The Effectiveness of Team Reflexivity Interventions: A Systematic Review and Meta-Analysis of Randomised Controlled Trials

Robin L. J. Lines[1,2], Simon Pietsch[1,2], Monique Crane[3], Nikos Ntoumanis[2,4], Philip Temby[5], Sally Graham[6], and Daniel F. Gucciardi[1,2]

[1]*School of Physiotherapy and Exercise Science, Curtin University*

[2]*Physical Activity and Well-being Research Group, Curtin University*

[3]*School of Psychology, Macquarie University*

[4]*School of Psychology, Curtin University*

[5]*Land Division, Defence Science and Technology Group*

[6]*Army People Capability Branch, Australian Army*

Author Notes

*Address correspondence to Daniel Gucciardi, School of Physiotherapy and Exercise Science, Curtin University, GPO Box U1987, Perth, Australia, 6845. Email: daniel.f.gucciardi@gmail.com

Funding statement: This research was supported by the Defence Science Centre, an initiative of the State Government of Western Australia.

Data availability statement: materials and data for this study are available on the project page located on the Open Science Framework (https://osf.io/ruzy4/).

Lines, R.L.J., Pietsch, S., Crane, M., Ntoumanis, N., Temby, P.T., Graham, S., & Gucciardi, D.F. (in press). The effectiveness of team reflexivity interventions: A systematic review and meta-analysis of randomised controlled trials. *Sport, Exercise, and Performance Psychology.* doi: 10.1037/spy0000251

Abstract

Cumulating evidence from 24 independent randomised controlled trials or experiments (N = 4,339), we meta-analytically examined questions regarding the effectiveness of team reflexivity on collective performance outcomes and behaviours, and the conditions under which such effects are strongest. We addressed these questions by testing the overall effect of team reflexivity on performance outcomes (i.e., indices or metrics that quantify goal attainment) and behaviours (i.e., those actions or states that precede or influence goal attainment), and assessing the robustness of this pooled effect across study (e.g., team size), outcome (e.g., measurement approach), and intervention characteristics (e.g., virtuality) via a series of sensitivity and moderator analyses. We found a positive and significant medium overall effect of team reflexivity interventions on performance outcomes (g = .549) and performance behaviours (g = .548). Moderator analyses indicated that the team reflexivity-performance outcome effect is contingent upon the measurement approach (in favour of self-reports over objective indices or researcher-assessed outcomes). With regard to performance behaviours, the effect of team reflexivity was strongest for cognitions and behaviours relative to affective dimensions, and when interventions were delivered to teams present physically relative to virtual teams. Collectively, our findings extend existing meta-analytic evidence regarding team reflexivity interventions in terms of resolution (i.e., inability to isolate unique effects), scope (i.e., primary studies missed via the systematic search), and methodological quality of primary evidence (i.e., incorporation of quasi-experimental designs).

Keywords: TIDieR guidelines; team debriefs; team reflections; three-level meta-analysis; virtuality.

The effectiveness of team reflexivity interventions: Systematic review and meta-analysis of randomised controlled trials

In modern society, teams are often the preferred or essential method of choice for accomplishing tasks in complex, dynamic, and challenging environments. Consider Deloittes' 2019 Human Capital Trends Report (Volini et al., 2019), for example, where 31% of respondents (N = 9,453) across 119 countries reported that most or almost all of their work is conducted within teams. In this report, significant performance improvements were reported by 53% of organisations as a key result of shifting towards team-based models. Despite this preference towards team-based work design, the empirical evidence to support the performance benefits of collective efforts is equivocal (Allen & Hecht, 2004). Thus, there is a need to enhance our knowledge of ways by which organisations are able to optimise team functioning. One key advantage of teams is that they enable organisations to organise work design in ways that produce outputs (e.g., performance) that would be unlikely or impossible for any individual alone to achieve (Shuffler, Diazgranados, Maynard, & Salas, 2018). In the medical domain, for example, the complexities of human health often preclude the effective treatment of illness or promotion of health by a single clinician, thus requiring the use of multidisciplinary teams (e.g., doctor, nurse, physiotherapist) with unique knowledge, skills, and abilities (Dobbins, Thomas, Stokes Melton, & Lee, 2016). Within the military context, where individuals are at times placed in extreme and dangerous conditions, teams can provide critical capabilities beyond those of individuals to cope with the dynamic and complex nature of military operations – both during warfare (e.g., combat) and peacetime (e.g., humanitarian) missions (Shuffler, Pavlas, & Salas, 2012). These examples and many others across numerous occupational contexts (e.g., aviation, sport, emergency services, security, business) underscore the criticality of teams for the safety, health, security, and success of societies and their citizens worldwide (e.g., see the 2018 special issue on teamwork in *American Psychologist*).

Given the importance of teams to countless aspects of life, it is unsurprising that interventions designed to enhance and/or maximise team performance are a key priority for

organisations and industries where collective efforts are critical for innovation, competitive advantage, and success. Formally defined, team development interventions (TDIs) represent "actions taken to alter the performance trajectories of organisational teams" in ways that foster returns to, maintenance of, and diversification of the healthy functioning of the unit (Shuffler et al., 2018, p. 689). In the broadest sense, TDIs can be characterised as addressing leader capabilities, team competencies, interpersonal processes, and/or team processes (Lacerenza, Marlow, Tannenbaum, & Salas, 2018). TDIs can be more narrowly characterised according to the classic input-process-output model (e.g., Ilgen, Hollenbeck, Johnson, & Jundt, 2005), which includes input (team task analysis, composition, work design, charters), process (performance monitoring and assessment), and output (team debriefs) programs, or some combination of these categories (team building, team training, team coaching, team leadership) (Shuffler et al., 2018). Numerous narrative and statistical synthesises support the effectiveness of targeted (e.g., shared leadership, D'Innocenzo, Mathieu, & Kukenberger, 2016) and multicomponent TDIs for enhancing collective outcomes that span affective (e.g., attitudes), cognitive (e.g., declarative knowledge), skill (e.g., display of teamwork behaviours), and performance dimensions (e.g., McEwan, Ruissen, Eys, Zumbo, & Beauchamp, 2017). Specifically, meta-analytic findings support the effectiveness of TDIs addressing leader capabilities (e.g., leadership training; Lacerenza, Reyes, Marlow, Joseph, & Salas, 2017), team competencies (e.g. communication; Marlow Lacerenza, Paoletti, Burke, & Salas, 2018), and interpersonal processes (e.g., trust; Breuer, Huffmeier, & Hertel, 2016; DeJong, Dirks, & Gillespie, 2016). Nevertheless, there exists a gap with regard to an integrative understanding of the effectiveness of interventions designed to address team processes and outcomes via collective reflections. Furthermore, our knowledge of the effectiveness of interventions designed to address team processes and outcomes via collective reflections is limited by the muddying of evidence between non-experimental and experimental designs. Our work looks to address this gap via a comprehensive systematic review, meta-analysis, and narrative synthesis of team reflexivity evidence from randomised controlled trials or experiments.

Team Reflexivity – Conceptual Foundations

Team reflexivity – defined as a conscious and intentional effort by which a collective evaluates and learns from experiences in devising and executing goal-directed pursuits (e.g., Otte, Konradt, Garbers, & Schippers, 2018; Schippers, West, & Edmondson, 2017) – is a key process by which collectives with a common purpose foster learning, adaptation, and performance. Team reflexivity interventions are routinely implemented in numerous occupational contexts. In military settings, for example, after-action reviews and debriefs have been used to classify non-technical skills for aviation operations (Tsifetakis & Kontogiannis, 2019) and optimise learning after stressful events (Moldjord & Hybertsen, 2015). Team reflections are used widely in medical and healthcare settings to structure conversations after critical events (Anderson, Sandars, & Kinnair, 2019), optimise interprofessional coordination among operating room teams (Boet et al., 2013), and improve communication in emergency resuscitation scenarios (Chamberland et al., 2018). The terms used to describe such interventions vary across studies and context, including after-action review (e.g., Cook & Kautz, 2016), crew resource management (e.g., Myers & Orndorff, 2013), after-event review (e.g., Ellis, Mendel, & Nir, 2006), debrief (e.g., Andersen, 2016), hot wash (e.g., Sinclair, Doyle, Johnston, & Paton, 2012), and huddle (e.g., Quinn & Bunderson, 2016). In addition to the assumption that different terms reflect the same concept, the variety of terms used to describe team reflection interventions over a diverse range of settings has led to fragmentation in literature (Tannenbaum & Cerasoli, 2013). Such fragmentation hinders the progress of research, with many papers reporting on similar findings despite drawing on and citing dissimilar research due to this diversity in terminology (Schippers et al., 2017). In essence, these interventions share the commonality of group members overtly reflecting upon, and communicating the collective's objectives, strategies, and processes following some team-level activity (Allen, Reiter-Palmon, Crowe, & Scott, 2018). Given the social nature of teams, our preference for the term 'reflexivity' to capture the essence of such interventions is informed by sociological theory because it characterises

the act of critically evaluating oneself with the view to understand cause and effect relationships (e.g., Archer, 2003).

Broadly speaking, reflective practice encompasses an iterative process whereby systems engage in some type of situated action, reflect on that experience, and plan how best to integrate those learnings into future acts (Kolb, 1984; West, 2000; Widmer, Schippers, & West, 2009). Within the context of teams, such an iterative process requires that members work collaboratively and openly to discuss and reflect upon unique and shared experiences from an event to develop goals which they can put into action to address in future situations. Tannenbaum and Cerasoli (2013) proposed four essential components for a reflexivity intervention. First, the intervention should foster active learning in which the system iteratively makes sense of the past with the view to inform the future; receiving feedback in isolation is insufficient for learning. Second, reflections should prioritise learning and development rather than offer a simple evaluation of what has occurred. Third, the benefits of reflections are maximised when they are contextualised or enacted with reference to specific events or situations rather than generalities that may have been gleaned from such experiences (e.g., individual member's strengths and weaknesses). Finally, reflections within the context of collectives, such as teams, require input from multiple information sources to maximise coverage and diversity of information, as well as interaction between members regarding their own perceptions of those experiences and the dynamics among members. Collectively, therefore, these indicators provide clarity on the necessary and sufficient conditions for the inclusion of team reflexivity interventions as part of this narrative and statistical synthesis.

Team reflexivity interventions are advocated as one of the best methods of producing desirable outcomes for individual members, teams, and organisations (Allen et al., 2018; Lacerenza et al., 2018; Schippers et al., 2017). A key expectation in this regard is that team reflexivity leads to better performance (West, 2000). Evidence to support this hypothesis comes in the form of various dimensions of performance, including team performance (e.g., Schippers, Homan, & van Knippenberg, 2013; Schmutz, Lei, Eppich, & Manser, 2018), team innovation (for a review, see

Schippers et al., 2017), and team effectiveness (Widmer et al., 2009). TDIs including team reflections as an active ingredient were considered in a recent meta-analysis, where it was found that programs (k = 22) encompassing team reflection had a moderate effect on team performance (ES = .64, 95% CI = .42, .86; McEwan et al., 2017). Focused specifically on debriefs, Tannenbaum and Cerasoli (2013) quantified the effect of team debriefs on performance as equating to a roughly 25% (k = 26, N = 2136; d = .67) improvement compared with control conditions, though this estimate included debriefs conducted at both the team and individual level. When considering samples specifically at the team level (k = 16; N = 546), a similar effect on performance was reported (25%, d = .66). Reflections focused at the team level whilst also assessing performance at the team level, rather than individual (k = 10; N = 176), produced the largest effect on performance (38%, d = 1.20). Thus, the available evidence supports the effectiveness of team reflexivity for optimising collective functioning.

There are two key reasons to justify the need for an updated or new meta-analysis of team reflexivity interventions. First, Tannenbaum and Cerasoli (2013) reported searching multiple databases for research "containing variants of the words *debrief* or *after-action review* coupled with *performance, effectiveness, ratings*, and similar terms" (p. 235); the exclusion of the exact details of the search strategy utilised in their meta-analysis creates uncertainty regarding the degree to which all relevant research was captured and summarised (e.g., the randomised controlled trial by Gurtner, Tschan, Semmer, & Nägele (2007) was absent from their meta-analysis). Relatedly, McEwan et al. (2017) used a comprehensive search strategy to capture the breadth of teamwork interventions (e.g., planning, coordination, situation awareness), yet specific terms for team reflexivity interventions were omitted (e.g., reflexivity OR reflection OR "after action" OR "after event" OR debrief*). This search strategy may have missed studies considered to be important for an assessment of the effectiveness of team reflexivity interventions. In addition, studies identified by McEwan et al. included reflection alongside other aspects of teamwork development (e.g., preparation and execution), which makes it impossible to isolate the unique effects of team reflexivity on team

performance. Second, the majority of primary studies that contributed to the pooled estimates in Tannenbaum and Cerasoli's (2013) meta-analysis were quasi-experimental and, therefore, cannot provide complete confidence regarding cause and effect. There is a need for a statistical synthesis of available studies that employed designs that provide evidence of causality and therefore the effectiveness of team reflexivity interventions. Therefore, we aimed to conduct an updated and extended meta-analysis in which we estimate the effectiveness of team reflexivity interventions using evidence from randomised controlled trials (RCT).

Hypothesis 1 (H1): *Team reflexivity interventions will foster better performance outcomes relative to control conditions.*

Moderators of the Team Reflexivity – Performance Effect

We expected team reflexivity interventions to foster positive performance outcomes, yet we also anticipated that this effect would be strengthened or weakened by characteristics of the team, the intervention, or context. Doing so speaks to an important consideration regarding the conditions in which team reflexivity interventions are most effective. In our protocol registration, we proposed to assess the robustness of the findings according to study design characteristics, including the occupational context (e.g., health, military), type of team (e.g., intact, new, virtual), intervention length, training type (e.g., virtual, face-to-face), measurement of performance (e.g., self- or informant-rated, objective), and degree of participant attrition. Nevertheless, the ability to test these moderator variables depends on the characteristics of the data retrieved from our systematic search process, that is, if there is sufficient variability in characteristics of the team, the intervention, or context. Thus, our final selection of moderator variables, presented below, strikes a balance between substantive guidance offered in past work on team development interventions (e.g., McEwan et al., 2018) and pragmatics based on the available data in the final pool of studies.

Occupational context. As teams are prominent across numerous contexts and come in many forms, it may be important to see how this factor may moderate the reflection – performance effect. Team debriefs are standard practice within certain occupational contexts (e.g., medical and

military), hence, teams situated in these contexts are likely to be more familiar with the process which may have implications for its effectiveness over teams for whom debriefings are novel (e.g., students). Meta-analytic evidence of TDIs has reported them to be effective in increasing performance across all sample contexts (i.e., health care, academia, laboratory experiment, military, aviation, and industry; d = .40 – 1.76), yet the effect size was substantially larger for industry teams (d = 1.76) when compared with other contexts (i.e., d = .40 – .76; McEwan et al., 2017). Looking specifically at reflexivity interventions, studies conducted within a medical environment reported similar effects to those carried out in other environments (k = 28, d = .66 and k = 18, d = .69; respectively; Tannenbaum & Cerasoli, 2013), though these estimates included both individual and team level reflection. Therefore, given the mixing of individual and team level outcomes, and the differences in findings across previous reviews, we will examine the differential effects of occupational context on the effectiveness of team reflexivity interventions, rather than propose a specific hypothesis.

Research Question 1 (RQ1): *Are team reflexivity interventions or training programs differentially effective for team performance depending on the type of occupational context?*

Team virtuality. The advent of flexible working environments and advancements in communication technology means that virtual teams are prevalent within contemporary organisations and work design (Marlow, Lacerenza, & Salas, 2017). Virtual teams contain distributed members working towards collective goals, communicating via technologies (e.g., instant messaging and e-mails) which allow for collaboration across multiple locations and time zones (Penarroja, Orengo, & Zornoza, 2017). The reliance on technology-mediated communication within virtual teams presents hurdles, such as the loss of important information available to face-to-face teams (e.g., verbal, non-verbal, and visual cues; Lacerenza et al., 2017; Penarroja et al., 2017), which can have a detrimental effect on the reflection processes (Konradt, Schippers, Garbers, & Steenfatt, et al., 2015). Meta-analytical findings have demonstrated that virtual teams exhibit lower levels of performance (e.g., Lacerenza et al., 2017). With regard to the benefits of collective

reflections within virtual teams, the findings are equivocal; in some instances, the effects of team reflections are stronger in face-to-face teams (Jarrett et al., 2016), yet in others team reflections improve performance in virtual teams (e.g., Gurtner, Tschan, Semmer, & Nagele, 2007; Konradt et al., 2015). For this reason, we will examine the differential effects of team virtuality on the effectiveness of team reflexivity interventions, rather than propose a specific hypothesis.

Research Question 2 (RQ2): *Are team reflexivity interventions or training programs differentially effective for team performance depending on the virtuality of the team?*

Facilitated team reflections. An important consideration regarding delivery mode is whether or not external facilitators trained in team reflexivity processes are used to administer the intervention. Facilitators help to guide a team's reflection discussions, ensuring that teams discuss relevant information, focus on what is important, and not stray from the task at hand (Allen et al., 2018). For these reasons, facilitators are considered among 'best practice' guidelines for the administration of team reflexivity interventions (Salas et al., 2008). Meta-analytical evidence has demonstrated that interventions involving facilitated reflections were around three times more effective than those utilising non-facilitated reflection (d = .75 and d = .27, respectively; Tannenbaum & Cerasoli, 2013). Nevertheless, it should be noted that only two studies (4.3%) reported a non-facilitated intervention as opposed to 34 (73.9%) using facilitation; the remaining ten studies (21.8%) did not report this information. To account for the possible advantages of a facilitator, self-guided reflection interventions tend to utilise detailed instructions or some kind of checklist to guide what is done (Allen et al., 2018). The effectiveness of structured self-guided reflections appear similar to those run by a facilitator (Boet et al., 2016), yet are more effective than unstructured self-guided reflection interventions (Eddy, Tannenbaum, & Mathieu, 2013).

Hypothesis 2 (H2): *Team reflexivity interventions will be most effective when they are administered by a facilitator or contain systematic guidelines for execution.*

Feedback. Feedback involves the communication of information relating to an event, behaviours, or actions concerning teamwork or the completion of a task (Gabelica et al., 2014).

Team feedback may be effective in improving performance as it serves as a guide, whereby effective behaviours are reinforced and ineffective behaviours are stopped (Penarroja et al., 2017). The evidence regarding the effectiveness of feedback for team functioning is mixed; in a review of the literature, findings were uniformly positive in roughly half of 59 studies reviewed, yet the positive effects of feedback varied in the other half of studies depending on the type of outcome variable (Gabelica, Van den Bossche, Segers, & Gijselaers, 2012). These findings suggest that feedback alone is insufficient to positively effect team functioning. Within team reflection interventions, teams may be able to use feedback to reflect upon an event critically and positively adapt to produce improved performance (Konradt et al., 2015). However, the findings regarding the effectiveness of integrating feedback with team reflections is equivocal. For example, teams who received feedback *and* engaged in guided reflection showed the greatest performance increase, when compared with teams who received feedback in isolation and teams that received neither feedback nor reflection (Gabelica et al., 2014). Similarly, teams who received feedback *and* guided reflection reported the highest quality reflections and subsequent performance when compared with teams who received only guided reflection or neither intervention (Konradt et al., 2015). These findings contrast with other research where the integration of feedback with reflections resulted in no meaningful benefits for team performance (Phielix, Prins, & Kirschner, 2010; Phielix, Prins, Kirschner, Erkens, & Jasper, 2011). Given these equivocal findings, we will examine the differential effects of reflections combined with feedback on the effectiveness of team reflexivity interventions, rather than propose a specific hypothesis.

Research Question 3 (RQ3): *Are team reflexivity interventions more effective when they are combined with feedback?*

Outcome measurement mode. Performance and other outcomes of interest can be measured via a number of subjective (e.g., questionnaires) and objective indicators (e.g., score on a test) that capture perceptions or direct assessments, respectively. There are conceptual and empirical reasons to expect that the relative strength of the effect of team reflexivity interventions on

performance outcomes is dependent on the method by which those outcomes are assessed. Objective assessments of performance are less contaminated by biases (e.g., social desirability), but may well have too narrow a scope to capture adequately the entire performance domain, thus often yield smaller effect sizes than subjective assessments (De Jong, Dirks, & Gillespie, 2016). With regard to TDIs, meta-analytic data indicates no meaningful differences in team performance outcomes between objective (d = .61) and third-party measures (d = .56), yet meaningful differences exist between third-party (d = .80) and self-reported (d = .38) measures of teamwork (McEwan et al., 2017). Focused specifically on team reflexivity interventions, meta-analytic data suggests that their effectiveness can be considered stronger when subjective criteria (d = 1.07) are used to assess outcomes, relative to objective criteria (d = .58; Tannenbaum & Cerasoli, 2013).

Hypothesis 3 (H3): *Team reflexivity interventions will yield larger effects when they are assessed using subjective indicators of performance rather than objective indices.*

Type of comparator. The use of a control group is an essential part of any experiment or RCT because they help manage potential threats to internal validity (e.g., regression to the mean, and response bias; Gold et al., 2017). In essence, the use of comparator groups enables researchers to answer one of three questions: (i) does the intervention work, (ii) how well does an intervention work relative to a relevant alternative, and (iii) how or why does an intervention work (Freelander et al., 2019). Despite the main function of the control group being to protect against potential threats to internal validity, they have been found to have unintentional effects on experimental outcomes (Mohr et al., 2009). In a meta-analysis (k = 125) examining the effect of control conditions on outcomes in RCT's, for example, meaningful effects in primary outcomes were observed for certain controls (e.g., waitlist, no treatment) but not in others (e.g., active comparator; Mohr et al., 2014). Although the type of comparator is an important factor within RCTs, this consideration has not yet been investigated within the context of team interventions. As such, we explored the differential effects of comparator type on the effectiveness of team reflexivity interventions, rather than proposed a specific hypothesis regarding this moderating effect.

Research Question 4 (RQ4): *Are team reflexivity interventions or training programs differentially effective for team performance depending on the type of comparator group?*

Methods

Consistent with best practice guidelines (Moher et al., 2015; Stewart, Moher, & Shekelle, 2012), we prospectively registered our study protocol via the Open Science Framework (OSF; https://osf.io/z3ehw) on September 24th 2019 using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-Protocol template (Shamseer et al., 2015).

Literature Search

We implemented a three-stage search strategy to locate relevant articles. The first stage included a comprehensive search of the following electronic databases from inception through until 25th September 2019: Web of Science (core collection), Scopus, Embase, Medline, PsycInfo, CINHAL Plus, and Business Source Complete. We used the following combination of search terms in each database: team* AND (reflexivity OR reflection OR "after action" OR "after event" OR debrief*) AND (intervention OR experiment* OR train* OR trial). The electronic database search was complemented via manual searching of the reference lists of all eligible studies (backward), and all articles which had cited eligible studies (forward). Finally, we manually searched reference lists of two relevant meta-analyses (McEwan et al., 2017; Tannenbaum & Cerasoli, 2013). [initials blinded for peer-review] executed these three stages of the search protocol.

Eligibility Criteria

We considered primary studies for inclusion in this meta-analysis if they: (i) tested the effectiveness of an intervention, training program, or experimental manipulation of team reflexivity; (ii) randomised teams to experimental or control conditions; and (iii) provided sufficient information in the published paper to compute an effect size for team-level outcomes, or this information was available by contacting the authors directly. We excluded papers if: (i) they did not employ a RCT design (e.g., cross-sectional, longitudinal, quasi-experiment such as non-random assignment), (ii) the article was written in any language other than English; (iii) the full-text was

unavailable via our university library subscriptions or directly from the corresponding author; (iv) information required to compute an effect size was unavailable in the document and following direct requests to the corresponding author; (v) the intervention was delivered to individuals in isolation from their team, rather than as a collective; and (vi) the results were published as a conference abstract rather than a full-text (e.g., dissertation, pre-print) because they are often reported inadequately (e.g., Hopewell & Clarke, 2005). Given our focus on interventions, it is important to specify the criteria we employed to characterise the population, intervention, comparison, and outcomes for primary studies.

Population. Teams are the focus of this systematic review and meta-analysis. For the purposes of this review, teams are defined as a collective of two or more individuals who work interdependently for a specified timeframe to achieve a common and valued outcome or objective (Sundstrom, de Meuse, & Futrell, 1990). We considered teams who are sampled naturistically from occupational contexts (e.g., sport, emergency services) or brought together for the purposes of an experiment (e.g., student teams).

Intervention. We prioritised interventions that aligned with the guiding definition of team reflexivity, that is, a conscious and intentional effort by which a collective evaluates and learns from experiences in devising and executing goal-directed pursuits (Otte et al., 2018; Schippers et al., 2018).

Comparators. We considered all types of comparators, including waitlist controls, no contact controls, placebo control, and active controls.

Outcomes. As teams exist primarily to achieve some type of valued objective that is typically unattainable by any individual alone, we focused on team performance as the primary outcome of this systematic review and meta-analysis. Of particular relevance in this regard is the distinction between performance behaviours and performance outcomes (Beal et al., 2003); the former refers to those actions that precede or influence goal attainment (e.g., coordination, communication), whereas the latter is characterised by indices or metrics that quantify goal

attainment (e.g., effectiveness, efficiency). Our main interest in this review was performance outcomes, though for completeness we also coded for performance behaviours. We categorised performance behaviours and performance outcomes as cognitive (i.e., information that reflects thinking, reasoning, and representation), affective (i.e., emotional dimensions characterised by feelings, expressions, and physiological factors), or behavioural (i.e., physical acts that are visible) in nature to minimise heterogeneity across studies.

Article Screening

The lead author screened all records retrieved from the literature search via a two-step process in which the title and abstract were first assessed against the eligibility criteria followed by a full-text review. A second reviewer [initials blinded for peer-review] screened a randomly selected portion (30%) of titles and abstracts at the first stage. When a decision was unable to be made based solely on the title and abstract, the full text was reviewed. Full text reviews were conducted independently by both reviewers. A third member of the research team adjudicated any disagreements regarding study eligibility between the two primary assessors. Reasons for study exclusion are summarised as part of the search and data extraction PRISMA flow diagram.

Data Extraction

One member of the research team [initials blinded for peer-review] extracted all data items from primary studies using a pre-determined form. When data was unavailable in the full text, we sent a request via e-mail to corresponding authors of eligible studies on two occasions, 1 month apart. We extracted data on the nature of the publication, type of team studied, participants characteristics, key details of the intervention as per the TIDieR guidelines (Hoffman et al., 2014), type of comparator, descriptive statistics of key study variables, source of ratings for moderator and outcome variables, and the statistical technique for the primary analysis. The full data extraction form is available on the OSF page (https://osf.io/ruzy4/). A second member of the research team [initials blinded for peer-review] assessed a random sample of 30% of data extraction forms to check accuracy and consistency.

Narrative and Statistical Analyses

Coding of Studies. A detailed coding system was used to record important characteristics of studies, interventions, samples, and outcome variables. Performance outcomes and behaviours were measured in numerous ways across eligible studies. If the variable quantified goal attainment in some way (e.g., effectiveness, efficiency) we considered it a performance outcome (yes vs no); all remaining performance outcomes were coded in terms of the categorisations of cognitive, affective, and behavioural. We also coded outcome variables in terms of the method of measurement, namely (i) objective (e.g., computer generated performance score), (ii) subjective (e.g., participants' self-rated perceptions), and (iii) researcher assessed (e.g., observers' subjective assessment of an outcome). In terms of study characteristics, we coded for publication type (peer-reviewed paper versus dissertation); journal impact factor at time of publication (continuous variable); type of comparator, where we categorised alternate interventions in which teams actively completed a task of some manner as an active control group, or waitlist or no intervention groups as no control; and, where relevant, the impact factor of the journal at the time of publication (continuous variable). The coded sample characteristics included occupational context (student vs employees), percentage of female participants (continuous variable)[1], mean age (continuous variable), number of teams in the experimental and control conditions (continuous variable), and team size (continuous variable; the mean was used when studies reported a range of team sizes). In terms of intervention characteristics, we coded team virtuality as either face-to-face or virtual (e.g., geographically dispersed teams that communicated via technology to accomplish a common goal), facilitation enhanced (facilitated vs self-led, where we defined facilitated as a facilitator being present and providing instruction or guidance), provision of handouts (present vs absent), total facilitation (combined presence of either a facilitator or handouts; present vs absent), provision of feedback

[1] We utilised the percentage of female participants in the total sample rather than the composition of teams (e.g., single gender, mixed) because this information was absent from most of the eligible studies. Doing so oversimplifies the operationalisation of gender for tests of the effectiveness of team reflexivity interventions and therefore should be considered a limitation of this study.

(feedback vs no feedback), and reflection time (continuous variable; total time spent reflecting as a team). For moderator analyses, continuous variables were mean centred and categorical variables were recoded using a binary dummy variable.

Calculation of effect sizes. We expected sufficient methodological heterogeneity to warrant the synthesis of pooled standardised mean differences. The standardised mean difference was computed to calculate a summary measure of effect size to quantify the effect of the intervention or experimental manipulation relative to comparators, thereby permitting synthesis of the same outcome variable (e.g., team-level communication) when measured using different scales or tools. We preferred Hedges' *g* as the effect size metric to account for the relative size of each sample (Lakens, 2013). Effect sizes were calculated from means, standard deviations (SD), and sample sizes of experimental or intervention groups at baseline and post-intervention using the formula's provided by Borenstein, Hedges, Higgins, and Rothstein (2009). We created an Excel calculator to compute these individual effects sizes, which is available on the OSF page (https://osf.io/ruzy4/). For a majority of performance outcomes, a higher score indicated a superior performance by the experimental group in comparison to the control group. Therefore, a positive effect size represents the beneficial effects of team reflexivity interventions. For those outcomes where a higher score was indicative of a worse performance (e.g., total time taken), the effect size direction was transformed so that a positive effect size represented a better performance for team reflection conditions.

Due to the variability in designs among eligible studies, we made a number of decisions prior to the calculation of effect sizes. This information was absent from our pre-registered protocol so we report them here for transparency. First, when outcome variables were assessed at multiple time points, we used the final measurement point in the analysis. Second, in cases where the sample size for each condition was unavailable ($K = 3$), we divided the total sample size by the number of conditions to provide an estimated sample size (e.g., Coles, Larsen, & Lench, 2019). Third, several studies ($K = 13$) compared multiple variants of team reflexivity interventions with a common

control group. Consistent with Cochrane group's guidelines (Higgins et al., 2019), we merged these conditions using the dmetar package in R (Harrer, Cuijpers, Furukawa, & Ebert, 2019) to avoid inappropriately counting groups twice and unit of analysis errors when including multiple groups from a single study. Similarly, in cases where there were multiple comparator groups (K = 3), we merged means and SDs to form a common control group from which a comparison could be made. In studies (K = 3) in which there was an experimental condition unrelated to the aims of the meta-analysis (e.g., individual reflection), the condition was excluded from the analysis. Fourth, two eligible studies were coded as containing two experiments (a and b); in one instance, two experiments were reported within the same study (pilot and experiment; Dyas, 2018), whereas experimental conditions were compared to control conditions in two different location conditions in the second paper (distributed and co-located; Jarrett et al., 2016).

Statistical synthesis of effect sizes. The majority of primary studies (91.7%) included two or more effect sizes and/or compared multiple treatments against the same comparator group (i.e., multiple treatment studies; Gleser & Olkin, 2009). In such cases, effect sizes from the same study are likely to be more alike than effect sizes from different studies (Wibbelink, Hoeve, Stams, & Oort, 2017). The inclusion of multiple effects sizes from a single study violates the assumption of independence in effect sizes in traditional meta-analyses (e.g., Cheung, 2014; Van den Noortgate, Lopez-Lopez, Martin-Martinez, & Sanchez-Meca, 2103). As such, we used a 3-level random effects model to account for interdependences in effect sizes from the same study (Cheung, 2019). Three-level meta-analysis enables analysts to account for the hierarchical nature of the data (e.g., effects sizes nested within studies) and, in so doing, the extraction of multiple effects from each eligible study preserves information facilitating maximum statistical power (Assink & Wibbelink, 2016; Van Dam et al., 2018). A 3-level random-effects meta-analysis enables us to decompose sampling variance of the observed effect sizes (level 1), and variance within studies (level 2) and between studies (level 3; Cheung, 2014). In other words, the 3-level approach allows effect sizes from primary studies to vary between participants, outcomes, and studies. We first estimated an overall

effect of team reflection on all outcomes within the same model using an intercept only model. Log-likelihood ratio tests were then carried out to ascertain if there was heterogeneity within (level 2) and between (level 3) studies (Assink & Wibbelink, 2016). If the variance at either of these levels is statistically significant ($p < .05$), then effect size distributions can be considered heterogeneous. In these cases, the model can be extended to include moderators to examine their effects on level 2 and 3 variance and the overall effect of team reflexivity interventions. We subsequently carried out three separate meta-analyses for performance outcomes, performance behaviours, and cognitive performance behaviours. We conducted all statistical analyses detailed below using the package *metafor* (Viechtbauer, 2010) in the R statistical platform (R Development Core Team, 2019) using guidelines provided by Assink and Wibbelink (2016). The R package *metaviz* (Kossmeier, Tran, & Voracek, 2020a) produces 'sunset' forest and funnel plots from the meta-analytic data that incorporate information on statistical power of each individual study included in the synthesis (Kossmeier, Tran, & Voracek, 2020b).

Moderator and sensitivity analyses. As forecasted in the literature review, we examined team virtuality, facilitation enhanced reflections, feedback, outcome measurement mode, and type of comparator as moderators of the effect of team reflexivity interventions. Sensitivity analyses were performed to assess the robustness of the findings according to study design characteristics, including sample size; team size; age of participants; impact factor of outlet in which the paper was published; percentage of female participants; total reflection time (continuous); publication type (i.e., published via peer-review versus unpublished data); occupational context (e.g., employed, student); training type (e.g., virtual, face-to-face); measurement of performance (e.g., self- or informant-rated, objective); type of comparator (e.g., active, control); feedback (e.g., present, absent); facilitation (e.g., present, absent); handouts (e.g., present, absent); total facilitation (e.g., present, absent). Continuous variables were mean centred.

Statistical heterogeneity. We calculated two indices to make an assessment of statistical heterogeneity: (i) I^2, which captures the proportion of total variance in effect estimates that is due to

heterogeneity rather than sampling error (0%-40% = might not be important; 30%-60% = may represent moderate heterogeneity; 50%-90% = may represent substantial heterogeneity; and 75%-100% = considerable heterogeneity; Higgins et al., 2003); and (ii) the degree of variance according to differences or similarities within (τ^2_{within}) studies and between studies ($\tau^2_{between}$), where a value of zero is indicative of no heterogeneity.

Confidence in cumulative evidence. The quality of evidence and strength of recommendations will be assessed using the GRADE approach (Guyatt et al., 2008; see https://gradepro.org/). The quality of evidence will be assessed across the domains of risk of bias, consistency, directness, precision, and publication bias. Additional domains may be considered where appropriate. The Cochrane quality assessment tool, the revised risk of bias tool (RoB 2; Sterne et al., 2019) was used to extract information regarding several dimensions of risk of bias: randomisation process, deviations from intended intervention, missing outcome data, measurement of the outcome, and selection of the reported result. Consistent with recommendations (Liberati et al., 2009), the results of this assessment are summarised in tabular format; full details are located on the OSF project page (https://osf.io/ruzy4/). Two members of the research team [initials blinded for peer-review] independently completed these assessments. Publication bias was assessed using Egger's test (Egger et al., 1997) and funnel plots, where asymmetry in the plot is interpreted as evidence of publication bias (Lau et al., 2006).

Narrative analysis of intervention content. In addition to a statistical synthesis, we planned to synthesise the findings of eligible studies narratively to summarise and explain the characteristics and findings of team reflexivity interventions according to the TIDieR guidelines (e.g., content, mode of delivery Hoffman et al., 2014). We were particularly interested in differences in the nature of interventions (e.g., content, mode, duration) between studies where team reflexivity training was found to be in/effective.

Deviations from Pre-Registered Protocol

We deviated from the pre-registered protocol in six ways. First, we planned to send "2 reminder emails each 7 days apart" requesting data from corresponding authors when information was unavailable in the full text. Due to the timing of the year (i.e., summer holidays), two emails were sent three weeks apart with a final reminder email sent 2 weeks later to authors who had replied but not yet sent the data. Second, we planned to use Zotero to import citations for screening, yet encountered issues when exporting citations from research databases (e.g., upload failed twice and the speed of upload was slow), so relied on Endnote for the final process. Third, our registered protocol included the Cochrane quality assessment tool (Higgins et al., 2011) for the risk of bias assessments, yet we ended up using the updated version (RoB 2; Sterne et al., 2019). Fourth, our pre-registered plan was to synthesise pooled standardised mean differences using a random effects meta-analysis model. However, it become clear that multilevel meta-analysis would be more appropriate to account for non-independence of effects and comparators within single studies (Cheung, 2019). Fifth, in two cases authors employed a pre-post design, yet team composition varied from pre to post intervention so treated them as a post design for calculating study effect sizes (Vashdi, 2013; Vashdi, Bamberger, & Erez, 2013). Finally, we planned to examine as secondary outcomes cognitive (e.g., team efficacy, team potency) and affective (e.g., trust, team passion) emergent states that characterise team functioning; however, the absence of data on such concepts meant we were unable to fulfil this intention.

Results

Literature Search Overview

An overview of search and study selection process is presented in the PRISMA diagram (https://osf.io/ruzy4/). In total, we identified 20 eligible studies from the electronic database search, and an additional 4 eligible studies via forward and backward scans. Of these 24 studies, the information required to compute effect sizes was unavailable in four cases, which resulted in a final sample of 20 studies included in the meta-analysis. These 20 studies were published between 2007 and 2019, and yielded 89 effect sizes (ES's), of which 66 were deemed relevant for inclusion. The

total sample included 4,339 participants who were members of 1,803 teams, where the mean team size was 3.8 members. On average, participants were 21.8 years of age and comprised 54.7% female participants. Table 1 displays an overview of the studies included within this meta-analysis.

Overall Effect of Team Reflexivity Interventions

The overall effect of team reflexivity was moderate in magnitude (g =.504, p < .001; see Table 2 and Figure 1). Egger's test (t (65) = 1.259, p = .213) indicated that funnel plot asymmetry was not significant (see Figure 2). A visual inspection of the funnel plot suggests a large study effect in that well powered studies (smaller standard errors) reported larger positive effects of team reflexivity. The log-likelihood ratio tests revealed significant variance between ES's within studies (level 2; p < .001) and between studies (level 3; p < .001), which explained 26.9% and 45.4% of the variance respectively (see Table 2). As there was substantial heterogeneity among effect sizes (I^2 = 72.3%; Higgins et al., 2003), we carried out moderator analyses to examine factors that may explain the variance within and/or between studies.

Moderator analyses of overall effect. Results of the moderator and sensitivity analyses of the overall effect of team reflexivity interventions are provided in Table 3. Of 14 potentially influential factors, only team virtuality moderated the overall effect of team reflexivity interventions (F(1, 64) = 8.914, p = .004), such that team reflexivity interventions were most effective for face-to-face teams (g = .678) when compared with teams who were virtual in nature (g = .166). All other intervention, study, and outcome characteristics were statistically inconsequential.

Effect of Team Reflexivity on Performance Outcomes

The effect of team reflexivity interventions on performance outcomes (H1) was moderate in magnitude (g =.549, p < .001; see Table 2 and supplementary Figure 2; https://osf.io/ruzy4/). Egger's test (t (26) = 0.143, p = .888) supported symmetry in the funnel plot (see Figure 2). Visual examination of the funnel plot suggests a slight large study effect, that is, well powered studies tended to report larger positive effects of team reflexivity. The I^2 for the model was 78.3%, suggesting considerable heterogeneity (Higgins et al., 2003). The log-likelihood ratio tests showed

within study (level 2) variance to be non-significant (p = .520), though variance between studies (level 3) was significant (p = .017), explaining 8.2% and 70.1% of the total variance, respectively (see Table 2). The meaningful between study variance supports heterogeneity among ES's and subsequent analyses to examine the moderating effect of outcome, study, and intervention characteristics.

Moderator analyses of the effect on performance outcomes. Results of the moderator and sensitivity analyses of the effect of team reflexivity interventions on performance outcomes are provided in Table 4. Measurement outcome mode (H3) was the only meaningful moderator of the direct effect (F(2, 24) = 5.253, p = .013), such that the effectiveness of team reflexivity on performance outcomes was salient only when the outcome variable was self-reported (g = 1.882) or objectively assessed (g = .559), but not for researcher assessed variables (g = .054). This finding should be interpreted with caution, as only one study used a self-reported measure of performance outcome.

Effect of Team Reflexivity on Performance Behaviours

The effect of team reflexivity on performance behaviours was moderate in magnitude (g =.548, p < .001; see Table 2 and supplementary Figure 3; https://osf.io/ruzy4/). Egger's test (t (38) = 1.703, p = .097) indicated that funnel plot asymmetry was not significant (see Figure 2). A visual inspection of the funnel plot indicates the presence of a large study effect, where well powered studies reported larger positive effects of team reflexivity. Log-likelihood ratio tests revealed significant variance both within studies (level 2; p = .002) and between studies (level 3; p = .039), explaining 34.8% and 34.9% of the variance, respectively (see Table 2). As there was substantial heterogeneity among effect sizes (I^2 = 69.7%; Higgins et al., 2003), we subsequently examined moderators of this heterogeneity in the effect.

Moderator analyses of the effect on performance behaviours. Results of the moderator and sensitivity analyses of the effect of team reflexivity interventions on performance behaviours are provided in Table 5. Analyses identified two meaningful moderators of this effect. First is type

of outcome ($F(2, 36) = 3.549$, $p = .039$), such that the effect of team reflexivity interventions on performance behaviours was strongest for cognitive behaviours ($g = .651$), when compared with behavioural outcomes ($g = .579$) and affective outcomes ($g = .070$). Second is team virtuality ($F(1, 37) = 7.807$, $p = .008$), whereby team reflexivity interventions were most effective for face-to-face teams ($g = .733$) when compared with teams who were virtual in nature ($g = .225$; RQ2).

Effect of Team Reflexivity on Cognitive Behaviours

The effect of team reflexivity on cognitive behaviours was moderate in magnitude ($g = .655$, $p < .001$; see Table 2 and supplementary Figure 4; https://osf.io/ruzy4/). Egger's test ($t(29) = 1.657$, $p = .109$) supported symmetry in the funnel plot (see Figure 2). A visual examination of the funnel plot reveals a clear large study effect, that is, well powered studies reported a larger positive effect of team reflexivity. The I^2 for the model was 64.1%, suggesting substantial heterogeneity (Higgins et al., 2003). The log-likelihood ratio tests revealed meaningful within study variance (level 2; $p = .001$), but not between study variance (level 3; $p = .475$), explaining 49.3% and 14.8% of the variance, respectively (see Table 2). The significant within study variance supported the need to examine moderators of the heterogeneity among ES's.

Moderator analyses of the effect on cognitive behaviours. Results of the moderator and sensitivity analyses of the effect of team reflexivity interventions on cognitive behaviours are provided in Table 6. No significant moderator variables were observed in the study and intervention characteristics.

Risk of Bias

We assessed risk of bias on all included outcomes ($N = 66$) using the RoB 2 framework and guidelines (Sterne et al., 2019). A summary of all primary studies is depicted in Figure 3, whereas individual assessments of each primary study are provided in a supplementary Table 1 (https://osf.io/ruzy4/). Overall bias decisions revealed that only four outcomes received a low risk of bias rating (all five RoB dimensions rated as low risk), with all four outcomes coming from the same study (Coppens et al., 2018). In terms of high risk of bias, four outcomes received the highest

rating (three or more dimensions rated as some concerns), with outcomes across three studies (Gurtner et al., 2007; Tesler et al., 2011; van Ginkel et al. 2009). The remaining 58 outcomes were rated as having some concerns (one or two dimensions rated as having some concerns). The main source of bias was from the reporting of the randomisation process, with only four outcomes receiving a low risk of bias on this dimension due to the paper in which they were reported (Coppens et al., 2018) including a detailed randomisation section within the studies methods. Of the 62 outcomes to receive a rating of some concerns for the randomisation process, for 22 it was the only dimension on which they received a non-low risk rating, meaning that adequate reporting of the randomisation process could have resulted in these additional 22 outcomes receiving an overall rating of a low risk of bias. The second main source of bias related to the selection of the reported results, with 34 of the 66 outcomes assessed to have some concerns. The major reason outcomes received this rating is due to the absence of a data analysis section in the paper. Thus, enhanced reporting of the randomisation process and articulating the statistical analyses performed could have resulted in the majority of outcomes being assessed as low risk rather than some concerns.

GRADE Assessment

We assessed the quality of evidence contributing to the analyses of the effects of team reflexivity interventions using the GRADE system (Guyatt et al., 2008). An overview of these findings is presented in Table 7. The overall level of evidence was downgraded by risk of bias, mainly due to poor reporting of the randomisation process, resulting in the majority of studies receiving a rating of some risk of bias. Level of evidence was also downgraded for inconsistency due to high levels of heterogeneity among effect sizes within each of the four meta-analyses ($I^2 >$ 64%). Overall, evidence was graded as low quality for each of the four outcomes.

Narrative Synthesis of Team Reflexivity Interventions

Full details of data extracted from each individual study according to the 12 TIDieR dimensions (Hoffman et al., 2014) is provided on the OSF project page (https://osf.io/ruzy4/). We summarise the findings of this review below, with a specific focus on dimensions that characterise

the nature of team reflexivity interventions or experimental manipulations within all 24 eligible papers (including a total of 25 studies). This narrative synthesis focuses on the core methodological items in the TIDieR checklist, items 3 - 9 (Dirven et al., 2020). None of the interventions involved any tailoring, so we exclude this item from our narrative synthesis. We also examined the nature of the team reflexivity interventions implemented in studies identified as highly effective (i.e., Hedges $g > .90$; n = 6) for performance outcomes (see Figure 4)[2]. We classified studies as highly effective when Hedges $g > .90$ based on commonly employed benchmarks for 'large' effects (Cohen, 1988).

Materials used to deliver team reflexivity interventions. All eligible studies provided an adequate description of materials used to administer interventions. Several different technologies were used within intervention delivery. Just over half of the studies (n = 14, 56%) used computers in some aspect of the intervention, of which ten used them for PC-based simulation tasks (e.g., battle, fire and rescue, quiz) where participants communicated via keyboard and/or headsets with microphones working together to complete the task. In the remaining four studies, an online environment/network was used in which participants could communicate electronically to take part in tasks (e.g., group essay). In one study (Calhoun et al., 2017), participants used a smart phone application to help guide them through the team reflexivity process. Written materials and handouts were provided to participants in eleven studies (44%) to guide or instruct teams regarding the reflexivity process. Studies conducted in medical settings (n = 4) utilised high/low fidelity manikins in task simulations. The six highly effective team reflexivity interventions utilised computers (67% vs 50%) and handouts (50% vs 29%) more often than all other studies. The utilisation of PC based simulations and online environments was largely similar across all interventions.

Procedures used in team reflexivity interventions. All eligible studies provided sufficient detail regarding the intervention procedures. Intervention procedures varied substantially across

[2] We also examined the nature of team reflexivity interventions implemented in studies identified as moderately effective (i.e., lower bound of 95% CI for Hedges $g \leq .25$; n = 12). The only meaningful difference was observed for the provision of feedback, with half of the moderately effective studies utilising this element as opposed to just over one in ten of the other studies (50% vs 13%).

studies, most likely due to the wide range of contexts and samples. The majority of interventions (72%, *n* = 18) required that participants complete a task followed by a period of team reflection before the completion of another task; this process occurred multiple times in eight studies (32%). Of the seven studies not to use a task-reflection-task design, three involved training/instruction on team reflection before assessing its effect on task performance, three utilised collaborative writing tasks over longitudinal periods that were interspersed with opportunities for team reflection, and the final study included reflection opportunities within a 72-hour training program (Gass & Priest, 2006). Several studies (32%, *n* = 8) involved complex computer simulation tasks in which teams would complete a period of training prior to task completion. Studies conducted in a medical setting (*n* = 4, 16%) familiarised participants with a simulation environment prior to completion of intervention tasks, or incorporated briefing and debriefing procedures in normal working life (*n* = 2, 8%; Vashdi, 2013; Vashdi et al., 2013). A greater proportion of highly effective studies involved training on complex simulation programs prior to task completion when compared with all other applications of team reflexivity (50% vs 29%). These simulation programs (e.g., Steel Beasts Pro, NeoCITIES) were conducted in teams and served as the basis for the team reflection intervention.

People involved in delivering team reflexivity interventions. The majority of eligible studies (68%, *n* = 17) reported details on who delivered the intervention, yet the information provided was often suboptimal. The primary reason for this interpretation is that limited information was provided on these individuals with regard to their suitability to deliver a team reflexivity intervention (e.g., expertise, training); for example, authors described the intervention provider as the principle investigator/lead experimenter (n = 3), experimenter(s)/researcher(s) (*n* = 4), external agent (*n* = 1), and a facilitator(s)/instructor (n = 7). The remaining two papers (Coppens et al, 2018; Diederich et al., 2019) provided adequate and detailed information on the providers of the intervention, including training and experience. The other eight studies (32%) excluded details on who provided the team reflexivity intervention. For highly effective studies, all six included information on who delivered the intervention, compared with 50% of the remaining studies.

Although this information was not detailed, describing providers only as; a facilitator(s) (n = 4), an experimenter (n = 1), and an instructor (n = 1).

Mode of delivery. Consistent with our eligibility criteria, all studies delivered the intervention in a group setting, with team size ranging from three to eight members (M = 4.5, SD = 4.1). The majority of studies provided interventions using a face-to-face delivery style (64%, n = 16), alternately studies utilised elements of virtuality in the delivery of interventions (28%, n = 7); two studies (Jarrett et al., 2016; Kondradt et al., 2015) utilised both face-to-face and virtual delivery styles. For interventions delivered in a virtual environment, team communication occurred through chat functions or e-mails in all but one study in which team members could communicate via headsets with a voice activated microphone. In seven studies (28%), a facilitator was present who provided instruction or guidance to teams during the reflection period. Finally, in several studies (n = 9, 36%) feedback was provided to teams to facilitate team reflexivity discussions. Team reflexivity interventions were delivered to teams of roughly equal sizes for both highly effective interventions and all other studies (M = 3.6 vs 4.0). The majority of highly effective interventions were delivered face-to-face when compared with two thirds of other studies (83.3% vs 64.3%), and a 19% lower proportion utilising a virtual element in the highly effective studies. Additionally, a greater proportion of the highly effective studies reported the use of a facilitator in intervention delivery (50% vs 29%). There was also a large difference in the provision of feedback, with two thirds of highly effective interventions utilising this element when compared with one if five other studies (67% vs 21%).

Location of delivery. Overall, the reporting of study location was relatively poor. Approximately half of the studies (44%, n = 11) failed to mention explicitly the geographical location of the study; six studies reported a vague area (e.g., geographical regions or states), whereas five reported no details of location. Of studies in which there was information on the geographical location, this detail was characterised by the nationality of participants (n = 7), geographical location (n = 4), and inferences via institutional ethical approval boards (n = 2) or

nationally reputable community computer system utilised (*n* = 1). Reporting of the physical location in which the intervention was delivered was also relatively poor. Approximately half of studies (56%, *n* = 14) provided clear details on the physical location (e.g., hospital simulation training centre, university laboratory, surgical wards), whereas the other half (44%, *n* = 44) excluded details on the physical location of the study or it was reported as an on/offline study with no detail on location for the offline component (Kim et al., 2011). Geographical reporting of intervention location was roughly equivalent across highly effective interventions and remaining studies (83% vs 79%), yet information regarding the physical location was reported less often (33% vs 57%).

Dosage of team reflexivity interventions. We assessed three criteria to characterise the dosage of interventions, namely time spent in team reflection, total study duration, and the number of sessions. All studies reported information for at least one of these areas, with thirteen (52%) reporting sufficient detail for all dimensions, nine (36%) reporting two of the three criteria, and three (12%) reporting on only one area. In terms of the total number of sessions for the intervention, all but three studies (88%, *n* = 22) provided information in this area. A majority of these interventions (81.8%, *n* = 18) occurred in a single session, with two studies conducted over two sessions and three studies implemented over three sessions. Information on the total study duration was reported in (or could be worked out from available details) in the majority of eligible studies (88%, *n* = 22), with time ranging from less than one hour (~ 45 minutes) to 6 months. The most commonly reported time was 5 hours (*n* = 5), with five studies reporting a longer time frame (72 hours, 3 weeks, 14 weeks, and two at 6 months). Finally, actual time spent reflecting was the most poorly reported of the criteria, with nine studies (36%) excluding details on the duration of team reflections. Of the 16 studies reporting time spent in team reflections, total reflection time ranged between 3 and 50 minutes (M = 22.6, SD = 19.8 minutes). Team reflection primarily occurred in a single block (*n* = 10), within the remaining eight studies the intervention consisted of multiple periods of reflection; one intervention utilised two periods, two interventions included three periods, three interventions consisted of four periods, and two interventions encompassed five periods of

reflection. Comparisons between highly effective interventions and remaining studies revealed clear differences in regard to delivery duration. The main difference was observed with regard to time spent in team reflection, with the six highly effective studies reporting this information more often than other studies (83.3% vs 64.3%). Highly effective interventions involved team reflections that were, on average, 9 minutes longer in duration (33 vs 24 minutes) and typically included multiple periods of reflection (50% vs 21.4%).

Discussion

Team reflexivity interventions – commonly referred to as debriefs, after action reviews, huddles, and hot washes – are core to the cultural fabric of many organisations worldwide (e.g., Allen et al., 2018; Schippers et al., 2018). Despite their popularity, evidence regarding the effectiveness of team reflexivity interventions is largely fragmented across diverse scientific disciplines (e.g., psychology, education) and occupational contexts (e.g., Defence, medical). Summary estimates of the effectiveness of team reflexivity interventions when integrated as part of a team development program (McEwan et al., 2017) or delivered in isolation (Tannenbaum & Cerasoli, 2013) are encouraging. However, these syntheses of the team reflexivity literature are limited in resolution (i.e., inability to isolate unique effects), scope (i.e., primary studies missed via the systematic search), and methodological quality of primary evidence (i.e., incorporation of quasi-experimental designs). We addressed these limitations in the current study by implementing a comprehensive search strategy to capture studies that employed designs capable of providing evidence of causality (i.e., randomised experiments or trials).

Implications for Theory and Practice

We sought to answer several key questions regarding the effectiveness of team reflexivity interventions. The answers to these questions and the degree of alignment of the findings with our expectations offer several important implications for theory and practice regarding team reflexivity. First, are team reflexivity interventions effective? We focused on team performance as the primary outcome of this systematic review and meta-analysis, hypothesising that team reflexivity

interventions would foster better performance outcomes relative to control conditions (i.e., indices or metrics that quantify collective goal attainment). Consistent with this expectation, we found a positive and significant medium overall effect of team reflexivity interventions on performance outcomes (g = .549). Our analyses supported a similar picture with regard to the magnitude and direction of the effect of team reflexivity interventions on performance behaviours (g = .548), that is, actions that precede or influence collective goal attainment (e.g., coordination, communication). We observed the largest effect of team reflexivity on cognitive behaviours (g =.655). These findings are consistent with previous estimates regarding the effectiveness of team reflexivity interventions in terms of both magnitude and direction (McEwan et al., 2017; Tannenbaum & Cerasoli, 2013) and therefore reinforce the centrality of reflections as a means by which to translate learnings of real-world experiences into future outcomes (Kolb, 1984; Ellis et al., 2016). Collectively, our findings support the effectiveness of team reflexivity interventions for optimising a broad range of performance outcomes, behaviours, and states.

Team reflexivity interventions share the commonality of group members overtly reflecting upon, and communicating the collective's objectives, strategies, and processes following some team-level activity (Allen, Reiter-Palmon, Crowe, & Scott, 2018), yet their content and structure varies across contexts. We extended past work by narratively synthesising the characteristics of team reflexivity interventions. We found that highly effective team reflexivity interventions (g >. 90), in general, were delivered in person, longer in duration (~Δ 9 min), incorporated multiple periods of reflection, utilised a facilitator, and combined reflections with feedback. Team reflexivity interventions delivered in person are likely more effective than technology-mediated methods because they offer greater richness in communication dimensions (e.g., audibility, contemporality, visibility; Clark & Brennan, 1991). In terms of intervention dosage, time spent reflecting was a statistically trivial moderator within the current review and past work, yet our narrative review indicated that the average time of the highly effective interventions involved nearly double the reflection time reported in a past review of team reflexivity studies (18 minutes; Tannenbaum &

Cerasoli, 2013). Ultimately, however, the key consideration is likely one of balancing quantity and quality (Otte et al., 2017; Otte et al., 2018); in other words, providing teams with additional time to reflect is likely most beneficial when they are supported to engage in high quality reflections (e.g., structured, combined with feedback). A second consideration of dosage was that highly effective interventions utilised multiple periods of reflection more often than less effective studies. By definition, reflection is characterised an iterative, cyclical process of action and reflection (e.g., Gibbs, 1988; Kolb, 1984), yet is one element that has received minimal attention in past work on team reflexivity. An iterative, cyclical process of action and reflecting is also consistent with widely accepted models of team effectiveness (Ilgen, Hollenbeck, Johnson, & Jundt, 2005). Therefore, teams may benefit most from reflection if it is incorporated into their routine creating a "metanorm of reflexivity" (Schippers, Edmondson, & West, 2014, p. 750). Of course, it is important to keep in mind that we classified interventions as 'highly effective' based solely on the effect size for that study (i.e., $g > .90$), which likely oversimplified the operationalisation of effective interventions (e.g., statistical power, adherence to pre-registered protocol and/or transparency regarding deviations when relevant).

Second, we also sought to examine the question – under what conditions are team reflexivity interventions or training effective? We conducted several sensitivity and moderator analyses to identify factors that might alter the effectiveness of team reflexivity interventions. Our analyses identified only one salient moderator for the primary outcome of this meta-analysis, such that team reflexivity interventions yielded the greatest effects when performance outcomes were assessed using self-reports ($g = 1.882$) and objective indices ($g = .555$) rather than researcher assessed variables ($g = .054$). This finding is consistent with our expectation and past meta-analytic estimates of team development interventions (McEwan et al., 2017; Tannenbaum & Cerasoli, 2013). Measurement techniques are an essential consideration for assessments of the validity of empirical research and the extent to which findings generalise. Self-reports are prone to social desirability bias (e.g., exaggerate positive characteristics; Budescu & Bruderman, 1995), yet can offer greater

breadth of the performance domain than objective methods (De Jong et al., 2016). Scholars have begun to prioritise objective and less obtrusive approaches to assessing psychological constructs in recent years (Weise, Shuffler, & Salas, 2015), which may lead to improvements in the quality of such measures when they cannot be captured via technological means (e.g., training raters on measurement systems or checklists). In contrast, the effectiveness of team reflexivity interventions seemed not to differ meaningfully according to characteristics of the study (e.g., team size, publication type) or the intervention itself (e.g., facilitation, type of comparator). A visual inspection of the differences in effect sizes suggests that feedback ($g_{present}$ = .881 versus g_{absent} = .357) and team virtuality ($g_{present}$ = .666 versus g_{absent} = .239) may be important considerations when making inferences regarding the extent to which the effectiveness of team reflexivity interventions generalise across conditions. The effect sizes for these moderators were noisy ($SE_{feedback}$ = .30 and $SE_{virtuality}$ = .29) so there is a need to consider these design factors in future research. Collectively, our sensitivity and moderator analyses indicated that the effectiveness of team reflexivity interventions for performance outcomes generalises across contextual and study conditions. The key theoretical and practical implications of these findings is that team reflexivity plays a critical role in optimising team performance.

With regard to performance behaviours, we identified the type of action that precedes or influences collective goal attainment and team virtuality as meaningful moderators of the effectiveness of team reflexivity interventions. Specifically, the effectiveness of team reflexivity interventions was strongest for cognitive behaviours (g = .651), when compared with behavioural outcomes (g = .579) and affective outcomes (g = .070), and when they were administered to teams present physically (g = .733) when compared with teams who were virtual in nature (g = .225). The nature of team reflexivity interventions offers one explanation for the differential effects according to the type of performance behaviour, given the focus is typically on cognitive and behavioural dynamics of collective action (Schmutz, Kolbe, & Eppich, 2018; Tannenbaum & Cerasoli, 2013). The prioritisation on cognitive dimensions of collective work is also evident in the focus on

cognitive dimensions for the evaluation of team reflexivity interventions (see Table 5). Individual emotions can influence team processes and outcomes, yet knowledge of their dynamics for collectives who share common experiences is limited (Menges & Kilduff, 2015; van Kleef & Fischer, 2016). Emotions can surface as emergent properties for collective in terms of the variability, type, magnitude, and time course of emotional responses (e.g., contagion or cascade effects; Goldenberg, Garcia, Halperin, & Gross, 2020). Thus, there is a need for future research to consider how these emotional dynamics might affect team processes and outcomes via collective reflection practices. The differential effects of team reflexivity on performance behaviours based on virtuality are consistent with a growing body of work that has shown collective processes and states (e.g., communication, coordination) are impeded when team members are dispersed via virtual means (Lacerenza et al., 2017; Marlow et al., 2018). Virtual teams utilise communication tools that inhibit certain elements of communication compared with face-to-face communication (e.g., verbal, non-verbal, and visual cues). Recent work underscores the importance of considering elements of work design to minimise the detrimental effects of virtuality on team effectiveness (e.g., knowledge characteristics, temporal stability, job resources; Handke, Klonek, Parker, & Kauffeld, 2020; Schaubroeck & Yu, 2017). No salient moderators existed for cognitive behaviours. With few exceptions, therefore, our moderator and sensitivity analyses suggest that the effectiveness of team reflexivity interventions for performance behaviours and cognitive outcomes are relatively robust to characteristics of the study, outcome, and intervention.

Our assessment of the methodological quality of eligible studies unearthed a pressing need for enhanced reporting of interventions or experimental manipulations of team reflexivity. Inadequate or incomplete descriptions of interventions pose a critical barrier to methodological practice (e.g., replication), knowledge accumulation (e.g., introduce bias into syntheses of evidence), and their application in the field by practitioners (Montgomery et al., 2018). Highlighting this point, our risk of bias assessment identified only one study as low risk. Two key issues stood out in this regard, namely the exclusion of explicit information to characterise the nature of the

randomisation process (e.g., sequence generation, allocation concealment) and the selection of reported results (e.g., no data analysis section to describe the analytical approach). Both of these elements of the scientific process are highlighted within the Consolidated Standards of Reporting Trials guidelines (CONSORT; Schulz, Altman, & Moher, 2010), which are an evidence-based checklist of the minimum standards of reporting required for specific types of research (Montgomery et al., 2018). The CONSORT guidelines have improved scientific reporting (e.g., Moher, Jones, & Lepage, 2001; Turner et al., 2012), and are now fully endorsed by over 600 journals (Montgomery et al., 2018). The Template for Intervention Description and Replication checklist (TIDieR) provides an important extension of CONSORT guidelines for studies involving an intervention (Hoffman et al., 2014). Developed in response to poor reporting quality of intervention studies (Hoffman et al., 2015), the TIDieR checklist includes 12 items considered as the bare minimum to describe an intervention sufficiently. Therefore, we encourage organisational behaviour and psychology scholars to engage with the CONSORT guidelines and TIDieR checklist to optimise reports of randomised controlled trials of team reflexivity interventions (and other areas of practice in the field). In so doing, researchers will be well positioned to reproduce intervention procedures, replicate results, effectively synthesise findings, and translate findings into practice.

Strengths, Limitations, and Future Research

Key strengths of this study include a focus on randomised trials or experiments as the primary source of evidence, robust methodological approach including statistical and narrative syntheses of intervention effectiveness and content, and adherence to a registered protocol and transparency regarding deviations from our intentions. Despite the comprehensive nature of our review and the multitude of benefits associated with meta-analyses and systematic reviews, it is important to keep in mind key limitations of our study when interpreting the findings. Our moderator tests relied on the characteristics available in the final set of eligible studies, which meant we were unable to examine several moderators proposed in our registered protocol. For example, we were unable to examine the differential effects of team reflexivity for newly formed versus

intact teams because the majority of interventions utilised newly formed teams (95.8%), which is an important consideration for teamwork training interventions (McEwan et al., 2017). There is a need for future research that tests the extent to which the effectiveness of team reflexivity training generalises across teams in varying stages of their developmental trajectory and other relevant characteristics (e.g., stability of team composition, human capital resources of individual members, horizontal versus vertical leadership). Occupational context is another important consideration for teamwork training interventions (McEwan et al., 2017). Due to the limited variability within eligible studies, we were only able to consider student samples against employees in general. The over reliance on student samples within eligible studies precluded us from being able to examine possible contextual variations in the effects of team reflexivity interventions. For example, team reflexivity interventions may be more or less effective in contexts where their use is commonplace (e.g., military and medical). We also were unable to extract data for each of the tested moderators from all eligible studies due to incomplete descriptions of interventions, thereby limiting the statistical power of these tests. As noted previously, researchers can utilise guidelines for the reporting of randomised controlled trials (e.g., CONSORT and TIDieR) to facilitate the planning and reporting of team development interventions. An additional limitation with regard to our moderation analyses was that some of the subgroups comprised unbalanced numbers, which meant that in some cases one of the groups included considerably more studies than the comparator (e.g., the student subgroup comprised 75% of studies for occupational context). These differences in subgroup sizes may affect the meta-analytical estimates as results may be subject to second-order sampling error (Schmidt & Hunter, 2015). Thus, we may have had reduced power to detect significant differences between groups due to the smaller sizes of some subgroups. Caution should be taken with the interpretation of some of the moderation results. Finally, sample sizes of primary studies varied considerably at both levels of analysis (individuals and teams), which meant that study effects were sometimes based on low numbers of teams (i.e., < 30). Statistical power for multilevel analyses, which are relevant for team factors, are complex because variance is

apportioned across two or more levels of analysis (e.g., time nested with individuals who are nested within teams). Simulation studies are considered the preferred approach for estimating statistical power for multilevel models (e.g., Arend & Schäfer, 2019; Lang, Bliese, & Runge, in press).

Conclusion

The findings of this systematic review and meta-analysis support the effectiveness of team reflexivity interventions and their robustness across most conditions and outcomes. There is evidence to suggest that the effectiveness of team reflexivity interventions can be enhanced with the use of feedback and facilitation, and when conducted face-to-face and over multiple time points. Collectively, these findings offer important practical implications for optimising team reflexivity interventions and recommendations for future research. As the environments in which teams operate are often highly pressurised and stressful, and demand high performance standards within a limited period, our findings suggest that organisations should integrate team reflexivity protocols within the experiential action and learning process. Although it may seem counterintuitive in an often time driven environment, providing teams with sufficient time and facilities to enable them to reflect effectively can help to maximise optimal performance.

References

Allen, N. J., & Hecht, T. D. (2004). The 'romance of teams': Toward an understanding of its psychological underpinnings and implications. *Journal of Occupational and Organizational Psychology*, 77, 439-461. Doi: 10.1348/0963179042596469

Allen, J. A., Reiter-Palmon, R., Crowe, J., & Scott, C. (2018). Debriefs: Teams learning from doing in context. *American Psychologist, 73,* 504-516. doi: 10.1037/amp0000246

Andersen, E. (2016). Enhancing the clinical reflective capacities of nursing students. *Nurse Education in Practice, 19,* 31–35. doi: 10.1016/j.nepr.2016.04.004

Anderson, E., Sandars, J., & Kinnair, D. (2019). The nature and benefits of team-based reflection on a patient death by healthcare professionals: A scoping review. *Journal of Interprofessional Care, 33,* 15-25. doi: 10.1080/13561820.2018.1513462

Archer, M. (2003). *Structure, agency and the internal conversation*. Cambridge: Cambridge University Press.

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1-19. doi: 10.1037/met0000195

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology, 88,* 989-1004. doi: 10.1037/0021-9010.88.6.989

Boet, S., Bould, M. D., Sharma, B., Revees, S., Naik, V. N., Triby, E., & Grantcharov, T. (2013). Within-team debriefing versus instructor-led debriefing for simulation-based education: a randomized controlled trial. *Annals of Surgery, 258*, 53-58. doi: 10.1097/SLA.0b013e31829659e4

Boet, S., Pigford, A. A., Fitzsimmons, A., Reeves, S., Triby, E., & Bould, M. D. (2016). Interprofessional team debriefings with or without an instructor after a simulated crisis scenario: An exploratory case study. *Journal of interprofessional care, 30,* 717-725. doi: 10.1080/13561820.2016.1181616

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley.

Breuer, C., Hüffmeier, J., & Hertel, G. (2016). Does trust matter more in virtual teams? A meta-analysis of trust and team effectiveness considering virtuality and documentation as moderators. *Journal of Applied Psychology*, *101*, 1151-1177. doi: 10.1037/apl0000113

Budescu, D. V., & Bruderman, M. (1995). The relationship between the illusion of control and the desirability bias. Journal of Behavioral Decision Making, 8, 109–125. doi: 10.1002/bdm.3960080204

*Calhoun, A. W., Sutton, E. R., Barbee, A. P., McClure, B., Bohnert, C., Forest, R., ... & Fallat, M. E. (2017). Compassionate Options for Pediatric EMS (COPE): Addressing communication skills. *Prehospital Emergency Care*, *21*, 334-343. doi: 10.1080/10903127.2016.1263370

Chamberland, C., Hodgetts, H. M., Kramer, C., Breton, E., Chiniara, G., & Tremblay, S. (2018). The critical nature of debriefing in high-fidelity simulation-based training for improving team communication in emergency resuscitation. *Applied Cognitive Psychology*, *32*, 727-738. doi: 10.1002/acp.3450

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211–229. doi: 10.1037/a0032968

Cheung, M. W. (2015). metaSEM: an R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*, 1521. doi:10.3389/fpsyg.2014.01521

Cheung, M. W. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*, 387-396. doi: 10.1007/s11065-019-09415-6

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), Perspectives on socially shared cognition (pp. 127–151). Washington, DC: American Psychological Association.

Cook, J. A., & Kautz, D. D. (2016). After action reviews in the emergency department: The positives of real-time feedback. *Journal of Emergency Nursing: JEN, 42,* 146–149. doi: 10.1016/j.jen.2015.10.008

*Coppens, I., Verhaeghe, S., Van Hecke, A., & Beeckman, D. (2018). The effectiveness of crisis resource management and team debriefing in resuscitation education of nursing students: A randomised controlled trial. *Journal of Clinical Nursing, 27,* 77-85. doi: 10.1111/jocn.13846

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology, 101*(8), 1134-1150. doi: 10.1037/apl0000110

*Diederich, E., Lineberry, M., Blomquist, M., Schott, V., Reilly, C., Murray, M., ... & Broski, J. (2019). Balancing deliberate practice and reflection: A randomized comparison trial of instructional designs for simulation-based training in cardiopulmonary resuscitation skills. *Simulation in Healthcare, 14,* 175-181. doi: 10.1097/SIH.0000000000000375

D'Innocenzo, L., Mathieu, J. E., & Kukenberger, M. R. (2016). A meta-analysis of different forms of shared leadership–team performance relations. *Journal of Management, 42,* 1964-1991. doi: 10.1177/0149206314525205

Dobbins, M. I., Thomas, S. A., Stokes Melton, S. L., & Lee, S. (2016). Integrated care and the evolution of the multidisciplinary team. *Primary Care, 43,* 177-190. doi: 10.1016/j.pop.2016.01.003

*Dyas, J. E. (2018). *Is deliberate knowledge sharing useful? Impacts of reflexivity, diversity, and motivation on team decision making* (Unpublished doctoral dissertation). The University of Alabama, Huntsville.

Eddy, E. R., Tannenbaum, S. I., & Mathieu, J. E. (2013). Helping teams to help themselves: Comparing two team-led debriefing methods. *Personnel Psychology, 66,* 975-1008. doi: 10.1111/peps.12041

Edmondson, A. C., Dillon, J. R., & Roloff, K. S. (2007). Three perspectives on team learning: Outcome improvement, task mastery, and group process. *Academy of Management Annals, 1*, 269–314.

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*: 629. doi: 10.1136/bmj.315.7109.629

Ellis, S., Mendel, R., & Nir, M. (2006). Learning from successful and failed experience: The moderating role of kind of after-event review. *Journal of Applied Psychology, 91,* 669–680. doi: 10.1037/0021-9010.91.3.669

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505. doi: 10.1126/science.1255484

Freedland, K.E., King, A.C., Ambrosius, W.T., Mayo-Wilson, E., Mohr, D.C., Czajkowski, S.M., … the National Institutes of Health Office of Behavioural and Social Science Research Expert Panel on Comparator Selection in Behavioural and Social Science Clinical Trials. (2019). The selection of comparators for randomised controlled trials of health-related behavioural interventions: Recommendations of an NIH expert panel. *Journal of Clinical Epidemiology, 110*, 74-81. doi: 10.1016/j.jclinepi.2019.02.011

*Gabelica, C., Van den Bossche, P., De Maeyer, S., Segers, M., & Gijselaers, W. (2014). The effect of team feedback and guided reflexivity on team performance change. *Learning and Instruction, 34*, 86-96. doi: 10.1016/j.learninstruc.2014.09.001

Gabelica, C., Van den Bossche, P., Segers, M., & Gijselaers, W. (2012). Feedback, a powerful lever in teams: A review. *Educational Research Review, 7*, 123-144. doi:10.1016/j.edurev.2011.11.003

Gabelica, C., Van den Bossche, P., Segers, M., & Gijselaers, W. (2014). Dynamics of Team Reflexivity after Feedback. *Frontline Learning Research, 2*, 64-91. doi: 10.14786/flr.v2i3.79

*Gass, M. A. & Priest, S. (2006). The effectiveness of metaphoric facilitation styles in corporate adventure training (CAT) programs. *Journal of Experiential Education, 29*, 78-94. doi: 10.1177/105382590602900107

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Gold, S.M., Enck, P., Hasselmann, H., Friede, T., Hegerl, U., … Otte, C. (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: A decision framework. *Lancet Psychiatry, 9*, 725–732. doi: 10.1016/ S2215-0366(17)30153-0

Goldenberg, A., Garcia, D., Halperin, E., & Gross, J. J. (2020). Collective emotions. *Current Directions in Psychological Science, 29*, 154-160. doi: 10.1177/0963721420901574

Gucciardi, D., Ntoumanis, N., Crane, M., Lines, R., & Pietsch, S. (2020, August 4). The effectiveness of team reflexivity interventions: Systematic review and meta-analysis of randomised controlled trials. Retrieved from https://osf.io/ruzy4/

*Gurtner, A., Tschan, F., Semmer, N. K., & Nägele, C. (2007). Getting groups to develop good strategies: Effects of reflexivity interventions on team process, team performance, and shared mental models. *Organizational Behavior and Human Decision Processes, 102*, 127-142. doi: 10.1016/j.obhdp.2006.05.002

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*: 924. doi: 10.1136/bmj.39489.470347.AD

Handke, L., Klonek, F. E., Parker, S. K., & Kauffeld, S. (2020). Interactive Effects of Team Virtuality and Work Design on Team Functioning. *Small Group Research, 51*, 3-47. doi: 10.1177/1046496419863490

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hands-on Guide*. DOI: 10.5281/zenodo.2551803.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., Welch, V. A. (Eds.) (2019). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0. Available from www.training.cochrane.org/handbook.

Higgins, J.P., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*: 557–560. doi: 10.1136/bmj.327.7414.557

Hoffmann, T.C., Glasziou, P.P., Boutron, I., Milne, R., Perera, R., Moher, D., et al. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*: g1687. doi: 10.1136/bmj.g1687

Hopewell, S., & Clarke, M. (2005). Abstracts presented at the American Society of Clinical Oncology conference: How completely are trials reported? *Clinical Trials, 2,* 265-268. doi: 10.1191/1740774505cn091oa

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From Input-Process-Output models to IMOI models. *Annual Review of Psychology, 56,* 517–543. doi:10.1146/annurev.psych.56.091103.070250

*Jarrett, S. M., Glaze, R. M., Schurig, I., Muñoz, G. J., Naber, A. M., McDonald, J. N., ... & Arthur Jr, W. (2016). The comparative effectiveness of distributed and colocated team after-action reviews. *Human Performance, 29*, 408-427. doi: 10.1080/08959285.2016.1208662

*Kim, P., Hong, J. S., Bonk, C., & Lim, G. (2011). Effects of group reflection variations in project-based learning integrated in a Web 2.0 learning space. *Interactive Learning Environments*, *19*, 333-349. doi: 10.1080/10494820903210782

Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.

*Konradt, U., Schippers, M. C., Garbers, Y., & Steenfatt, C. (2015). Effects of guided reflexivity and team feedback on team performance improvement: The role of team regulatory processes and cognitive emergent states. *European Journal of Work and Organizational Psychology*, *24*, 777-795. doi: 10.1080/1359432X.2015.1005608

Kossmeier, M., Tran, U. S., Voracek, M. (2020a). Conducting visual inference with funnel plots using *R* package metaviz. Retrieved from: https://cran.r-project.org/web/packages/metaviz/vignettes/funnelinf.html

Kossmeier, M., Tran, U. S., & Voracek, M. (2020b). Power-enhanced funnel plots for meta-analysis: The sunset funnel plot. *Zeitschrift für Psychologie, 228*, 43–49. https://doi.org/10.1027/2151-2604/a000392

*Kündig, P., Tschan, F., Semmer, N. K., Morgenthaler, C., Zimmermann, J., Holzer, E., ... & Marsch, S. (2020). More than experience: a post-task reflection intervention among team members enhances performance in student teams confronted with a simulated resuscitation task—a prospective randomised trial. *BMJ Simulation and Technology Enhanced Learning*, *6*, 1-6. doi: 10.1136/bmjstel-2018-000395

Lacerenza, C. N., Marlow, S. L., Tannenbaum, S. I., & Salas, E. (2018). Team development interventions: Evidence-based approaches for improving teamwork. *American Psychologist*, *73*, 517-531. doi: 10.1037/amp0000295

Lacerenza, C. N., Reyes, D. L., Marlow, S. L., Joseph, D. L., & Salas, E. (2017). Leadership training design, delivery, and implementation: A meta-analysis. *Journal of Applied Psychology*, *102*, 1686-1718. doi: 10.1037/apl0000241

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*: 418 doi:10.3389/fpsyg.2013.00863

Lang, J. W., Bliese, P. D., & Runge, J. M. (2019). Detecting Consensus Emergence in Organizational Multilevel Data: Power Simulations. *Organizational Research Methods*, 1-23. doi: 10.1177/1094428119873950

Lau, J., Ioannidis, J.P.A., Terrin, N., Schmid, C.H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ, 333*: 597. doi: 10.1136/bmj.333.7568.597

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*:e1000100. doi:10.1371/journal.pmed.1000100

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes, 144,* 145-170. doi: 10.1016/j.obhdp.2017.08.001

Marlow, S. L., Lacerenza, C. N., & Salas, E. (2017). Communication in virtual teams: A conceptual framework and research agenda. *Human Resource Management Review, 27,* 575-589. doi: 10.1016/j.hrmr.2016.12.005

McEwan, D., Ruissen, G. R., Eys, M. A., Zumbo, B. D., & Beauchamp, M. R. (2017). The effectiveness of teamwork training on teamwork behaviors and team performance: A systematic review and meta-analysis of controlled interventions. *PLoS ONE,* 12: e0169604. doi:10.1371/journal. pone.0169604

Menges, J. I., & Kilduff, M. (2015). Group emotions: Cutting the Gordian knots concerning terms, levels of analysis, and processes. The Academy of Management Annals, 9, 845–928. doi: 10 .1080/19416520.2015.1033148

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews, 4,* 1. doi: 10.1186/2046-4053-4-1

Mohr, D. C., Ho, J., Hart, T. L., Baron, K. G., Berendsen, M., Beckner, V., ... & Schroder, K. E. (2014). Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. *Translational Behavioral Medicine, 4,* 407-423. doi: 10.1007/s13142-014-0262-3

Moher, D., Jones, A., Lepage, L., & Consort Group. (2001). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA, 285*, 1992-1995.

Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Arean, P., Hollon, S. D., Ockene, J., & Kaplan, R. (2009). The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychotherapy and Psychosomatics, 78*, 275-284. doi: 10.1159/000228248

Moldjord, C., & Hybertsen, I. D. (2015). Training reflective processes in military aircrews through holistic debriefing: the importance of facilitator skills and development of trust. *International Journal of Training and Development, 19*, 287-300. doi: 10.1111/ijtd.12063

Myers, C., & Orndorff, D. (2013). Crew resource management: Not just for aviators anymore. *Journal of Applied Learning Technology, 3,* 44–48.

Otte, K. P., Konradt, U., Garbers, Y., & Schippers, M. C. (2017). Development and validation of the REMINT: a reflection measure for individuals and teams. *European Journal of Work and Organizational Psychology, 26*, 299-313. doi: 10.1080/1359432X.2016.1261826

*Peñarroja, V., Orengo, V., & Zornoza, A. (2017). Reducing perceived social loafing in virtual teams: The effect of team feedback with guided reflexivity. *Journal of Applied Social Psychology, 47*, 424-435. doi: 10.1111/jasp.12449

*Phielix, C., Prins, F. J., & Kirschner, P. A. (2010). Awareness of group performance in a CSCL-environment: Effects of peer feedback and reflection. *Computers in Human Behavior, 26*, 151-161. doi: 10.1016/j.chb.2009.10.011

*Phielix, C., Prins, F. J., Kirschner, P. A., Erkens, G., & Jaspers, J. (2011). Group awareness of social and cognitive performance in a CSCL environment: Effects of a peer feedback and reflection tool. *Computers in Human Behavior, 27*, 1087-1102. doi: 10.1016/j.chb.2010.06.024

*Pieterse, A. N., Van Knippenberg, D., & van Ginkel, W. P. (2011). Diversity in goal orientation, team reflexivity, and team performance. *Organizational Behavior and Human Decision Processes, 114,* 153-164. doi:10.1016/j.obhdp.2010.11.003

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86,* 207-236. doi: 10.3102/0034654315582067

Quinn, R. W., & Bunderson, J. S. (2016). Could we huddle on this project? Participant learning in newsroom conversations. *Journal of Management, 42,* 386–418. doi: 10.1177/0149206313484517

R Development Core Team. (2019). *R: A language and environment for statistical computing*. Vienna: Austria. Retrieved from: http://www.r-project.org/

Salas, E., Klein, C., King, H., Salisbury, M., Augenstein, J. S., Birnbach, D. J., ... & Upshaw, C. (2008). Debriefing medical teams: 12 evidence-based best practices and tips. *The Joint Commission Journal on Quality and Patient Safety, 34,* 518-527. doi: 10.1016/S1553-7250(08)34066-5

Schaubroeck, J. M., & Yu, A. (2017). When does virtuality help or hinder teams? Core team characteristics as contingency factors. *Human Resource Management Review, 27,* 635-647. doi: 10.1016/j.hrmr.2016.12.009

Schippers, M. C., Edmondson, A. C., & West, M. A. (2014). Team reflexivity as an antidote to team information-processing failures. *Small Group Research, 45,* 731-769. doi: 10.1177/1046496414553473

Schippers, M. C., Homan, A. C., & van Knippenberg, D. (2013). To reflect or not to reflect: Prior team performance as a boundary condition of the effects of reflexivity on learning and final team performance. *Journal of Organizational Behavior, 34,* 6-23. doi: 10.1002/job.1784

Schippers, M. C., West, M. A., & Edmondson, A. C. (2017). Team reflexivity and innovation. In E. Salas, R. Rico, & J. Passmore (Eds.), *The Wiley Blackwell handbook of the psychology of team working and collaborative processes* (pp. 459-478). Wiley.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks: Sage.

Schmutz, J. B., Kolbe, M., & Eppich, W. J. (2018). Twelve tips for integrating team reflexivity into your simulation-based team training. *Medical Teacher*, *40*, 721-727. Doi: 10.1080/0142159X.2018.1464135

Schmutz, J., Lei, Z., Eppich, W., & Manser, T. (2018). Reflection in the heat of the moment: The role of in-action team reflexivity in healthcare emergency teams. *Journal of Organizational Behavior, 39(6),* 749–765. doi: 10.1002/job.2299

*Schurig, I. A. (2012). *An investigation of the effect of after-action reviews on teams' performance-efficacy relationships* (Unpublished doctoral dissertation). Texas A&M University, Texas.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, *349,* g7647. doi: 10.1136/bmj.g7647

Shuffler, M. L., Diazgranados, D., Maynard, M. T., & Salas, E. (2018). Developing, sustaining, and maximizing team effectiveness: An integrative, dynamic perspective of team development interventions. *Academy of Management Annals*, *12*, 688-724. doi: 10.5465/annals.2016.0045

Shuffler, M. L., Pavlas, D., & Salas, E. (2012). Teams in the military: A review and emerging challenges. In J. H. Laurence & M. D. Matthews (Eds.), *The Oxford handbook of military psychology* (pp. 282-310). Oxford: Oxford University Press.

Sinclair, H., Doyle, E. E., Johnston, D. M., & Paton, D. (2012). Assessing emergency management training and exercises. *Disaster Prevention and Management: An International Journal, 21,* 507–521. doi: 10.1108/09653561211256198

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews, 1:* 7. doi: 10.1186/2046-4053-1-7

Sundstrom, R., de Meuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45,* 120–133. doi: 10.1037/0003-066X.45.2.120

Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., ... & Emberson, J. R. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. doi: 10.1136/bmj.l4898

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors*, *55*, 231-245. doi: 10.1177/0018720812448394

*Tesler, R. (2010). *The effects of storytelling and reflexivity on team mental models and performance in distributed decision-making teams* (Unpublished master's thesis). Pennsylvania State University, Pennsylvania.

Tsifetakis, E., & Kontogiannis, T. (2019). Evaluating non-technical skills and mission essential competencies of pilots in military aviation environments. *Ergonomics*, *62*, 204-218. doi: 10.1080/00140139.2017.1332393

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., ... & Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews, 11*. doi: 10.1002/14651858.MR000030.pub2

*van der Kleij, R., & Hoeppermans, M. (2011). How performance feedback and reflection affect transactive memory. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55, 311-315. doi: 10.1177/1071181311551065

*van Ginkel, W., Tindale, R. S., & van Knippenberg, D. (2009). Team reflexivity, development of shared task representations, and the use of distributed information in group decision

making. *Group Dynamics: Theory, Research, and Practice*, *13*, 265-280. doi: 10.1037/a0016045

*Vashdi, D., R. (2013). Teams in public administration: A field study of team feedback and effectiveness in the Israeli public healthcare system. *International Public Management Journal*, *16*, 275-306. doi: 10.1080/10967494.2013.817255

*Vashdi, D. R., Bamberger, P. A., & Erez, M. (2013). Can surgical teams ever learn? The role of coordination, complexity, and transitivity in action team learning. *Academy of Management Journal*, *56*, 945-971. doi: 10.5465/amj.2010.0501

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

*Villado, A. J. (2008). *The after-action review training approach: An integrative framework and empirical investigation* (Unpublished doctoral dissertation). Texas A&M University, Texas.

*Villado, A. J., & Arthur Jr, W. (2013). The comparative effect of subjective and objective after-action reviews on team performance on a complex task. *Journal of Applied Psychology*, *98*, 514-528. doi: 10.1037/a0031510

Volini, E., Schwartz, J., Roy, I., Hauptmann, M., Van Durme, Y., Denny, B., & Bersin, J. (2019). Deloitte Global Human Capital Trends

*Weger, K. (2017). *A classification system of behavior indicators exhibited in team communication: The intricate role of team reflexivity with team cognition, regulatory processes and performance* (Unpublished doctoral dissertation). Otto-Friedrich-Universität Bamberg, Bamberg.

West, M.A. (2000). Reflexivity, revolution and innovation in work teams. In M.M.Beyerlein, D.A. Johnson, & S.T. Beyerlein (Eds.), *Product development teams* (Vol. 5, pp. 1–29). Stamford, CT: JAI Press.

Widmer, P. S., Schippers, M. C., & West, M. A. (2009). Recent developments in reflexivity research: A review. *Psychology of Everyday Activity*, *2*, 2-11.

Wiese, C. W., Shuffler, M. L., & Salas, E. (2015). Teamwork and team performance measurement. In J. Wright (Ed.), International encyclopedia of the social & behavioral sciences (pp. 96–103). Oxford: Pergamon.

Table 1. *Characteristics of Studies Included in the Meta-Analysis and Narrative Review.*

Study	Sample Size	Total Teams	Percentage Female	Mean Age	Country	Occupational Context	Outcomes	Type of Measurement	Type of Outcome	Effect Size (Hedges *g*)
Calhoun et al. (2017)	148	38	0.26	-	USA	EMS providers	Team Performance - communication	Subjective	Performance	1.88
Coppens, Verhaeghe, Van Hecke, & Beeckman (2018)	116	30	0.78	-	Belgium	Medical Students	Team-Efficacy	Subjective	Cognitive	0.64
							Teamwork	Researcher Assessed	Behavioural	0.09
							Self-Efficacy	Subjective	Cognitive	0.31
							Technical Skills	Objective	Performance	-0.09
Diederich et al. (2019)	131	26	-	-	USA	Pre-residency Medics	Compression Quality	Objective	Performance	0.36
							Use of compression adjuncts	Researcher Assessed	Performance	-0.66
							Defibrillation	Objective	Performance	-0.13
							Airway management	Researcher Assessed	Performance	-0.30
Dyas (2018) a[1]	42	14	0.45	39.14	USA	Employees	Task Time	Objective	Performance	0.12
							Task Accuracy	Objective	Performance	-0.07
							Knowledge Sharing	Subjective	Cognitive	-0.26
Dyas (2018) b[1]	216	72	0.54	20.64	USA	University Students	Task Time	Objective	Performance	-0.09
							Task Accuracy	Objective	Performance	-0.20
							Knowledge Sharing	Subjective	Cognitive	-0.07
*Gabelica et al. (2014)	212	106	0.48	22.3	Netherlands	University Students	-	-	-	-
*Gass & Priest (2006)	106	5	-	-	-	Banking Workers	-	-	-	-
Gurtner, Tschan, Semmer, & Nägele (2007)	147	49	0.66	25.4	Switzerland	University Students	Performance	Objective	Performance	0.19
							Team interaction mental models	Subjective	Cognitive	1.24
Jarrett et al. (2016) a[2]	251	63	0.52	18.63	USA	University Students	Team Performance	Objective	Performance	1.41
							Team Efficacy	Subjective	Cognitive	1.04

							Openness of communication	Subjective	Cognitive	1.55
							Cohesion	Subjective	Cognitive	1.43
Jarrett et al. (2016) b[2]	240	60	0.43	19.06	USA	University Students	Team Performance	Objective	Performance	0.57
							Team Efficacy	Subjective	Cognitive	0.48
							Openness of communication	Subjective	Cognitive	0.44
							Cohesion	Subjective	Cognitive	0.53
Kim, Hong, Bonk, & Lim (2011)	38	12	-	-	-	University Students	Team Effectiveness	Subjective	Cognitive	0.35
							Performance	Objective	Performance	1.61
							Participation	Objective	Behavioural	0.09
*Konradt, Schippers, Garbers, & Steenfatt (2015)	294	98	0.68	22.76	-	University Students	-	-	-	-
Kündig et al. (2020)	168	56	0.7	24.44	-	Medical Students	Hands-on Performance	Objective	Performance	0.51
							Coordination performance	Researcher Assessed	Performance	0.39
							Defibrillation performance	Researcher Assessed	Performance	-0.10
Peñarroja, Orengo, & Zornoza (2017)	212	54	0.8	23.91	-	University Students	Perceived Social Loafing	Subjective	Cognitive	0.38
							Group Cohesion	Subjective	Cognitive	0.55
							Satisfaction with team	Subjective	Affective	-0.07
							Satisfaction with the result	Subjective	Affective	-0.37
Phielix, Prins, Kirschner (2010)	39	10	0.51	15.54	Netherlands	High School Students	Cognitive Performance	Researcher Assessed	Performance	-0.37
Phielix et al. (2011)	108	38	0.46	15.85	Netherlands	High School Students	Cognitive Performance	Researcher Assessed	Performance	0.35
							Team Development	Subjective	Cognitive	0.19
							Group-Process Satisfaction	Subjective	Affective	0.34

							Intra-Group conflicts	Subjective	Cognitive	0.23
							Collaborative problem solving	Subjective	Affective	0.23
*Pieterse, Van Knippenberg, & van Ginkel (2011)	147	49	0.33	20	Netherlands	University Students	-	-	-	-
Schurig (2012)	492	123	0.47	18.84	USA	University Students	Team Performance	Objective	Performance	0.97
							Team-efficacy	Subjective	Cognitive	0.77
Tesler (2010)	321	107	0.49		USA	University Students	Positive affect	Subjective	Affective	0.20
							Negative Affect	Subjective	Affective	-0.10
van der Kleij & Hoeppermans (2011)	102	34	0.58	23.6	Netherlands	University Students	Team Performance	Objective	Performance	0.65
							Transactive memory	Subjective	Cognitive	0.67
Van Ginkel, Tindale, & van Knippenberg (2009)	252	84	0.72	18.8	USA	University Students	Decision Quality	Objective	Performance	0.92
							Task representations	Subjective	Cognitive	0.87
							Information elaboration	Researcher Assessed	Behavioural	1.12
Vashdi (2013)	-	174	-	-	Israel	Surgical teams	Relative Duration	Objective	Performance	0.44
							Attention to Detail	Subjective	Cognitive	1.30
							Cooperation	Subjective	Cognitive	1.02
							Psychological Safety	Subjective	Affective	0.88
Vashdi, Bamberger, & Erez (2013)	-	362	-	-	Israel	Surgical teams	Relative Duration	Objective	Performance	0.36
							Team Workload Sharing	Subjective	Cognitive	0.54
							Team Helping	Subjective	Cognitive	0.02
Villado (2008)	120	30	0.39	19.08	USA	University Students	Team Performance	Objective	Performance	0.03
							Declarative knowledge	Subjective	Cognitive	0.79
							Team Efficacy	Subjective	Cognitive	0.72
Villado & Arthur (2013)	188	47	0.45	18.93	USA	University Students	Team Performance	Objective	Performance	1.96
							Declarative knowledge	Subjective	Cognitive	-0.25

							Team Efficacy	Subjective	Cognitive	1.27
							Openness of communication	Subjective	Cognitive	1.32
							Cohesion	Subjective	Cognitive	1.29
Weger (2017)	248	62	0.56	22.2	USA	University Students & Employees	Team Performance	Objective	Performance	0.50

Note: - = information not provided; * Studies included in narrative review only; [1] Study included two samples within the same paper; [2] Study was split based upon virtuality (assessed reflection vs control in both distributed and co-located teams).

Table 2. *Results for Overall Mean Effect Sizes.*

Outcome	#Studies	#ES	Mean *g* (SE)	95% CI	t-statistic	*p*	Level 1 variance (%)	Level 2 variance (%)	Level 3 variance (%)
Performance Total	20	66	.504 (.102)	0.301, 0.707	4.956	< .001	27.71	26.92***	45.37***
Performance Outcomes	18	27	.549 (.148)	0.246, 0.853	3.719	< .001	21.76	8.18	70.06*
Performance Behaviours	15	39	.548 (.111)	0.324, 0.772	4.956	< .001	30.33	34.81**	34.86*
Cognitive Outcomes	14	30	.655 (.103)	0.444, 0.866	6.351	< .001	35.84	49.33**	14.83

Note: #Studies = number of studies; #ES = number of effect sizes; Mean *g* = mean effect size; SE = standard error; CI = confidence interval; *p* = significance of mean effect size; Level 2 variance = percentage variance between effect sizes from the same study; Level 3 Variance = percentage variance in effect sizes between studies; *** = p < .001, ** = p < .01, * = p < .05.

Table 3. *Moderator and Sensitivity Analyses of the Overall Effect of Team Reflexivity Interventions.*

Moderator	#Studies	#ES	Mean *g* (95%CI)	β (95% CI)	Overall[a]	*p*[b]	Level 2 variance	Level 3 variance
Study Characteristics								
Sample Size	20	66	.505 (0.298, 0.713)***	-0.000 (-0.004, 0.004)	F(1, 64) = 0.005	.943	27.29***	45.73***
Team Size	19	60	.519 (0.306, 0.732)***	-0.052 (-0.290, 0.186)	F(1, 58) = 0.188	.666	29.70***	43.84***
Age	13	45	.532 (0.297, 0.768)***	-0.014 (-0.052, 0.024)	F(1, 43) = 0.529	.471	35.56***	35.30*
Impact Factor	13	47	.582 (0.294, 0.870)***	-0.010 (-0.234, 0.213)	F(1, 45) = 0.009	.925	32.66***	45.90**
Percentage Female	16	52	.553 (0.323, 0.782)***	-1.224 (-2.821, 0.373)	F(1, 50) = 2.368	.130	29.22**	43.33**
Publication type					F(1, 64) = 0.927	.339	27.29***	44.97***
Peer-Reviewed (RC)	15	52	.561 (0.326, 0.796)***					
Dissertation	5	14	.337 (-0.065, 0.738)	-0.224 (-0.689, 0.241)				
Outcome Characteristics								
Outcome measurement					F(2, 63) = 1.004	.372	29.93***	41.92***
Self-report (RC)	16	37	.556 (0.328, 0.784)***					
Objective	15	21	.507 (0.249, 0.764) ***	-0.049 (-0.309, 0.210)				
Research assessed	6	8	.257 (-0.149, 0.663)	-0.299 (-0.721, 0.123)				
Performance outcome					F(1, 64) = 0.080	.778	27.92***	44.64***
No (RC)	15	39	.519 (0.289, 0.750)***					
Yes	18	27	.484 (0.238, 0.730)***	-0.035 (-0.282, 0.212)				
Type of outcome					F(3, 62) = 1.550	.210	30.09***	39.58**
Cognitive (RC)	14	30	.599 (0.366, 0.831)***					
Behavioural	3	3	.530 (-0.069, 1.128)	-0.069 (-0.674, 0.537)				
Affective	3	6	.103 (-0.348, 0.553)	-0.496 (-0.961, -0.031)*				
Performance	18	27	.507 (0.272, 0.742)***	-0.092 (-0.344, 0.160)				
Intervention Characteristics								
Reflection Time	13	44	.539 (0.244, 0.834)***	0.005 (-0.009, 0.020)	F(1, 42) = 0.526	.472	25.28**	53.02***
Facilitator					F(1, 64) = 0.028	.868	25.79***	47.61***
Present (RC)	7	31	.482 (0.142, 0.823)**					
Absent	13	35	.518 (0.252, 0.784)***	0.036 (-0.396, 0.468)				
Handouts					F(1, 64) = 0.010	.920	25.97***	47.38***
Present (RC)	7	24	.518 (0.183, 0.852)**					
Absent	13	42	.496 (0.228, 0.764)***	-0.022 (-0.450, 0.407)				
Total Facilitation					F(1, 64) = 0.574	.451	26.15***	46.78***
Present (RC)	12	47	.448 (0.194, 0.703)***					
Absent	8	19	.614 (0.258, 0.970) ***	0.166 (-0.271, 0.603)				
Feedback					F(1, 64) = 2.791	.100	29.94***	40.14**
Present (RC)	8	29	.695 (0.396, 0.993)***					

Absent	12	37	.371 (0.124, 0.618)**	-0.324 (-0.711, 0.063)				
Occupational Context					F(1, 63) = 0.013	.910	25.97***	48.05***
Students (RC)	15	50	.500 (0.260, 0.739)***					
Employed	5	15	.525 (0.117, 0.933)*	0.025 (-0.422, 0.472)				
Team Virtuality					F(1, 64) = 8.914	**.004**	16.76*	55.28***
Present (RC)	8	22	.166 (-0.145, 0.476)					
Absent	13	44	.678 (0.436, 0.920)***	0.512 (0.169, 0.855)**				
Comparator					F(1, 64) = 0.032	.858	25.92***	47.43***
Active (RC)	9	27	.483 (0.168, 0.799)**					
Control	11	39	.521 (0.242, 0.801)	0.038 (-0.384, 0.459)				

Note: #Studies = number of studies, #ES = number of effect sizes, Mean *g* = mean effect size, CI = confidence interval, β = estimated regression coefficient, Level 2 variance = variance in effect sizes from the same study, Level 3 variance = variance in effect sizes between studies, RC = reference category, *** = p < .001, ** = p < .01, * = p < .05, [a] Omnibus test of all regression coefficients in the model, [b] p-value of omnibus test.

Table 4. *Moderator and Sensitivity Analyses of the Effect of Team Reflexivity on Performance Outcomes.*

Moderator	# Studies	#ES	Mean *g* (95%CI)	β (95% CI)	Overall[a]	*p*[b]	Level 2 variance	Level 3 variance
Study Characteristics								
Sample Size	18	27	.551 (0.229, 0.873)**	0.000 (-0.005, 0.005)	F(1, 25) = 0.000	.987	8.15	71.54*
Team Size	17	24	.549 (0.243, 0.855)**	-0.135 (-0.457, 0.187)	F(1, 22) = 0.752	.395	18.47	58.75
Age	12	18	.554 (0.199, 0.909)**	-0.005 (-0.053, 0.042)	F(1, 16) = 0.056	.815	15.06	59.97
Impact Factor	12	16	.644 (0.158, 1.131)*	0.079 (-0.302, 0.461)	F(1, 14) = 0.199	.662	19.00	66.80
Percentage Female	14	20	.598 (0.260, 0.936)**	-2.235 (-4.685, 0.215)	F(1, 18) = 3.674	.071	9.05	67.11*
Publication type					F(1, 25) = 0.572	.457	8.18	71.11*
Peer-Reviewed (RC)	14	20	.616 (0.258, 0.973)**					
Dissertation	4	7	.350 (-0.277, 0.978)	-0.265 (-0.987, 0.457)				
Outcome Characteristics								
Outcome measurement					F(2, 24) = 5.253	**.013**	4.04	65.90*
Self-report (RC)	1	1	1.882 (0.670, 3.094)**					
Objective	15	20	.559 (0.283, 0.835)***	-1.323 (-2.566, -0.081)*				
Research assessed	4	6	.054 (-0.418, 0.526)	-1.828 (-3.129, -0.528)**				
Intervention Characteristics								
Reflection Time	12	21	.606 (0.162, 1.051)*	0.005 (-0.018, 0.028)	F(1, 19) = 0.211	.652	5.95	78.74**
Facilitator					F(1, 25) = 0.113	.739	7.57	72.09*
Present(RC)	6	10	.627 (0.068, 1.186)*					
Absent	12	17	.516 (0.135, 0.898)**	-0.111 (-0.788, 0.566)				
Handouts					F(1, 25) = 0.160	.693	7.37	72.22*
Present (RC)	6	11	.631 (0.114, 1.149)*					
Absent	12	16	.505 (0.109, 0.901)*	-0.126 (-0.778, 0.525)				
Total Facilitation					F(1, 25) = 0.007	.936	7.67	72.07*
Present (RC)	10	19	.562 (0.150, 0.975)**					
Absent	8	8	.537 (0.047, 1.027)*	-0.025 (-0.666, 0.615)				
Feedback					F(1, 25) = 3.406	.077	12.12	62.34
Present (RC)	7	8	.881 (0.410, 1.351)***					
Absent	11	19	.357 (0.011, 0.704)*	-0.524 (-1.108, 0.061)				
Occupational Context					F(1, 24) = 0.004	.949	8.29	72.47*
Students (RC)	13	17	.551 (0.177, 0.924)**					
Employed	5	9	.569 (0.030, 1.109)*	0.019 (-0.571, 0.608)				
Team Virtuality					F(1, 25) = 2.452	.130	1.80	80.11**
Present (RC)	6	6	.239 (-0.289, 0.766)					
Absent	13	21	.666 (0.315, 1.018)***	0.428 (-0.135, 0.990)				
Comparator					F(1, 25) = 0.000	.983	7.65	72.07*

Active (RC)	8	14	.548 (0.082, 1.014)*	
Control	10	13	.555 (0.126, 0.983)*	0.007 (-0.627, 0.640)

Note: #Studies = number of studies, #ES = number of effect sizes, Mean *g* = mean effect size, CI = confidence interval, β = estimated regression coefficient, Level 2 variance = variance in effect sizes from the same study, Level 3 variance = variance in effect sizes between studies, RC = reference category, *** = p < .001, ** = p < .01, * = p < .05, [a] Omnibus test of all regression coefficients in the model, [b] p-value of omnibus test.

Table 5. *Moderator and Sensitivity Analyses of the Effect of Team Reflexivity on Performance Behaviours.*

Moderator	# Studies	#ES	Mean *g* (95%CI)	β (95% CI)	Overall[a]	*p*[b]	Level 2 variance	Level 3 variance
Study Characteristics								
Sample Size	15	39	.554 (0.319, 0.788)***	-0.001 (-0.008, 0.006)	F(1, 37) = 0.058	.811	34.05**	36.67*
Team Size	14	36	.574 (0.339, 0.809)***	0.081 (-0.197, 0.359)	F(1, 34) = 0.349	.559	37.93***	32.46
Age	10	27	.575 (0.290, 0.860)***	-0.029 (-0.089, 0.031)	F(1, 25) = 0.992	.329	47.94***	23.91
Impact Factor	10	31	.620 (0.343, 0.898)***	-0.046 (-0.244, 0.151)	F(1, 29) = 0.232	.634	38.91***	33.70
Percentage Female	12	32	.549 (0.284, 0.813)***	-0.383 (-2.248, 1.482)	F(1, 30) = 0.176	.678	37.97**	34.38
Publication type					F(1, 37) = 1.399	.245	36.23**	32.68
Peer-Reviewed (RC)	11	32	.620 (0.369, 0.872)***					
Dissertation	4	7	.323 (-0.121, 0.766)	-0.298 (-0.808, 0.212)				
Outcome Characteristics								
Outcome measurement					F(2, 36) = 0.189	.829	36.63***	33.83*
Self-report (RC)	15	36	.552 (0.320, 0.784)***					
Objective	1	1	.129 (-1.345, 1.602)	-0.424 (-1.910, 1.062)				
Research assessed	2	2	.622 (-0.096, 1.339)	0.069 (-0.655, 0.794)				
Type of outcome					F(2, 36) = 3.549	**.039**	42.77**	19.75
Cognitive (RC)	14	30	.651 (0.444, 0.857)***					
Behavioural	3	3	.579 (-0.021: 1.180)	-0.071 (-0.692, 0.549)				
Affective	3	6	.070 (-0.341, 0.481)	-0.580 (-1.022, -0.138)*				
Intervention Characteristics								
Reflection Time	9	23	.562 (0.281, 0.844)***	0.010 (-0.004, 0.024)	F(1, 21) = 2.016	.170	53.37**	17.02
Facilitator					F(1, 37) = 0.025	.876	33.47**	37.52*
Present(RC)	6	21	.569 (0.217, 0.920)**					
Absent	9	18	.532 (0.224, 0.841)**	-0.036 (-0.504, 0.431)				
Handouts					F(1, 37) = 0.009	.927	33.46**	37.60*
Present (RC)	6	13	.561 (0.196, 0.926)**					
Absent	9	26	.539 (0.239, 0.840)***	-0.022 (-0.495, 0.452)				
Total Facilitation					F(1, 37) = 0.239	.628	33.08**	37.93*
Present (RC)	10	28	.511 (0.233, 0.789)***					
Absent	5	11	.633 (0.210, 1.057)**	0.122 (-0.384, 0.628)				
Feedback					F(1, 37) = 1.436	.238	34.71**	34.52*
Present (RC)	8	21	.682 (0.365, 1.000)***					
Absent	7	18	.419 (0.109, 0.730)**	-0.263 (-0.707, 0.182)				
Occupational Context					F(1, 37) = 0.001	.971	33.77**	37.18*
Students (RC)	13	33	.546 (0.294, 0.798)***					
Employed	3	6	.557 (0.014, 1.100)*	0.011 (-0.581, 0.602)				

Team Virtuality					F(1, 37) = 7.807	**.008**	28.17*	35.93
Present (RC)	6	16	.225 (-0.087, 0.537)					
Absent	10	23	.733 (0.485, 0.980)***	0.508 (0.140, 0.876)**				
Comparator					F(1, 37) = 0.236	.630	33.83**	36.90*
Active (RC)	6	13	.625 (0.229, 1.022)**					
Control	9	26	.509 (0.226, 0.791)***	-0.117 (-0.604, 0.370)				

Note: #Studies = number of studies, #ES = number of effect sizes, Mean *g* = mean effect size, CI = confidence interval, β = estimated regression coefficient, Level 2 variance = variance in effect sizes from the same study, Level 3 variance = variance in effect sizes between studies, RC = reference category, *** = p < .001, ** = p < .01, * = p < .05, [a] Omnibus test of all regression coefficients in the model, [b] p-value of omnibus test.

Table 6. *Moderator and Sensitivity Analyses of the Effect of Team Reflexivity on Cognitive Outcomes.*

Moderator	# Studies	#ES	Mean *g* (95%CI)	β (95% CI)	Overall[a]	*p*[b]	Level 2 variance	Level 3 variance
Study Characteristics								
Sample Size	14	30	.657 (0.437, 0.876)***	-0.001 (-0.008 0.007)	F(1, 28) = 0.034	.854	49.12**	16.37
Team Size	13	28	.698 (0.482, 0.915)***	0.034 (-0.217, 0.285)	F(1, 26) = 0.079	.781	57.69***	5.20
Age	10	22	.654 (0.405, 0.903)***	-0.031 (-0.090, 0.027)	F(1, 20) = 1.238	.279	66.56***	2.00
Impact Factor	10	24	.735 (0.514, 0.957)***	-0.071 (-0.217, 0.076)	F(1, 22) = 1.003	.327	60.25***	3.84
Percentage Female	11	24	.669 (0.432, 0.906)***	-0.344 (-2.079, 1.390)	F(1, 22) = 0.169	.685	58.60**	7.70
Publication type					F(1, 28) = 1.153	.292	53.42**	10.36
Peer-Reviewed (RC)	11	25	.711 (0.486, 0.937)***					
Dissertation	3	5	.432 (-0.051, 0.915)	-0.280 (-0.813, 0.254)				
Intervention Characteristics								
Reflection Time	8	19	.595 (0.284, 0.905)***	0.011 (-0.004, 0.027)	F(1, 17) = 2.396	.140	66.15***	8.38
Facilitator					F(1, 28) = 0.600	.445	51.78**	12.55
Present(RC)	6	17	.736 (0.443, 1.029)***					
Absent	8	13	.578 (0.280, 0.876***	-0.158 (-0.576, 0.260)				
Handouts					F(1, 28) = 0.003	.959	47.54**	18.37
Present (RC)	5	10	.644 (0.275, 1.013)**					
Absent	9	20	.656 (0.380, 0.932)***	0.012 (-0.449, 0.473)				
Total Facilitation					F(1, 28) = 0.000	.990	47.56**	18.30
Present (RC)	9	21	.653 (0.384, 0.922)***					
Absent	5	9	.650 (0.263, 1.037)**	-0.003 (-0.474, 0.468)				
Feedback					F(1, 28) = 2.823	.104	54.47**	6.78
Present (RC)	8	18	.803 (0.548, 1.057)***					
Absent	6	12	.483 (0.188, 0.778)**	-0.320 (-0.709, 0.070)				
Occupational Context					F(1, 28) = 0.126	.726	47.85**	17.83
Students (RC)	12	24	.672 (0.425, 0.918)***					
Employed	3	6	.580 (0.108, 1.052)*	-0.092 (-0.621, 0.438)				
Team Virtuality					F(1, 28) = 3.106	.089	35.77*	28.34
Present (RC)	5	9	.379 (-0.003, 0.761)					
Absent	10	21	.746 (0.490, 1.002)***	0.368 (-0.060, 0.795)				
Comparator					F(1, 28) = 1.735	.198	50.49**	12.20
Active (RC)	5	10	.858 (0.487, 1.229)***					
Control	9	20	.572 (0.328, 0.816)	-0.286 (-0.730, 0.159)				

Note: #Studies = number of studies, #ES = number of effect sizes, Mean g = mean effect size, CI = confidence interval, β = estimated regression coefficient, Level 2 variance = variance in effect sizes from the same study, Level 3 variance = variance in effect sizes between studies, RC = reference category, *** = p < .001, ** = p < .01, * = p < .05, [a] Omnibus test of all regression coefficients in the model, [b] p-value of omnibus test.

Table 7. *GRADE Summary of Findings Table.*

Outcomes	Certainty Assessment						Summary of Findings			
	Number of studies (#ES)	Risk Of Bias	Inconsistency	Indirectness	Imprecision	Other Considerations	Team Reflexivity Condition	Comparator Condition	Effect (95% CI)	Certainty
Overall Effect	20 (66)	Serious [a]	Serious [b]	Not Serious	Not Serious	None	2047/3572 (57.3%)	1525/3572 (42.7%)	.504 (.301 to .707)	LOW
Performance Outcomes	18 (27)	Serious [a]	Serious [c]	Not Serious	Not Serious	None	792/1444 (54.8%)	652/1444 (45.2%)	.549 (.246 to .853)	LOW
Performance Behaviours	15 (39)	Serious [a]	Serious [d]	Not Serious	Not Serious	None	1255/2128 (59%)	873/2128 (41%)	.548 (.324 to .772)	LOW
Cognitive Behaviours	14 (30)	Serious [a]	Serious [e]	Not Serious	Not Serious	None	966/1606 (60.1%)	640/1606 (39.9%)	.655 (.444 to .866)	LOW

Note: #ES = Number of Effect Sizes; CI = Confidence Interval.

[a] A majority of the eligible studies had a risk of bias rating of some concerns (see supplementary table 1; https://osf.io/ruzy4/).

[b] Substantial heterogeneity among effect sizes (I2 = 72.3%).

[c] Substantial heterogeneity among effect sizes (I2 = 78.3%).

[d] Substantial heterogeneity among effect sizes (I2 = 69.7%).

[e] Substantial heterogeneity among effect sizes (I2 = 64.1%).

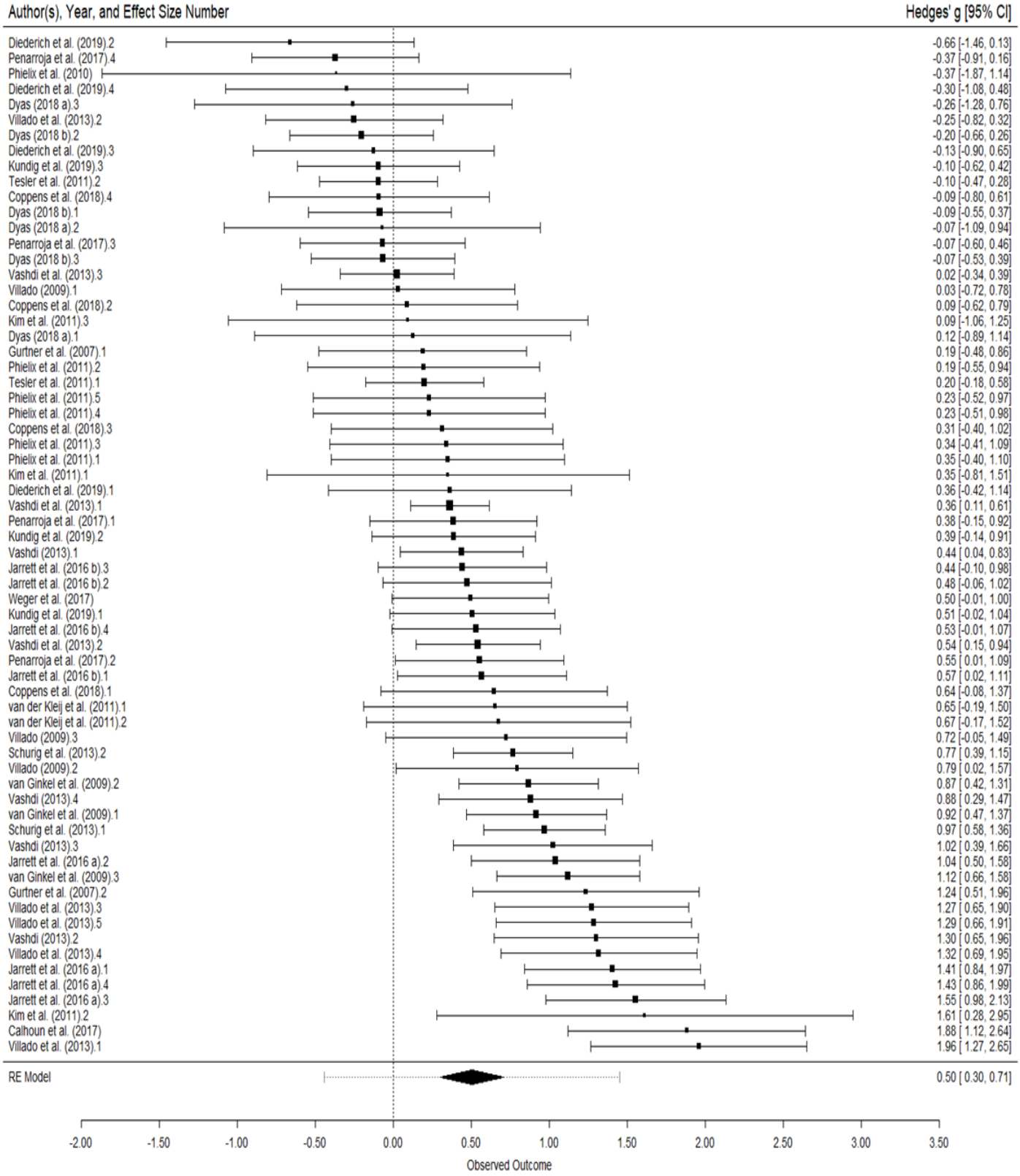

Figure 1. Forest plot of the overall effect of team reflexivity intervention.
Note: Dotted interval around the random effects model represents 95% prediction interval.

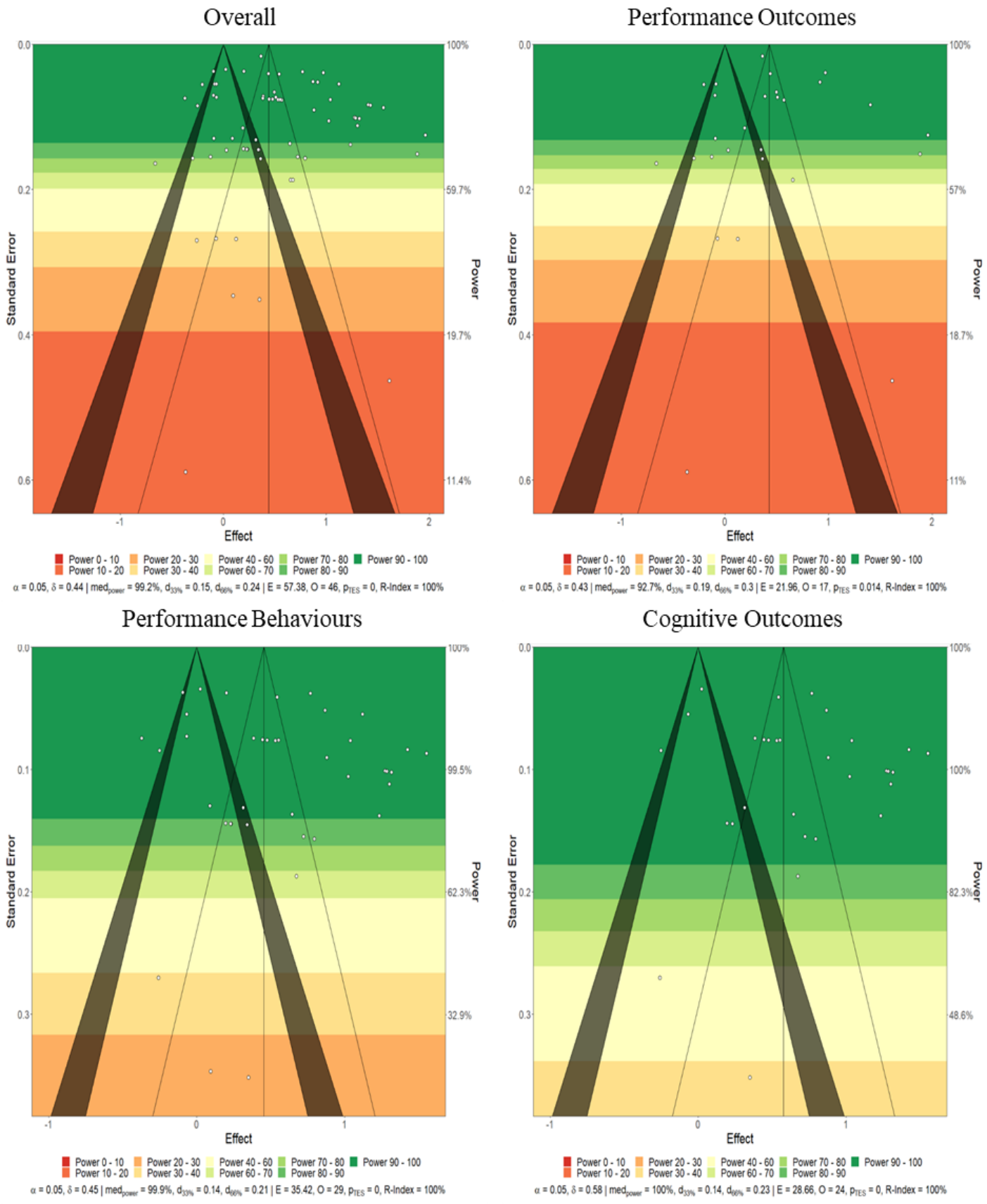

Figure 2. Sunset funnel plots for all meta-analyses.

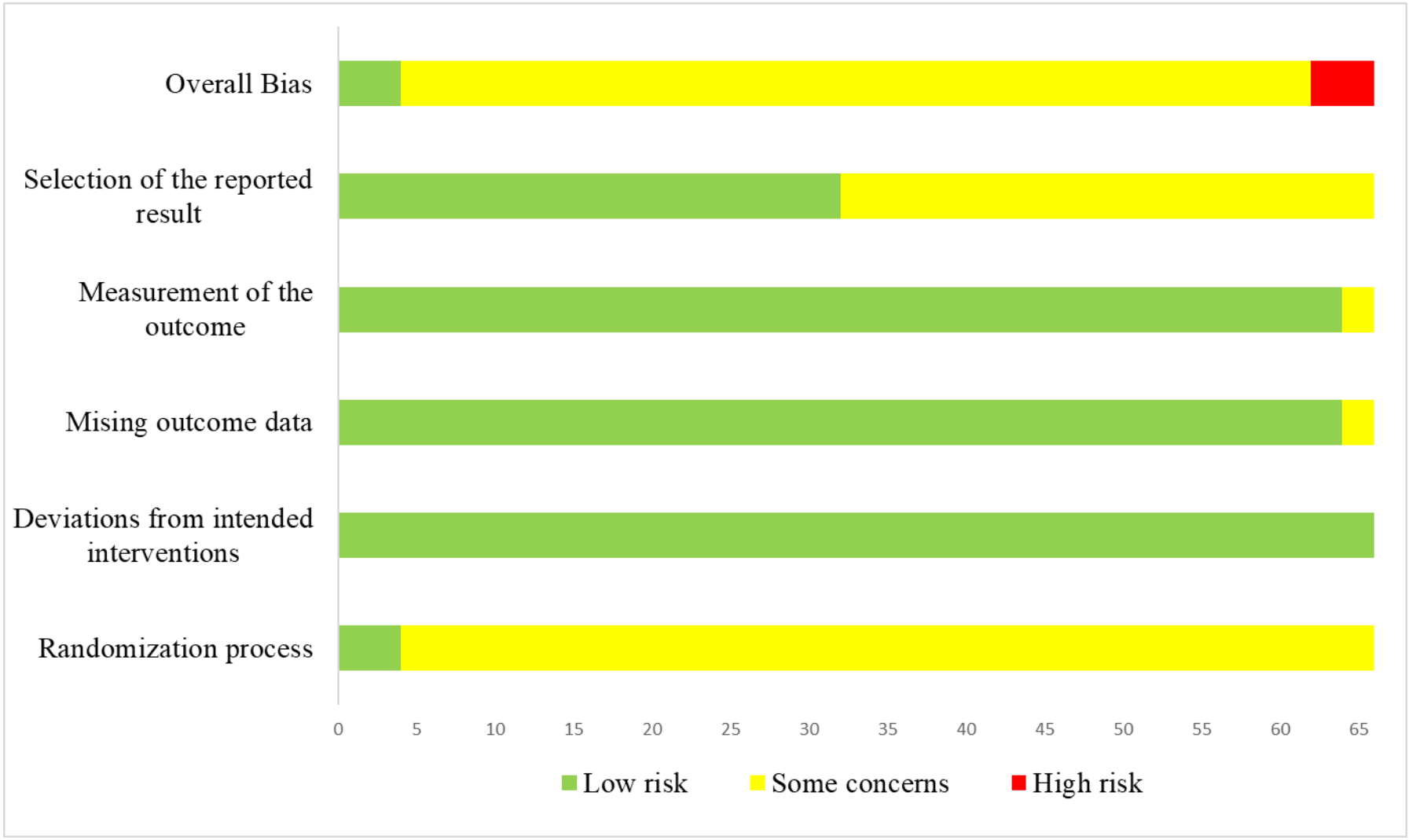

Figure 3. Risk of bias summary.