Review title:

Group-Based Community Interventions to Support the Social Reintegration of Marginalized Adults with Mental Illness: A Systematic Review and Meta-Analysis

Authors and affiliations

Nina Thorup Dalgaard, VIVE The Danish Center for Social Science Research
Jakob Kaarup Jensen, VIVE The Danish Center for Social Science Research
Jasmin Sami Adada, VIVE The Danish Center for Social Science Research
Maya Christiane Flensborg Jensen, VIVE The Danish Center for Social Science Research
Elizabeth Bengtsen, VIVE The Danish Center for Social Science Research
Mikkel Helding Vembye, VIVE The Danish Center for Social Science Research

Contact person

Nina Thorup Dalgaard, VIVE The Danish Center for Social Science Research

Abstract

Background

Adults suffering from mental illness constitute a vulnerable population with an increased risk of experiencing co-morbidity. Common co-morbid conditions include personal and social problems such as substance or alcohol abuse, self-harming behavior, criminal behavior, homelessness, long-term unemployment, poverty and social isolation. A co-morbidity that increase the risk of (social)marginalization. In order to support the social reintegration of marginalized adults with mental illness and related problems, a number of interventions exist. For example, occupational therapy, intensive case management, psycho-education, supportive psychotherapy or mentoring are targeting people with mental disorder and related problems (e.g. substance or alcohol abuse, criminal behavior, homelessness and marginalization). These interventions are costly and time consuming, and the evidence regarding their efficacy is far from unequivocal. Therefore, more recently, the use of group-based interventions has expanded as an alternative to individual therapy or other individually delivered interventions.

Objectives

The main objective was first:

- To explore the general efficacy of group-based community interventions aimed at supporting marginalized adults with mental illness and related problems on outcomes related to social reintegration, such as subjective well-being, alcohol and substance use, loneliness, homelessness, poverty, and employment.
- To explore the general efficacy of group-based community interventions aimed at supporting marginalized adults with mental illness and related problems on outcomes related to mental health, such as anxiety, depression, and symptoms of psychosis.
- A secondary objective was to explore the potential advantages/disadvantages of using a group-based versus an individual intervention when targeting specific problems or when using specific types of interventions.

Search methods

The following databases were searched electronically: MEDLINE,EMBASE (OVID) 1974 – 2022, APA PsycINFO (EBSCO), CINAHL (EBSCO), Sociological Abstracts (ProQuest) Social Services Abstracts (ProQuest), SocINDEX (EBSCO), Academic Search Premier (EBSCO), International Bibliography of the Social Sciences (IBSS) (ProQuest), Science Citation Index (Web of Science Core Collection), Social Sciences Citation Index (Web of Science Core Collection) , Cochrane Central Register of Controlled Trials (CENTRAL). The electronic database searches were completed in September 2022 and other resources were completed in July 2024. We searched to identify both published and unpublished literature. The searches were international in scope. The searches were limited to publications published from 2000. In addition to electronic searches we implemented a wide range of search methods and strategies to maximise coverage of relevant references

Selection criteria

The population of this review are adults in the OECD countries with at least one psychiatric diagnosis who are experiencing any kind of personal and social problems in addition to their mental health problems. Social or personal problems are defined broadly and include problems such as loneliness, homelessness or alcohol or substance use. This review includes all interventions delivered in a group format, meaning that more than one participant receive the intervention at the same time and place and by the same therapists/case workers. In addition, included interventions had to be based in a community or out-patient setting. In order to summarize what is known about the possible causal effects of group-based community interventions, we included all study designs (including RCTS) that use a well-defined control group. Control interventions consist of individually delivered interventions.

Data collection and analysis

The total number of potential relevant records was 18.832 after excluding duplicates (database: 8534; grey, hand search, snowballing and other resources: 10.098). All records were screened based on title and abstract; 18.009 were excluded for not fulfilling the screening criteria. including 35 records that were unobtainable despite efforts to locate them through libraries and

searches on the Internet. 623 records were ordered, retrieved and screened in full text. Of these, 561 did not fulfil the screening criteria and were excluded. 62 studies were included in the review.

One study was deemed to be at critical risk of bias due to high/serious risk of bias in multiple domains. Based on the review protocol, this study was excluded from the meta-analyses. Nine studies did not report results in a manner that allowed us to calculate effect sizes, and these were also not used in the data synthesis. 49 studies were used in the meta-analysis.

Main results

The 48 studies used in the meta-analyses were published between 2000 and 2022. Participants in the included studies suffered from a wide range of different types of mental illness and indicators of social marginalization. The most common mental disorder reported in primary studies was schizophrenia or other primary psychotic disorders (36 studies) and the most common co-morbid condition was substance use (13 studies) in the majority of studies, the participants faced more than one indicator of social marginalization. The group-based interventions were diverse in nature. However, the most common type of intervention was Group-based Cognitive Behavioral Therapy (11 studies). The total number of participants across studies were XX

A series of 10 meta-analyses were conducted. All the meta-analyses showed a weighted average effect that favored group-based interventions, although not all results were statistically significant. Results of the two main analyses using outcomes measuring all social reintegration outcomes and all mental health outcomes were both statistically significant and favoured group-based interventions. The random effects weighted standardised mean difference was 0.18 (95% confidence interval (Cl):0.11-0.24 for the meta-analysis using all social reintegration outcomes and the random effects weighted standardised mean difference was 0.18 (95% confidence interval (Cl):0.06- 0.3 for the meta-analysis using all mental health outcomes.

Authors' conclusions

Finding based on the series of meta-analyses suggest that for adults who suffer from both mental illness and face indicators of social marginalization, group-based interventions are a promising type of intervention. Our findings suggest that on measures of all types of mental health symptoms and all social reintegration outcomes, group-based interventions have larger average effects than usual care if delivered as an individual intervention. All though not all meta-analyses were statistically significant all average effect sizes favoured group-based interventions indicating that there were no adverse effects of group-based interventions compared with individually delivered control interventions.

Plain language summary

Group-based Interventions for Marginalized Adults with Mental Illness are more effective than Individually delivered Interventions

The review in brief

For adults who in addition to mental illness face problems such as loneliness, homelessness, substance and alcohol abuse or other indicators of social marginalization, group interventions outperform individually delivered interventions on a number of outcomes. Results from the meta-analyses suggest that for adults who suffer from both mental illness and social marginalization group interventions are either more effective or have similar effects compared with individually delivered interventions.

What is this review about?

Adults with mental illness are a vulnerable group at higher risk of experiencing additional challenges, such as substance abuse, self-harm and social isolation. Various interventions, including occupational therapy, intensive case management, and mentoring, aim to support their social reintegration. However, these approaches can be costly and time-intensive. As a result, group-based interventions have gained popularity as a more feasible alternative to individual therapy, e.g. because it is possible to treat many patient simulatnously and the group offer a social platform. The aim of this review was to explore if interventions delivered in a group to adults who suffer from both mental illness and social marginalization are more effective than individually delivered interventions on measures of both mental health and social reintegration.

This review includes studies of interventions delivered in a group format, where multiple participants receive support simultaneously from the same therapists or caseworkers. Additionally, the interventions had to take place in a community or outpatient setting. Control interventions are individually delivered interventions. Participants in the studies were adults who suffer from both mental illness and social marginalization.

This systematic review and meta-analyses summarises evidence from 62 studies.

What is the aim of this review?

This systematic review and series of meta-analyses summarises evidence from 62 primary studies. The aim of the review was to explore if interventions delivered in a group to adults who suffer from both mental illness and social marginalization are more effective than individually delivered interventions on measures of both mental health and social reintegration.

What are the main findings of this review?

.

What at studies are included?

This review includes studies that evaluate the effects of group-based interventions to support the social reintegration of marginalized adults who suffer from both mental illness and other problems when compared with an individually delivered intervention. A total of 62 studies were identified. However, only 57 of these were assessed to be of sufficient methodological quality to be included in the data synthesis, and only 48 studies could be used in at least one meta-analysis. The studies spanned the period from 2000 to 2022 and were carried out in the OECD countries.

The included studies were mostly randomised control trials and all included studies had a well defined control group.

Main Findings

The 48 studies included in the meta-analyses were published between 2000 and 2022 and involved participants with various mental illnesses and indicators of social marginalization. Schizophrenia and other psychotic disorders were the most common mental health conditions (36 studies), while substance use was the most frequently reported co-occurring issue (13 studies). Most participants faced multiple challenges related to social exclusion. The interventions varied but were most commonly group-based Cognitive Behavioral Therapy (11 studies).

Ten meta-analyses were conducted, all showing overall positive effects of group-based interventions, though not all results were statistically significant. The two main analyses, one examining social reintegration outcomes and the other focusing on mental health outcomes both showed statistically significant benefits of group-based interventions. A unique characteristic of this body of literature is that more than 50% of the studies represent preregistered studies. Although prespecified generally were not able to explain the differences between effect size estimates, our review suggests that preregistered studies yield lower but still substantial effects relative to non-registered studies. However, we did not find this relationship when adding focal theoretical and methodological moderators to the subgroup model, and we did not find any further clear evidence for publication and/or small study biases.

What do the findings of this review mean?

The findings suggest that group-based interventions are a promising approach for adults experiencing both mental illness and social marginalization. Across all measures of mental health symptoms and social reintegration, group-based interventions showed greater overall benefits compared to usual care delivered individually. While not all meta-analyses reached statistical significance, the consistent trend in favor of group-based interventions indicates that they are at least as effective as individual interventions, with no evidence of harmful effects

How up-to-date is this review?

The review authors searched for studies up to 2024.

Background

The problem, condition, or issue

Adults suffering from mental illness constitute a vulnerable population with an increased risk of experiencing co-morbidity. Common co-morbid conditions include personal and social problems such as substance or alcohol abuse, self-harming behaviour, criminal behaviour, homelessness, long-term unemployment, poverty and social isolation. These problems increase the risk that

mental illness leads to (social) marginalization, stigmatization and increased welfare costs (Draine et al., 2002; Lai et al., 2015; Nielsen et al., 2011; Schreiter et al., 2017).

Several studies suggest that mental illness, discrimination and (self-) stigmatization may become part of a vicious cycle. A cycle in which adults who suffer from mental illness abstain from engaging in social activities, which may lead to further marginalization and sometimes to a further deterioration in mental health (Brouwers, 2020; Feldman & Crandall, 2007). For example, in a qualitative study based on interviews with 46 adults suffering from a wide range of mental health diagnoses, Dinos, Stevens, Serfaty, Weich, & King (2004) found that participants described experiencing stigma even in the absence of overt discrimination by others or within society. In the study, participants describe how their experiences of stigma often cause stress, anxiety and rumination, and how this fear of being stigmatised leads to self-isolating and self-limiting behaviours. Many adults suffering from mental illness thereby have to cope with both their mental illness and the risk of social marginalization at the same time ,(Dinos et al., 2004).

In order to support the social reintegration of marginalized adults with mental illness and related problems, a number of interventions exist. For example, occupational therapy, intensive case management, psycho-education, supportive psychotherapy or mentoring are targeting people with mental disorder and related problems (e.g. substance or alcohol abuse, criminal behaviour, homelessness and marginalization). These interventions are costly and time consuming, and the evidence regarding their efficacy is far from unequivocal(Dutra et al., 2008; Sledge et al., 2011; Ziguras & Stuart, 2000). Therefore, more recently, the use of group-based interventions has expanded as an alternative to individual therapy or other interventions.

The growing demand for and use of group-based interventions happen in a context where most high-income countries' mental health services have been transformed from hospital-centred to community based services. A transformation that leave more responsibility and/or the cost of treatment and interventions to community-based services (Wahlbeck et al., 2011). From a community-based service perspective, the implementation of group-based interventions is increasingly celebrated as a way to bridge the gap between a growing demand for treatment and limited budgets for outpatient interventions (Ruesh et al., 2015).

Group interventions have the advantage of being able to treat many patients simultaneously. Therefore, the costs are low (Ruesh et al., 2015). In addition, Ruesh et al. (2015) find that group-based interventions in relation to depression treatment are marginally inferior or have similar effects as individual therapy. For patients with co-morbid mental illness group-based intervention may also be beneficial because the group offer social benefits through the reduction of the individual's feelings of loneliness and social isolation (Ruesh et al., 2015).

The high prevalence of personal and social co-morbidities for psychiatric patients, the changed institutional setting in mental healthcare, and the popularity of group-based community interventions (partly driven by budget concerns) create a demand for a thorough literature review in the field. Hence, the purpose of our review is to provide insights regarding efficacy of group-based community interventions for marginalized adults with mental illness.

The intervention

Group-based interventions can be adapted for different (mental) disorders, age groups and diverse communities and settings. Group-based interventions will often be provided in a small, selected group of individuals who meet regularly with a therapist or case worker (Fehr, 2019).

This review includes all interventions targeting adults who suffer from mental illness and related social and personal problems if the intervention is delivered in a group format, meaning that more than one participant receive the intervention at the same time and place and by the same therapists/case workers/mentors etc. In addition, included interventions had to be based in a community or out-patient setting. We excluded psychiatric interventions based on psychopharmacological treatment alone and interventions taking place in hospital settings while patients are receiving around the clock care.

In order to be eligible for the present review, included group-based intervention had to be aimed at supporting the social reintegration of participants. This means that interventions with the sole focus of reducing symptoms of a specific mental health diagnosis were not eligible. The review includes interventions for participants with all types of mental illness symptoms as long as the intervention also targets other aspects of the participants' lives and well-being. Examples of personal/social problems, which the interventions target are:

- Alcohol/substance abuse,
- Self-harming behaviour,
- Criminal behaviour,
- Homelessness,
- Poverty,
- Unemployment,
- Hospital admissions,
- Participants' subjective well-being and quality of life,
- Social isolation ,
- Feelings of loneliness

This list is not exhaustive, and thus the review defines personal and social problems very broadly in order to include all relevant studies.

How the intervention might work

Theoretically, group-based interventions for adults suffering from mental illness aimed at supporting social reintegration may be understood through a *recovery* lens. The concept of recovery in mental health can be traced to the early 1980s, when personal accounts of individuals living with mental illness were published, describing their ability to live and cope with their mental illness (Gibson et al., 2011). As described by Anthony (1993), recovery is:

"a deeply personal, unique process of changing one's attitudes, values, feelings, goals, skills, and/or roles. It is a way of living a satisfying, hopeful, and contributing life, even with limitations caused by the illness. Recovery involves the development of new meaning and purpose in one's life as one grows beyond the catastrophic effects of mental illness"

(Anthony, 1993 cited in Gibson et al in p. 248)

Recovery can also be described as a process in which the individual may or may not experience a reduction in symptoms but in which the ability to cope with symptoms is improved enabling the individual to participate in social or occupational activities and to lead a meaningful life despite the mental illness. Thus, interventions, included in the present review have a broader aim than to simply reduce the symptoms of mental illness. In essence, the aims are to help participants form new relationships, develop coping and social skills enabling the participants to subsequently participate in more social and occupational contexts and to increase their general well-being and quality of life. Theoretically, group-based interventions may also be seen thorugh a *social identity* lens in which becoming members of a group may affect the social identity of marginalized individuals positively. According to Tarrant, Hagger & Farrow (2011) health-promoting behaviors are affected by social identity through the individual's adoption of norms of the group, and this may be seen as one of the central mechanisms of change in group based interventions (Tarrant et al., 2011) .

Advantages of group-based interventions: Focus on interpersonal and (social) support factors

Adults suffering from mental illness and indicators of social marginalization constitute a highly diverse population with a multitude of challenges in terms of both mental and physical health. It is beyond the scope of the present review to present the specific risk and protective factors associated with each diagnosis, but what many of the diagnoses and conditions have in common is that interpersonal functioning and support constitute major predictive factors when studying relapse prevention and recurrence of symptoms following treatment (Brown & Moran, 1994; Hammen, 1991; Keitner & Miller, 1990a). In addition, interpersonal and support factors are also one of the few changeable predictors in the course of illness (Keitner et al., 1992). This has high relevance for this review since, compared with individual therapy, the interpersonal and social support factor is an inherent part of group-based interventions (Ford et al., 2009; Keitner & Miller, 1990b; McDermut et al., 2001). Thus, group interventions may address important factors in long-term outcome of treatment of mental illness in ways that individual treatments may not e.g. individual's feelings of loneliness and social isolation (Ruesh et al., 2015). Thus, it can be suggested that group-based interventions may add benefits to individual interventions, as the context of group processes are proposed to encourage social functioning and provide buffering effects of social support. Furthermore, previous studies suggest that when compared to individual interventions for psychiatric patients with bipolar disorder group-based interventions may offer advantages in terms of self-confidence, behaviour and social functioning but not on symptom reduction (Castle et al., 2007).

Furthermore, a study carried out by Colom & Vieta (2004) indicate that group-based interventions offer advantages beyond the supportive effects of being placed in a group. Colom & Vieta (2004) compared a 21-session group based psycho-education intervention based at a hospital incorporating a number of key approaches of other interventions, including stress management techniques, problem-solving, establishment of routines and strategies for managing warning signs with a befriending group (to control for the supportive effect of the group itself). The intervention group experienced a significant reduction in the number of participants who relapsed and number of recurrences per person. The number and length of hospitalizations were also lower for those in the intervention group (Colom & Vieta, 2004).

Deteriorating effects of (group-)based interventions

The potential adverse effects of group psychotherapy or group interventions more broadly have not been the subject to the same scientific scrutiny as individual therapy (Roback, 2000). However, the research into adverse outcomes and or *deterioration effects* in individual psychotherapy are well-established and documented in several trials and systematic reviews. While we have argued that group and individual therapy are different types of treatment, they also share common characteristics. This makes the well-established knowledge about the pitfalls of individual-based therapy interesting from a group intervention perspective.

Based on Strupp, Hadley & Gomes-Schwartz (1977), the negative outcomes of individual psychotherapy that may occur during the course of treatment or following the end of treatment may include:

1. Exacerbation of presenting symptoms, e.g., generalization of symptoms
2. Misuse/abuse of therapy, e.g., patient substituting intellectualized insights for other obsessional thoughts
3. Undertaking unrealistic goals or tasks, e.g., pursuing goals that one is ill equipped to achieve in an attempt to please the therapist
4. Loss of trust in therapy or the therapist, e.g., patient's disillusionment prevents him or her from seeking out necessary therapy in the future
5. Appearance of new symptoms (suicide would be an extreme example)

(Strupp et al., 1977)

Regarding this last point, it should be noted, that it is often very difficult to determine if these negative outcomes were therapy-induced or merely occurred at the time when the patient was receiving an ineffective treatment (Roback, 2000). In explaining these negative outcomes in individual psychotherapies, a number of studies document associations between characteristics of both therapist and patients and negative outcomes (eg. some therapists appear be unsuitable or ineffective for patients with certain characteristics such as specific diagnoses, personality traits or underlying undiagnosed conditions). These effects are likely to be similar for group interventions (eg. some patients and therapists are likely to be unfit for certain therapies when delivered in a group format). However, group interventions may also fail patients for reasons associated with the group. According to Roback (2000):

" A group is often more than the sum of its parts. At times, however, it may be less than the sum of its parts. Ideally, therapeutic groups develop a work culture under the skillful direction of a leader knowledgeable not only in the areas of psychopathology and psychodiagnostics, but also in group dynamics and interpersonal communication. That is, characteristics of the group itself become critical in treatment outcomes. Dynamic properties of therapeutic groups include factors such as intragroup cohesion, group norms, group roles, group pressure, conformity, communication structure, social comparison, and self-disclosure."

(Roback 2000, p. 117)

Theoretically, it is thus possible, that for some marginalized adults suffering from mental illness, group interventions may not bring about the expected positive change or they may even have negative effects. These potential negative effects may happen if the group lacks cohesion, if confidentiality is breached by participants in the group, or if participants feel rejected or invalidated by other participants during the intervention (Fehr, 2019). These negative characteristics or intra-group dynamics may increase rather than decrease the participants' feeling of isolation, rejection and sense of self-worth (Fehr, 2019). Thus, it is also possible that group interventions may be less effective than individual treatment for some.

In summary, group-based interventions aimed at recovery and social reintegration of participants are proposed to offer advantages to participants when compared with both no treatment and with individual interventions in terms of psychosocial support, which is then proposed to lead to increased social and interpersonal functioning. The experience of social support and increased social and interpersonal functioning may subsequently constitute a prospective protective factor, and thus it is proposed that group-based treatment may lead to more sustainable treatment results. However, previous research also points to the potential negative effects of group therapeutic interventions. Theoretically, it is possible that participants with certain characteristics (such as specific diagnoses, co-morbidities or personality traits) will experience negative effects of group interventions and that for some participants individual interventions may be more effective.

Why it is important to do this review

A large body of reviews explore the efficacy of psychiatric group interventions targeting specific mental health disorders such as group psychotherapy for anxiety or personality disorders(Barkowski et al., 2020; McLaughlin et al., 2019). However, most reviews focus on symptom reduction as the only outcome, and are thus not relevant to the present review, in which we aim to explore the efficacy on a more broad range of outcomes associated with social reintegration and not just symptom reduction e.g. experience of a meaningful and social life *despite* the mental illness.

For the purpose of this review, we have identified six existing reviews, which include outcomes other than symptom reduction. The first two reviews that we present focus on the effects of outpatient psychiatric group interventions for a specific mental health diagnosis (psychosis and post-traumatic stress disorder). In contrast, the remaining four reviews focuses on treatment for respectively illicit drug dependence, homelessness, substance abuse disorder and alcohol use disorder, which are examples of central comorbidities, which are often experienced by adults suffering from mental illness.

In a review on the effects of group programs for recovery from psychosis, Segredou, Livaditis, Liolios, & Skartsila (2008) identified 20 studies, and concluded that findings suggest positive effects on participants' social and vocational functioning in addition to symptom reduction. However, they also conclude, that findings are uncertain, as many studies lack appropriate control groups, follow-up and standardised measures of symptoms and diagnosis (Segredou et al., 2008). The review which was presented as a conference poster provides a very limited description of the search process, no risk of bias assessment of included studies and they do not conduct a meta-analysis

Bøg, Filges, Brännström, Jørgensen & Fredriksson (2015) conducted a systematic review and meta-analysis on the effectiveness of 12-step interventions for participants with illicit drug dependence based on 10 randomized controlled trials and quasi-experimental studies (N =1071). In addition to the primary outcome of drug use the review included outcomes such as criminal behavior, prostitution, psychiatric symptoms, social functioning, employment status and homelessness. The review concludes that there is no difference in the effectiveness of 12-step interventions compared to alternative psychosocial interventions in reducing drug use during treatment, post treatment, and at 6- and 12-month follow-ups (Bøg et al., 2017). Furthermore, the review found no statistically significant differences between 12-step and another psychosocial interventions post-treatment on measures of psychiatric symptoms, social functioning, and employment.

Munthe-Kaas, Berg & Blaasvær (2018) conducted a systematic review and meta-analysis on the effectiveness of interventions to reduce homelessness based on 43 samples. The review concludes that the included interventions; High intensity case management, Housing First, Critical Time Intervention (CTI), Abstinence-contingent housing, Non-abstinence-contingent housing with high intensity case management, Housing vouchers and Residential treatment perform better than the usual services at reducing homelessness or improving housing stability in all comparisons. Furthermore it was concluded that group living arrangements may be better than individual apartments at reducing homelessness (Munthe-Kaas et al., 2018).

Mahoney, Karatzias, & Hutton (2019) conducted a systematic review and meta-analysis on the effects of group treatments for adults with symptoms associated with complex post-traumatic stress disorder based on 36 randomized controlled trials. Outcomes included four types of symptoms and substance misuse. Medium to large significant effect sizes favouring group-based trauma interventions were found for four of the outcome domains with only substance misuse resulting in a small non-significant effect size (Mahoney et al., 2019).

In a systematic review and meta-analysis on the effectiveness of group treatment for substance use disorder in adults based on 33 randomized clinical trials (N= 3951), Lo Coco et al. (2019) compared group psychotherapy to no treatment control groups, individual psychotherapy, medication, self-help groups, and other active treatments applying no specific psychotherapeutic techniques for patients with substance use disorder. The primary outcome was abstinence, and the secondary outcomes were frequency of substance use and symptoms of substance use disorder, anxiety, depression, general psychopathology, and attrition. Significant small effects of group therapy were found on abstinence compared to no treatment, individual therapy, and other treatments. Effects on substance use frequency and symptoms of substance use disorder were not significant, but significant moderately sized effects emerged for mental state when group therapy was compared to no treatment. There were no differences in abstinence rates between group therapy and control groups (Lo Coco et al., 2019).

Group-based interventions targeting comorbidities relevant for our population of interest have proven to be effective in general populations. A noticeable and recent example is a Cochrane review (Kelly et al. 2020) on the effect of Alcohol Anonymous (AA) and other 12-step programs against alcohol use disorder (AUD). In its original form, AA works through a social fellowship (meetings with peers) and a 12-step program. Hence, AA is considered group intervention/therapy. Kelly et al. (2020) review 27 studies (N= 10 565) and compare AA with motivational enhancement therapy (MET), cognitive behavioral therapy (CBT), variants of 12-step programs and no treatment. Outcomes consists of a range of drinking related outcomes

(abstinence, intensity, consequences and addiction severity) and healthcare cost offsets. Kelly et al. (2020) report evidence that AA results in longer periods of abstinence and AA perform as good as other treatments with respect to intensity, consequences and addiction severity. In addition, Kelly et al. (2020) report that 4 out of 5 studies found cost saving benefits, which in turn probably leads to reduced healthcare costs (Kelly et al., 2020).

Our review adds to the existing body of reviews by exploring the efficacy of group interventions on a more broad range of outcomes, than what is seen in the existing reviews. Secondly, we will review interventions targeting a larger population (e.g. adults suffering from any kind of mental illness) and we will in include both community-based and outpatient psychiatric interventions. Finally, we will provide a thorough risk of bias assessment of the included studies and if possible conduct meta-analyses on outcomes, which are not included in the existing reviews

The number of people with mental illness is growing in the Western world and both direct and indirect costs are expected to rise (Bloom et al., 2011). This growth force policymakers to reconsider how they can meet the increasing demand. Especially local governments, since psychiatric institutional care (hospital beds), is increasingly being replaced by out-patient care ((Wahlbeck et al., 2011).

The effects of psychiatric interventions aimed at reducing symptoms for patients with specific diagnoses have been extensively explored in a large number of reviews and meta-analyses, but only a much smaller number of existing reviews have explored the effects of interventions on a broader range of outcomes. The present review contributes to the knowledge base by including a broader range of outcomes.

As pointed out by McDaid et al (2015) the economic cost of comorbidities have been remarkably neglected by health economists in health in general but also across mental and physical health. The relative increase in costs for comorbid diabetes is for example in the range of 1.8-2.0 for patients diagnosed with schizophrenia or depression. In addition, McDaid and Park (2015) point out that the costs of non-health related comorbid conditions have been even more neglected despite clear evidence of much higher prevalence of non-health related comorbidities among physical and mental health patients. As example, McDaid and Park (2015) points out that patients with major depressive disorder in Australian data have been found to have higher adjusted odds of 4.0 in difficulty of day to day work and higher adjusted odds of 1.7 in number of days unable to work. This underline the importance of considering a broader range of outcomes when assessing costs of mental health disorders (McDaid & Park, 2015). A further underlining of this, is the finding by Stant et al (2007) where group differences in the treatment of schizophrenia only revealed itself when using multiple health outcomes including the preference-based QALY (Quality-Adjusted Life Years). This led the authors to issue a caution when assessing the results of economic studies only using a single and specific outcome(Stant et al., 2007).

The cost of group-based interventions can be less than half the cost of individual therapy (Ruesch et al., 2015). Yet, when policymakers choose group-based community interventions they do so without having a solid knowledge base. Knowledge about the efficacy of group-based community interventions in general, and when compared to individually delivered interventions, is thus crucial for policy makers in charge of deciding which interventions to fund.

Objectives

The main objective is to explore the general efficacy of group-based community interventions aimed at supporting adults with mental illness and related personal and social problems on outcomes such as problem behavior, subjective well-being, homelessness, poverty and employment.

Furthermore, the objective is to explore the potential advantages/disadvantages of using a group-based versus an individual intervention when targeting specific problems or when using specific types of interventions.

Methods

We followed the modernized Campbell's Methodologic Expectations for Campbell Collaboration Intervention Reviews (MECCIR; Aloe et al., 2024) reporting guidelines and conducted our analyses in accordance with our pre-registered protocol (Dalgaard et al., 2022) to the greatest extent possible..

Transparency and openness

To ensure the transparency and openness of the review, we have shared all parts for review, including all risk of bias (RoB) assessments, effect size calculation, data extraction schemes, as well as the final meta-analysis code and data. The RoB assessments, effect size calculations, and meta-analysis code are available at https://osf.io/s2j9a/files/osfstorage, whereas the data extraction schemes and the final meta-analysis data are provided alongside the publication of the review. We have followed the FAIR data sharing principle (Findable, Accessible, Interoperable, and Reusable; Logan et al., 2021; Wilkinson et al., 2016) to maximize the use of the open-sourced data.

Criteria for considering studies for this review

Types of studies

The review includes randomized controlled trials. In order to summarize what is known about the possible causal effects of group-based community interventions, we also included study designs that used a well-defined control group and aimed to control for important confounding factors. To be included, non-randomized studies, where participants are assigned to conditions outside the researcher's control, had to demonstrate pre-treatment group equivalence, for example, via matching, statistical controls, or evidence of equivalence on key risk variables and participant characteristics. These factors are outlined later on in this section. The final assessment of the studies methodological appropriateness was assessed via Cochrane's RoB2 and ROBINS-I risk of bias tools. See the 'Assessment of risk of bias in included studies ' section.

Our study design definitions are heavily inspired by Shadish et al. (2002). The specific study designs included in the review were:

1.) Randomized controlled trials (RCTs) and clustered versions thereof (CRCTs).

2) Quasi-randomized controlled trial designs (QRCTs), where participants are allocated by means that are not expected to influence outcomes, for example, alternate allocation, participants' birth data, case number, or alphabetic name order.

3) Quasi-experimental studies (QES)/non-randomized studies (NRS). This category refers to both studies, where participants are allocated by other actions controlled by the researcher, or where allocation to the intervention and control groups is not controlled by the researcher (for example, allocation according to time differences or policy rules). The latter amounts to what Shadish et al. (2002, p. 12) define as a natural experiment.

Studies using single-group pre-post comparisons were excluded. Furthermore, in accordance with the aims of the present review, studies exploring two different group-based interventions without a control group were excluded. A list including these studies can be found in the supplementary material accompanying this present review.

Types of participants

The population of this review consists of adults in OECD countries with at least one psychiatric diagnosis who are experiencing personal and social problems in addition to their mental health condition. We included participants with any type of psychiatric diagnosis, drawing on both studies in which patients self-reported their diagnosis and studies in which diagnoses were established by a mental health professional. Social and personal problems are defined broadly and may include one or more of the following:

- Alcohol/substance abuse
- Self-harming behaviour
- Criminal behaviour
- Homelessness
- Poverty
- Unemployment
- Hospital admissions
- Participants' subjective well-being and quality of life
- Social isolation
- Feelings of loneliness

We excluded studies of interventions targeting youth under the age of 18. Psychiatric patients, without any co-morbid personal and social problems, who received outpatient treatment for their specific mental disorder with symptom reduction as the primary aim, were also not eligible for this review.

Types of interventions

In this review, we applied a broad definition of group-based interventions. This included all interventions targeting adults who suffer from mental illness and related social and personal problems that received an intervention delivered in a group format. Specifically, this means that more than one participant should have received the intervention simultaneously, in the same setting, and from the same therapist, caseworker, mentor, etc. In addition, interventions were required to take place in a community or outpatient setting, as outlined in the section entitled *'The intervention'*.

Types of outcome measures

In this section, we describe all the included outcome measures of the review. In the protocol, we describe that we do not clearly distinguish between primary and secondary outcomes. However, as all of the outcomes highlighted in the protocol can be said (directly or indirectly) to belong to social reintegration, we consider this to be our primary outcome. Meanwhile, studies that assess reintegrational outcomes commonly include measures of mental health as well. Originally, these types of outcomes were not the main focus of the review. Therefore, we consider mental health outcomes to represent secondary outcomes. Because this distinction was not explicitly defined in the protocol, the analyses of mental health outcomes should be considered as exploratory. That said, throughout the review, we separate analyses between the primary (social reintegrational) and secondary (mental health) outcomes.

Primary outcomes: social reintegration

The relevant outcomes for this review broadly concern problem behaviors and social difficulties associated with social marginalization. Broadly, we characterized these types of outcomes as reintegrational outcomes where we interpret a decrease on the given scale to be an indirect indicator of increased reintegration into society. The specific outcomes include:

- Alcohol/substance abuse
- Self-harming behaviour
- Criminal behaviour
- Loneliness
- Self-efficacy
- Homelessness
- Poverty
- Unemployment
- Hospital admissions
- Participants' subjective well-being and quality of life

In order for a study to be included, the study needed to include at least one measure of social reintegration. Below, we list the exact outcomes that were included in the review:

Social functioning (impairment)

- Areas of Change Index (ACI)
- Clinician-rated Global Assessment of Functioning Scale (GAF)
- Global Assessment Scale (GAS)

- Global Assessment of Functioning scale (GAF)
- Inventory of Interpersonal Problems 64
- Life Skills Profile 16 (LSP-16)
- Macay et al. (2007) Global Assessment of Functioning Scale (GAF)
- Personal and Social Performance Scale (PSP)
- Social Adaptation Self-Evaluation Scale (SASS)
- Social Function Questionnaire
- Sheehan Disability Scale (Functional status)
- The Disability Assessment Schedule (DAS-II)
- The Social Skills Performance Assessment (SSPA)
- Satisfaction with changes in overall daily functioning (single-item self-report: SCI)

Loneliness

- The Loneliness Scale
- 11-item De Jong Gierveld scale,
- The UCLA Loneliness Scale

Hope, Empowerment & Self-efficacy

- 13-item General Help-Seeking Questionnaire
- 6-item Job Search Self-Efficacy Scale
- Beck's Hopelessness Scale
- The Core Components of Treatment Scale
- Coping with Stress Self-efficacy (CSSE)
- The Dutch Empowerment Scale
- The Empowerment Scale
- The Herth Hope Index
- The Integrative Hope Scale (IHS)
- The Recovery Assessment Scale
- The Self-Efficacy Scale (SES)
- The Self-Identified Stage of Recovery Scale
- The Work Hope Scale, The Work Motivation Scale
- Miller Hope Scale (MHS)
- Morton et al. (2012) The Beck Hopelessness Scale
- Netherlands Empowerment List
- Rogers Empowerment Scale (RES)
- Rosenberg Self-Esteem Scale (RSES)

Subjective Wellbeing and Quality of Life

- General Wellbeing Scale Five-item EQ-5D

- Physical health related quality of life; Mental health related quality of life
- PSYCHOLOPS (consists of four questions, three domains: problems, functions, and well-being)
- Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q)
- Quality of Life: Short Form Health Survey (SF-36)
- Retrospective quality of life scale (RQOL)
- Short-Form Health Survey (SF-36)
- The Manchester Short Assessment of Quality of Life
- The Perceived Stress Scale (PSS)
- The Quality of Life Scale
- The Quality of Life index
- The Warwick Edinburgh Mental Wellbeing Scale (WEMWBS)
- WHO Quality of Life-BREF

Self-esteem (positive views of the self, decreased self-stigma)

- 5-item self-concurrence subscale of the Self-Stigma of Mental Illness Scale-Short Form
- Internalized Stigma of Mental Illness scale
- Link Perceived Stigma Questionnaire (LPSQ)
- Modified Engulfment Scale (MES)
- Tennessee Self-Concept Scale (TSCS)
- The Rosenberg Self-Esteem Scale (RSES)
- The State Self-Esteem Scale (SSES)

Homelessness

- Housing Status in Past 60 days

Unemployment

- Employment outcome (self-reported)
- The Employment and Vocational Activities Checklist

Alcohol and Substance Abuse

- Adaptation of the Addiction Severity Index plus Saliva testing
- Alcohol Use Disorder Identification Test
- Drug Abuse Screening Test
- Opiate Treatment Index
- Substance Use in Past 30 days
- The Addiction Severity Index
- The Addiction Severity Index-lite (ASI-lite)
- The Alcohol Use Disorders Identification test (AUDIT)
- The Severity of Dependence Scale

Secondary outcomes: mental health

When a study reported at least one social reintegrational outcome, we also extracted psychiatric symptoms/mental health measures as secondary outcomes if these were reported. Specifically, we extracted the following mental health outcome:

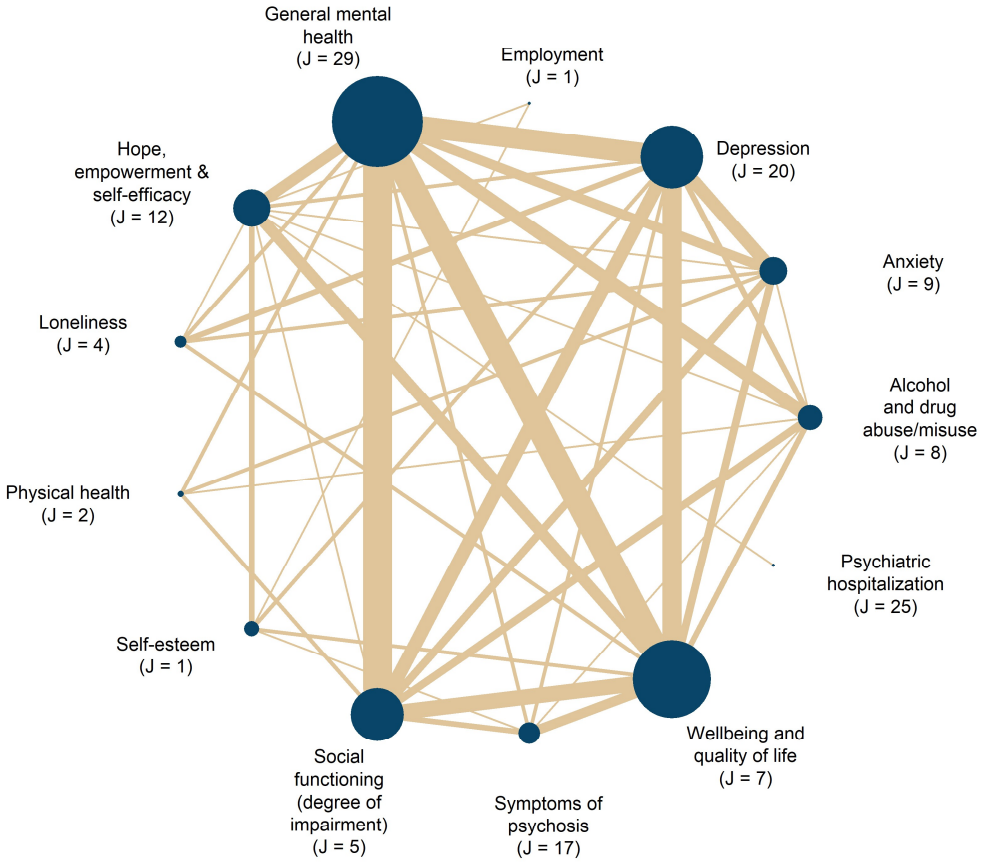
General mental health, Anxiety, Depression, and Symptoms of Psychosis.

- Beck's Anxiety Inventory (BAI)
- Beck's Depression Inventory (BDI)/Beck Depression Inventory II (BDI-II)
- Brief Assessment of Cognition in Schizophrenia
- Brief Psychiatric Rating Scale (BPRS)
- Brief Symptom Inventory (BSI)
- Calgary Depression Scale for Schizophrenia (CDSS)
- Clinical Global Impression Scale for Bipolar Disorder
- Clinical Global Impression Scale (CGI)
- Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM)
- Current Mental Health Symptoms
- Depression, Anxiety, and Stress Scale-21 (DASS 21)
- The Depression Anxiety Stress Scale (DASS)
- The Difficulties in Emotion Regulation Scale (DERS)
- The Expanded Brief Psychiatric Rating Scale
- The Hamilton Depression Rating Scale (HAM-D), BDI
- The Hopkins Symptoms Checklist (HSCL-25)
- The Liebowitz Social Anxiety Scale
- The Mental Health Confidence Scale
- The Mini Social Phobia Inventory
- The Modified Scale for the Assessment of Negative Symptoms
- The Overall Anxiety Severity and Impairment Test (OASIS)
- The Patient Health Questionnaire (PHQ-9)
- Perceived Stress Scale (PSS)
- The Positive and Negative Syndrome Scale (PANSS)
- The Recovery Assessment Scale
- The 15-item German version of the Center for Epidemiologic Studies-Depression Scale (CES-D)
- The Scale for the Assessment of Negative Symptoms (SANS)
- The Scale for the Assessment of Positive Symptoms (SAPS)
- The Short Borderline Symptom List (BSL-23)
- The Symptom Checklist 90-Revised (SCL-90-R)
- The Toronto Alexithymia Scale – 20 (TAS-2)
- The PTSD Symptom Scale Interview (PSS-I)
- The Borderline Evaluation of Severity over Time (BEST)
- The Young Mania Rating Scale (YMRS)
- Self-reported recovery

- The Brief Fear of Negative Evaluation Scale
- Short version of the Social Phobia Inventory (mini-SPIN)
- The Positive and Negative Syndrome Scale
- The 53-item Brief Symptom Inventory
- The Anxiety Disorder (GAD-7) measure
- Personal Health Questionnaire Depression Scale (PHQ-9)

To recap, Figure 1 shows the interconnections between all reported outcomes.

FIGURE 1 Network structure of contrasts between primary and secondary outcome constructs



Duration of follow-up

We included any given time point for the measurement of treatment effects. As per protocol, we characterized follow-up measures as follows:

- Effects measured 0-1 year after the end of the intervention were defined as posttest effects
- Effects measured 1-2 years after the end of the intervention were defined as medium-term follow-up effects
- Effects measured more than 2 years after the end of the intervention were defined as long-term follow-up effects

That said, we only detected posttest effects in the included literature. Therefore, all the analyses of this review concern the posttest effects of group-based interventions only.

Types of settings

In order to be eligible for the present review, interventions had to be based in a community or outpatient setting and must be aimed at supporting the social reintegration of participants.

We excluded interventions taking place in hospital settings where patients are receiving around-the-clock care. However, if patients are admitted to in-hospital treatment and subsequently receive out-patient group-based services or interventions in a psychiatric or hospital setting, the study was also included in the review.

Search methods for identification of studies

Relevant studies were identified through electronic searches of bibliographic databases, governmental and grey literature repositories, hand searched in specific targeted journals, attempts to contact experts, and Internet search engines. The electronic database searches were completed in September 2022 and other resources were completed in July 2024. We searched to identify both published and unpublished literature. The searches were international in scope. The searches were limited to publications published from 2000 onward to maximise contemporary relevance of the review. Reference lists of included studies used in the meta-analysis were also searched.

We implemented a wide range of search methods and strategies to maximise coverage of relevant references, while simultaneously attempting to reduce different types of bias related to publication and dissemination systems.

Subject terms in the facets were selected according to the thesaurus or index of each database. Keywords were supplied if the search technique provided additional results. Use of truncation and wildcards were used to address English spelling variants.

The different strategies and methods are presented below and detailed documentation of the searches is available in Supporting Information.

Electronic searches

The following databases were searched electronically:

MEDLINE (OVID) 1966 - 2022

EMBASE (OVID) 1974 - 2022

APA PsycINFO (EBSCO) 1800 - 2022

CINAHL (EBSCO) 1981 - 2022

Sociological Abstracts (ProQuest) 1952 - 2022

Social Services Abstracts (ProQuest) 1979 - 2022

SocINDEX (EBSCO) 1908 - 2022

Academic Search Premier (EBSCO) 1975 - 2022

International Bibliography of the Social Sciences (IBSS) (ProQuest) 1951 - 2022

Science Citation Index (Web of Science Core Collection) 1900 - 2022

Social Sciences Citation Index (Web of Science Core Collection) 1990 - 2022

Cochrane Central Register of Controlled Trials (CENTRAL) (1996) – 2022

Searching other resources

In addition til electronic databases, we also searched the following venues:

- Google Scholar—https://scholar.google.com
- Google —https://www.google.com/
- Searches in Google and Google scholar
- Social Science Research Network
  — https://papers.ssrn.com/sol3/DisplayAbstractSearch.cfm
- CORE – https://core.ac.uk Internationale repositorier
- Danish National Research Database—http://www.forskningsdatabasen. dk/en
- NORA—Norwegian Open Research Archives—http://nora. openaccess.no/

- Cristin - Current Research Information SysTem In Norway - https://wo.cristin.no/as/WebObjects/cristin.woa/wa/fres?la=no
- SwePub—Academic publications at Swedish universities—http:// swepub.kb.se/
- DIVA - https://www.diva-portal.org/smash/search.jsf?dswid=69

Searches for working papers and conference proceedings in English

SHS Web of Conferences (www.shs-conferences.org) Open Access proceedings in Humanities and Social Sciences

The Social Care Institute for Excellence (SCIE) : www.scie.org.uk/publications/index.asp

Searches for Government Documents

NICE National Institute for Health and Care Excellence www.nice.org.uk

Searches for Dissertations

EBSCO Open Dissertations (https://biblioboard.com/opendissertations/)

Open Access Theses and Dissertations (oatd.org)

Hand Searches

Hand searches were carried out in selected journals:

- BMC Public health
- BMC psychiatry
- Journal of Psychosocial Rehabilitation and Mental Health
- Psychiatric Quarterly
- Community Mental Health Journal
- Disability and rehabilitation
- International journal of mental health systems
- Sociology of health and illness

Searches included all issues published in 2022, 2023 and until may 2024

Citation-tracking

We also checked the references for all identified existing systematic reviews and meta-analyses and of all included primary studies.

Contacting experts in the field

We did not contact international experts, as we did not identify anyone with a specific area of expertise central to the present review.

Language restrictions

We reviewed studies published in English, Danish, Swedish, and Norwegian.

Data collection and analysis

All data extraction schemes are either enclosed with this publication or can be found at https://osf.io/s2j9a/files/osfstorage. We used EPPI-reviewer, Excel, and R to extract data.

Selection of studies

Under the supervision of review authors, two review team assistants first independently screened titles and abstracts to exclude studies that were clearly irrelevant. Studies considered eligible by at least one assistant or studies where there was insufficient information in the title and abstract to judge eligibility, were retrieved in full text. Two review team assistants under the supervision of the review authors subsequently screened the full texts independently. The review authors resolved any disagreement about eligibility.

Data extraction and management

Data extraction was done in collaboration between NTD, JSA, JKJ, and MHV, with minor support from two research assistants. All coding and data extraction were done independently by at least two reviewers. Before initiating the final extraction, our extraction scheme was piloted to ensure a standardized use thereof. Throughout the entire process, any extraction disagreements were resolved by NTD and/or MHV. To make the extraction as theoretically relevant as possible, we aligned the data extraction with the factors we described in the protocol as factors potentially explaining differences in effect sizes. Among other things, this included extracted data on the characteristics of the participants in the sample, characteristics of the type of intervention and control groups, preregistration, research design, sample size, type of outcomes, and results.

Effect sizes were primarily calculated by JKJ and MHV, and quality checked by JSA in accordance with the prescripted procedures for ensuring reproducible research in statistics developed by Hofner et al. (2016). All effect size issues were resolved by MHV.

We extracted all covariates and background information in MS Excel, whereas all effect size calculations were done in R. These two different datasets were then combined using a row ID (the variables are termed vary_id in the covariate data and varifier in the effect size data), ensuring that the covariates were correctly combined with the adjacent effect sizes. All coding schemes and effect size computations can be found at https://osf.io/s2j9a/files/osfstorage.

Assessment of risk of bias in included studies

As we extracted result data (i.e., effect size estimates) from various research designs and studies with varying quality, we conducted comprehensive risk of bias (RoB) assessments. We did this to: 1) prevent the inclusion of flawed results in our meta-analyses, 2) investigate how different

levels of risk of bias influence our final meta-analytical results, and 3) provide an overview of the general quality of the existing literature. All RoB assessments were conducted independently by at least two review authors, and all disagreement was resolved by either NTD or MHV. For all studies, we assessed the risk of bias individually for each calculated effect size estimate, meaning that studies contributing with multiple effect sizes (e.g., due to reported results across multiple eligible outcomes) underwent multiple and potentially different RoB assessments. Yet, we never experienced that RoB assessment varied across within-study effect sizes. All RoB assessments can be at https://osf.io/s2j9a/files/osfstorage in the 'Risk of bias assessments' folder. In this folder, one can find an Excel file for each study containing the specific domain assessments.

As prescribed in our protocol, we used the Cochrane's revised RoB 2 (Risk of Bias) tools (Eldridge et al., 2021; Sterne et al., 2019) for assessing the risk of bias in randomized and cluster-randomized trial studies, respectively. For non-randomized studies, such as quasi- and natural experiments as well as observational studies, we assessed the risk of bias using the 2016 version of the ROBINS-I tool (Risk Of Bias In Non-randomized Studies - of Interventions; Sterne, Hernán, et al., 2016). Since the RoB 2 tools and ROBINS-I tool used different rating schemes, we will present the assessments separately for each tool set. We used the ggplot2 R package (Wickham, 2016) to visualize our risk of bias results.

Assessing randomized studies (#)

To be specific, we assessed individually randomized controlled trial (RCT) studies using the five main domains from the RoB 2 tool which aim to cover the most common factors usually biasing trial results. The five domains covered in our assessment were

1. bias arising from the randomization process;
2. bias due to deviations from intended interventions (separate signaling questions for the effect of assignment and adhering to intervention);
3. bias due to missing outcome data;
4. bias in measurement of the outcome;
5. bias in selection of the reported result.

Assessing cluster-randomized studies

For cluster-randomized trials, an additional domain is included (i.e., 1b Bias arising from identification or recruitment of individual participants within clusters). In the cluster randomized (CRCT) template (Eldridge et al., 2021), however, only the risk of bias due to deviation from the intended intervention (effect of assignment to intervention; intention to treat ITT) is present and the signaling question concerning the appropriateness of the analysis used to estimate the effect is missing. Therefore, for cluster-randomized trials we only used the signaling questions concerning the bias arising from identification or recruitment of individual participants within clusters from the template for cluster-randomized parallel-group trials; otherwise, we used the template and signaling questions for individually randomized parallel-group trials.

Assessing non-randomized studies

To assess the risk of bias in non-randomized studies, we used the seven domains below covered by The ROBINS-I tool:

1. bias due to confounding
2. bias in selection of participants
3. bias in classification of interventions
4. bias due to deviations from intended interventions;
5. bias due to missing outcome data;
6. bias in measurement of the outcome;
7. bias in selection of the reported result.

As there is no universally correct way to construct counterfactuals for non-randomized designs, we looked for evidence that identification was achieved and that the authors of the primary studies convincingly justified their choice of method by discussing the assumption(s) leading to identification (the assumption(s) that make it possible to identify the counterfactual). Preferably the authors should make an effort to justify their choice of method and convince the reader that the only difference between a treated individual and a non-treated individual is the treatment.

Assessing important pre-specified confounding factors

An important part of the risk of bias assessment of non-randomized studies is the consideration of how the studies deal with confounding factors. Systematic baseline differences between groups can compromise comparability between groups. Baseline differences can be observable (e.g., age and gender) and unobservable (to the researcher; e.g. motivation and 'ability'). There is no single non-randomized study design that always solves the selection problem. Different designs represent different approaches to dealing with selection problems under different assumptions and consequently require different types of data. There can be particularly great variations in how different designs deal with selection on unobservables. The "adequate" method depends on the model generating participation, i.e., assumptions about the nature of the process by which participants are selected into a programme.

We identified the following observable confounding factors to be most relevant: *age*, *gender,* and *risk indicators* as described in the 'Type of participants*'* section. The prevalence of different types of behavioural and psychological problems, coping skills, cognitive and emotional abilities vary throughout human development through puberty and into adulthood, and therefore we consider age to be a potential confounding factor.

Furthermore, there are substantial gender differences in behavior problems, coping and risk of different types of adverse outcomes which is why we also include gender as a potential confounding factor (Card et al., 2008; Cook et al., 1992; Hampel & Petermann, 2005).

Pre-treatment group equivalence on mental illness such as primary diagnosis and comorbid conditions/problems such as alcohol/substance use, homelessness, poverty, etc. are indisputable important confounders as the magnitude and severity of pre-existing conditions and problems within the target population is very likely to be associated with treatment effects (Compton Iii et al., 2003) . Therefore, the accuracy of the estimated effects of group-based interventions will likely depend crucially on how well these factors are controlled for.

In each study, we assessed whether these factors had been considered, and in addition, we looked for other factors likely to be a source of confounding within the individual included studies. If studies did not ensure baseline equivalence among intervention groups, they had to provide pretest or baseline measures or covariate-controlled results from which we can calculate pretest-/baseline-adjusted effect sizes, otherwise, non-equivalent group-designed studies were excluded due to a critical risk of confounding.

Assessing effects of primary interest and important co-interventions

We are mainly interested in the effect of starting and adhering to the intended intervention, i.e. the treatment on the treated (TOT) effect. The risk of bias assessments were therefore carried out in relation to this specific effect. Important co-interventions may include psychopharmacological treatment or other active treatments such as individual psychotherapy, mentoring, or counseling.

General decision rules across all risk of bias tools

To align the RoB assessment across the three used tools, we required (as for the studies assessed with the RoB 2 tools) that non-randomized studies either provide the raw data or a pre-registered protocol to be considered to have a low risk of bias in selective reporting. Furthermore, to align the RoB assessment to the standards of psychological and social science research, we did not consider questions about blinding and double-blinding to have any consequential impact on the overall RoB assessment. Across all assessment tools, we also applied the decision rule that whenever a study received four non-low judgments, we moved up the overall RoB judgment to the next level. For example, if a study received four moderate judgments, we considered the overall risk of bias to be serious alternatively of high concern (see McCay et al. 2006 & Smith et al. 2021). Whenever all effect size estimates within a non-randomized study were judged as 'Critical' in one domain, we stopped the RoB assessment. Consequently, studies that received a critical RoB assessment were thus excluded from our meta-analyses, as prescribed by the tool guidance (Sterne, Higgins, et al., 2016, p. 17).

Measures of treatment effect

Measuring the treatment effect of group-based interventions involves comparing the intervention to an eligible control condition. To measure a potential treatment effect, we used standardized mean differences (SMD) and odds ratios (OR) as our effect size metrics. Since ORs were calculated for only one study (i.e., Bond et al., 2015) and were not used in further analyses, we focus primarily on presenting SMDs in this section.

All effect size calculations were conducted using R 4.5.1 (R Core Team, 2022) and RStudio (RStudio Team, 2015). Specifically, we drew substantially on the `tidyverse` packages for data manipulations, visualization, and summary statistics of various kinds (Wickham et al., 2019). We calculated effect sizes from medians and quantiles by using the `estmeansd` (version 1.0.1; McGrath et al., 2019), and we used the `metafor` package (version 4.0-8; Viechtbauer, 2010) to aggregate within-study effect sizes not relevant to our moderator analyses.

Moreover, all effect size calculations can be found in the HTML document entitled 'Effect size calculation for Group-based community interventions study' in the 'Effect size calculation' folder at https://osf.io/s2j9a/files/osfstorage. In this scheme, we have calculated effect sizes

individually for each study. This means that one can press on each study to see the exact methods used to calculate effect sizes for the particular study. We developed this scheme in response to widespread critiques indicating that effect sizes reported in (psychological) meta-analyses are generally difficult to reproduce (Maassen et al., 2020).

Effect size calculation

We calculated SMDs using the Hedges' *g* estimator, which corrects for the small-sample bias inherent in Cohen's *d* (Hedges, 1981). For sensitivity analyses, we also calculated Cohen's *d*. All SMDs were coded so that a positive value indicated a beneficial effect of the group-based treatment. Accordingly, effect sizes were reversed for test scales where lower scores indicated better outcomes (e.g., improved condition). For an illustration of this procedure, see Lim et al. (2020). Across all interventions and outcomes, we calculated 349 SMD estimates clustered within 49 studies. As can be seen from Table 1, 205 effect sizes from 46 studies captured social reintegration outcomes, while 144 effect sizes from 42 studies captured mental health outcomes. All but two effect sizes represent pretest-adjusted effect sizes. For all effect sizes, we used the PRIMED workflow to test for computational errors.

TABLE 1 Descriptive statistics for effect size estimates

Outcome	J	Multi-treatment studies	Multi-control studies	K	n_j	Min	Median	Max	Participants
Reintegration	45	3	0	202	4.5	1	3	28	5390
Mental health	41	3	0	141	3.4	1	2	18	4663

Note: J = Number of studies; K = Number of effect sizes; n_j = average number of effect sizes per study; min = minimum number of effect sizes per study; median number of effect sizes per study; Max = maximum number of effect sizes per study.

As we computed effect sizes from the result data deduced from various research designs, estimation techniques, and reporting standards, we applied a wide range of different methods to obtain the relevant statistics for effect size calculation (Borenstein & Hedges, 2019; Fitzgerald & Tipton, 2024; Hedges et al., 2023; Higgins et al., 2019; Pustejovsky, 2016; Wilson, 2016; WWC, 2021). Specifically, to increase the internal validity and precision of Hedges' *g*, we prioritized calculating pretest-/baseline- and/or covariate-adjusted versions of the *g* metric (Hedges et al., 2023; Morris, 2008; Pustejovsky, 2016; WWC, 2021). As a further attempt to increase the statistical power of our analyses, we reduced (artificial[1]) within-study variability by aggregating all study results that were reported across subgroups that were not pre-specified in our protocol, as recommended by Vembye, Pustejovsky et al. (2025). See Sacks et al. (2011) for an example of this procedure.

[1] Within-study variability can be artifically inflated when small sample studies report large amount of effect sizes, as sample error in this case can make the variabilty look more extreme as it truly is (see Vembye, Pustejovsky et al., 2024)

To ensure comparability between effect sizes (Taylor et al., 2021), all effect sizes and the corresponding variance estimates were standardized by the *total variance*. That is, we computed effect sizes and variance estimates that incorporate both the variation arising from the participant/individual level as well as the cluster level, which means the group-based treatment level. To do so, we cluster-bias adjusted all effect sizes from studies, ignoring the nesting of participants in the given group-based treatment. Only two of the included studies accounted for this issue in the statistics we used from effect size computation (i.e., Haslem et al. 2019; Michalak et al. 2015).[2]

Although participants in the majority of studies had been individually randomized to treatment and control groups, the fact that the intervention was provided in a group format at the same time and space creates dependence among members of the same group, as they share common traits such as receiving treatment from the same therapist/professional, etc. If not accounted for, this yields effect size standard errors that are incorrect. In this regard, we followed the recommendation from the Cochrane Handbook (section 23.1.8; Higgins, Eldridge, et al., 2019) to adjust for clustering arising from this type of clustering caused by group treatment.

In this review, all studies represent so-called *partially clustered* studies, with clustering arising in the treatment group only. Therefore, we used the cluster-bias methods developed by Hedges and Citkowicz (2015) that specifically account for this design issue. Because our priority was the covariate-adjusted effect size, and given that Hedges and Citkowicz primarily developed cluster formulas for posttest-only study designs, we integrated their formulas with those presented by WWC (2021) and Hedges et al. (2023).[3] Vembye (2024) provides an overview of the exact formulas used can be found at https://mikkelvembye.github.io/VIVECampbell/reference/vgt_smd_1armcluster.html.

A common challenge with cluster-bias adjustments is that they are premised upon intraclass correlation (ICC) values, which are rarely reported in practice. Among the included studies, only three studies (Crawford et al., 2012; Haslem et al., 2019; van Gestel-Timmermans et al., 2012) reported ICC values; otherwise, ICC values were imputed, as suggested by Hedges (2007). We imputed ICC values of 0.1 for the main analyses and conducted sensitivity analyses imputing ICC equal to 0.05 and 0.2. For further details, see section 'Unit of analysis issues' below.

The final distributions of effect sizes for reintegrational and mental health outcomes, respectively, in presented in Figures 2 and 3.

FIGURE 2 Empirical distribution of reintegrational (primary outcome) effect size estimates.

[2] van Gestel-Timmermans et al., (2012) and Crawford et al., (2012) accounted for the multi-level structure of their data but we were not capable/unsure on how to calculate effect size from the reported model results.
[3] For this matter, we consulted Larry Hedges who confirmed the appropiateness of this approach.

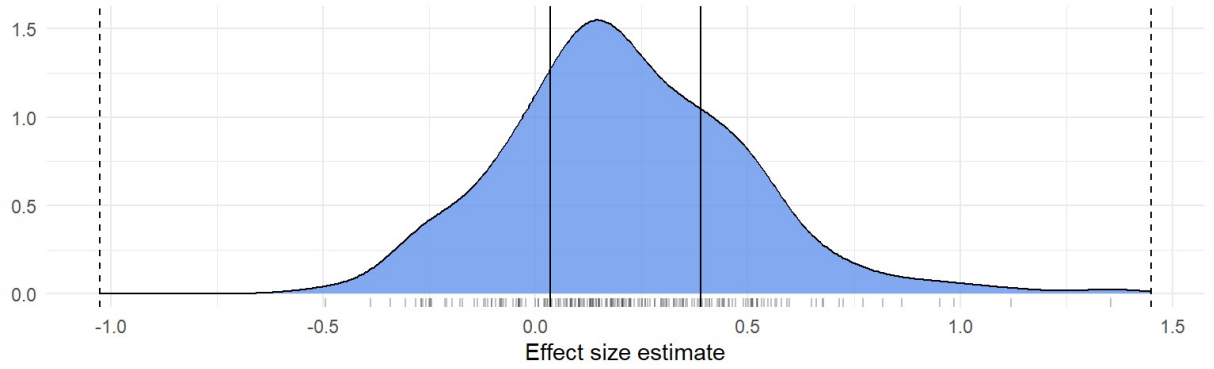
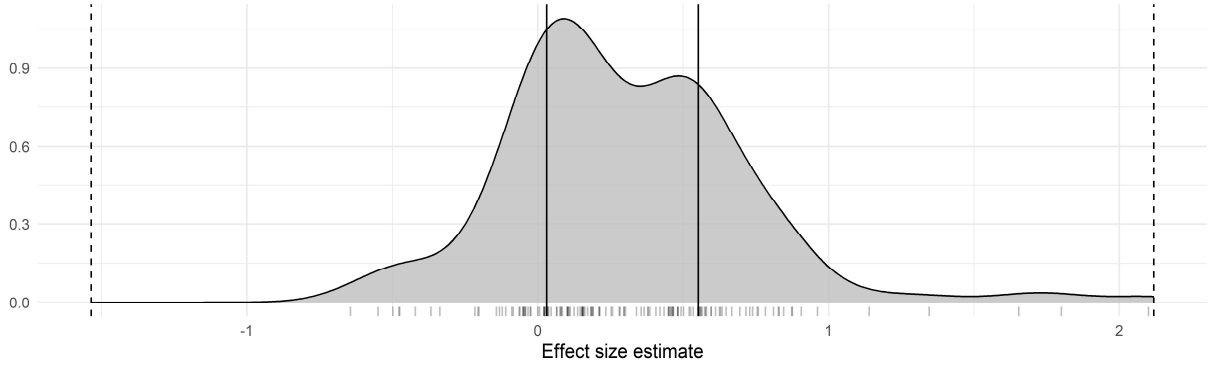


FIGURE 3 Empirical distribution of mental health (secondary outcome) effect size estimates.



For more detailed visualization across more specific outcomes, see the PRIMED Figures 95 and 96.

Technical description of the effect size calculation (#)

To describe the above procedure more formally, the Hedges' g estimator we used can be written as

$$g_t = \omega \times \left(\frac{b}{S}\right) \times \gamma \tag{1}$$

The subscripted t in (1) indicates that the g metric was standardized by the total variance. ω is the small study corrector, $1 - \frac{3}{4*df-1}$, where df is the degrees of freedom, which typically equals N . That is the total sample size of the study. Yet, as all included studies are partially cluster-designed studies, we calculated cluster-adjusted degrees of freedom (df_{cl}) as (c.f. Hedges & Citkowicz, 2015, Equation 7)

$$df_{cl} = \frac{[(N-2)(1-\rho_{ICC}) + (N_T - n)\rho_{ICC}]^2}{(N-2)(1-\rho_{ICC})^2 + (N_T - n)n\rho_{ICC}^2 + 2(N_T - n)(1-\rho_{ICC})\rho_{ICC}} \tag{2}$$

Where N is given as above, ρ_{ICC} is the intraclass correlation, N_T is the sample size of the treatment group, and n is the average cluster size. That is the average sample size of the group-based treatment format. As previously described, we could only obtain empirical values of ρ_{ICC} from three studies. Thus, we most often imputed this value. Whenever possible, we empirically obtained values of n from the included studies. If the average sample size of the group treatment was not reported, we assumed $n = 8$. This value closely resembled the average group size (which was 7.65) found in the studies that empirically reported this value.

Next, b in Equation (1) denotes the mean difference between the group-based treatment group and the individual control group. Most commonly, we computed b as a difference-in-differences estimate, that is, $(M_T^{post} - M_T^{pre}) - (M_C^{post} - M_C^{pre})$, where, M_T^{pre}, M_C^{pre}, M_T^{post}, and M_C^{post} represent the pre- and posttest means for the treatment and control groups, respectively. We define this type of effect size as a difference-in-difference (DID) effect size. For 40 studies, we calculated the nominator of Equation (1) from the raw pre- and posttest
means. For five studies (van Gestel-Timmermans et al., 2012; Gonzalez & Prihoda, 2007; McCay et al., 2007; Smith et al., 2021; Wuthrich & Rapee, 2013), b was obtained either from repeated ANOVAs or estimated marginal means estimates. Moreover, b was obtained as regression estimates from multi-level regression models for two studies (Haslam et al., 2019; Michalak et al., 2015)[4]. One study (Gutman et al., 2019) provided the raw data. From this data, we estimated pretest-adjusted effects using standardized linear regression. Finally, for one study only (Bond et al., 2015), we calculated the posttest-only version of b, that is $b = (M_T^{post} - M_C^{post})$.

S in Equation (1) represents the pooled standard deviation (SD). Sometimes referred to as the pooled within-group SD (WWC, 2022). This was most frequently calculated (i.e., for 318 effect sizes and 43 studies) as $S = \sqrt{\frac{(N_T-1)SD_T^2+(N_C-1)SD_C^2}{N_T+N_C-2}}$, where N_C is the sample size of the control group, whereas SD_T and SD_C represent the SD of the treatment and control group, respectively. Four studies (Haslam et al., 2019; Smith et al., 2021; Wojtalik et al., 2022; Wuthrich & Rapee, 2013) did not report posttest SDs. In these cases, we used the pretest standard deviation instead. For one study (McCay et al., 2007), where we extracted Cohen's d estimates calculated by the authors, it was unclear how S is calculated, but we assumed that a pooled SD was used as prescribed for Cohen's d. From two studies (Gonzalez & Prihoda, 2007; Michalak et al., 2015), we extracted the total SD and not the pooled within-study SD. To align this SD measure, we used Equation [E.14] from the What Works Clearinghouse's standard and procedure handbook (version 5; 2022, p. 168) to convert them to the pooled within-group SD.

For effect sizes based on the GAF (Global Assessment of Functioning), PHQ-9 (Patient Health Questionnaire), and BDI (Beck Depression Inventory) scales, we calculated S as a population-based standard deviation, following the recommendation of Fitzgerald and Tipton (2024). Specifically, S was estimated by pooling all control group standard deviations. We adopted this approach to increase the generalizability/external validity of the results, as the population-based version of S provides an estimate closer to the population we want to generalize to. We applied this method only to these three scales, as each was reported in at least five studies—the minimum

[4] Note, from Michalak et al. (2015), we 4 effect size from the raw means and 10 effect size from multi-level models. The standard deviation used to standardize the mean effect difference also varied within this study.

number required for adequately estimating population-based standard deviations (Fitzgerald & Tipton, 2024).

Finally, γ in Equation (1) is a small number of clusters adjustment factor (WWC, 2021) that is computed as $1 - \frac{(N_C+n-2)\rho_{ICC}}{N-2}$ with all elements given as previously defined.

A general formula expressing the sample variance (V_{gt}) of our Hedges' g estimator can be written as

$$V_{gt} = W \times \xi + P \tag{3}$$

Here W is the scaled/standardized sampling variance of b from Equation (1), i.e., $\left(\frac{se_b}{S}\right)^2$, which expresses the contribution of the variability of b, whereas P "captures the contribution of the variance of [S]—that is, how precisely estimated is the *scale* [*/standard deviation*] of the outcome" (Pustejovsky & Rodgers, 2019, p. 59). ξ either represents the small number of clusters adjustment factor, γ, or the design effect, η, depending on whether a study adequately accounted for clustering or not. η is given by $1 + \left(\frac{nN_C}{N} - 1\right)\rho_{ICC}$.

Again, as we extracted data from various research designs, estimation techniques, and reporting standards, we computed W from Equation (3) in several ways. As we most frequently calculated difference-in-differences (DID) effect sizes, we obtained W as $2(1-r)\left(\frac{1}{N_T} + \frac{1}{N_C}\right)$, where r is the pre-posttest correlation. Although r is usually not reported in primary studies, this can be calculated from commonly reported measures. For nine studies (Acarturk et al., 2022; Bækkelund et al., 2022; Craigie & Nathan, 2009; van Gestel-Timmermans et al., 2012; Gutman et al., 2019; McCay et al., 2006; Popolo et al., 2019; Rabenstein et al., 2016; Somers et al., 2017), we could obtain r directly from the study or calculate r either using the equation from the Cochrane Handbook (section 6.5.2; Higgins, Li, et al., 2019) or Equation 31 from Wilson (2016). As described by Pustejovsky (2016), an equivalent alternative to estimating W for covariate-adjusted effect sizes is to use t of F test values, thus $W = \frac{g_t^2}{F} = \left(\frac{g_t}{t}\right)^2$. We used either t or F values from seven studies (Craigie & Nathan, 2009; Dyck et al., 2000; Gatz et al., 2007; Gordon et al., 2018; Michalak et al., 2015; Rüsch et al., 2019; Schrank et al., 2016) reporting difference-in-difference and one study (Gonzalez & Prihoda, 2007) reporting repeated ANOVA estimates. If either the *t/F* test seemed flawed or yielded a larger variance estimate relative to the DiD variance estimator, we used the DID variance estimator and imputed r if necessary (see Druss et al., 2010; James et al., 2004).

On this matter, Hedges et al. (2023) suggest using test-retest reliability values of a given outcome scale whenever it is impossible to compute r. We imputed test-retest reliability measures as proxies for r in two studies (Cano-Vindel et al., 2021; Morton et al., 2012). For studies where

we were not able to apply any of the above methods, we imputed $r = 0.5$.[5] Thus, W reduces to $\left(\frac{1}{N_T} + \frac{1}{N_C}\right)$, which equals W computed for posttest-only effect sizes. For three studies, we calculated $W = \left(\frac{se_b}{S}\right)^2$, where se_b is the standard error of b, typically representing the covariate-adjusted mean difference between the treatment and control group.

In three studies using ANCOVA-like estimation (i.e., repeated ANOVA and estimated marginal means) methods (McCay et al., 2007; Smith et al., 2021; Wuthrich & Rapee, 2013), we estimated $W = (1 - r^2)\left(\frac{1}{N_T} + \frac{1}{N_C}\right)$.

For all covariate-adjusted effect size estimates, we calculated P in Equation (3) as $\frac{g_t}{2(df-q)}$, g_t and df are defined in Equation (1) and (2), and q is the number of covariates. For posttest-only effect sizes $q = 0$, whereas for pretest/baseline-only adjusted effect sizes $q = 1$. We did not adjust V_{gt} for small sample bias (i.e., multiplying ω^2 with Equation 3), as Hedges et al. (2023, p. 12) recommend not doing so.

Unit of analysis issues

As also mentioned in the previous section, this review focuses on the treatment effect of group-based interventions at the individual level. In other words, the unit of analysis of this review was the effects of group-based intervention on the individual participants. However, the fact that group-based interventions are provided to the participant in a group format at the same time and space likely creates dependence/clustering among/of participants. That is, the treatment effects for participants in the same group are more similar compared to other individuals who are not a part of the particular group. This breaks the classical statistical assumption of independence (Raudenbush & Bryk, 2002). Thus, if this type of clustering is not accounted for, the average individual treatment effects will appear to be more certain than they actually are (i.e., the standard errors will be underestimated). As mentioned in the previous section, we, therefore, cluster-biased corrected all studies to account for this dependence of participants from the same group, which in this case is only an issue in the treatment group.

Although cluster-bias correction of treatments received in groups is recommended by Higgins, Eldridge, et al. (2019), "Weiss et al. (2016) indicate that both adjusting and not adjusting are likely to yield biased standard errors in primary studies (over- and underestimated, respectively)" (Dietrichson et al., 2025). Therefore, to accommodate this issue, we conducted a sensitivity analysis, using a non-cluster-adjusted version of the *g* estimator. That is, we did not add the γ and ξ factors to Equations (1) and (3), respectively.

As the statistical method we used for cluster adjustments (i.e., Hedges & Citkowicz, 2015) has not been implemented in any standard software, we developed our own R package `VIVECampbell`

[5] The raw average pre-posttest correlation, r, for reintegrational and mental health outcomes is .623 and .51, respectively. We calculated these measures from studies that either reported the pre-posttest correlation or where we were able compute this value from the reported statistics.

(Vembye, 2024b), in which we implement cluster-bias adjustments when there is clustering in one treatment group only across various effect size estimation techniques.

Other unit of analysis issues

Our protocol states that we intended to analyze posttest (i.e., effects measured 1 year or less after the end of the intervention) and follow-up (i.e., effects measured more than 1 year after the end of the intervention) effects separately to avoid unit-of-analysis errors (Higgins, Li, et al., 2019). However, we did not detect any follow-up effects. As an exploratory sensitivity analysis, we investigated whether the measurement timing within the first year could explain differences between effect sizes.

Criteria for the determination of independent findings (dependent effect sizes)

As the majority of the included studies contribute multiple effect sizes, the final datasets contain dependencies among the computed effect sizes. The primary dependency structure we detected in our datasets pertained to the *correlated effects dependency structure*, where studies report multiple outcome results from the same or partially the same sample of participants. Most commonly, studies (30 for social reintegration outcomes and 36 for mental health outcomes) reported results for the same participants across different outcomes and/or time points. As noted in Table 1, three studies also created correlated dependencies by comparing multiple treatment groups to a single control group.

We did not have any studies creating the so-called *hierarchical effects dependency structure*, in which the dependency arises from outcomes reported across distinct, non-overlapping samples. Finally, 12 (social reintegration) and 8 (mental health) studies reported a single effect size only.

This type of complex dependency requires advanced meta-analytical techniques to avoid the production of overly optimistic results. To address it, we used the *correlated hierarchical effects* (CHE) model family (Pustejovsky & Tipton, 2025). These models account simultaneously for both hierarchical and correlational dependence in meta-analytic data. To guard against model misspecification, we employed either robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010; Tipton, 2015; Tipton & Pustejovsky, 2015) or cluster bootstrap methods (Joshi et al., 2022; Pustejovsky, 2023; Pustejovsky, Citkowicz, et al., 2025). RVE was used when estimating standard errors for single coefficients in the main analyses, whereas bootstrap methods were applied for Wald test statistics and for publication bias analyses. These applications are described in more detail in later sections.

A challenge when working with dependent effect size data is that the true dependence is largely unknown, requiring the meta-analytist to make arbitrary guesses about the true correlation among the dependent effect size estimates. For the three multi-treatment, however, we were able to asymptotically estimate the correlation among the effect sizes using the `vcalc` function from the `metafor` (Viechtbauer, 2025), which implements formulas from Gleser and Olkin (2009) and Wei and Higgins (2013). For all other cases, we assumed a constant within-study correlation of 0.8, as specified in our protocol.

To capture this combination of empirically estimated and assumed correlations, all fitted models are denoted with the prefix *PE* (Partially Empirical), following Pustejovsky & Tipton (2022). For example, the CHE model is denoted *PECHE*, standing for *Partially Empirically Correlated Hierarchical Effects*.

To provide a more profound understanding of the dependency structure of our data, we applied the Preliminary Data Analysis for meta-analysis of dependent effect sizes (PRIMED) workflow: On the one hand, this helped us to illustrate and visualize both the hierarchical and correlational structure of the dataset, with effect sizes nested within studies and with studies reporting on multiple eligible outcomes. On the other hand, it helped us detect any coding errors that were not caught during our quality checks. All the PRIMED analyses can be found at https://osf.io/s2j9a/files/osfstorage

Dealing with missing data

According to Pigott (2019), missing data in meta-analyses arises for three main reasons: (1) missing studies, meaning that some studies cannot be detected for various reasons; (2) missing effect sizes within a study, for example, because certain outcomes are not reported or because statistical measures needed to calculate effect sizes are unavailable; and (3) missing predictor variables, that is, study, sample, or outcome characteristics that researchers wish to use to predict differences in effect sizes but which are not reported. As the first two reasons primarily reflect publication and reporting biases, we describe how these issues are addressed in the section entitled 'Assessment of reporting biases' below.

For the latter issue concerning missingness on the predictor variables described in our protocol, we used mean imputation to recover missing values. Although this was not prespecified in our protocol and is not considered to be a state-of-the-art management of the missingness, we did so because we experienced having a maximum of one missing study across all moderator variables prescribed in the protocol. Specifically, we could not obtain information about the average number of males in the study sample for Gordon et al. (2018), prompting us to impute means on the 'average percent males in sample' variable related to three reintegrational effect size estimates and one mental health effect size estimate, respectively. Further, we could not back out the number of intervention sessions for Somers et al. (2017), causing us to impute means on the 'total number of sessions' variable related to four reintegrational effect size estimates and one mental health effect size estimate, respectively.

As the number of missing values was so few, we found it unnecessary to use more sophisticated techniques such as multiple imputation. This would moreover unnecessarily complete the reliable computation of the cluster robust standard errors and the embedded degrees of freedom (see Vembye et al., 2024, for a description of this problem).

A final type of missingness pertains to attrition of participants in the study sample. The consequences of this type of issue are assessed in our risk of bias assessment.

Assessment of heterogeneity

We primarily assessed heterogeneity with the measures of between-study (τ) and within-study (ω) SDs, along with the total variation between true effects. That is $\sigma_T = \sqrt{(\tau^2 + \omega^2)}$. We used the restricted maximum likelihood versions of τ and ω. For the overall average effect size estimations, we reported I^2 (Viechtbauer, 2021). However, this should not be interpreted as a key indicator of heterogeneity, as this is just a measure of the ratio between the true variation between effects and the sampling error (Borenstein et al., 2017). For the overall average effect size estimates, we also report Q statistics and 67% prediction intervals. With inspiration from Treves et al. (2025), we chose to use a 67% prediction interval because this interval covers the most typical range of observations that one would expect to see in a new study. In fact, it represents 2/3 of the most likely effect sizes expected to be seen in a new study.

Assessment of reporting biases

We conducted a range of complementary publication bias and/or small study effects tests (henceforth publication bias tests[6]), as no publication bias test clearly outperforms all other methods. Also, this follows the general recommendations for publication bias testing in meta-analysis (c.f. Carter et al., 2019; Chen & Pustejovsky, 2025; Hedges & Vevea, 2005; McShane et al., 2016). Specifically, we conducted the following tests separately for the reintegrational (primary outcome) and mental health (secondary outcome) data:

1. Bootstrap versions of the newly developed hybrid extended meta-analysis (HYEMA) tests (van Aert, 2025)
2. Worst-case sensitivity analysis tests (Mathur & VanderWeele, 2020),
3. p-uniform* tests (van Aert & van Assen, 2025).
4. Robust and adjusted version of the partially empirically correlated-hierarchical effects model that incorporates inverse sampling covariance weights (PECHE-ISCW; Chen & Pustejovsky, 2025; Rodgers & Pustejovsky, 2021).
5. Robust and adjusted versions of PET/PEESE, incorporating ISCW (Chen & Pustejovsky, 2025),
6. The newly developed bootstrapped step-function selection models for meta-analysis of dependent effect sizes (Pustejovsky, Citkowicz et al., 2025).

On the one hand, we chose these tests because they have shown the most promising statistical properties in the simulation studies embedded in the cited studies above, as well as shown in independent evaluations of publication bias methods (Carter et al., 2019; McShane et al., 2016). On the other hand, Chen and Pustejovsky (2025) and Pustejovsky, Citkowicz et al. (2025) have shown that regression-based methods (such as PECHE-RVE-ISCW and PET/PEESE-RVE-ISCW) are generally better at detecting publication biases when selection is weak relative to selection models (such as the 3PSM and 4PSM models), whereas the selection models clearly outperform regression-based methods in the presence of moderate to strong selection in effect sizes. We,

[6] We acknowlede that small study, reporting, and publication bias tests are not the same and that many statistical test cannot distinguish between these types of bias. For simplicity, however, For simplicity, however, we subsumed all these tests under the heading of publication bias tests, similar to Rothstein et al. (2005).

therefore, aimed to include a mixture of publication bias tests that function well under different levels of selective reporting.

To differentiate between publication bias and systematic and substantial differences between effect sizes, we conducted two types of tests, where we adjusted subgroup effects for publication bias. Specifically, we adjusted subgroup effects for publication bias for preregistration status and outcome type. As not all publication bias tests can be used for correcting subgroup effects, we only used a subset of the above-listed tests. Find information on the exact models used for this type of publication bias adjustment in the next section.

Across all models, despite the worst-case meta-analysis models, we used a modified version of the standard error and variance presented in Equation (3). More precisely, we defined the modified version of the standard error and variance as follows:

$$V_{gt}^{mod} = W \times \xi \text{ and } SE_{mod} = \sqrt{V_{gt}^{mod}}. \quad (4)$$

As described by Pustejovsky and Rodgers (2019), this approach is necessary to avoid the artificial correlation between SMD effect size estimates and their sampling variance, induced by the fact that SMD estimates are used to calculate P in Equation (3). For a deeper understanding of the relative difference between V_{gt}^{mod} and V_{gt}, see Figures 91 and 92 in the PRIMED workflow. We did not use modified standard errors in the worst-case meta-analysis, as it is intended to represent a sensitivity analysis that replicates the main analysis, just without including effect sizes that affirm one's non-null hypothesis and are statistically significant.

For visualization purposes, we applied contour-enhanced funnel plots (Peters, 2005), which aim to depict the relationship between the effect sizes and their estimated standard errors. In these plots, we included the estimated slope from the robust Egger's regression tests (van Aert, 2025) and colored the effect size estimate by their overall risk of bias assessment. We both visualize funnel plots of the study and effect size levels. For the study-level plots, we averaged[7] all within-study effect sizes, assuming a constant between-effects correlation of 0.8. As all of our risk of bias assessments were conducted at the effect size level, we did not color the average effect sizes in study-level plots. As a supplementary analysis, we also visualized funnel plots across different types of reintegrational as well as mental health outcomes.

Selective reporting and preregistration

A special feature of our data is that approximately half of the included studies (22 of 45 studies in the social integrational data and 20 out of 41 studies in the mental health data) were preregistered. In preregistered studies, one could expect that publication bias is either completely absent or at least much less pronounced compared to conventional/non-preregistered studies. For this type of effect size data, it has recently been suggested not to correct preregistered studies for publication bias (van Aert, 2025) or to model this factor (Pustejovsky, Citkowicz et al., 2025). In line with these recommendations, we assessed publication bias by only adjusting non-preregistered studies

[7] We used the aggregate.escalc() function from the metafor package for the aggregation.

or by adding a centered dummy variable for preregistration status to the given publication bias model. Centering of binary variables follows the recommendation forwarded by Fisher and Tipton (2015). Moreover, we depicted funnel plots separately for preregistered and non-preregistered studies.

Model-specific details (#)

Cluster bootstrap HYEMA

As the only test, HYEMA provides average effect size estimates where only effect sizes from non-preregistered studies are adjusted for publication bias. Although this test has shown promising performance (van Aert, 2025), it has only been evaluated under the assumption of independence among effect sizes (i.e., assuming all studies contribute one effect size only). To overcome this issue and to control the nominal Type I error rate, we cluster-boostrapped this model (Pustejovsky & Joshi, 2023). For the bootstrap models, we calculated the percentile confidence intervals, as these have shown the most promising performance in other applications (Pustejovsky, Citkowicz, et al., 2025).

Furthermore, we used the HYEMA model to adjust outcome moderator effects for publication bias. For multi-constrast tests, investigating whether effect sizes differed across different types of outcome, we computed bootstrap Wald test *p*-values defined as

$$p = \frac{1}{R}\sum_{r=1}^{R} I(F^r > F)$$

Where R is the total number of cluster bootstrap replications, and F is the naïve F-test. We used similar formulas to calculate the naïve F-test, similar to those implemented in the rma() function in metafor. Across all of these tests, we used $R = 1999$.

*p-uniform**

The p-uniform* method intends to adjust for publication bias by estimating the overall average effect size that makes the distribution of *p*-values in the data as uniform as possible. Apparently, a downside of the p-uniform* method is that it is based on the assumption of independence among effect sizes, which by design makes it miscalibrated when applied with dependent effect sizes. Nonetheless, Chen and Pustejovsky (2025) showed that the method performs well even in dependent effect size data, which is the main reason why we included this method. Yet, this method has only been developed to adjust the overall average mean effect, and we therefore only used it for this purpose.

Worst-case meta-analysis

The worst-case meta-analyses are a sensitivity analysis in which all positive and statistically significant effect sizes were excluded, under the extreme assumption that they represent false positives. If the remaining effect(s) is/are still statistically significant and substantial in size, this provides strong evidence that publication bias is not the primary factor driving the effect(s). We used this type of analysis to reestimate the overall average effect size as well as subgroup effects

across preregistration status and types of outcome. For the latter analysis, we controlled for the preregistration status.

Partly to ease comparison between publication bias tests and partly to ease the presentation of these tests, we only fitted these models using the PECHE-RVE model as presented in Equation (5) below (Pustejovsky & Tipton, 2021; Vembye et al., 2023). For the subgroup test, this model type slightly differs from the one used in the main analysis. To align the models, we also fitted moderator models using the same model as in the main subgroup analyses. These tests can be found in the publication bias script following this review.

(PE)CHE-RVE-ISCW

Next, we used the newly developed PECHE-RVE-ISCW model to test for publication bias (Chen & Pustejovsky, 2025). The key feature of this model relative to the original CHE-RVE model (see Equation 5) is that it incorporates weights that are based on inverse sampling variance estimates and the assumed covariance between these estimates. When using the modified version of the variance described in Equation (4), this corresponds to fixed-effect weighting, or more precisely, weighting by the inverse of the effective sample sizes and the assumed covariance among these.

To guard against model mis-specification and adjust for small sample issues, we used RVE, or more precisely, the robust HTZ for single-contrast tests (Tipton, 2015). This type of test was used for the overall average effect size and single-subgroup effects estimations, following the recommendation by Joshi, Pustejovsky et al. (2022). When estimating multiple-contrast hypothesis tests, we estimated cluster wild bootstrap Wald test *p* values (Joshi, Pustevjosky et al., 2022).

We used this test to adjust for publication bias in both the overall average effect size and the moderator effects across preregistration status and outcome types

PET/PEESE-RVE-ISCW

The PET/PEESE-RVE-ISCW models we used had the same shape as the PECHE-RVE-ISCW model, with the only exception that we either added the modified sampling variance (PEESE; precision-effect estimator with standard error) or the modified standard error (PET; precision-effect test) presented in Equation (4) as a predictor to the models. As with the PECHE-CHE-ISCW model, we used this test to adjust the overall average effect size as well as the moderator effects across preregistration status and types of outcome. Differently, and in addition to the PECHE-CHE-ISCW model, we also controlled for the preregistration status when estimating the overall average effect.

It has been recommended by Standley and Doucouliagos (2014) that "[w]hen the PET test is not rejected, meaning that the average effect is not statistically distinguishable from zero, then the PET intercept is used for estimating the adjusted average effect. However, if the PET test is rejected and the average effect is statistically distinct from zero, the PEESE is used for estimating the adjusted average effect" (Chen & Pustevjosky, 2025, p. 6). We followed this decision rule when reestimating the overall average effect with the PET/PEESE models. For subgroup models, we used the PET-RVE-ISCW only, as the decision rule was not clearly defined in this context.

Cluster bootstrap selection models

As selection models have shown promise across a range of simulations (Carter et al., 2019; Chen & Pustejovsky, 2025; Pustejovsky, Citkowicz, et al., 2025; Rodgers & Pustejovsky, 2021), we applied two versions of these types of models. That is, we used the three-parameter selection model (3PSM) and the four-parameter selection model (4PSM). The former included a single parameter, describing the "likelihood of nonaffirmative effect sizes being observed relative to affirmative effect sizes." (Chen & Pustejovsky, 2025, p. 5). For this model, we set the step parameter to 0.025, amounting to the threshold of a classical one-sided *p*-value. The latter model included two-step parameters, which we set to .025 and 0.50. This allowed for modeling different probabilities for non-significant effect sizes, conditional on the direction of the effect.

To account for dependencies in our effect size data, we used cluster bootstrap selection models (Pustejovsky, Citkowicz, et al., 2025). Specifically, we fitted these models using the composite marginal likelihood (CML) estimator with two-stage bootstrapped percentile confidence intervals based on 1999 re-sampled replicates, as this model has shown the best performance relative to other estimators and confidence intervals (see Figures 4 & 5 in Pustejovsky, Citkowicz, et al., 2025).

As with the PECHE-RVE-ISWC model, we used this test to adjust for publication bias in both the overall average effect size and the moderator effects across preregistration status and outcome types. In the outcome model, we controlled for the preregistration status.

R package used for publication bias testing

For publication bias testing, we used the `wildmeta` (version 0.3.2; Joshi & Pustejovsky, 2022), `boot` (Canty & Ripley, 2017)[8], and `metaselection` (version 0.1.5; Pustejovsky et al., 2025) packages to integrate bootstrapping techniques. We use `puniform` (version 0.2.7; van Aert, 2023) for the p-uniform* and HYEMA estimations. For the single-constrast test (i.e., *t*-test), we used the `metafor` (version 4.8-0; Viechtbauer, 2010) and `clubSandwich` packages.

Data synthesis

All main meta-analyses were conducted using R 4.5.1 (R Core Team, 2022) and RStudio (RStudio Team, 2015). Specifically, we applied the `metafor` package (version 4.8-0; Viechtbauer, 2010) together with its integrated sandwich estimators (version 0.6.0; Pustejovsky, 2020). For multiple-contrast hypothesis (i.e., Wald test *p*-values), we used the `wildmeta` package (version 0.3.2; Joshi & Pustejovsky, 2022).

Overall average effect – PECHE-RVE

As we experienced having various types of dependencies in our data, we derived the overall average effects for social integrational and mental health outcomes, using the partially empirical correlated-hierarchical effects (PECHE-RVE) model (Pustejovsky & Tipton, 2022), as this model can handle various types of dependencies among effect sizes. Specifically, this model accounts for the hierarchical data structure, with effect sizes nested within studies, allowing us to obtain

[8] Test drawing on the boot package was heavily inspired by Pustejovsky and Joshi (2023).

heterogeneity measures at both the study-level and the effect size-level. This provides valuable diagnostic information about at what levels unexplained heterogeneity might be explained. We used restricted maximum likelihood to estimate the between-study (τ^2) and within-study (ω^2) heterogeneity (Viechtbauer, 2005, 2007).

Moreover, the model accounts for correlation among the effect size standard errors, either by imputing or estimating the covariance between within-study effect sizes. In our data, we were able to estimate the covariance for some within-study effects. However, for the majority of the effect sizes, we estimate the study's variance-covariance matrix by imputing $\rho = 0.8$, as specified in our protocol.

As previously described in the 'Criteria for the determination of independent findings' section, we used robust variance estimation (RVE) to guard against model misspecifications.

Further model details and notation (#)

To explicate the statistical models used in this review, consider that our data represent a collection of J studies, each contributing. $k_j \geq 1$ effect size estimate. Let T_{ij} be the effect size estimate i ($i = 1 \ldots, k_j$) from study j ($j = 1, \ldots, J$) with know and fixed sampling variance s_{ij}^2. Then let x_{ij} denote a row vector of p covariates and $\boldsymbol{\beta}$ denote a vector of p regression coefficients. In formal parlance, the PECHE working model can be written as:

$$T_{ij} = x_{ij}\boldsymbol{\beta} + u_j + v_{ij} + e_{ij} \quad (5)$$

For the intercept-only model, estimating the overall average effect, x_{ij} reduces to a row vector of 1's and $\boldsymbol{\beta}$ represents the overall average effect. In this model, it is assumed that the study-level and within-study random effects follow and normal distribution with $u_j \sim N(0, \tau^2)$ and $v_{ij} \sim N(0, \omega^2)$. Also, the sampling errors $e_{ij} = T_{ij} - \boldsymbol{\beta}$ are assumed to follow a normal distribution with $e_{ij} \sim N(0, s_{ij}^2)$.

To account for correlated dependency, the model assumes that all within-study effects are equally correlated, thus that $Cov(e_{ij}, e_{hj}) = \rho s_{ij} s_{hj}$ for effect sizes $i \neq h$ within study j. As the correlation ρ among within-study effect sizes is usually unknown, it the often imputed and assumed to be constant across studies. However, when studies included multiple treatment groups compared to the same control group, the covariance between within-study effects can be asymptotically approximated. We used this approach when estimating the variance-covariances for the three studies with multiple treatment groups. We thereby used a mixture of empirically known and unknown covariance estimates.

Subgroup analysis and investigation of heterogeneity

To investigate whether focal moderators could explain the true difference in effect sizes, we conducted a series of subgroup and moderator analyses. These analyses fall into three categories: 1) analyses regarding theoretically relevant categorical moderators, 2) analyses of methodological and bias-related categorical moderators, and 3) analyses concerning theoretically relevant continuous moderators. For the first and second sets of analyses, we used the partially empirical

subgroup correlated-effects plus (PESCE+[RVE]) model (Pustejovsky & Tipton, 2022). For the third category, we relied on the PECHE-RVE model, as presented in Equation (5).

Specifically, we applied the PESCE+ model combined with RVE and cluster wild bootstrapping when investigating binary or categorical moderators. Generally, this amounts to conducting individual PECHE meta-analyses[9] across each subgroup while accounting for studies contributing $k_j \geq 1$ effects to $C \geq 1$ subgroups (i.e., the model handles dependent effect sizes). Importantly, this allows us to estimate reliable Wald test *p-values* with adequate statistical properties. To test whether single-subgroup effects were statistically distinct from null, we used RVE (i.e., the HTZ test), whereas we used cluster wild bootstrapping to estimate Wald test *p*-values comparing differences between the subgroup effects.[10] In line with the estimation of the overall average effect size, we fitted this model to include heterogeneity at both the effect and study levels (indicated by the + sign).

List of moderators

As per protocol, the moderator analyses included the following factors:

Theoretical factors:

- The measured outcome (e.g., alcholhol and drug use, social functioning, well-being and quality of life)
- The sample characteristics including:
  - samples with vs. without participants with schizophrenia,
  - the average percentage og males in the sample,
  - the average age of the participants.

- The type of intervention (cognitive-behavioral therapy [CBT] vs. other types of interventions).
- The number of intervention sessions
- The duration of the intervention

Methodological factors

- The preregistration status
- The type of test (clinician-rate vs. self-reported)
- The analytical strategy (ITT vs. TOT)
- The research design ([C]RCT vs. QES)
- The type of control group (individual treatment vs. treatment as usual (TAU and waitlist)

[9] Although the weighting scheme slightly varies across the SCE and CHE models (see Pustejovsky, 2020b; Supplementary material in Pustejovsky & Tipton, 2022). However, the heterogeneity measures would be the same as if one fitted the CHE model on each subgroup category.

[10] This follows the recommendation from Joshi, Pustejovsky et al. (2022), which is: "we recommend using CWB rather than HTZ for tests of multiple-contrast hypotheses in meta-analyses conducted using RVE. Tests of single-contrast hypotheses (i.e., t-tests) can be conducted using either the CWB or HTZ test because both methods have very similar power." (p. 473)

- The overall risk of bias status (high vs. low/moderate overall risk of bias
- The timing of the measurement after the end of the intervention.

Across all of the above-listed meta-regression analyses, we both fitted models, with and without controlling for other covariates than the independent subgroup variable. In the covariate-adjusted models, we controlled for 1) the type of outcome, 2) whether the sample included participants with schizophrenia, 3) whether the group-based intervention was based on CBT, 4) preregistration status, 5) type og test, 6) the analytical strategy, 7) the research design ([C]RCT vs. QES), and finally 8) type of control group. To guard against multicollinearity, we only added covariates that did not show intercorrelations above 0.5. Consequently, we did not add the total number of sessions as it correlated with the duration of the intervention, and we did not control for the effects' overall risk of bias, as it correlated with the research design of the studies, with non-randomized generally yielding effects with a higher risk of bias. Find the full correlation matrix used for the selection of variables in the PRIMED Tables 33 and 38.

To ease and ensure correct interpretation of overall average subgroup means from the PESCE+ model (Fisher & Tipton, 2015), we centered all bivariate and continuous control variables. For the former set of variables, we used mean-centering, whereas we used rounded median-centering for the continuous variables. When using the type of outcome measure as control, we added a dummy variable individually for each outcome to the model.

Further model details and notation (#)

The PESCE+[RVE] model accounts for dependent effect sizes by assuming that effects from the same studies falling into the same subgroup are correlated, whereas effects falling into different subgroups are assumed to be uncorrelated. Despite more advanced models being available that account for correlations between effects from the same study appearing across different subgroups, these models have only been shown to perform adequately under specific data structures, which were not present in our dataset. Therefore, we used the PESCE+[RVE], which formally takes the following form

$$T_{ij} = \sum_{c=1}^{C} m_{ij}^{c}\left(x_{ij}\boldsymbol{\beta_c} + u_{cj} + v_{cij}\right) + e_{ij} \quad (6)$$

Here, m_{ij}^{c} $\left(m_{ij}^{1} \ldots m_{ij}^{C}\right)$ is an indicator of whether the effect size i from study j falls in subgroup c for all $c = 1 \ldots C$. If so, $m_{ij}^{c} = 1$, otherwise $m_{ij}^{c} = 0$. The model is based in the following assumptions: $u_{cj} \sim N(0, \tau_c^2)$, $v_{cij} \sim N(0, \omega_c^2)$, and $e_{ij} \sim N(0, s_{ij}^2)$, $Cov\left(u_{cj}, u_{bj}\right) = 0$, $Cov\left(v_{cij}, v_{bij}\right) = 0$, and $Cov\left(e_{ij},\ e_{lj}\right) = \rho s_{ij} s_{lj} \sum_{c=1}^{C} m_{ij}^{c} m_{lj}^{c}$, thereby $Cov\left(e_{ij},\ e_{hj}\right) = 0$ when $m_{ij}^{c} \neq m_{hj}^{c}$.

Additional modeling (#)

As the majority of studies both reported on reintegrational and mental health outcomes, we used the partially correlated and multivariate effects plus model (PECVME; Pustejovsky & Tipton,

2022, first version on OSF) to statistically examine whether group-based interventions with larger impacts on reintegrational outcomes also tended to have greater impacts on mental health outcomes. The major difference between PECMVE and PESCE is that the PECMVE model allows effects from the same studies falling in different subgroups to be correlated. The model takes the following form

$$T_{ij} = x_{ij}\boldsymbol{\beta} \sum_{c=1}^{C}\left(m_{ij}^{c}u_{cj} + m_{ij}^{c}v_{cij}\right) + e_{ij} \quad (7)$$

where $u_{cj} \sim N(0, \tau_c^2)$, $v_{cij} \sim N(0, \omega_c^2)$, and $e_{ij} \sim N(0, s_{ij}^2)$, $Cov\left(u_{cj}, u_{bj}\right) = \psi_{bc}\tau_b\tau_c$, $Cov\left(v_{cij}, v_{bij}\right) = \zeta_{bc}\omega_b\omega_c$, and $Cov\left(e_{ij},\ e_{lj}\right) = \rho s_{ij}s_{lj}$. ψ_{bc} and ζ_{bc} are the correlations between the random effects at the study and effect size level, respectively.

Sensitivity analysis

To test the robustness of our results, we conducted a range of sensitivity analyses. First, we tested if our results were sensitive to the assumed correlation among within-study effects, as this assumption can impact some parts of the model estimation. Specifically, we reestimated all results by setting $\rho = 0,\ .2,\ .4,\ .6$. Note, the first specification amounts to fitting a classical multi-level meta-analysis model (MLMA-RVE; Fernández-Castilla et al., 2020; Van den Noortgate et al., 2013).

Second, we ran sensitivity analyses across different types of effect size metrics. That is, we reestimated all results using a covariate/pretest-adjusted version of both Cohen's d and Hedges' g, g_t (c.f. Equation 1) standardized by its original standard deviation and not the population-based version, a posttest-only version of g_t. We also tested the sensitivity across different assumptions of the interclass correlation (ICC) used to cluster-bias-adjust effect sizes, setting $ICC = 0.05, 0.15$.

Finally, we investigated whether the impact of removing high-risk of bias effect and non-randomized studies on our results. We did not examine the impact of outliers, as we did not find any one as defined as effect sizes falling more than three times the interquartile range below the first quartile or above the third quartile (Pustejovsky, Zhang, et al., 2025; Tukey, 1977)

Treatment of qualitative research

We did not include qualitative research in this review.

Deviations from the preregistered protocol

In a few instances, we have deviated from our protocol. The main reason for deviation from the protocol was that new and better-performing methods were developed since we submitted the protocol. This included the methods developed by Chen and Pustejovsky (2025), Fitzgerald and Tipton (2024), Pustejovsky, Citkowitz et al. (2025), Pustejovsky, Zhang et al. (2025), van Aert

(2025), and Wu, Duan et al. (2025). As all of these methods (or advice) show more appropriate statistical performance than the methods we originally described in the protocol, we found it reasonable to implement these methods. Of particular note, the majority of these method developments were developed by the statisticians who had developed most of the methods we describe in the protocol. Thus, we felt confident in updating our methods to keep up with the state-of-the-art. Moreover, we provide open data, allowing others to replicate our work but also to conduct the original suggested analyses, if desired.

A more questionable method deviation from the protocol, however, was that we used mean imputation to handle the missingness of focal moderating factors. Meanwhile, we only used this approach as we only had one missing study for two moderators. Therefore, we considered the advances of mean imputation to outweigh the downside of this approach for the following reasons. Firstly, using more advanced methods to handle missing values would unnecessarily complicate Wald test estimation (see Vembye, Weiss et al., 2024 for a discussion of this issue). Secondly, using list-wise deletion would make us lose important information on other moderating factors that were fully reported within the given study. Therefore, we found the mean imputation to be an acceptable compromise between these two alternative strategies for handling missing values.

In the protocol, we originally wrote that we would add a critical risk of bias judgment to the RoB2 tools. Yet, since these tools clearly state that reviewers are not allowed to modify or extend the tools, we did not follow this practice. In addition, we wrote that we would include Eklund et al. (2017). However, after scrutinizing this study, we did not include it as it contained group-based interventions in both the treatment and control groups. Thus, it fell outside the inclusion criteria of the review.

Finally, with our improved understanding of the literature, we now distinguish more clearly between primary and secondary outcomes than was originally described in the protocol.

Results

Description of studies

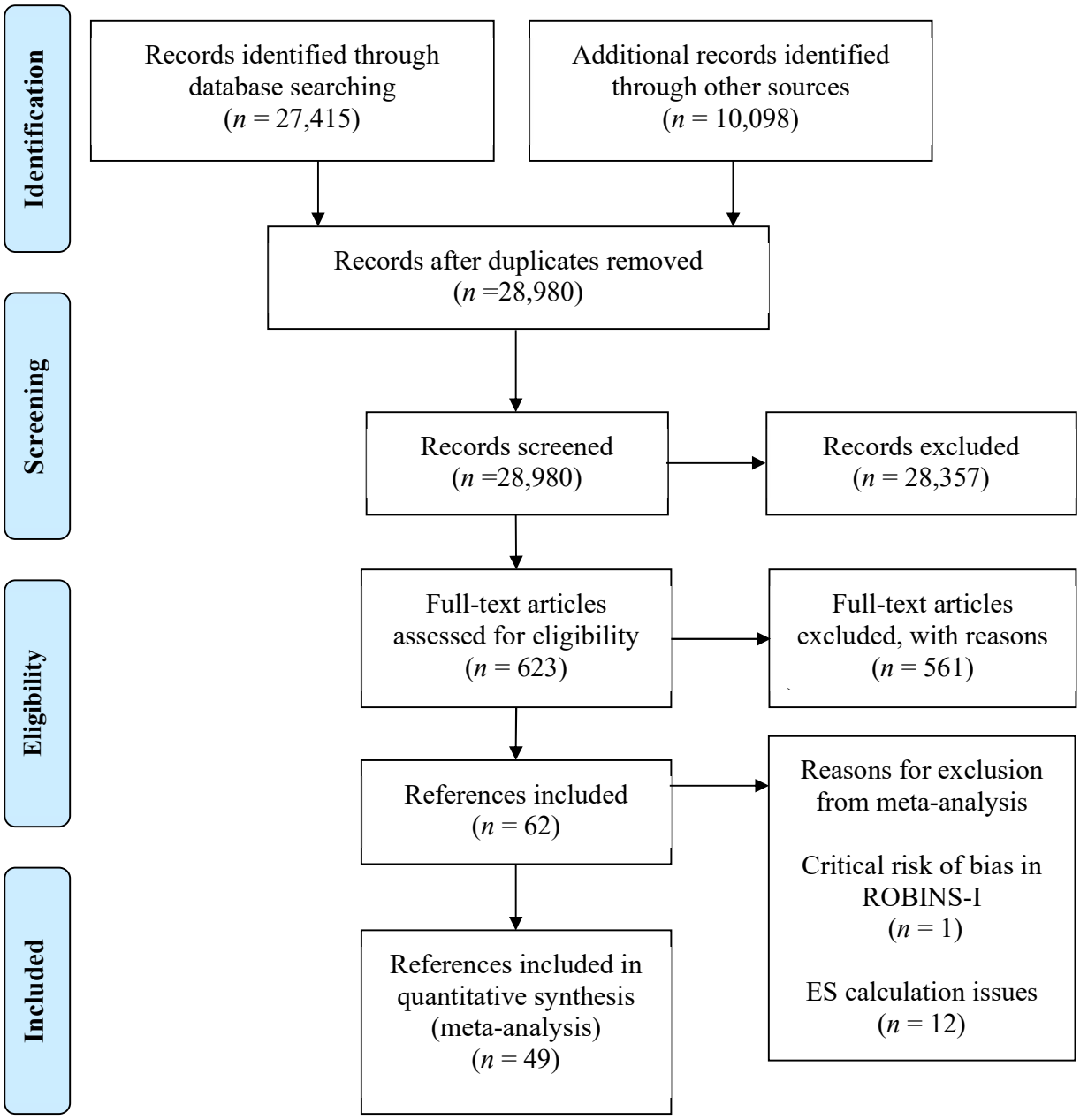
In the following sections, we describe and present the most essential descriptive statistics and figures. Yet, we have followed the preliminary data analysis workflow for meta-analysis of dependent effect sizes (PRIMED), as suggested by Pustejovsky, Zhang, et al. (2025). These expanded analyses can be found in the supplementary files accompanying this review.

Results of the search

The PRISMA flow chart (Moher et al., 2009), presented in Figure 4, documents our search process and the criteria for exclusion of references. The total number of potentially relevant records was 28,980 after excluding duplicates (database: 18,882; grey, hand search, snowballing, and other resources: 10,098; duplicates: 8533). All records were screened based on title and abstract; 28,357 were excluded for not fulfilling the screening criteria, including 35 records that were unobtainable despite efforts to locate them through libraries and internet searches. 623 records were ordered, retrieved, and screened in full text. Of these, 561 did not fulfil the screening criteria and were excluded. 62 studies were included in the review, of which 49 studies could be used in the data

synthesis/meta-analysis. The main reason for not including studies in the meta-analysis was due to insufficient reporting of the results, which hindered valid effect size calculation. We excluded 12 studies for this reason. Moreover, we excluded one study (i.e., Bond et al., 1991) due to critical risk og bias, as it received serious risk of bias in four domains.

FIGURE 4 Flow chart of the screening process PRISMA template



Included studies

In this section, we first describe the full body of literature, independently of whether the study was included in the meta-analysis. Afterwards, we describe and provide descriptive statistics for the studies entering the meta-analysis. All details about the included studies can be found in the supplementary material accompanying the review.

Included studies (Overall)

The search resulted in a final selection of 62 studies, which met the inclusion criteria for this review. 51 studies were RCTs and 11 studies were non-randomised studies, with a comparison of two or more groups of participants, that is, at least a treated group and a control group. Descriptions of the intervention and control conditions within each included study were extracted in as much detail as possible and can be found in the supplementary descriptive table. The 62 studies analysed data from 62 different samples. In one case, we identified a large number of publications from the same trial (Trial registration: ISRCTN57595077; ISRCTN66721740), but data were only extracted from one publication (Sommers et al., 2017).

Participants in the included studies

Participants in the included studies suffered from a wide range of different types of mental illness diagnoses and other problems. Table 1 summarizes the different categories of participants based on the reported diagnosis and indicators of social marginalization.

TABLE 2

Characteristics of the participants	Number of studies (N=62)
Mental health disorder	
Schizophrenia or other primary psychotic disorders	36
Borderline personality disorder	8
Personality disorders[a]	10
Depressive disorders	25
Bipolar disorders	20
Anxiety disorders	15
Obsessive-compulsive disorder	3
Post-traumatic stress disorder	8

Other mental health disorders[b]	9
Not specified	5
Social Marginalization	
Homelessness	2
Substance abuse[c]	13
Unemployment	5
Criminal justice involvement	1
Chronic medical conditions[d]	6
Psychiatric comorbidity	1
Loneliness	2
Trauma experience	3
Other social problems[e]	5
Not clearly specified	28

Note.

[a] If specified, primarily DSM cluster C personality disorders.

[b] Including dissociative identity disorder, autism spectrum disorder, eating disorders, other psychotic disorders, suicidal ideation, and elevated levels of psychological stress

[c] Including both alcohol and drugs

[d] Including both psychiatric and somatic chronic conditions

[e] Including living in residential group homes, being older than 60 years old (?), and self-reported social problems such as having experienced disruptive periods, elevated internalized stigma, and experiencing social, or cognitive disability

Interventions in included studies

Interventions in the included studies were as follows.[11] The parentheses indicate the number of studies included in the meta-analysis.

- Group-based cognitive behavioral therapy (CBT): 12 (10)
- Illness management: 8 (8)
- Addiction management: 2 (2)
- Illness and addiction management: 4 (2)
- Cognitive-behavioral social skills training: 6 (3)
- Social cognition and interaction training: 6 (6)
- Residential treatment: 2 (2)
- Group psychoeducation & social skill training: 9 (6)
- Group psychoeducation: 5 (3)
- Seeking safety: 2 (2)
- Social network training: 2 (2)

[11] The numbers might sum to more than 62 studies (49 studies included meta-analysis), as three studies contributed with two treatments each.

- Vocational training: 3 (2)
- Positive psychology group intervention: 2 (2)
- Art therapy: 1 (1)
- Reading group intervention: 1 (1)

In studies where the intervention is explicitly named as cognitive behavioral therapy with group elements, they are categorized as Group-based Cognitive Behavioral Therapy. This includes a total of 12 studies: Beames et al. (2020), Cano-Vindel et al. (2022), Craige & Nathan (2009), Hagen & Nordahl (2005), Halperin et al. (2000), Himle et al. (2004), Madigan et al. (2012), Michalak et al. (2015), Rabenstein et al. (2015), Smith et al. (2020), Wuthrich & Rapee (2013), and Yanos et al. (2012).

Several studies combine teaching about the participants' psychiatric diagnosis while also facilitating social interactions among participants as part of the intervention, aimed at developing social skills. These interventions are referred to as Group Psychoeducation & Social Skill Training. This includes nine studies: Acaturk et al. (2022), Barbic et al. (2009), Burnam et al. (1995), Gonzalez & Prihoda (2007), Gutman et al. (2019), Haslam et al. (2019), Munroe & Marziali (1995), Popolo et al. (2019), and Vallina-Fernández (2001). Of note, the intervention investigated by Acaturk et al. (2022) also included elements of cognitive behavioral therapy.

Group interventions that use cognitive behavioral therapy but focus on providing participants with practical training in engaging in social interactions are referred to as Cognitive-Behavioral Social Skills Training. One study directly uses this designation for their group intervention (Rajji et al. 2021). The five other studies mention it indirectly by stating that they use cognitive behavioral therapy in a group format with a special focus on teaching participants social skills (Daniels & Roll 1998, Izquierdo et al. 2021, Michalak et al. 2015, Jacob et al. 2010, Patterson et al. 2003).

Studies where the treatment effort is named Social Cognition and Interaction Training and specifically deals with training participants' social skills. These interventions include both training in understanding social contexts and concrete training in engaging in interpersonal relationships. With varying emphasis on cognition and interpersonal interaction, the six studies by Crawford et al. (2012), Gordon et al. (2018), Hilden et al. 2021, Kanie et al. (2019), Lim et al. (2020), and Wojtalik et al. (2022) fall under this category.

Five studies (Bond et al., 1991; Bækkelund et al., 2022; Kallestad et al., 2006; Morton et al., 2012; van Gestel-Timmermans et al., 2012) focus exclusively on the educational aspect, where the group intervention involves participants collectively learning about their diagnoses and how to manage them. These interventions are classified under Group Psychoeducation. Of note, the intervention investigated by Bækkelund et al. (2022) also included elements of cognitive behavioral therapy.

In four studies (Ball et al., 2005; Morley et al., 2014; Rosenblum, 2013; Weiss et al.,2000), the group intervention involves managing participants' substance abuse alongside their psychiatric illness. The overarching category for these interventions is called Illness and Addiction Management.

Illness Management pertains to interventions that in a group format focus on providing participants with tools to manage their mental illness and related symptoms. Eight studies fall under this

category: Druss et al. (2010), Druss et al. (2018), Dyck et al. (2000), McCay et al. (2006), McCay et al. (2014), Rüsch et al. (2019), Sajatovic et al. (2008), Saloheimo et al. (2016).

Addiction Management pertains to interventions that in a group format focus on providing participants with tools to manage their addiction and related symptoms. Two studies fall under this category: James et al., 2004; Schäfer et al., 2019.

Vocational training was used in three studies: Bond et al. (2015), Bozzer et al. (1999) and Russinova et al. (2018). Here, participants in the interventions practiced their verbal skills and presentation to facilitate reintegration and employment.

In several studies, the group element in the treatment-intervention involves participants living together in some form of residential setting. The category is called Residential Treatment and two studies fall under this category: Patterson et al. (2014) / Somers et al. (2017) and Sacks et al. (2011).

In the studies by Gatz et al. (2007) and Schäfer et al. (2019), the intervention called Seeking Safety is used, which is a group-intervention aimed at mentally ill and vulnerable individuals.

Two studies, Lloyed-Evans et al. (2020) and Tjaden et al. (2021), aimed to develop and broaden participants' social networks through group interventions, which are classified as Social Network Training.

In two studies (Schrank et al. 2016 and Valiente et al. 2022), the treatment interventions are based on positive psychology but are applied in a group format. These are referred to as Positive Psychology Group Intervention.

One study, Crawford et al. (2012), utilizes art-based group therapy to address participants' comorbid conditions, why this intervention is classified as Art Therapy.

A single study by Volpe et al (2015) used books and their stories to help people with mental health issues. In this Reading Group Intervention, participants discussed the stories they read in a group setting, which helped them understand their own emotions and those of others.

Excluded studies (Overall)

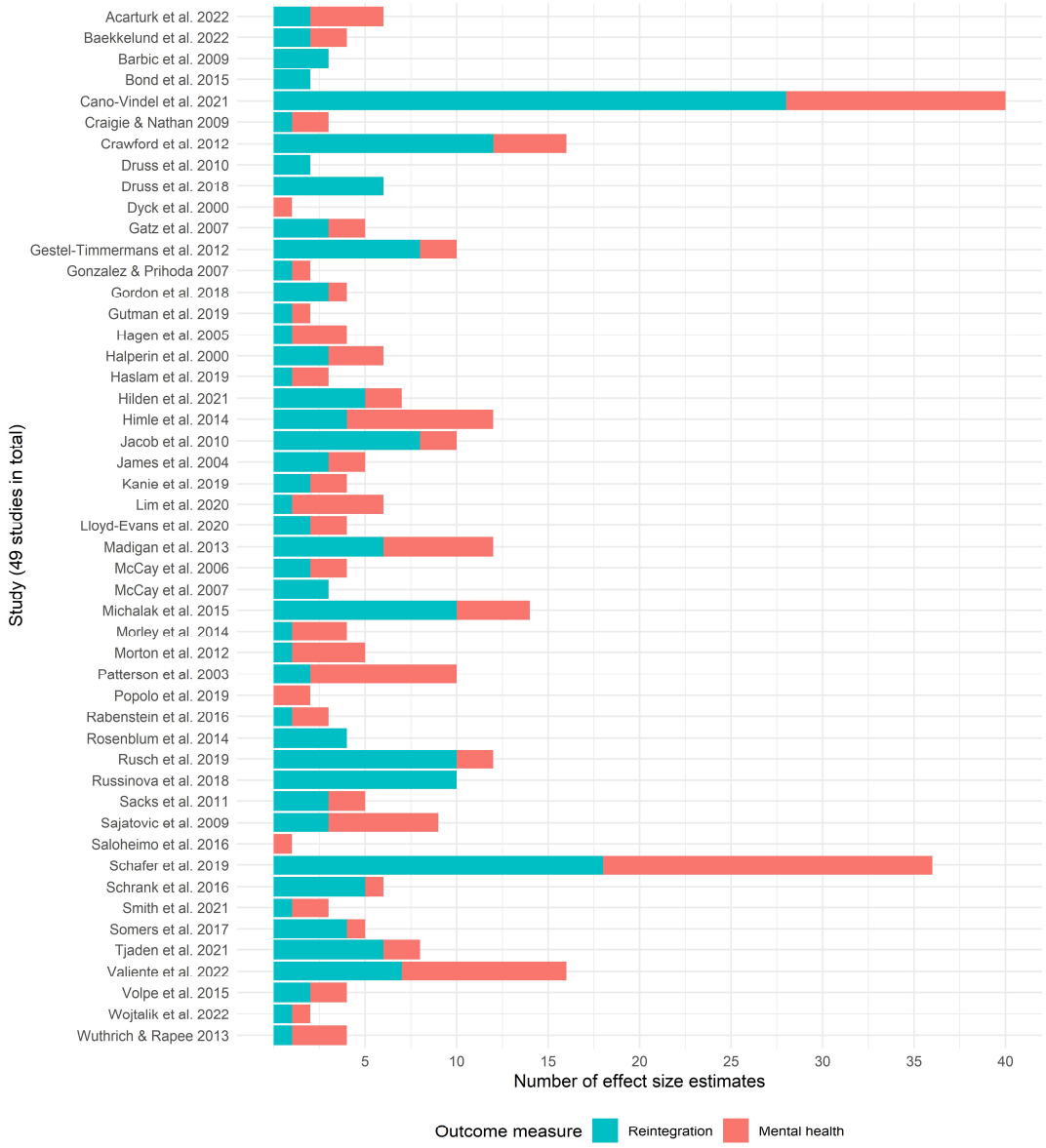
118 studies were excluded after initial inclusion, as both the intervention and control groups received interventions in a group. A list of these studies is available from the authors upon request.

Included studies (meta-analysis)

In this section, we present descriptive statistics for all studies included in the final meta-analysis. Since we both extracted and calculated effect sizes for primary and secondary outcomes (i.e., social reintegrational and mental health outcomes), we present all descriptive statistics separately for these two outcomes. We will follow this strategy in the remaining sections. The Descriptive statistics for studies reporting on reintegrational outcomes are presented in Tables 3 and 4, while the descriptive statistics for studies reporting mental health outcomes are presented in Tables 5 to

6. This section primarily focuses on the descriptive statistics for social reintegration, as similar trends emerge due to the considerable overlap between studies reporting reintegration and mental health outcomes. Figure 5 illustrates how the two types of outcomes are distributed across studies.

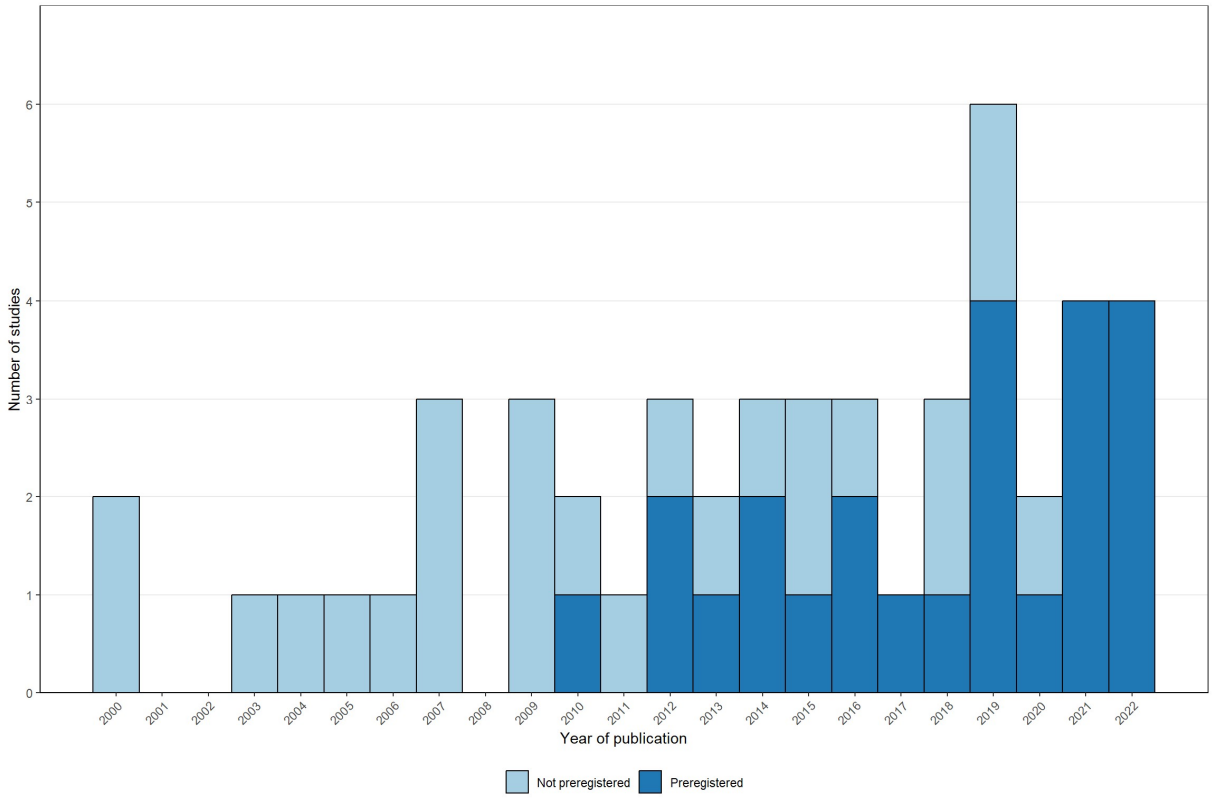
FIGURE 5 Number of reported effect size estimates across studies by type of outcome



Note: This plot is based on 349 effect sizes from 49 studies.

On a general note, Figure 6 presents the year of publication and preregistration status across all included studies. Here, it appears that the number of published studies has significantly increased since 2009, with 40 studies (82%) published in this period. Since 2012, preregistration has been a common feature in this field of study. Between 2012 and 2022, 23 out of 34 studies (68%) were preregistered. In our view, this indicates a major improvement in the quality of the literature.

FIGURE 6 Number of studies included in meta-analysis by year



Descriptive statistics for studies reporting social integrational outcomes

The final meta-analysis of reintegrational outcomes was based on 46 independent studies and 205 effect sizes, encompassing a total of 5,390 individual participants. All studies included one sample and were published in an academic journal. Also, all studies represented short-term effects. That is, all outcomes were measured less than a year after the end of the intervention. As an exploratory analysis, we investigated whether variation in effect size could be explained by the measurement timing within the first year after the end of the intervention. However, we did not detect any relationship in this regard, see PRIMED Figures 75 and 76.

The included studies were conducted in a broad range of countries. Most commonly, studies were conducted in the U.S. (13), Europe (15), or Commonwealth nations (16). We only found two studies from Asia – one from Japan and one from Korea.

The most commonly reported outcomes of social reintegration were well-being and quality of life (25 studies, 68 effect sizes), social functioning (17 studies, 48 effect sizes), hope, empowerment, and self-efficacy (12 studies, 32 effect sizes), and alcohol and drug abuse (eight studies, 32 effect sizes), respectively. Less frequently reported measures were self-esteem (five studies, 14 effect sizes), loneliness (four studies, five effect sizes), physical health (two studies, three effect sizes), employment (one study, two effect sizes), and psychiatric hospitalization (one study, one effect). For our later analyses, we merged the physical health, employment, and psychiatric hospitalization outcomes into the 'Other' category, as these outcomes were reported too infrequently to be used meaningfully in the subgroup analysis.

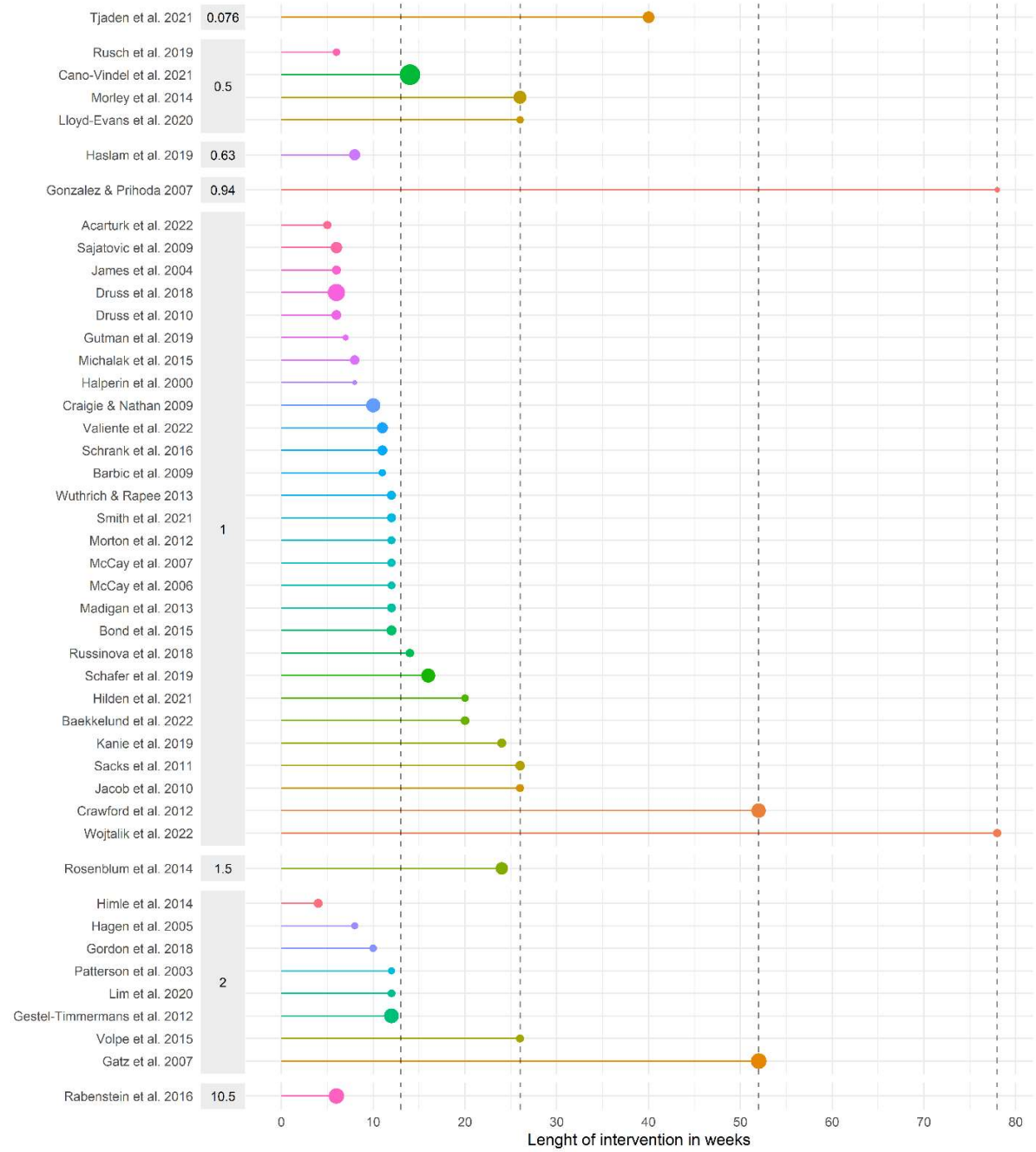
Of further measurement characteristics, 12 studies contributed 36 effect sizes derived from clinician-rated outcomes, while 39 studies contributed 168 effect sizes derived from self-reported outcomes. A pile of 24 studies reported treatment-on-the-treated (TOT) effects, with the other pile (22) reporting intent-to-treat (ITT) effects. As stated in the protocol, we prioritized TOT estimates. Accordingly, we excluded 45 effect sizes from Cano-Vindel et al. (2022), Craige and Nathan (2009), and Wojtalik et al. (2022), as these studies reported both TOT and ITT estimates, resulting in redundant data.

Of the included samples, 9 samples/studies included participants with schizophrenia. The average age of the samples was ~ 41 years, ranging from 25 to 67 years. In addition, the average percentage of males in the sample was ~ 47%, ranging from 0 to 81%. As seen in Table 4, the average raw sample size was 118, ranging from 16 to 631, and the average effective sample size was 70, varying from 10 to 351. The treatment and control sample sizes were generally well-balanced across studies, with averages of 58 and 54 participants for the treatment and control groups, respectively.

The vast majority (38) of the studies were randomized controlled trials (RCTs). Of the 38 RCT studies, 22 were preregistered. Interestingly, preregistered studies yielded 2.13 (140/65) times more effect sizes relative to nonpreregistered studies, despite accounting for less than half of the total studies included. The remaining 8 studies were quasi-experimental. Across all types of research designs, only three studies accounted for the multilevel structure of the data prompted by the clustered group-intervention. The most common control group used across studies was treatment as usual with or without a waiting-list (39), with the remaining studies using a waiting-list only (4) or an individual version of the group-based treatment (3).

The average length of the included interventions for studies reporting reintegrational outcomes was ~ 18 weeks, ranging from 4 to 72 weeks. Most interventions (49 out of 52) were administered at a rate of 0.5 to 2 sessions per week. Figure 7 depicts the number of sessions per week and the length of the interventions across studies that reported on social reintegrational outcomes (see the PRIMED Figure 72 for mental health outcomes). One study (Somers et al., 2017) is missing in Figure 2, as the duration and length of the intervention were not reported. Of the included interventions, 12 studies contained cognitive behavioral therapy (CBT). Although not mainly categorized under group-based CBT, we included Bækkelund et al. (2022) and Acarturk et al. (2022) in this category, as CBT was a major component of the given interventions.

FIGURE 7 Duration and intensity (number of sessions per week)



Note: Point sizes are weighted by the total sample size of the study. The gray left facets indicate the average number of sessions per week. Dashed lines indicate the length of three months, six months, one year, and one and a half years, respectively.

Descriptive statistics for studies reporting mental health outcomes

The overall descriptive statistical patterns were proportionally similar between studies reporting reintegrational and mental health outcomes. The specific descriptive statistics can be found in Tables 5 and 6.

The dataset included 144 effect sizes derived from 42 studies focusing on mental health outcomes. Specifically, nine studies contributed 14 effect sizes for anxiety, 20 studies contributed 37 effect sizes for depression, 29 studies contributed 72 for general mental health, and seven studies contributed 21 effect sizes for psychotic symptoms.

Of note, three studies (Dyck et al., 2000; Popolo et al., 2019; Saloheimo et al., 2016) were included in this data only, as they did not report on any eligible reintegrational data, which can also be seen in Figure 5.

Risk of bias assessment of included studies in meta-analysis

As previously mentioned, we excluded one study (Bond et al., 1991) due to a critical risk of bias, as it received high risk of bias judgments in four domains. Figures 8 and 9 depict the risk of bias in raw numbers (i.e., studies and effect sizes) for the included randomized studies across preregistration status for reintegrational and mental health outcomes, respectively. We visualize the risk of bias plot across preregistered and non-preregistered studies, as different bias mechanisms might operate in these types of studies (van Aert, 2025). As we only found RCTs to be preregistered, we only make this distinction for studies assessed with RoB2. Figures 10 and 11 display the corresponding risk of bias assessments for non-randomized studies for reintegrational and mental health outcomes, respectively. Additional plots displaying percentages, weighted percentages, and risk of bias across various outcome types are presented in PRIMED Figures 1–18.

RoB2 results

Generally, the majority of effect sizes from RCTs (~90%; 162 out of 187 effect sizes) received an overall risk of bias judgment of low or some concern. Only nine studies contained minimum one effect size judged to be of high risk of bias. The main reason for effect size from RCTs to receive high risk of bias was due to selective reporting when studies were not preregistered, and missing data when studies were preregistered. As shown in Figure 8, only the 10 preregistered studies received a low risk of bias rating. One preregistered study (Himle et al., 2014) received 'some concern' in risk of bias in the selection of reported results-domain, as the authors stated that the protocol was available upon request. However, we contacted the authors for this information without success. Yet, we did not find any clear evidence of selective reporting within the given studies. By contrast, the absence of a preregistration protocol in non-preregistered studies was the main factor contributing to an overall risk of bias judgment of 'some concern.'

Similar patterns between studies reporting reintegrational and mental health outcomes, see Figure 9.

FIGURE 8 RoB2 judgments for reintegrational outcomes

Note: The plot is based on 187 effect sizes from 38 studies. The reason why not all bars have the same length is that Acarturk et al. (2022) contained effect sizes that received differential risk of bias assessments.

FIGURE 9 RoB2 judgments for mental health outcomes

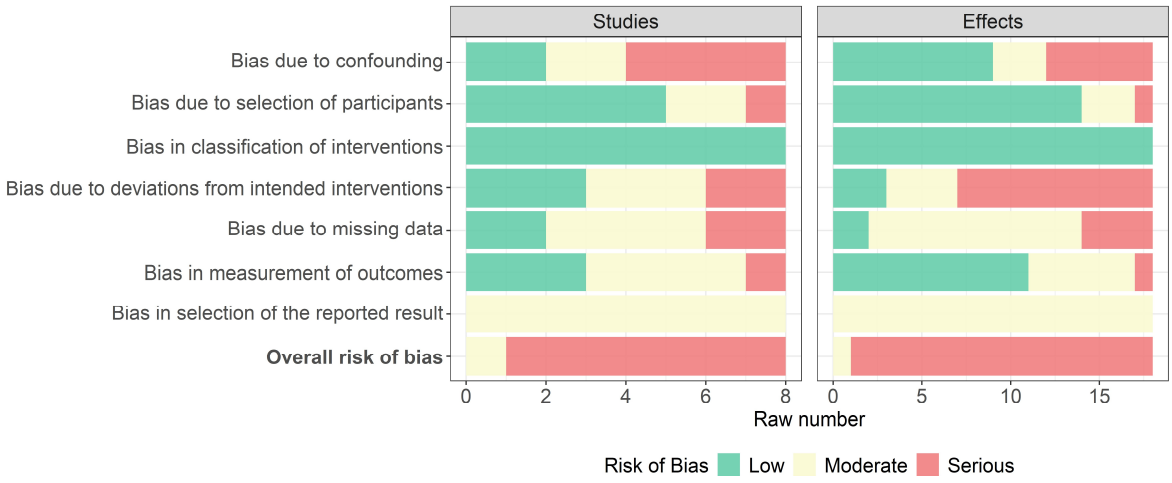
Note: The plot is based on 127 effect sizes from 34 studies. The reason why not all bars at the study-level have the same length is that Acarturk et al. (2022) contained effect sizes that received differential risk of bias assessments.

ROBINS-I results

As shown in the ROBINS-I Figures 10 and 11, the overall risk of bias was generally higher for non-randomized studies, with seven out of eight studies (add effects) receiving a "serious" risk of bias rating. Across all outcome types, the primary reason for this higher risk among non-randomized studies was confounding, with four of the eight studies judged to have a high risk in this domain. Minor reasons for high risk of bias included deviation from the intended intervention missing data. We did not find any non-randomized studies with a prespecified protocol. Therefore, all non-randomized studies were never rated to have a low overall risk of bias.

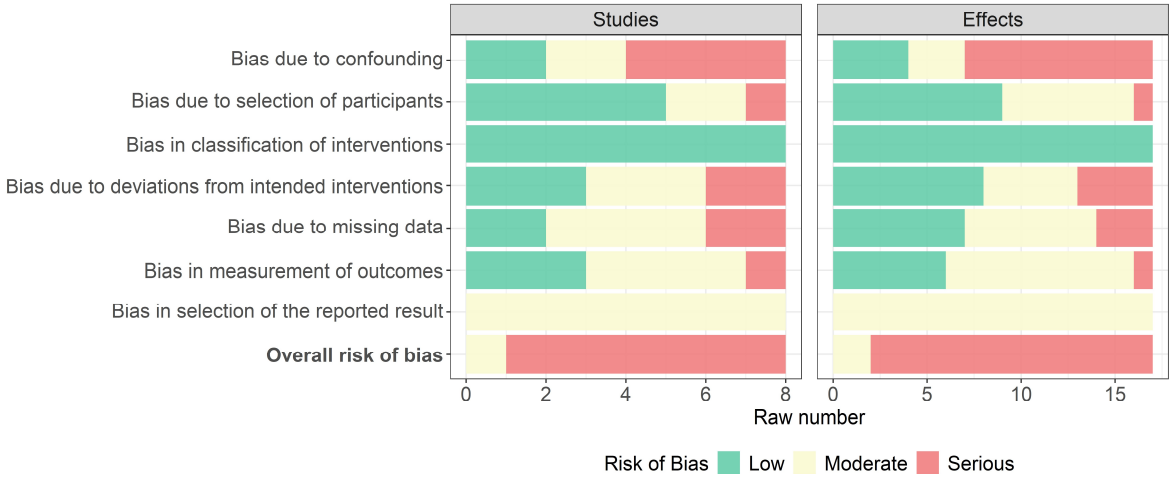
In sum, when including non-randomized studies, careful consideration must be given to the high risk of bias associated with their effect size estimates when incorporating them into the subsequent meta-analysis. Accordingly, we included risk of bias as a control variable in our covariate-adjusted moderator analyses (Scherer & Emslander, 2025). However, across all analyses, we found no systematic or substantial differences in effect sizes between studies with high/serious and those with low/moderate risk of bias (see Tables 8 and 11).

FIGURE 10 ROBINS-I assessment of reintegrational outcomes



Note: The plot is based on 18 effect sizes from 8 studies.

FIGURE 11 ROBINS-I assessment of mental health outcomes



Note: The plot is based on 17 effect sizes from 8 studies.

Synthesis of results

Main analysis

In this section, we describe the overall average effect size estimates (i.e., the main analysis) separately for reintegrational and mental health outcomes, respectively.

Overall average effects on social reintegration

The overall average effect size estimate from the meta-analysis of reintegrational outcomes summarizes a total of 205 effect sizes from 46 studies published between 2000 to 2022. We excluded one effect size from Barbic et al. (2009) because the means appeared to be flawed, as they were heavily disproportional to the reported standard deviations, yielding an effect size above four. Figure 12 depicts the distribution of all effect sizes across studies and shows the specific weight attributed to each effect size within the given study. As can also be seen from Figure 12, we found a positive, statistically significant overall standardized mean difference for reintegrational outcomes of 0.195 standard deviation (SD), $t(24.9) = 5.51$, $p < 0.001$, 95% *CI*[0.122, 0.268]. In other words, on average, participants in the group-based treatment conditions had better reintegrational outcome scores relative to individual treated control and waitlist conditions.

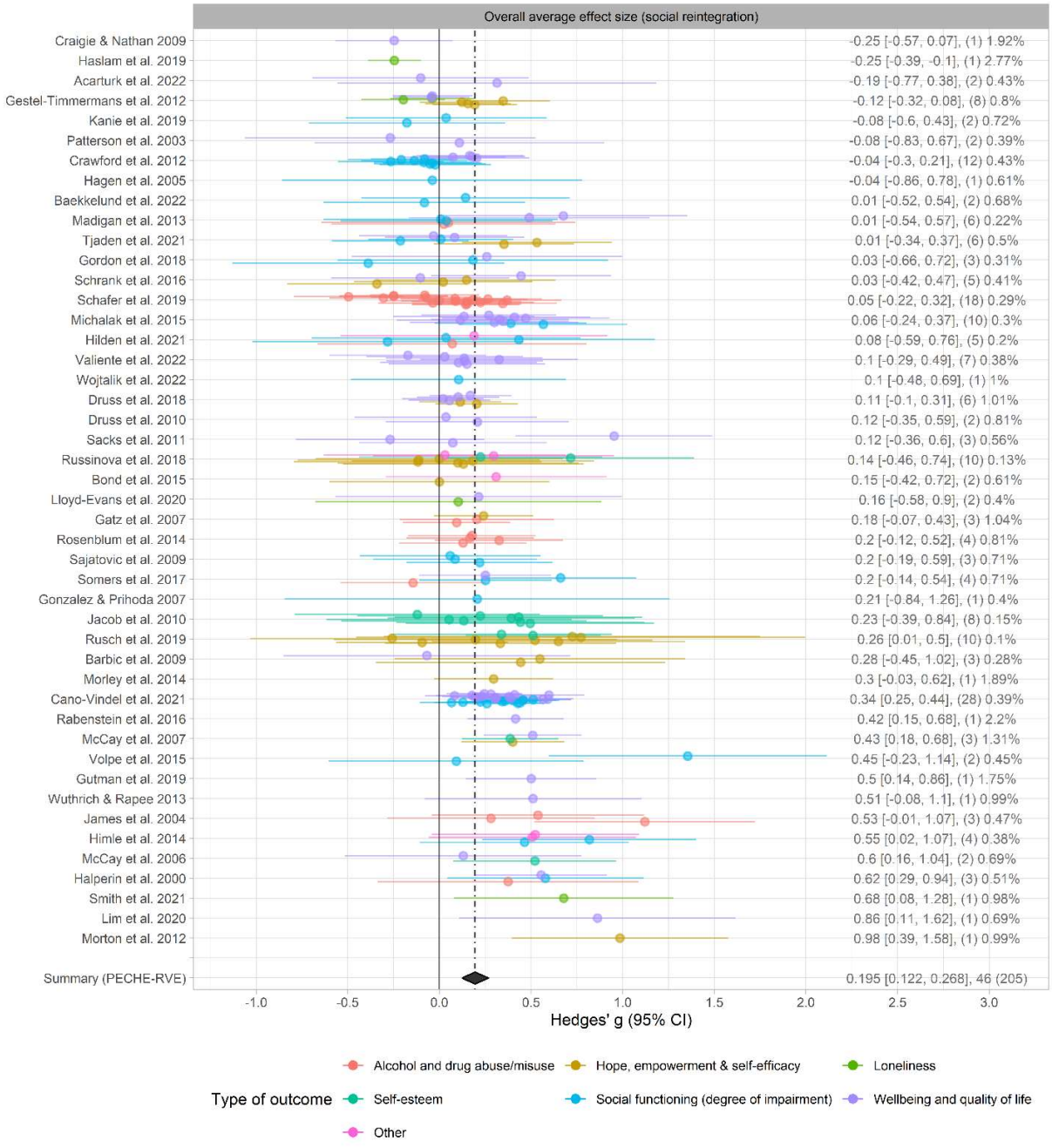
Due to the considerable cost-effectiveness of group-based interventions relative to individual treatments (NICE, 2025; see Figures 8, 10, 12, and 14), we consider this overall average effect size to be substantial. Using Cohen's U_3, which expresses the percentage of one group exceeding the mean of another, the overall average effect size of 0.195 SD amounts to 57.7 percent of the treatment participants having a better score than the average control participant, with the 95% confidence interval ranging from ~54.9 to ~60.6 percent of the treatment participants scoring better than the average control participant. Assuming constant effects, on the individual level, this translates to an expectation that a typical participant from the control group would have had a

percentile gain of 7.7% had they instead been exposed to a group-based intervention. For alternative interpretation metrics, see our main analysis codes accompanying the review.

In this context, another metric that can express the substantial impact of group-based intervention on the social reintegration of the participants is the comparison between the treatment effect and the usual improvement/development of participants receiving individual treatment as usual (TAU). As the majority of studies (i.e., 35 out of 39 studies) using TAU both reported baseline and posttest effects, we were able to calculate the typical development in the TAU control group adequately. Based on 35 studies and 173 effect sizes, we found the typical improvement on social reintegration within the TAU control group to be 0.083 SD, $t(28.04) = 2.01$, $p = 0.055$, 95% CI[-0.002, 0.168].[12] Seen in this light, the treatment effect can be said to be 2.3 times larger (0.195/0.083) than the typical improvement when receiving TAU.

[12] We calculated these pre-posttest effect sizes using the formulas from Borenstein and Hedges (2019) and when used the CHE-RVE model (with $\rho = 0.8$) to derive the overall average effect. For interpretation, we drew on the recommendation from Valentine, Aloe et al. (2019).

FIGURE 12 Effect size forest plot with dependent effect sizes for reintegrational outcomes

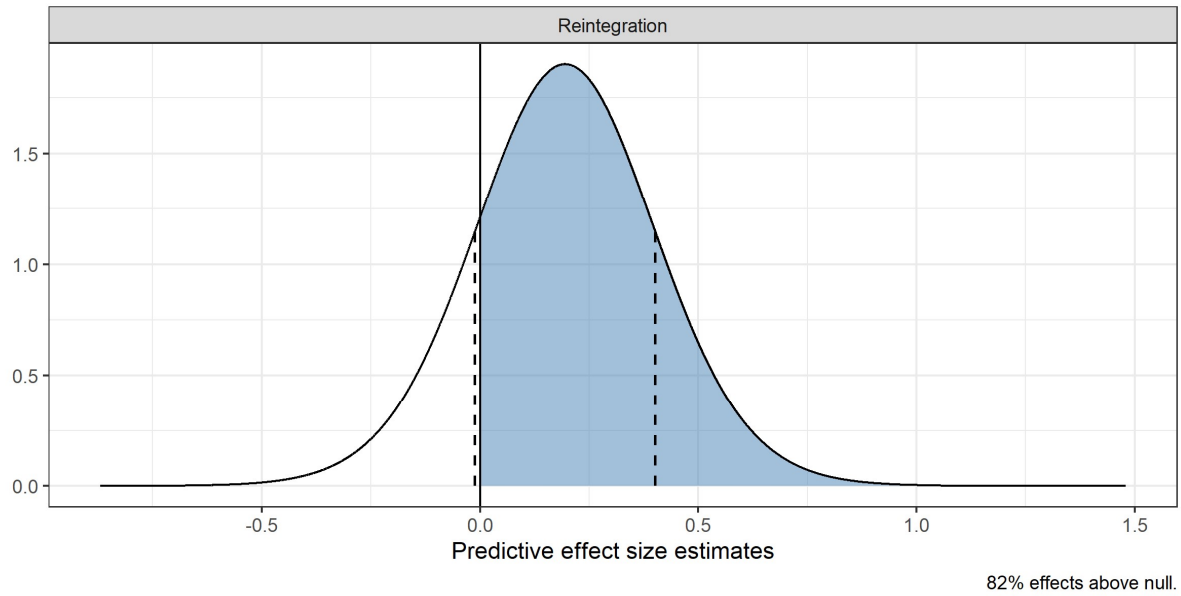


Note: The number of effect size estimates is in parentheses. The percentages indicate the weight given to each effect size within the given study. Studies are ordered by the study mean effect size obtained from fitting the within-study effect sizes to a univariate meta-analysis model, as suggested by Fernández-Castilla, Declercq, et al. (2020). The dashed line indicates the overall average effect size (g = .195), and the diamond indicates the 95% confidence interval from the fitted PECHE-RVE model.

We also found substantial amounts of heterogeneity among the effect sizes, with $Q(204) = 1220.8$, $p < 0.001$, $I^2 = 88.9$, and a total heterogeneity $\sigma_T = 0.204$ SD. The total amount of heterogeneity

compresses variation from study and effect size levels, respectively, with between-study SD $\hat{\tau}$ = 0.067 95% *CI*[0.000, 0.173] and within-study SD $\hat{\omega}$ = 0.193 95% *CI*[0.167, 0.224]. Based on these estimates, Figure 13 illustrates the predictive distribution of future effect size estimates and the 67% PI of [-0.012, 0.401]. It can also be seen that 82% of future studies would be expected to yield effect sizes above zero. Although the 67% PI slightly overlaps zero, we do not interpret this as evidence of an ineffective intervention, given the lower cost of group-based approaches. By this, we mean that even if individual and group-based treatments are equally effective (e.g., with an effect size of 0.083), group-based interventions would allow the same number of individuals to be treated at a substantially lower cost, or enable treatment of many more people for the same cost as individual treatment as usual. Consequently, we interpret the predictive distribution to indicate that the majority of future studies will yield meaningful and efficacious effect sizes.

FIGURE 13 Predictive distribution for reintegrational outcomes



Note: Dash lines indicate the 67% prediction interval.

Overall average effects on mental health

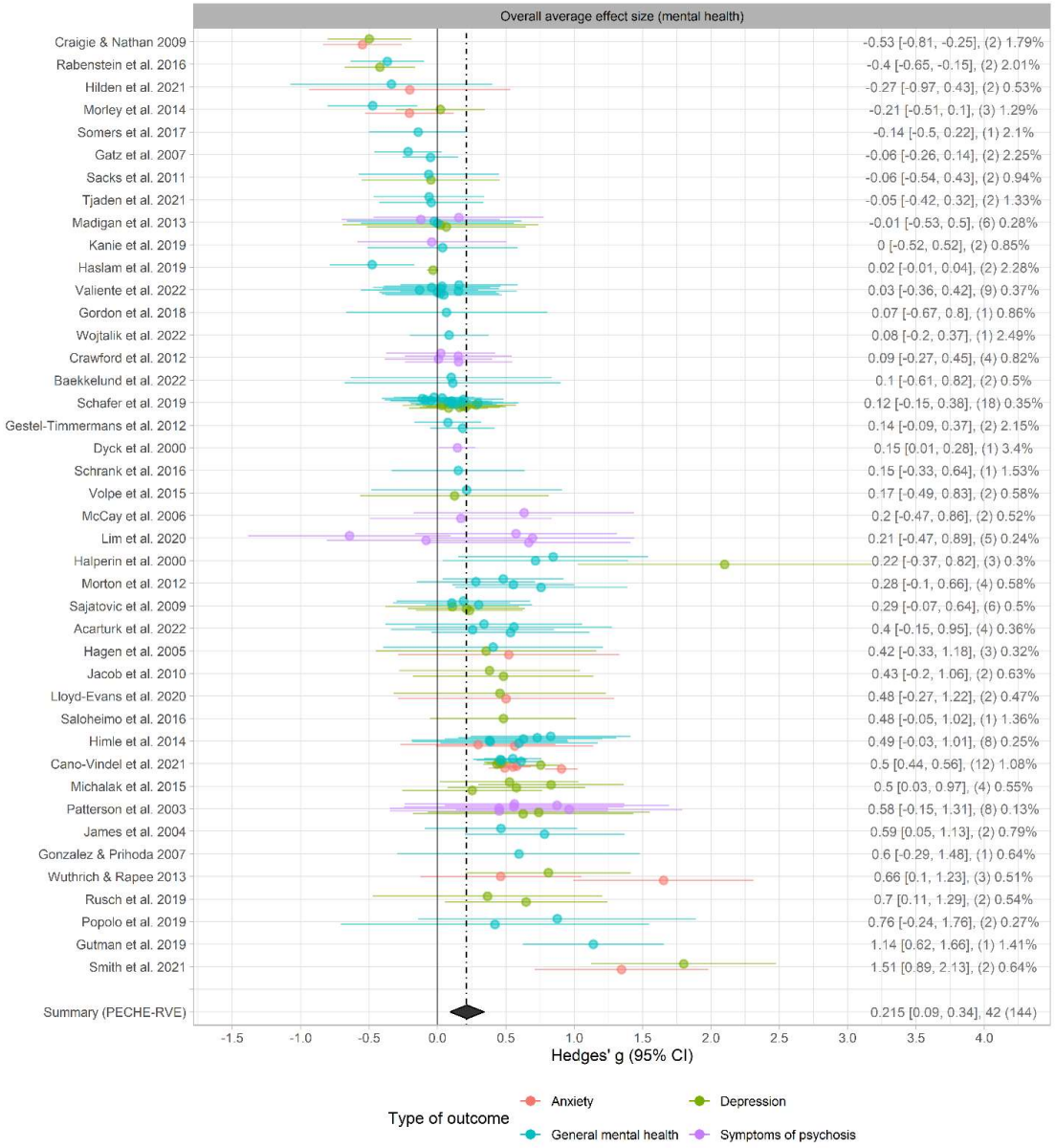
The overall average effect size estimate for mental health outcomes summarizes a total of 144 effect sizes and 42 studies published from 2000 to 2022. As shown in Figure 14, we found a positive, statistically significant overall standardized mean difference for mental health outcomes of 0.215 SD, $t(37.5) = 3.48$, $p = .001$, 95% CI[0.090, 0.340]. This means that on average, participants in group-based interventions score higher on mental health outcomes relative to participants exposed to individual treatment or a wait-list.

This effect is similar to the overall average effect size of social reintegration. Again, seen in light of the significant reduction in the incremental cost per treated individual, we also consider this effect size to be substantial. The overall average effect size of 0.215 amounts to 58.5 percent of the treatment participants having a better score than the average control participant, with the 95% confidence interval ranging from ~53.6 to ~63.29 percent of the treatment participants scoring better than the average control participant. On the individual level, when assuming constant

effects, this translates to an expectation that a typical participant from the control group would have had a percentile gain of 8.5% had they instead been exposed to a group-based intervention. For alternative interpretation metrics, see our main analysis codes accompanying the review.

When comparing the average growth in mental health in the control group, when receiving TAU, the treatment effect can be said to be 1.4 times larger than the typical improvement in TAU, which we found to be 0.157 SD, $t(31.25) = 2.65$, $p = 0.012$, 95% CI[0.036, 0.277].

FIGURE 14 Effect size forest plot with dependent effect sizes for mental health outcomes.

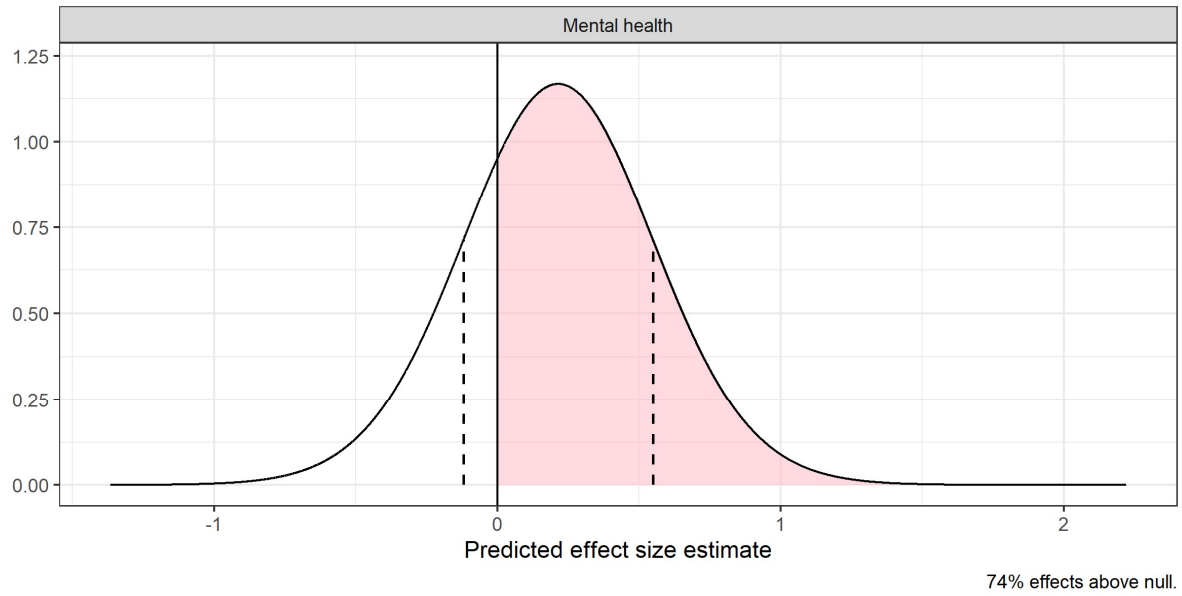


Note: The number of effect size estimates is in parentheses. The percentages indicate the weight given to each effect size within the given study. Studies are ordered by the study mean effect size obtained from fitting the within-study effect sizes to a univariate meta-analysis model, as suggested by Fernández-Castilla, Declercq, et al. (2020). The

dashed line indicates the overall average effect size (g = .195), and the and the diamond indicates the 95% confidence interval from the fitted PECHE-RVE.

Also, for mental health outcomes, we found a substantial amount of heterogeneity among the effect sizes, with $Q(143) = 174072.2$, $p < .001$, $I^2 = 99.94$, and a total heterogeneity $\sigma_T = 0.333$. As with social reintegrational outcomes, the total heterogeneity estimate comprises heterogeneity from the between-study and within-study levels, respectively, with between-study SD $\hat{\tau} = 0.298$ 95% *CI*[0.201, 0.425] and within-study SD $\hat{\omega} = 0.149$ 95% *CI*[0.121, 0.186]. Figure 15 depicts the predictive distribution of future effect size estimates and the 67% PI of [-0.120, 0.549]. In addition, it can be seen that 74 % of future studies would be expected to yield effect sizes above zero. Put another way, we would expect new studies to yield effects indicating slight reductions in mental health to large improvements. Based on the predictive distribution, we consider the effect of group-based intervention on participants' mental health to be more fragile relative to the effect on social reintegrational outcomes. In sum, a majority of individuals will likely experience a positive effect of group intervention on their mental health, while we can not exclude that a minor group of participants will possibly experience a small reduction in their mental health when exposed to group-based interventions.

FIGURE 15 Predictive distribution for mental health outcomes



Note: Dash lines indicate the 67% prediction interval.

That said, by using the PECMVE model, we found a substantial covariance between reintegrational and mental health outcomes, indicating that group-based interventions with larger impacts on reintegrational outcomes also show larger impacts on mental health outcomes. This suggests that when group-based interventions are effective, they tend to have substantial impacts on both the social reintegration of the participants and their mental health. See PRIMED Figures 97 and 98 for an overview of how reintegrational and mental health outcomes are distributed within and between studies.

Sensitivity analysis of main effects

We found the overall average effect size ($\bar{g}_t$) was largely agnostic to changes in the assumed correlation (ρ) among effect sizes within studies for reintegrational outcomes. The total SD remained generally constant across values of ρ, whereas the between-study SD (τ) and within-study SD (ω) decreased and increased, respectively, as a function of ρ (see Sensitivity Analysis Figure 1). In contrast to reintegrational outcomes, $\bar{g}_t$ slightly decreased as a function of ρ for mental health outcomes, ranging from 0.259 SD 95% CI[0.130, 0.388] when ρ = .0 to 0.209 SD 95% CI[0.084, 0.334] when ρ = .9. In addition, both the total SD and ω decreased when ρ increased, while the τ was relatively constant across all assumed values of ρ (see Sensitivity Analysis Figure 5).

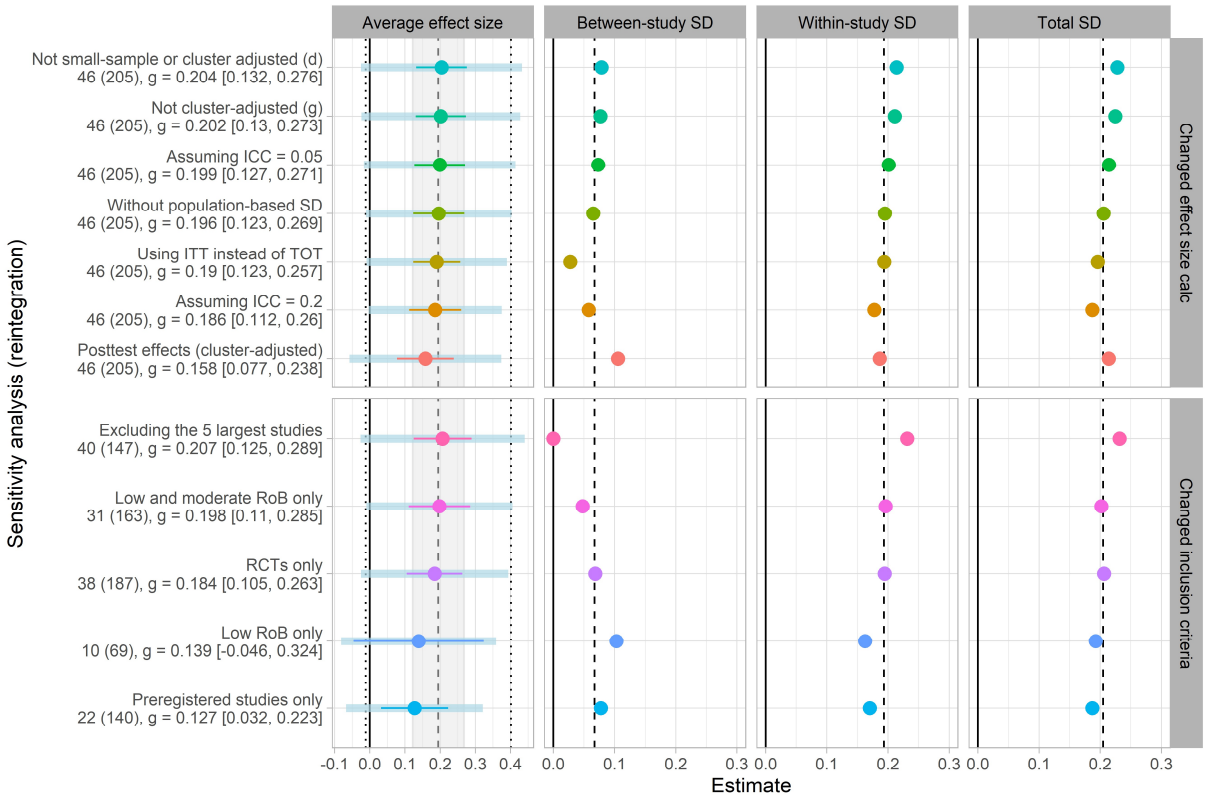
Overall, we found $\bar{g}_t$ to be insensitive to leaving any single study out of the analysis (see Sensitivity Analysis Figures 2 and 6). For reintegrational outcomes, $\bar{g}_t$ ranged from 0.173 SD 95% CI[0.102, 0.244] when leaving out Cano-Vindel et al. (2021) to 0.209 SD 95% CI[0.136, 0.282] when leaving out Crawford et al. (2012). For mental health outcome, $\bar{g}_t$ ranged from 0.182 SD 95% CI[0.066, 0.298] when leaving out Smith et al. (2021) to 0.234 SD 95% CI[0.114, 0.354] when excluding Craigie & Nathan (2009).

With one exception, all heterogeneity estimates were also agnostic to leaving any single study out. Specifically, τ dropped to null for reintegrational outcomes, when Cano-Vindel et al. (2021) was omitted. This finding suggests that the estimation of τ may be fragile, and that the primary moderators explaining variability in true effect sizes likely operate at levels that vary within, rather than between, studies.

Figures 16 and 17 illustrate the sensitivity analyses conducted by varying the effect size calculation assumptions and inclusion criteria for reintegrational and mental health outcomes, respectively. As shown in both figures, changes in these computational assumptions and inclusion criteria led to only minor variations in the overall average effect size compared to the main analyses. This pattern held for both reintegrational and mental health outcomes.

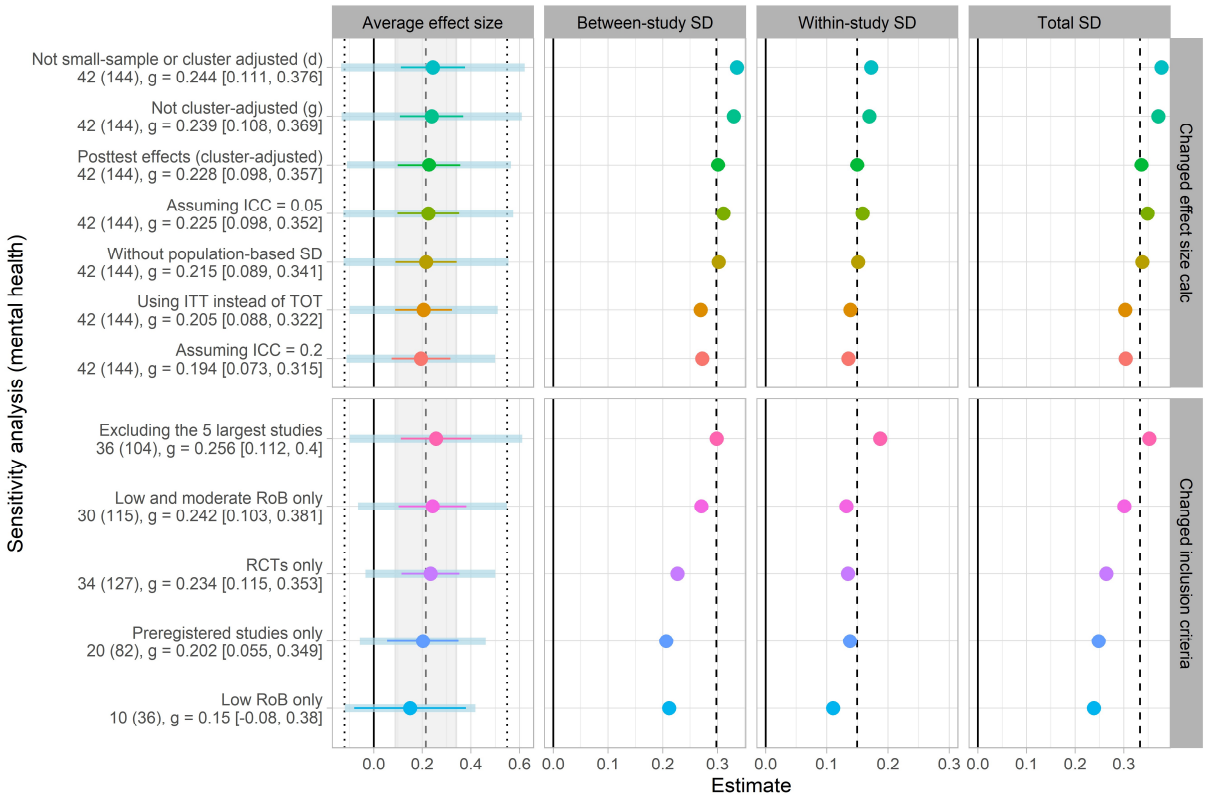
For both outcome types, $\bar{g}_t$ was no longer statistically different from zero when analyses were restricted to effect sizes rated as having a low risk of bias. However, this likely reflects reduced statistical power, as the magnitude of $\bar{g}_t$ remained close to the estimates from the main analyses. Notably, the 67% prediction interval remained highly similar to the original estimates across all sensitivity analyses, and heterogeneity estimates also differed only slightly.

FIGURE 16 Sensitivity analyses changing effect size calculation assumptions and inclusion criteria for reintegrational outcomes



Note: Dashed lines indicate the estimated values from the main analysis. The gray shading in the first column represents the width of the confidence interval from the main analysis, while the point lines indicate the reestimated confidence intervals. Dotted lines indicate the 67% prediction interval estimated from the main analysis, and the gray bands in the first column represent the re-estimated 67% prediction intervals.

FIGURE 17 Sensitivity analyses changing effect size calculation assumptions and inclusion criteria for mental health outcomes



Note: Dashed lines indicate the estimated values from the main analysis. The gray shading in the first column represents the width of the confidence interval from the main analysis, while the point lines indicate the reestimated confidence intervals. Dotted lines indicate the 67% prediction interval estimated from the main analysis, and the gray bands in the first column represent the re-estimated 67% prediction intervals.

Exploratory sensitivity analysis

Stanley et al. (2022) suggest that small-sample studies tend to exhibit greater heterogeneity than large-sample studies, and that this correlation between sample size and heterogeneity may violate the assumptions of random-effects models. Hence, as an exploratory analysis, we examined whether heterogeneity decreased when small studies were excluded (see Sensitivity Analysis Figures 3 and 7). However, we found no relationship between effective sample size and the total SD (σ_t), with σ_t = 0.166 for reintegrational outcomes and σ_t = 0.366 for mental outcomes, when estimated using only the five largest studies.

Moderator analysis

While the results summarized the mean difference between group-based interventions and no or individual treatments for reintegrational and mental health outcomes, respectively, it did not take into account the potential differences in effect sizes across moderating factors. As we detected substantial heterogeneity both within and between studies, this indicates that moderator factors operate at various levels, which in turn justifies the conduct of meta-regression analysis. As per protocol, we, therefore, conducted a comprehensive set of moderator analyses, informed by both theoretical considerations and methodological concerns. As in the previous section, we present the results separately for reintegrational and mental health outcomes.

Moderator effects of group-based interventions on social reintegration

Based on reintegrational outcomes, Tables 7 and 8 report the results for theoretically and methodologically informed categorical moderator analysis, while Table 9 presents analyses based on continuous moderators. All tables provide marginal subgroup effect sizes (i.e., without adding covariates to the model) as well as covariate-adjusted subgroup effect sizes. All moderator analyses related to social reintegration were based on 46 studies and 205 effect sizes. We generally found similar patterns between the moderator effects derived from unconditional and covariate-adjusted models. The only minor difference was that the covariate-adjusted subgroup effects were generally larger. Therefore, we will mainly comment on the unconditional moderator effects in this section.

The overall pattern that can be deduced from these tables is that most moderators neither explain differential effects nor the variation of true effect size at the study and effect size levels. Most of the subgroup analyses based on categorical moderators were dispersed relatively closely around the overall mean effect, varying from 0.12 to 0.42 SD. The only instances where we found substantially lower or higher subgroup effects relative to the overall average mean of 0.195 SD were when we subgroup effects were based on a few studies and effect sizes. For example, we found no effect of group-based intervention on participants' alcohol consumption, with the marginal average subgroup effect size of 0.02, 95% *CI*[-0.63, 0.66]. Yet, this estimate was based on four studies and five effect sizes only. As can also be seen from the confidence interval, we cannot exclude that this subgroup effect is statistically different from the main effect.

As mentioned above, we generally did not find that moderators were able to substantially explain variation among true effect sizes. Again, we only saw a discernible reduction in the total heterogeneity when subgroup effects were estimated with few studies and effect sizes, as was the case with the 'Other' category under outcome type, which was based on four studies and six effect sizes only. This yielded a total heterogeneity σ_T = .0.

Most subgroup effect sizes did not yield statistically significant estimates. However, we primarily considered this to be an issue caused by the obvious reduction of statistical power when estimating the effects with few studies and effect sizes. For many individual subgroup effects, we were hesitant to make hard inferences as the degrees of freedom were often low, either due to having few studies or being estimated from models with multiple covariates.

For most subgroup analyses, the average group means differed by only a small amount, and the differences between the subgroup dimensions were usually not statistically distinct from one another. Specifically, we did not find any statistical difference between different types of reintegrational outcomes, $F(6, 7.97) = 2.44$, $p = 0.120$. Across the types of reintegrational outcomes, the group effects ranged from 0.02, 95% *CI*[-0.63, 0.66] for loneliness to 0.42, 95% *CI*[0.31, 0.54] for self-esteem outcomes. We also did not find any difference in subgroup effects across different types of samples (samples with and without participants with schizophrenia) and interventions (group-based CBT vs. other types of interventions). Moreover, we did not find any subgroup differences between studies rated as having high versus not high risk of bias.

Across all moderator analyses with continuous moderators (i.e., the percent of males in the sample, the number of sessions per week, duration, and follow-up timing) did not predict differences among effects.

That said, we did observe some trends that aligned with usual methodological expectations (for similar examples in education, see Cheung & Slavin, 2016), but which should be interpreted cautiously, as we could not infer statistically significant differences between most of the subgroup effects.

First, as the only subgroup dimension, we found a statistical difference between preregistered and nonpreregistered studies, with $F(1, 29.58) = 7.68$, $p = 0.010$, with $\bar{g}_t = 0.13$, 95% *CI*[0.03, 0.22] for preregistered studies and $\bar{g}_t = 0.31$, 95% *CI*[0.21, 0.42] for nonpreregistered studies. Although the effect of group-based intervention is estimated to be lower in preregistered studies, it remains statistically significant with $p = 0.013$. Moreover, we still consider this effect size to be substantial, as this treatment effect is 1.56 times bigger than the usual improvement from TAU. One might think that this difference can be due to publication bias issues. However, we did not find clear evidence of this in our publication bias testing, see Table 14 and Figure 16 in the 'Publication bias assessment' section.

Second, as expected, we both found that ITT (intention-to-treat) effect sizes yielded a smaller average effect size relative to TOT (Treatment-On-the-Treated) and that effect sizes from RCT were lower than effect sizes from QES. Also, we found that effect sizes based on pure waitlist control groups, as commonly expected (Laws et al., 2022), yielded larger effects than effect sizes that were computed using TAU or the individual version of the group-based intervention control. In this regard, we, furthermore, found that the overall average effect for effect sizes with the individual version of the group-based intervention control group yielded a slightly negative effect size with $\bar{g}_t$ = -0.06 95% *CI*[-0.82, 0.70]. Although this effect was close to null, we due not consider the as hard evidence against the effectiveness of group-based interventions due to the relative cost-effectiveness of group-based interventions relative to individual treatments. Finally, we found that, on average, clinician-reported measures were lower than self-reported outcome measures. This might reflect the more objective nature of clinician-reported measures.

Moderator effects of group-based interventions on mental health

Based on the mental health effect sizes, Tables 10 and 11 report the results for theoretically and methodologically informed categorical moderator analysis, while Table 12 presents analyses based on continuous moderators. All tables provide marginal subgroup effect sizes (i.e., without adding covariates to the model) as well as covariate-adjusted subgroup effect sizes. All moderator analyses related to social reintegration were based on 42 studies and 144 effect sizes. We generally found similar patterns between the moderator effects derived from unconditional and covariate-adjusted models. The only minor difference was that the covariate-adjusted subgroup effects were generally larger. Therefore, we will mainly comment on the unconditional moderator effects in this section.

Overall, we found the same pattern for the moderator analyses based on mental health outcomes. Yet, we found minor deviation from the reintegrational results. Counter to the reintegrational results, we did not find any statistical difference between preregistered and nonpreregistered

studies, with $\bar{g}_t$ = 0.20 SD, 95% *CI*[0.06, 0.35] for preregistered studies and $\bar{g}_t$ = 0.25 SD, 95% *CI*[0.03, 0.48] for nonpreregistered studies. Furthermore, by contrast to the reintegrational analysis, we found that statistically different effect sizes ITT and TOT estimates, with $F(1, 32.38) = 7.08$, $p = 0.012$, and $\bar{g}_t$ = 0.05 SD, 95% *CI*[-0.08, 0.18] for ITT estimates and and $\bar{g}_t$ = 0.35 SD, 95% *CI*[0.15, 0.56] for TOT. However, we could not confirm these results in the covariate-adjusted model. Across the unconditional and conditional models, we also found that mental health effect sizes tended to statistically significantly increase as a function of the age of the participant, with $\beta = 0.025$, SE = 0.007 in the unconditional model.

Yet, we also identified some patterns that were not consistent with the results for reintegrational outcomes and ran counter to common methodological expectations. As we could not conclude that these trends differed statistically, the results should be interpreted with caution. First, clinician-reported measures of mental health generally yielded larger effects than self-reported measures. Second, we found that QESs tended to yield smaller effects on mental health than RCTs. Third and finally, we found that low or moderate risk of bias mental health estimates yielded smaller effects than high risk of bias estimates.

Sensitivity analysis of moderator effects

WILL BE ADDED WHEN HAVING ASSESS TO UCLOUD

Publication bias assessment

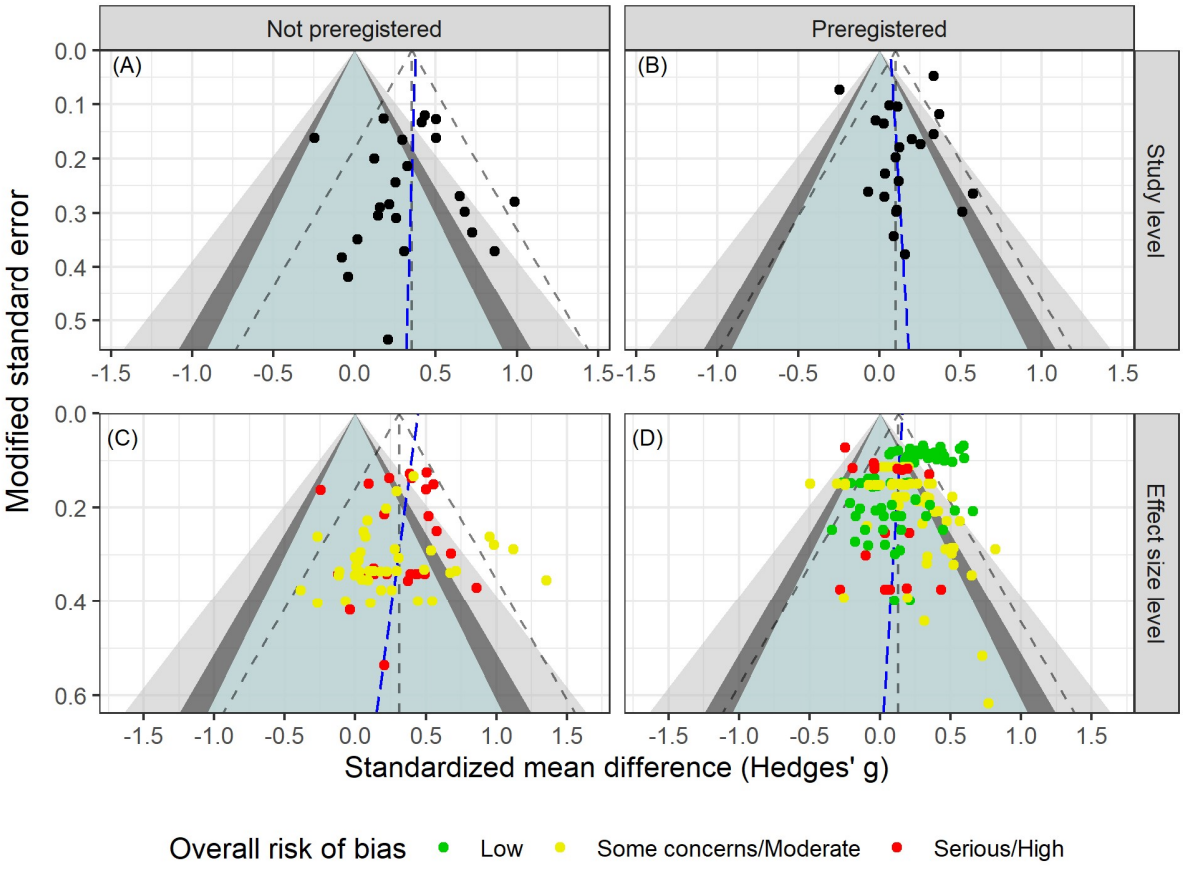
All the results of our publication bias testing are presented in Figures 16-17 and Tables 13-17, below. As can be seen from the results, we generally found no evidence of publication or small study bias across a wide range of publication bias tests. When correcting for publication bias, we more often found the overall average effect tended to increase. This counted for subgroup means as well, meaning that we did not find any outcome-specific indications of publication bias.

Figures 16 and 17 illustrate contour-enhanced funnel plots across preregistration status and the study and effect size levels for reintegrational and mental health outcomes, respectively. The dashed lines show the slope of regressing effect sizes on modified standard errors. We generally found no association between modified standard errors and effect sizes at the study level, indicated by the almost vertical slopes. However, counter to typically expected selective reporting processes, we found that larger standard errors were associated with smaller effect size estimates (though not statistically significant) at the effect size level for reintegrational outcomes (see Figures 18C and 18D). As the only instance, we found that larger standard errors predicted larger effect sizes to a larger extent for non-preregistered outcomes for the mental health outcome, indicated by the positive increasing slope in Figure 19C. Yet, this slope was not statistically significant, and we did not find consistent evidence of publication bias for this subgroup of outcomes. As can be seen from Table 16, the overall average group mean declined to $\bar{g}_t$ = 0.02, 95% CI[-0.18, 0.22] for nonpregistered mental health outcomes, when assuming that all positive, statistically significant results were false positives. Conversely, the three-parameter selection model estimates indicated selection *in favor* of non-affirmative findings, with λ_1= 2.93, 95% CI[1.00, 12.84]. When correcting for this selection process $\bar{g}_t$ = 0.44, 95% CI[0.11, 0.81].

That said, we found hard evidence against publication bias. For example, when correcting non-preregistered studies for publication bias (HYEMA) and removing[13] all affirmative effect sizes (worst-case scenario), we still found the overall average effect size to be statistically different from zero for reintegrational outcomes, with $\bar{g}_t = 0.19$, 95% CI[0.09, 0.20] and $\bar{g}_t = 0.078$, 95% CI[0.03, 0.13], respectively. For mental health outcomes, we found $\bar{g}_t = 0.29$, 95% CI[0.13, 0.44] and $\bar{g}_t = 0.069$, 95% CI[-0.03, 0.17], indicating that group-based intervention is at least as effective on mental health as no or individual treatments. Furthermore, across all used selection models, the point estimates stayed statistically different from zero and substantial in size, while all selection parameters were above one or close to one for non-significant effect sizes.

To recap, we believe these results clearly suggest that publication bias (entirely excluded studies), *p*-hacking (not reported effect size within studies), small-study effects, and related issues are not a major concern in this field of studies. This conclusion might not be a surprise, as close to 50% of the studies drew on preregistration.

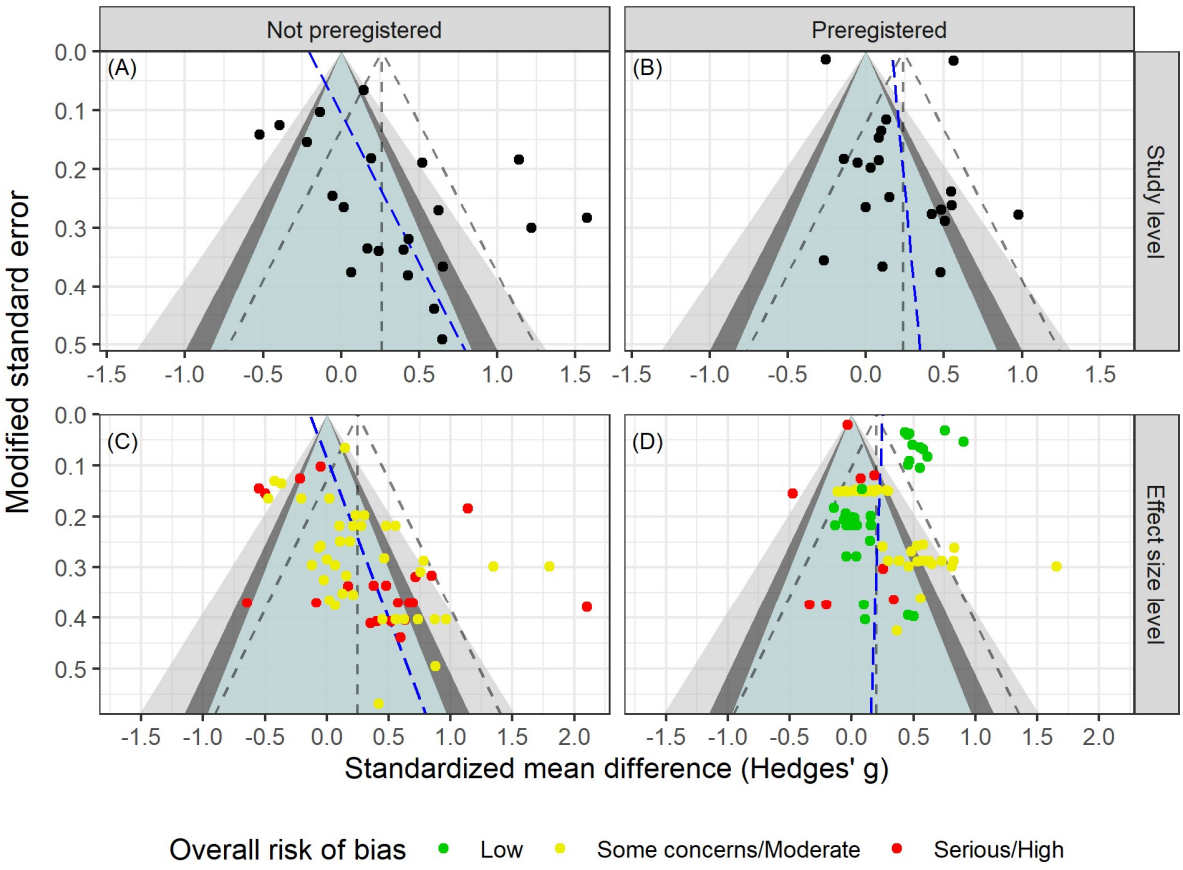
FIGURE 18 Funnel plot for reintegration effect sizes (primary outcome)



Note: Add description of figure.

[13] Of note, only nonpreregistered studies were fully removed in the worst-case meta-analysis of reintegrational outcomes.

FIGURE 19 Funnel plot for mental health effect sizes (secondary outcome)



Note: We did not use the CHE-ISCW-RVE test for the preregistered effect size funnel plots, as it seemed to heavily overestimate the relationship between standard errors and effect sizes. Instead, we used a regression test similar to the one developed by Rodgers and Pustejovsky (2021).

Discussion

Summary of main results

Results from the two primary meta-analyses (using all social reintegration and all mental health outcomes respectively) were both statistically significant and both favored group-based interventions over the control interventions, which consist of individually delivered interventions. All average effect sizes within the meta-analyses favored group-based interventions over the control interventions. However, not all meta-analyses performed reached statistical significance. When outcomes were divided further, results revealed significant results favoring group based interventions on outcomes: Subjective Wellbeing and Quality of Life, Hope, Empowerment and Self-efficacy and Depression. Results of the meta-analyses using the outcomes: Social functioning (impairment), Alcohol and Substance use, General mental health, Anxiety and Psychotic Symptoms analyses were not statistically significant and thus the findings

do not support the notion that group-based interventions outperform individually delivered interventions on these outcomes.

Overall completeness and applicability of evidence

The studies reported a very high number of usable outcomes covering a broad range of outcomes measuring indicators of social marginalization. The outcomes used to measure mental health were all well-known clinical rating scales/interviews or self-assessment tools. Thus, we believe the evidence on outcomes of both social marginalization and mental health are trustworthy.

However only three studies reported on an intervention in which the control condition consisted of the same intervention delivered individually. The most common control/comparison condition consisted of usual clinical care, which was mostly less intense, less well implemented and organized than the group interventions. The most common type of control intervention consisted of access to usual mental health care. Therefore, some of the benefits of the group-based interventions identified in this review may be due to the increased amount time and frequency of face-to-face interactions between therapists/case workers and participants.

In order to isolate the effects of group interventions per se future studies should compare the same intervention delivered individually and in a group.

Quality of the evidence

Generally, the quality of the included studies was good and only five studies were excluded from the data-synthesis because of to the risk of bias assessment.

Potential biases in the review process

We performed a comprehensive electronic database search, combined with grey literature searching, and hand searching of key journals. All citations were screened by two independent screeners from the review team, and one review author (NTD) assessed all included studies against inclusion criteria.

We believe that all the publicly available studies on the effects of group-based interventions for marginalized adults suffering from both mental illness and social problems within the OECD countries published after 2000 were identified during the review process. However, references 35 were not obtained in full text.

- A selective sample of mental health outcomes, as we primarily included studies reporting mental health outcomes if they also reported results related to social reintegration.

Agreements and disagreements with other studies or reviews

The effects of psychiatric interventions aimed at reducing symptoms for patients with specific diagnoses have been extensively explored in a large number of reviews and meta-analyses, but only a much smaller number of existing reviews have explored the effects of group interventions on a broader range of outcomes. The present review contributes to the knowledge base by exploring the efficacy of group interventions on a more broad range of outcomes, than what is seen in the existing reviews.

Findings from our review thus expands the knowledge base, but agrees with previous reviews in which positive effects of group-based interventions for adults with both mental illness and indicators of social marginalization have been identified.

Authors' conclusions

Implications for practice and policy

The number of people with mental illness is growing in the Western world, which force policy makers to reconsider how they meet the increased demands (Bloom et al., 2011). In this review we have focused on group-based interventions in out-patient community centered care, because hospitals beds are being replaced by community care at increasing rates. As theorized in the introduction, the present review was motivated by the fact that group-based interventions generally reduce the cost of interventions. As previously noted, group-based interventions are an appealing solution because the cost of group-based interventions can be less than half the cost of individual therapy (Ruesch et al., 2015).

Findings based on the series of meta-analyses suggest that for adults who suffer from both mental illness and face indicators of social marginalization, group-based interventions are a promising type of intervention. Findings suggest that in addition to reducing symptoms of mental health, group-interventions are also more effective than individually delivered control interventions at reducing social marginalization.

Our findings suggest that on measures of all types of mental health symptoms and all social reintegration outcomes, group-based interventions have larger average effects than usual care if delivered as an individual intervention. All though not all meta-analyses were statistically significant all average effect sizes favoured group-based interventions indicating that there are, no adverse effects of group-based interventions compared with individually delivered control interventions.

In a policy context it is, however, important to emphasize that although none of the included studies reported average deterioration effects, this does not mean that all individual participants benefitted and thus it is still possible, that some participants may have negative experiences with group-based interventions as hypothesized in the introduction. In addition, it is important to highlight that the most common control/comparison condition/intervention did not consist of individual therapy (see Ruesch et al., 2015) but rather the less expensive usual community outpatient-based care (usual mental health care). In fact the usual mental health care is mostly described as less intense, less well implemented and organized than the group interventions in the reviewed studies.

As part of our review we also specified and sub-grouped the group-interventions in terms of focus/topic. We identified 14 different group-intervention types. We label the four most used intervention 'Group based Cognitive Behavioral Therapy' (11), Group psychoeducation & Social skill training (9), Illness Management (8), Cognitive-Behavioral Social Skills Training (8). These prototypes of interventions may have practical implications, because policy makers may use our categorization to get a better insight into the different group intervention types that exist for this specific target group.

Implications for research

Previous research suggests that interpersonal and support factors are one of the few changeable predictors in the course of mental illness (see e.g. Keitner et al., 1992). Consequently, a large body of studies have explored the efficacy of psychiatric group interventions targeting a specific mental health disorder. However, the primary focus has been on symptom reduction as the only outcome. In the protocol (Dalgaard et al., 2022), we identified only six existing reviews, which include outcomes other than symptom reduction, and thus the evidence regarding the efficacy of group interventions on outcomes beyond symptom reduction was far from unequivocal.

Our review adds to the existing body of reviews by showing a rather consistent result that favor group-based interventions over the control interventions. On average we find that the effects of group-based interventions is beneficial compared with individual control interventions for adults suffering from both mental illness and social marginalization on outcomes measuring social reintegration *and* mental health.

Findings from the present review suggest that on average the effects of group-based interventions is beneficial compared with individual control interventions for adults suffering from both mental illness and social marginalization on outcomes measuring social integration and mental health. However, more research is needed in order to explore what works for whom within this vulnerable population group. Furthermore, future studies should explore the effects of group interventions beyond the end of the intervention period, as the included studies in the present review did not allow us to conduct meta-analysis on the long-term effects of interventions on any outcome.

Only three studies reported on an intervention in which the control condition consisted of the same intervention delivered individually. The most common control/comparison condition consisted of usual clinical care, which was mostly less intense, less well implemented and organized than the group interventions. The most common type of control intervention consisted of access to usual mental health care. Therefore, some of the benefits of the group-based interventions identified in this review may be due to the increased amount of time and frequency of face-to-face interactions between therapists/case workers and participants.

In order to isolate the effects of group interventions per se future studies should compare the same intervention delivered individually and in a group.

Most of the included studies reported on a heterogeneous group of participants with different diagnoses and co-morbid conditions, who also faced multiple social or personal problems. In order to refine the knowledge on efficacy of group-interventions more, studies focused on

participants facing more narrowly defined problems or with similar diagnoses would improve the knowledge base.

None of the included studies reported average deterioration effects, suggesting that both group-based interventions and the control interventions had beneficial average effects for the participants. However, this does not mean that all individual participants benefitted and thus it is still possible, that some participants may have negative experiences with group-based interventions as hypothesized in the introduction. Future research should explore this issue by conducting qualitative research into the experiences of group-based interventions from the perspective of participants.

Acknowledgements

The authors would like to thank research assistants Rune Klitgård and Julie Mulla Reich

Contributions of authors

- Content: Nina Thorup Dalgaard, Jakob Kaarup Jensen, Maya Christiane Flensborg Jensen, Mikkel Helding Vembye,
- Systematic review methods: Nina Thorup Dalgaard, Jacob Kaarup Jensen, Jasmin Sami Adada, Mikkel Helding Vembye,
- Statistical analysis: Jakob Kaarup Jensen, Jasmin Sami Adada, Mikkel Helding Vembye,
- Information retrieval: Elizabeth Bengtsen

Declarations of interest

The authors have no conflicts of interests

Plans for updating this review

If funding is available, the first and last author will update the review.

Differences between protocol and review

In a few instances, we have deviated from our protocol. The main reason for deviation from the protocol was that new and better-performing methods were developed since we submitted the protocol. This included the methods developed by Chen and Pustejovsky (2025), Fitzgerald and Tipton (2024), Pustejovsky, Citkowitz et al. (2025), Pustejovsky, Zhang et al. (2025), van Aert (2025), and Wu, Duan et al. (2025). As all of these methods (or advice) show more appropriate statistical performance than the methods we originally described in the protocol, we found it reasonable to implement these methods. Of particular note, the majority of these method

developments were developed by the statisticians who had developed most of the methods we describe in the protocol. Thus, we felt confident in updating our methods to keep up with the state-of-the-art. Moreover, we provide open data, allowing others to replicate our work but also to conduct the original suggested analyses, if desired.

A more questionable method deviation from the protocol, however, was that we used mean imputation to handle the missingness of focal moderating factors. Meanwhile, we only used this approach as we only had one missing study for two moderators. Therefore, we considered the advances of mean imputation to outweigh the downside of this approach for the following reasons. Firstly, using more advanced methods to handle missing values would unnecessarily complicate Wald test estimation (see Vembye, Weiss et al., 2024 for a discussion of this issue). Secondly, using list-wise deletion would make us lose important information on other moderating factors that were fully reported within the given study. Therefore, we found the mean imputation to be an acceptable compromise between these two alternative strategies for handling missing values.

In the protocol, we originally wrote that we would add a critical risk of bias judgment to the RoB2 tools. Yet, since these tools clearly state that reviewers are not allowed to modify or extend the tools, we did not follow this practice. In addition, we wrote that we would include Eklund et al. (2017). However, after scrutinizing this study, we did not include it as it contained group-based interventions in both the treatment and control groups. Thus, it fell outside the inclusion criteria of the review.

Finally, with our improved understanding of the literature, we now distinguish more clearly between primary and secondary outcomes than was originally described in the protocol.

Tables

Characteristics of included studies in meta-analysis

See appendix descriptive table

Characteristics of excluded studies in meta-analysis

See appendix Double-group based interventions

Summary of findings tables

Additional tables

1 Table example: Included studies by document type

Footnotes

References

References to included studies

Acarturk C, Uygun E, Ilkkursun Z, Yurtbakan T, Kurt G, Adam-Troian J, Senay I, Bryant R, Cuijpers P, Kiselev N, McDaid D, Morina N, Nisanci Z, Park AL, Sijbrandij M, Ventevogel P, & Fuhr DC. (2022). Group problem management plus (PM plus) to decrease psychological distress among Syrian refugees in Turkey: A pilot randomised controlled trial. *BMC PSYCHIATRY*, *22*(1). https://doi.org/10.1186/s12888-021-03645-w

Ball SA, Cobb-Richardson P, Connolly AJ, Bujosa CT, & O'Neall TW. (2005). Substance abuse and personality disorders in homeless drop-in center clients: Symptom severity and psychotherapy retention in a randomized clinical trial. *COMPREHENSIVE PSYCHIATRY*, *46*(5), 371–379. https://doi.org/10.1016/j.comppsych.2004.11.003

Barbic S, Krupa T, & Armstrong I. (2009). A Randomized Controlled Trial of the Effectiveness of a Modified Recovery Workbook Program: Preliminary Findings. *PSYCHIATRIC SERVICES*, *60*(4), 491–497. https://doi.org/10.1176/ps.2009.60.4.491

Beames L, Strodl E, Dark F, Wilson J, Sheridan J, & Kerswell N. (2020). A Feasibility Study of the Translation of Cognitive Behaviour Therapy for Psychosis into an Australian Adult Mental Health Clinical Setting. *BEHAVIOUR CHANGE*, *37*(1), 22–32. https://doi.org/10.1017/bec.2020.1

Bond Gary R & McDonel Elizabeth C. (u.å.). Assertive community treatment and reference groups: An. *Psychosocial Rehabilitation Journal*, *15*(2), 31.

Bond GR, Kim SJ, Becker DR, Swanson SJ, Drake RE, Krzos IM, Fraser VV, O'Neill S, & Frounfelker RL. (2015). A Controlled Trial of Supported Employment for People With Severe Mental Illness and Justice Involvement. *PSYCHIATRIC SERVICES*, *66*(10), 1027–1034. https://doi.org/10.1176/appi.ps.201400510

Bozzer M, Samsom D, & Anson J. (1999). An evaluation of a community-based vocational rehabilitation program for adults with psychiatric disabilities. *Canadian journal of community mental health = Revue canadienne de sante mentale communautaire*, *18*(1), 165–179.

Burnam M Audrey, Morton Sally C, McGlynn Elizabeth A, Petersen Laura P, Stecher Brian M, Hayes Charles, & Vaccaro Jerome V. (1995). An Experimental Evaluation of Residential and Nonresidential Treatment for Dually Diagnosed Homeless Adults. *Journal of Addictive Diseases*, *14*(4), 111–134.

Bækkelund Harald, Ulvenes Pål, Boon-Langelaan Suzette, & Arnevik Espen Ajo. (2022). Group treatment for complex dissociative disorders: A randomized clinical trial. *BMC psychiatry*, *22*(1), 338.

Cano-Vindel A, Munoz-Navarro R, Moriana JA, Ruiz-Rodriguez P, Medrano LA, & Gonzalez-Blanch C. (2021). Transdiagnostic group cognitive behavioural therapy for emotional disorders in primary care: The results of the PsicAP randomized controlled trial. *Psychological medicine*, 1-13. https://doi.org/10.1017/S0033291720005498

Craigie Mark A & Nathan Paula. (2009). A Nonrandomized Effectiveness Comparison of Broad-Spectrum Group CBT to Individual CBT for Depressed Outpatients in a Community Mental Health Setting. *Behavior Therapy*, *40*(3), 302–314.

Crawford MJ, Killaspy H, Barnes TRE, Barrett B, Byford S, Clayton K, Dinsmore J, Floyd S, Hoadley A, Johnson T, & et al. (2012). Group art therapy as an adjunctive treatment for people with schizophrenia: Multicentre pragmatic randomised trial. *BMJ (online)*, *344*(7847). https://doi.org/10.1136/bmj.e846

Daniels L & Roll D. (u.å.). Group treatment of social impairment in people with mental illness. *Psychiatric Rehabilitation Journal*, *21*(3), 273–278. https://doi.org/10.1037/h0095302

Druss Benjamin G, Singh Manasvini, von Esenwein Silke A, Glick Gretl E, Tapscott Stephanie, Tucker Sherry Jenkins, Lally Cathy A, & Sterling Evelina W. (2018). Peer-Led Self-Management of General Medical Conditions for Patients With Serious Mental Illnesses: A Randomized Trial. *Psychiatric Services*, *69*(5), 529–535. https://doi.org/10.1176/appi.ps.201700352

Druss BG, Zhao L, von Esenwein SA, Bona JR, Fricks L, Jenkins-Tucker S, Sterling E, Diclemente R, Lorig K, Druss Benjamin G, Zhao Liping, von Esenwein Silke A, Bona Joseph R, Fricks Larry, Jenkins-Tucker Sherry, Sterling Evelina, Diclemente Ralph, & Lorig Kate. (2010). The Health and Recovery Peer (HARP) Program: A peer-led intervention to improve medical self-management for persons with serious mental illness. *Schizophrenia Research*, *118*(1–3), 264–270. https://doi.org/10.1016/j.schres.2010.01.026

Dyck DG, Short RA, Hendryx MS, Norell D, Myers M, Patterson T, McDonell MG, Voss WD, & McFarlane WR. (2000). Management of negative symptoms among patients with schizophrenia attending multiple-family groups. *Psychiatric Services*, *51*(4), 513–519. https://doi.org/10.1176/appi.ps.51.4.513

Gatz Margaret, Brown Vivian, Hennigan Karen, Rechberger Elke, O'Keefe Maura, Rose Tara, & Bjelajac Paula. (2007). Effectiveness of an integrated, trauma-informed approach to treating women with co-occurring disorders and histories of trauma: The Los Angeles site experience. *Journal of Community Psychology*, *35*(7), 863–878.

Godoy Izquierdo, Débora, Vázquez Pérez, María Luisa, Lara Moreno, Raquel, Godoy García, & Juan F. (2021). Training coping skills and coping with stress self-efficacy for successful

daily functioning and improved clinical status in patients with psychosis: A randomized controlled pilot study. *Science Progress*, 1–22.

Gonzalez Jodi M & Prihoda Thomas J. (2007). A Case Study of Psychodynamic Group Psychotherapy for Bipolar Disorder. *American Journal of Psychotherapy (Association for the Advancement of Psychotherapy)*, *61*(4), 405–422.

Gordon A, Davis PJ, Patterson S, P ng CA, Scott JG, Salter K, & Connell M. (2018). A randomized waitlist control community study of Social Cognition and Interaction Training for people with schizophrenia. *BRITISH JOURNAL OF CLINICAL PSYCHOLOGY*, *57*(1), 116–130. https://doi.org/10.1111/bjc.12161

Gutman Sharon A, Barnett Sara, Fischman Lauren, Halpern Jamie, Hester Genni, Kerrisk Colleen, McLaughlin Travis, Ozel Ezgi, & Wang Haisu. (u.å.). Pilot Effectiveness of a Stress Management Program for Sheltered Homeless Adults With Mental Illness: A Two-Group Controlled Study. *Occupational Therapy in Mental Health*, *35*(1), 59–71.

Hagen Roger, Nordahl Hans M, Kristiansen Lena, & Morken Gunnar. (2005). A Randomized Trial of Cognitive Group Therapy vs Waiting List for Patients with Co-Morbid Psychiatric Disorders: Effect of Cognitive Group Therapy after Treatment and Six and Twelve Months Follow-Up. *Behavioural and Cognitive Psychotherapy*, *33*(1), 33–44. https://doi.org/10.1017/S1352465804001754

Halperin Stephen, Nathan Paula, Drummond Peter, & Castle David. (2000). A cognitive-behavioural, group-based intervention for social anxiety in schizophrenia. *Australian and New Zealand Journal of Psychiatry*, *34*(5), 809–813. https://doi.org/10.1046/j.1440-1614.2000.00820.x

Haslam Catherine, Cruwys Tegan, Chang Melissa X-L, Bentley Sarah V, Haslam S Alexander, Dingle Genevieve A, & Jetten Jolanda. (2019). GROUPS 4 HEALTH Reduces Loneliness and Social Anxiety in Adults With Psychological Distress: Findings From a Randomized Controlled Trial. *Journal of consulting and clinical psychology*, *87*(9), 787. https://doi.org/10.1037/ccp0000427

Hilden H M, Rosenstrom T, Karila I, Elokorpi A, Torpo M, Arajarvi R, & Isometsa E. (2021). Effectiveness of brief schema group therapy for borderline personality disorder symptoms: A randomized pilot study. *Nordic Journal of Psychiatry*, *75*(3), 176–185. https://dx.doi.org/10.1080/08039488.2020.1826050

Himle Joseph A, Bybee Deborah, Steinberger Edward, Laviolette Wayne T, Weaver Addie, Vlnka Sarah, Golenberg Zipora, Levine Debra Siegel, Heimberg Richard G, & O'Donnell Lisa A. (2014). Work-related CBT versus vocational services as usual for unemployed persons with social anxiety disorder: A randomized controlled pilot trial. *Behaviour Research & Therapy*, *63*, 169–176.

Jacob Gitta A, Gabriel Susanne, Roepke Stefan, Stoffers Jutta M, Lieb Klaus, & Hammers Claas-Hinrich. (2010). Group therapy module to enhance self-esteem in patients with borderline personality disorder: A pilot study. *International Journal of Group Psychotherapy*, *60*(3), 373–387. https://doi.org/10.1521/ijgp.2010.60.3.373

James W, Preston N J, Koh G, Spencer C, Kisely S R, & Castle D J. (2004). A group intervention which assists patients with dual diagnosis reduce their drug use: A randomized controlled trial. *Psychological Medicine*, *34*(6), 983–990.

Kallestad Håvard, Wullum Elin, Scott Jan, Stiles Tore C, & Morken Gunnar. (2016). The long-term outcomes of an effectiveness trial of group versus individual psychoeducation for

bipolar disorders. *Journal of Affective Disorders*, *202*, 32–38. https://doi.org/10.1016/j.jad.2016.05.043

Kanie A, Kikuchi A, Haga D, Tanaka Y, Ishida A, Yorozuya Y, Matsuda Y, Morimoto T, Fukuoka T, Takazawa S, Hagiya K, Ozawa S, Iwata K, Ikebuchi E, Nemoto T, Roberts DL, & Nakagome K. (2019). The Feasibility and Efficacy of Social Cognition and Interaction Training for Outpatients With Schizophrenia in Japan: A Multicenter Randomized Clinical Trial. *FRONTIERS IN PSYCHIATRY*, *10*. https://doi.org/10.3389/fpsyt.2019.00589

Lim JE, Kwon YJ, Jung SY, Park K, Lee W, Lee SH, P Horan W, & Choi KH. (2020). Benefits of social cognitive skills training within routine community mental health services: Evidence from a non-randomized parallel controlled study. *Asian journal of psychiatry*, *54*, 102314. https://doi.org/10.1016/j.ajp.2020.102314

Lloyd-Evans Brynmor, Frerichs Johanna, Stefanidou Theodora, Bone Jessica, Pinfold Vanessa, Lewis Glyn, Billings Jo, Barber Nick, Chhapia Anjie, Chipp Beverley, Henderson Rob, Shah Prisha, Shorten Anna, Giorgalli Maria, Terhune James, Jones Rebecca, & Johnson Sonia. (2020). The Community Navigator Study: Results from a feasibility randomised controlled trial of a programme to reduce loneliness for people with complex anxiety or depression. *PLoS ONE*, *15*(5), 1–18.

Madigan K, Brennan D, Lawlor E, Turner N, Kinsella A, O'Connor JJ, Russell V, Waddington JL, O'Callaghan E, Madigan Kevin, Brennan Daria, Lawlor Elizabeth, Turner Niall, Kinsella Anthony, O'Connor John J, Russell Vincent, Waddington John L, & O'Callaghan Eadbhard. (2013). A multi-center, randomized controlled trial of a group psychological intervention for psychosis with comorbid cannabis dependence over the early course of

illness. *Schizophrenia Research*, *143*(1), 138–142. https://doi.org/10.1016/j.schres.2012.10.018

McCay Elizabeth A, Beanlands Heather, Zipursky Robert, Roy Paul, Leszcz Molyn, Landeen Janet, Ryan Kathy, Conrad Gretchen, Romano Donna, Francis Daphene, Hunt Jennifer, Constantini Lucia, & Chan Eugene. (2007). A randomised controlled trial of a group intervention to reduce engulfment and self-stigmatisation in first episode schizophrenia. *Australian e-journal for the advancement of mental health*, *6*(3). https://www.proquest.com/scholarly-journals/randomised-controlled-trial-group-intervention/docview/37000639/se-2?accountid=27042

Michalak J, Schultze M, Heidenreich T, & Schramm E. (2015). A randomized controlled trial on the efficacy of mindfulness-based cognitive therapy and a group version of cognitive behavioral analysis system of psychotherapy for chronically depressed patients. *Journal of consulting and clinical psychology*, *83*(5), 951-963. https://doi.org/10.1037/ccp0000042

Morley Kirsten C, Sitharthan Gomathi, Haber Paul S, Tucker Peter, & Sitharthan Thiagarajan. (2014). The efficacy of an opportunistic cognitive behavioral intervention package (OCB) on substance use and comorbid suicide risk: A multisite randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *82*(1), 130–140. https://doi.org/10.1037/a0035310

Morton J, Snowdon S, Gopold M, & Guymer E. (2012). Acceptance and commitment therapy group treatment for symptoms of borderline personality disorder: A public sector pilot study. *Cognitive and behavioral practice*, *19*(4), 527-544. https://doi.org/10.1016/j.cbpra.2012.03.005

Munroe-Blum H & Marziali E. (1995). A controlled trial of short-term group treatment for borderline personality disorder. *Journal of personality disorders*, *9*(3), 190-198.

Patterson Michelle L, Moniruzzaman Akm, & Somers Julian M. (2014). Community Participation and Belonging Among Formerly Homeless Adults with Mental Illness After 12 months of Housing First in Vancouver, British Columbia: A Randomized Controlled Trial. *Community mental health journal*, *50*(5), 604–611. https://doi.org/10.1007/s10597-013-9672-9

Patterson TL, McKibbin C, Taylor M, Goldman S, Davila-Fraga W, Bucardo J, & Jeste DV. (2003). Functional adaptation skills training (FAST): A pilot psychosocial intervention study in middle-aged and older patients with chronic psychotic disorders. *American journal of geriatric psychiatry*, *11*(1), 17-23.

Popolo R, MacBeth A, Canfora F, Rebecchi D, Toselli C, Salvatore G, & Dimaggio G. (2019). Metacognitive Interpersonal Therapy in group (MIT-G) for young adults with personality disorders: A pilot randomized controlled trial. *Psychology and psychotherapy*, *92*(3), 342-358. https://doi.org/10.1111/papt.12182

Rabenstein R, Pintzinger N, Knogler V, Kirnbauer V, Lenz G, & Schosser A. (2015). Effectiveness of a cognitive-behavioral rehabilitation day clinic program—A waiting list controlled trial. *Verhaltenstherapie*, *25*, 192-200.

Rajji Tarek K, Mamo David C, Holden Jason, Granholm Eric, & Mulsant Benoit H. (2022). Cognitive-Behavioral Social Skills Training for patients with late-life schizophrenia and the moderating effect of executive dysfunction. *Schizophrenia Research*, *239*, 160–167. https://doi.org/10.1016/j.schres.2021.11.051

Rosenblum Andrew, Matusow Harlan, Fong Chunki, Vogel Howard, Uttaro Thomas, Moore Thomas L, & Magura Stephen. (2014). Efficacy of dual focus mutual aid for persons with

mental illness and substance misuse. *Drug & Alcohol Dependence*, *135*(1), 78–87. https://doi.org/10.1016/j.drugalcdep.2013.11.012

Russinova Zlatka, Gidugu Vasudha, Bloch Philippe, Restrepo-Toro Maria, & Rogers E Sally. (2018). Empowering Individuals With Psychiatric Disabilities to Work: Results of a Randomized Trial. *Psychiatric Rehabilitation Journal*, *41*(3), 196–207.

Rüsch Nicolas, Staiger Tobias, Waldmann Tamara, Dekoj Marie Christine, Brosch Thorsten, Gabriel Lisa, Bahemann Andreas, Oexle Nathalie, Klein Thomas, Nehf Luise, & Becker Thomas. (2019). Efficacy of a peer-led group program for unemployed people with mental health problems: Pilot randomized controlled trial. *The International journal of social psychiatry*, *65*(4), 333–337. https://doi.org/10.1177/0020764019846171

Sacks Stanley, McKendrick Karen, Vazan Peter, Sacks JoAnn Y, & Clelanda Charles M. (2011). Modified therapeutic community aftercare for clients triply diagnosed with HIV/AIDS and co-occurring mental and substance use disorders. *AIDS Care*, *23*(12), 1676. https://doi.org/10.1080/09540121.2011.582075

Sajatovic M, Davies MA, Ganocy SJ, Bauer MS, Cassidy KA, Hays RW, Safavi R, Blow FC, & Calabrese JR. (2009). A comparison of the life goals program and treatment as usual for individuals with bipolar disorder. *Psychiatric services (Washington, D.C.)*, *60*(9), 1182-1189. https://doi.org/10.1176/ps.2009.60.9.1182

Saloheimo HP, Markowitz J, Saloheimo TH, Laitinen JJ, Sundell J, Huttunen MO, Aro TA, Mikkonen TN, & Katila HO. (2016). Psychotherapy effectiveness for major depression: A randomized trial in a Finnish community. *BMC PSYCHIATRY*, *16*. https://doi.org/10.1186/s12888-016-0838-1

Schrank B, Brownell T, Jakaite Z, Larkin C, Pesola F, Riches S, Tylee A, & Slade M. (2016). Evaluation of a positive psychotherapy group intervention for people with psychosis: Pilot randomised controlled trial. *Epidemiology and psychiatric sciences*, *25*(3), 235-246. https://doi.org/10.1017/S2045796015000141

Schäfer Ingo, Lotzin Annett, Hiller Philipp, Sehner Susanne, Driessen Martin, Hillemacher Thomas, Schäfer Martin, Scherbaum Norbert, Schneider Barbara, & Grundmann Johanna. (2019). A multisite randomized controlled trial of Seeking Safety vs. Relapse Prevention Training for women with co-occurring posttraumatic stress disorder and substance use disorders. *European Journal of Psychotraumatology*, *10*(1), 1577092. https://doi.org/10.1080/20008198.2019.1577092

Smith Ronald, Wuthrich Viviana, Johnco Carly, & Belcher Jessica. (2021). Effect of Group Cognitive Behavioural Therapy on Loneliness in a Community Sample of Older Adults: A Secondary Analysis of a Randomized Controlled Trial. *Clinical Gerontologist*, *44*(4), 439–449. https://doi.org/10.1080/07317115.2020.1836105

Tjaden C, Mulder CL, den Hollander W, Castelein S, Delespaul P, Keet R, van Weeghel J, & Kroon H. (2021). Effectiveness of Resource Groups for Improving Empowerment, Quality of Life, and Functioning of People With Severe Mental Illness A Randomized Clinical Trial. *JAMA PSYCHIATRY*, *78*(12), 1309–1318. https://doi.org/10.1001/jamapsychiatry.2021.2880

Valiente C, Espinosa R, Contreras A, Trucharte A, Caballero R, Peinado V, Calderón L, & Perdigón A. (2022). A multicomponent positive psychology group intervention for people with severe psychiatric conditions; a randomized clinical trial. *Psychiatric rehabilitation journal*, *45*(2), 103-113. https://doi.org/10.1037/prj0000509

Vallina-Fernandez O, Lemos-Giraldez S, Roder V, Garcia-Saiz A, Otero-Garcia A, Alonso-Sanchez M, & Gutierrez-Perez AM. (2001). Controlled study of an integrated psychological intervention in schizophrenia. *EUROPEAN JOURNAL OF PSYCHIATRY*, *15*(3), 167–179.

van Gestel-Timmermans H, Brouwers EP, van Assen MA, & van Nieuwenhuizen C. (2012). Effects of a peer-run course on recovery from serious mental illness: A randomized controlled trial. *Psychiatric Services*, *63*(1), 54–60. https://doi.org/10.1176/appi.ps.201000450

Volpe Umberto, Torre Fabiana, De Santis Valeria, Perris Francesco, & Catapano Francesco. (u.å.). Reading Group Rehabilitation for Patients with Psychosis: A Randomized Controlled Study. *Clinical Psychology & Psychotherapy*, *22*(1), 15–21. https://doi.org/10.1002/cpp.1867

Weiss Roger D, Griffin Margaret L, Greenfield Shelly F, Najavits Lisa M, Wyner Dana, Soto Jose A, & Hennen John A. (2000). Group therapy for patients with bipolar disorder and substance dependence: Results of a pilot study. *The Journal of Clinical Psychiatry*, *61*(5), 361–367. https://doi.org/10.4088/JCP.v61n0507

Wojtalik J, Eack S, & Keshavan M. (2019). Confirmatory efficacy of cognitive enhancement therapy for early schizophrenia: Results from a multi-site randomized trial. *2019 Congress of the Schizophrenia International Research Society, SIRS 2019. Orlando, FL United States.*, *45*(Supplement 2), S184. http://dx.doi.org/10.1093/schbul/sbz021.235

Wuthrich, V. M., & Rapee, R. M. (2013). Randomised controlled trial of group cognitive behavioural therapy for comorbid anxiety and depression in older adults. *Behaviour Research and Therapy*, *51*(12), 779–786. https://doi.org/10.1016/j.brat.2013.09.002

Yanos Philip T, Roe David, West Michelle L, Smith Stephen M, & Lysaker Paul H. (2012). Group-based treatment for internalized stigma among persons with severe mental illness: Findings from a randomized controlled trial. *Psychological Services*, *9*(3), 248–258. https://doi.org/10.1037/a0028048

References to excluded studies

References to studies awaiting classification

References to ongoing studies

Additional references

Aloe, A. M., Dewidar, O., Hennessy, E. A., Pigott, T., Stewart, G., Welch, V., Wilson, D. B., & Group, C. M. W. (2024). Campbell Standards: Modernizing Campbell's Methodologic Expectations for Campbell Collaboration Intervention Reviews (MECCIR). *Campbell Systematic Reviews*, *20*(4), e1445. https://doi.org/https://doi.org/10.1002/cl2.1445

Borenstein, M., & Hedges, L. V. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–242). Russell Sage Foundation West Sussex.

Chen, M., & Pustejovsky, J. E. (2024). Adapting Methods for Correcting Selective Reporting Bias in Meta-Analysis of Dependent Effect Sizes. *OSF*. https://doi.org/10.31222/osf.io/jq52s

Eldridge, S., Campbell, M. K., Campbell, M. J., Drahota, A. K., Giraudeau, B., Reeves, B. C., Siegfried, N., & Higgins, J. P. (2021). *Revised Cochrane risk of bias tool for randomized trials (RoB 2): Additional considerations for cluster-randomized trials (RoB 2 CRT)*. Cochrane Bias Methods Group. https://drive.google.com/file/d/1yDQtDkrp68_8kJiIUdbongK99sx7RFI-/view

Fitzgerald, K. G., & Tipton, E. (2024). Using Extant Data to Improve Estimation of the Standardized Mean Difference. *Journal of Educational and Behavioral Statistics*, 10769986241238478. https://doi.org/10.3102/10769986241238478

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*(4), 1295–1308. https://doi.org/10.3758/s13428-014-0538-z

Hedges, L. V, Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12296

Higgins, J. P. T., Eldridge, S., & Li, T. (2019). Including variants on randomized trials. In J. P. T. Higgins, J. Thomas, J. Chandler, M. S. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (2nd ed., pp. 569–593). Wiley Online Library. https://doi.org/10.1002/9781119536604

Higgins, J. P. T., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. *Cochrane Handbook for Systematic Reviews of Interventions*, 143–176.

Joshi, M., & Pustejovsky, J. E. (2022). *wildmeta: Cluster wild bootstrapping for meta-analysis*. https://github.com/meghapsimatrix/wildmeta

Logan, J. A. R., Hart, S. A., & Schatschneider, C. (2021). Data sharing in education science.

AERA Open, *7*, 23328584211006476.

Maassen, E., van Assen, M., Nuijten, M., Olsson Collentine, A., & Wicherts, J. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PloS One*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

McGrath, S., Zhao, X., Steele, R., & Benedetti, A. (2019). estmeansd: Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. *CRAN: Contributed Packages*.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Pustejovsky, J. E. (2016). *Alternative formulas for the standardized mean difference*. https://www.jepusto.com/alternative-formulas-for-the-smd/

Pustejovsky, J. E. (2020). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (0.5.5). cran.r-project.org. https://cran.r-project.org/web/packages/clubSandwich/index.html

Pustejovsky, J. E., Joshi, M., & Citkowicz, M. (2025). *metaselection: Meta-analytic selection models with cluster-robust and cluster-bootstrap standard errors for dependent effect size estimates* (0.1.5). https://github.com/jepusto/metaselection/tree/main

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

RStudio Team. (2015). *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. https://www.rstudio.com/

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M.,

Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Higgins, J. P., Elbers, R. G., & Reeves, B. C. (2016). *Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance*. http://www.riskofbias.info

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., & Eldridge, S. M. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Taylor, J. A., Pigott, T. D., & Williams, R. (2021). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher*, *51*(1), 72–80. https://doi.org/10.3102/0013189X211051319

van Aert, R. C. M. (2023). *puniform: Meta-analysis methods correcting for publication bias* (0.2.7). CRAN. https://cran.r-project.org/package=puniform

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes,

A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, D. B. (2016). *Formulas used by the "Practical Meta-Analysis Effect Size Calculator."* https://mason.gmu.edu/~dwilsonb/downloads/esformulas.pdf

WWC. (2021). *Supplement document for Appendix E and the What Works Clearinghouse procedures handbook, version 4.1*. Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-41-Supplement-508_09212020.pdf

Data and analytic code

Include a statement indicating if data will be available and how to access it. Data coding sheets and analytic codes could be submitted as supplementary material or add a link to an external repository (if applicable) and cite it.

Figures

Figures should be prepared following this guidance and inserted in the body of the text. They should also be submitted as separate files in their original format if possible.

Sources of support

Internal sources

Click or tap here to enter text.

External sources

Click or tap here to enter text.

Appendices

Appendices should be submitted as supplementary material.