hypomanic symptoms on performance. It is important to note that none of the studies included in the meta-regression involved patients with a current manic episode, and as such, their scores on the YMRS are indicative of subclinical or residual symptoms. Though full-blown manic episodes are known to have a negative effect on cognitive performance (68), *symptoms* of (hypo)mania may have a different impact on individual functioning. Previous studies have identified 'positive' dimensions of hypomania, associated with symptoms such as increased activity and energy, which may actually have a beneficial effect on cognition (69, 70). Given the small sample size of the current study, however, and the methodological heterogeneity in the literature, these observations must be interpreted with caution. Another interesting finding from the moderator analysis has to do with the cognitive effects of education. In measures of working memory and verbal fluency,

highly educated patient samples were associated with smaller effect sizes upon comparison with control participants. The effect of education-related variables, such as cognitive reserve, on these two particular functions has been reported before. In a study by Grande et al. (26), although cognitive reserve had a significant influence on several cognitive functions, working memory and verbal fluency were among the only domains where these effects remained significant after controlling for clinical variables. These findings highlight the importance of cognitive reserve as a protective factor against cognitive impairment in BD.

The present findings should be interpreted in light of some limitations, such as the fact that patients in the studies analyzed were not strictly euthymic. This methodological choice was made in an attempt to provide a more comprehensive picture of the current literature and to ensure a more naturalistic sample. The comprehensiveness issue stems from the fact that several observational studies do not use euthymia as an inclusion criterion, or choose to admit patients with subsyndromal mood symptoms. The inclusion of these patients in clinical trials has actually been recommended in some cases, in the interest of recruitment feasibility and generalization of results (71). As such, since one aim of this study was precisely to provide a more comprehensive review of the literature than has been previously attempted, it was important to include studies which evaluated non-euthymic as well as euthymic samples. The need to ensure a more naturalistic sample was also important, so that these findings may apply to a larger set of patients. Interepisodic mood symptoms are common in BD, as is the occurrence of full-fledged mood episodes (72). To ensure our findings would be representative of patients seen in clinical practice, it was important to include unremitted individuals in this analysis. Previous meta-analyses have followed a similar approach (12), and in one recent study, mood status (euthymic/not euthymic) was not found to moderate the cognitive effects of BDI and BDII (9). These finding were largely replicated in our meta-regression, where euthymia was only associated with effect sizes for one cognitive function.

Another possible limitation is the correlation between test scores, which may have raised the issue of statistical dependence when results were combined across studies for the calculation of composite effect sizes. However, in this regard, it is important to note that, though cognitive constructs may be related to one another (73), the scores on cognitive tests themselves are not reliably correlated (74, 75). Also, all effect sizes per cognitive component per study were collapsed into a single value, so that each study only contributed one effect size to each calculation.

Two additional concerns pertain to the inclusion criteria for the meta-analysis and to the cognitive abilities examined. When selecting articles from the literature, the authors of the present study specified that only those published since 2008 would be screened for inclusion in the meta-analysis. Though this time period does capture the majority of studies of cognition in BD, it may have led to the exclusion of important studies conducted prior to 2008. The last issue which may be viewed as a limitation is the fact that the present analysis was restricted to the EF and memory, rather than encompassing other cognitive abilities such as attention and processing speed. Though memory and the EF have indeed been cited as potential endophenotypes of BD and are associated with high rates of functional impairment, they are by no means the only cognitive functions affected in this disorder. Impairments of varying severity in attention and processing speed have also been reported in the literature (7). As such, though this was beyond the scope of the present investigation, future studies may wish to conduct a similar analysis of the literature on cognitive skills including processing speed and attention.

Despite these limitations, the present study makes some important contributions to the literature and suggestions for future investigations. The present study shed light on the scarcity of comparative studies between BDI and BDII, and the methodological heterogeneity in the cognitive study of these disorders. This reflects the findings of previous authors, who have identified heterogeneity as a major obstacle to drawing significant and clinically applicable conclusions from reviews and meta-analyses (1, 8). We echo the suggestions of researchers who call for stricter adherence to current methodological recommendations in the study of cognition in BD (76).

This was also one of the largest existing meta-analyses of cognition in BDI and BDII, comparing these conditions to one another and to control groups, in order to verify whether these disorders are associated with cognitive impairments relative to the general population. The approach adopted in this investigation provided both a comprehensive picture of cognition, by analyzing multiple cognitive domains, but also a detailed look at individual assessment instruments and scores. The fact that effect sizes were calculated for individual neuropsychological tests may help researchers develop test batteries which are more sensitive to the cognitive impairments in different disorders.

Instruments such as the TMT B, HSCT B, Digit Span Total, and Category fluency, specifically, should be included in the neuropsychological assessment batteries administered to patients with BD. Another important finding pertains to the significance of executive dysfunction relative than memory impairment in BDII; patients with BDI, on the other hand, showed moderate to severe impairments in most of the cognitive functions examined. These findings may outline a pattern of neurocognitive performance which may help differentiate between BDI and BDII. Lastly, findings regarding moderator variables may also help investigators comprehend the association between BD and the presence of impairments in different cognitive abilities, highlighting the potential effects of education on cognition. These and other moderating factors identified in the present study may be candidate variables for longitudinal studies of cognition in these populations.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pesssoal de Nível Superior – Brasil (CAPES) - Finance Code 001.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

Scripts and datasets are available from the authors upon request.

References

1. Cullen B, Ward J, Graham NA et al. Prevalence and correlates of cognitive impairment in euthymic adults with bipolar disorder: a systematic review. J Affect Disord 2016;**205**:165–181.
2. Martino DJ, Samamé C, Ibáñez A, Strejilevich SA. Neurocognitive functioning in the premorbid stage and in the first episode of bipolar disorder: a systematic review. Psychiatry Res 2015;**226**:23–30.
3. Burdick KE, Russo M, Frangou S et al. Empirical evidence for discrete neurocognitive subgroups in bipolar disorder: clinical implications. Psychol Med 2014;**44**:3083–3096.
4. Martino DJ, Marengo E, Igoa A et al. Neurocognitive and symptomatic predictors of functional outcome in bipolar disorders: a prospective 1 year follow-up study. J Affect Disord 2009;**116**:37–42.
5. Raust A, Daban C, Cochet B, Henry C, Bellivier F, Scott J. Neurocognitive performance as an endophenotype for bipolar disorder. Front Biosci (Elite Ed) 2014;**6**:89–103.
6. Kessler U, Schoeyen HK, Andreassen OA et al. Neurocognitive profiles in treatment-resistant bipolar I and bipolar II disorder depression. BMC Psychiatry 2013;**13**:105.
7. Cotrena C, Branco LD, Shansis FM, Fonseca RP. Executive function impairments in depression and bipolar disorder: association with functional impairment and quality of life. J Affect Disord 2016;**190**:744–753.
8. Dickinson T, Becerra R, Coombes J. Executive functioning deficits among adults with Bipolar Disorder (types I and II): a systematic review and meta-analysis. J Affect Disord 2017;**218**:407–427.
9. Bora E. Neurocognitive features in clinical subgroups of bipolar disorder: a meta-analysis. J Affect Disord 2018;**229**:125–134.
10. Bora E, Yücel M, Pantelis C, Berk M. Meta-analytic review of neurocognition in bipolar II disorder. Acta Psychiatr Scand 2011;**123**:165–174.
11. Bora E, Pantelis C. Meta-analysis of cognitive impairment in first-episode bipolar disorder: comparison with first-episode schizophrenia and healthy controls. Schizophr Bull 2015;**41**:1095–1104.
12. Bo Q, Mao Z, Li X, Wang Z, Wang C, Ma X. Use of the MATRICS consensus cognitive battery (MCCB) to evaluate cognitive deficits in bipolar disorder: a systematic review and meta-analysis. PLoS ONE 2017;**12**: e0176212.
13. Soraggi-Frez C, Santos FH, Albuquerque PB, Malloy-Diniz LF. Disentangling working memory functioning in mood states of bipolar. Disorder: A Systematic Review. Front Psychol 2017;**8**. https://doi.org/10.3389/fpsyg.2017.00574
14. Miskowiak KW, Carvalho AF, Vieta E, Kessing LV. Cognitive enhancement treatments for bipolar disorder: a systematic review and methodological recommendations. Eur Neuropsychopharmacol 2016;**26**:1541–1561.
15. Lee RSC, Hermens DF, Naismith SL et al. Neuropsychological and functional outcomes in recent-onset major depression, bipolar disorder and schizophrenia-spectrum disorders: a longitudinal cohort study. Transl Psychiatry 2015;**5**:e555–e510.
16. Romero E, Holtzman JN, Tannenhaus L et al. Neuropsychological performance and affective temperaments in Euthymic patients with bipolar disorder type II. Psychiatry Res 2016;**238**:172–180.
17. Nguyen TT, Kovacevic S, Dev SI, Lu K, Liu TT, Eyler LT. Dynamic functional connectivity in bipolar disorder is associated with executive function and processing speed: a preliminary study. Neuropsychology 2017;**31**:73–83.
18. Malloy-Diniz LF, Neves FS, Abrantes SSC et al. Suicide behavior and neuropsychological assessment of type I bipolar patients. J Affect Disord 2009;**112**:231–236.
19. Oertel-Knöchel V, Reuter J, Reinke B et al. Association between age of disease-onset, cognitive performance and cortical thickness in bipolar disorders. J Affect Disord 2015;**174**:627–635.
20. Bowie CR, Best MW, Depp C et al. Cognitive and functional deficits in bipolar disorder and schizophrenia as a function of the presence and history of psychosis. Bipolar Disord 2018;**20**:604–613.
21. Maalouf FT, Klein C, Clark L et al. Impaired sustained attention and executive dysfunction: bipolar disorder versus depression-specific markers of affective disorders. Neuropsychologia 2010;**48**:1862–1868.
22. Vrabie M, Marinescu V, Talaşman A, Tăutu O, Drima E, Micluţia I. Cognitive impairment in manic bipolar patients: important, understated, significant aspects. Ann Gen Psychiatry 2015;**14**:41.
23. Solé B, Jiménez E, Torrent C et al. Cognitive variability in bipolar II disorder: who is cognitively impaired and who is preserved. Bipolar Disord 2016;**18**:288–299.
24. Volkert J, Kopf J, Kazmaier J et al. Evidence for cognitive subgroups in bipolar disorder and the influence of subclinical depression and sleep disturbances. Eur Neuropsychopharmacol 2015;**25**:192–202.
25. Anaya C, Torrent C, Caballero FF et al. Cognitive reserve in bipolar disorder: relation to cognition, psychosocial

functioning and quality of life. Acta Psychiatr Scand 2016;**133**:386–398.
26. Grande I, Sanchez-Moreno J, Sole B et al. High cognitive reserve in bipolar disorders as a moderator of neurocognitive impairment. J Affect Disord 2017;**208**:621–627.
27. Hinrichs KH, Easter RE, Angers K et al. Influence of cognitive reserve on neuropsychological functioning in bipolar disorder: findings from a 5-year longitudinal study. Bipolar Disord 2017;**19**:50–59.
28. Frieden TR. Evidence for health decision making — Beyond randomized. Controlled Trials. N Engl J Med 2017;**377**:465–475.
29. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;**16**:495.
30. Bora E, Özerdem A. Meta-Analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia. Psychol Med 2017;**47**: 2753–2766.
31. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. Front Psychol 2015;**6**:1–24. https://doi.org/10.3389/fpsyg.2015.00328
32. Cotrena C, Branco LD, Ponsoni A, Shansis FM, Fonseca RP. Neuropsychological clustering in bipolar and major depressive disorder. J Int Neuropsychol Soc 2017;**23**: 584–593.
33. Cankorur VŞ, Demirel H, Atbaşoğlu C. Cognitive functioning in euthymic bipolar patients on monotherapy with novel antipsychotics or mood stabilizers. Noropsikiyatri Ars 2017;**54**:244–250.
34. Miskowiak KW, Kjærstad HL, Meluken I et al. The search for neuroimaging and cognitive endophenotypes: a critical systematic review of studies involving unaffected first-degree relatives of individuals with bipolar disorder. Neurosci Biobehav Rev 2017;**73**:1–22.
35. Moher D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;**151**:264.
36. Stroup DF. MOOSE statement: meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA 2000;**283**:2008.
37. Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. Psychol Bull 2013;**139**:81–132.
38. Diamond A. Executive functions. Annu Rev Clin Psychol 2014;**64**:135–168.
39. Scott JC, Matt GE, Wrocklage KM et al. A quantitative meta-analysis of neurocognitive functioning in posttraumatic stress disorder. Psychol Bull 2015;**141**: 105–140.
40. Cooper H, Hedges L, Valentine J. eds. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation, 2009.
41. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, FL: Academic Press, 1985.
42. Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. Res Synth Methods 2010;**1**:112–125.
43. Jackson D, Turner R. Power analysis for random-effects meta-analysis. Res Synth Methods 2017;**8**:290–302.
44. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester, UK: John Wiley & Sons Ltd; 2009.
45. Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: the inverse variance heterogeneity model. Contemp Clin Trials 2015;**45**:130–138.
46. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;**315**:629–634.
47. Barendregt JJ, Doi SAR. MetaXL user guide. 2014;Version 2.:1–52.
48. Duval S, Tweedie R. A nonparametric, "trim and fill" method of accounting for publication bias in meta-analysis. J Am Stat Assoc 2000;**95**:89–98.
49. Barendregt JJ.Doi SAR. MetaXL. 2016.
50. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw 2010;**36**:1–48.
51. DelRe AC, Hoyt WT. MAd: Meta-Analysis with Mean Differences. 2014.
52. Lin L, Chu H. altmeta: Alternative meta-analysis methods. 2016. https://cran.r-project.org/package = altmeta
53. Team RC. R: A language and environment for statistical computing; 2017. http://www.r-project.org/
54. Zhang Q, Shen Q, Xu Z et al. The effects of CACNA1C gene polymorphism on spatial working memory in both healthy controls and patients with schizophrenia or bipolar disorder. Neuropsychopharmacology 2012;**37**:677–684.
55. Carrus D, Christodoulou T, Hadjulis M et al. Gender differences in immediate memory in bipolar disorder. Psychol Med 2010;**40**:1349–1355.
56. Vaskinn A, Sundet K, Simonsen C, Hellvin T, Melle I, Andreassen OA. Sex differences in neuropsychological performance and social functioning in schizophrenia and bipolar disorder. Neuropsychology 2011;**25**:499–510.
57. Bücker J, Popuri S, Muralidharan K et al. Sex differences in cognitive functioning in patients with bipolar disorder who recently recovered from a first episode of mania: data from the Systematic Treatment Optimization Program for Early Mania (STOP-EM). J Affect Disord 2014;**155**:162–168.
58. Gur RC, Richard J, Hughett P et al. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. J Neurosci Methods 2010;**187**:254–262.
59. Pålsson E, Figueras C, Johansson AG et al. Neurocognitive function in bipolar disorder: a comparison between bipolar I and II disorder and matched controls. BMC Psychiatry 2013;**13**:165.
60. Miguélez-Pan M, Pousa E, Cobo J, Duño R. Cognitive executive performance influences functional outcome in euthymic type I bipolar disorder outpatients. Psicothema 2014;**26**:166–173.
61. Schretlen DJ, Pena J, Aretouli E et al. Confirmatory factor analysis reveals a latent cognitive structure common to bipolar disorder, schizophrenia, and healthy adults. Bipolar Disord 2013;**15**:422–433.
62. Nelson HE. A modified card sorting test sensitive to frontal lobe defects. Cortex 1976;**12**:313–324.
63. Carmody TJ, Rush AJ, Bernstein IH, Brannan S, Husain MM, Trivedi MH. Making clinicians lives easier: guidance on use of the QIDS self-report in place of the MADRS. J Affect Disord 2006;**95**:115–118.
64. Carmody TJ, Rush AJ, Bernstein I et al. The Montgomery Äsberg and the Hamilton ratings of depression: a comparison of measures. Eur Neuropsychopharmacol 2006;**16**:601–611.
65. King S, Stone JM, Cleare A, Young AH. A systematic review on neuropsychological function in bipolar disorders type I and II and subthreshold bipolar disorders—something to think about. CNS Spectr 2019;**24**:127–143.

66. López-López JA, Marín-Martínez F, Sánchez-Meca J, Van denNoortgate W, Viechtbauer W. Estimation of the predictive power of the model in mixed-effects meta-regression: a simulation study. Br J Math Stat Psychol 2014;**67**:30–48.
67. Malhi GS, Mcaulay C, Gershon S et al. The Lithium Battery: assessing the neurocognitive profile of lithium in bipolar disorder. Bipolar Disord 2016;**18**:102–115.
68. Basso MR, Lowery N, Ghormley C et al. Neuropsychological impairment and psychosis in mania. J Clin Exp Neuropsychol 2009;**31**:523–532.
69. Glaus J, VanMeter A, Cui L, Marangoni C, Merikangas KR. Factorial structure and familial aggregation of the Hypomania Checklist-32 (HCL-32): results of the NIMH family study of affective spectrum disorders. Compr Psychiatry 2018;**84**:7–14.
70. Angst J, Meyer TD, Adolfsson R et al. Hypomania: a transcultural perspective. World Psychiatry 2010;**9**:41–49.
71. Miskowiak K, Burdick K, Martinez-Aran A et al. Methodological recommendations for cognition trials in bipolar disorder by the International Society for Bipolar Disorders Targeting Cognition Task Force. Bipolar Disord Published Online First 2017;**19**:614–626.
72. McKnight RF, Bilderbeck AC, Miklowitz DJ, Hinds C, Goodwin GM, Geddes JR. Longitudinal mood monitoring in bipolar disorder: Course of illness as revealed through a short messaging service. J Affect Disord 2017;**223**:139–145.
73. Karr JE, Areshenkoff CN, Rast P, Hofer SM, Iverson GL, Garcia-Barrera MA. The unity and diversity of executive functions: a systematic review and re-analysis of latent variable studies. Psychol Bull 2018;**144**:1147–1185.
74. Jantscher S, Willinger U, Schmoeger M, Mueller C, Auff E. P01–417 - validation of the hayling sentence completion test – German version & Stroop-Test. Eur Psychiatry 2011;**26**:420.
75. Bock O, Haeger M, Voelcker-Rehage C. Structure of executive functions in young and in older persons. PLoS ONE 2019;**14**:e0216149.
76. Miskowiak KW, Burdick KE, Martínez-Arán A et al. Assessing and addressing cognitive impairment in bipolar disorder: the International Society for Bipolar Disorders Targeting Cognition Task Force recommendations for clinicians. Bipolar Disord 2018;**20**:184–194.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite effect sizes of inhibition.
Figure S2. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite scores of working memory.
Figure S3. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite effect sizes for flexibility.
Figure S4. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite fluency scores.
Figure S5. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite effect sizes of episodic memory.
Figure S6. Forest plot showing the main effect of group (bipolar disorder type I *versus* control subjects) for composite effect sizes of planning.
Figure S7. Forest plots showing the main effect of group (bipolar disorder type II *versus* control subjects) for composite effect sizes of cognitive flexibility, working memory, inhibition, verbal fluency and episodic memory.
Figure S8. Forest plots showing the main effect of group (bipolar disorder type I *versus* bipolar disorder type II) for composite scores of cognitive flexibility, working memory, inhibitory control, verbal fluency and episodic memory.