Supplementary Materials

Compressing Wallas: Does Improvisation-Based Instruction Transfer to Executive Functions and Creative Thinking? A Systematic Review and Meta-Analysis Guided by a Temporal-Regime Model of Creativity

Kutup Ata Tuncer [1,*] and Seçil Soytok Nalçacı [1]

[1] Department of Music Education, Faculty of Education, Manisa Celal Bayar University, Manisa, Türkiye; secil.nalcaci@cbu.edu.tr

Contents: S1, review protocol as written before screening, search documentation, and departures from the protocol; S2, effect-size derivations (Table S2); S3, coding sheet, record lists, analysis code and verification log (separate archive, S3_code_and_data.zip); S4, moderator analyses for creative-thinking outcomes (Table S4); S5, reports excluded at full-text assessment, reports not retrieved, and risk-of-bias ratings (Tables S5a and S5b); S6, sensitivity to the assumed pre-test–post-test correlation (Table S6); S7, PRISMA 2020 checklist; S8, GRADE-style appraisal of the certainty of the evidence (Table S8).

Supplementary Material S1. Review protocol (as written on 14 October 2026) and search documentation

Editorial note (added 15 October 2026, before analysis of the final data set). The protocol below is reproduced verbatim as it was written before screening began, so that departures from it can be audited. Three citations in its Section 3 could not be verified against bibliographic databases after the protocol was written; they are corrected here rather than in the protocol text: "Degé et al., 2022" refers to Degé and Frischen (2022); "Jouin et al., 2026" corresponds to Phipps (2025); "Hua, 2026" could not be identified and was not used. The prior reviews used for backward citation chasing are those listed in Section 2.3 of the article. Departures from the protocol are reported in Section 2.7 of the article and are summarized in Part B of this supplement.

Part A. Protocol

Protocol — Improvisation-based instruction, executive functions and creative thinking: a meta-analysis

Working title: *Compressing Wallas: Does real-time creative production transfer to executive functions and divergent thinking? A meta-analysis of improvisation-based instruction*

Prepared 14 Oct 2026. Not pre-registered (time constraints); deviations from this protocol will be reported in the manuscript.

1. Rationale and hypotheses

Improvisation is creative production under a compressed temporal window (the real-time regime of the temporal-regime model). The model predicts that the real-time regime exercises (i) retrieval of compiled domain patterns and (ii) attention control (inhibition of prepotent/learned material, flexible switching between retrieved options), with a smaller load on working-memory capacity per se. Hypotheses:

- H1. Improvisation-based instruction produces positive transfer to performance-based measures of executive functions and of creative thinking relative to control conditions ($g > 0$).
- H2. Within executive functions, effects are larger for inhibition and cognitive flexibility than for working-memory capacity.
- H3. Within creative thinking, effects are larger for flexibility/originality indices than for fluency.
- H4. Effects are larger when the comparison is a passive control than when it is an active (matched non-improvisational) control; the active-control contrast isolates the real-time ingredient.
- H5 (exploratory). Dose (number of sessions, total minutes) moderates effects; acute single-session effects are smaller than multi-session training effects.
- H6 (exploratory). Music-based (including rhythm-based) and other-arts (theatre, dance, verbal) improvisation do not differ in transfer.

2. Eligibility criteria (PICOS)

Population. Healthy/typically developing participants of any age; primary analyses on learners in educational settings (preschool to university). Studies with clinical samples (e.g., ASD, ADHD, dementia, MCI, psychiatric diagnoses) are excluded. Healthy older-adult samples are included and coded as a separate age group (sensitivity analysis).

Intervention. Structured activity or program in which participants generate novel musical, rhythmic, movement/dance, dramatic, or verbal material in real time (improvisation) and in which improvisation is the defining or manipulated component (explicitly labelled as improvisation/improv; ≥50% of session content or the only ingredient differing from the comparison condition). Both multi-session programs (training) and single-session activities (acute) are eligible and coded separately. Mixed programs in which improvisation is one minor component among many are excluded unless the improvisation component is isolated by design.

Comparison. No-treatment, wait-list, business-as-usual, or active comparison (matched non-improvisational activity: e.g., repertoire/notation-based music instruction, scripted drama, learned choreography, sports, non-creative games). Within-subject counterbalanced designs with a control activity are eligible for acute studies.

Outcomes. Performance-based (objective) measures of: (a) executive functions: inhibition/inhibitory control, working memory/updating, cognitive flexibility/set-shifting, attention control, EF composites; (b) creative cognition: divergent thinking (fluency, flexibility, originality, elaboration, composite), convergent creative problem solving (RAT, insight), domain-specific creativity measures (e.g., Webster's MCTM; expert-rated creative products or performances). Excluded: self-report measures (creative self-efficacy, creative personality, self-rated creativity), affect/well-being, social/communication skills, academic achievement, motor skill without a creativity criterion.

Study design. Randomized or non-randomized controlled designs with a comparison group or counterbalanced comparison condition; pre–post or post-only measurement; sufficient statistics to compute a standardized mean difference (means/SDs, t/F with df, exact p with n, or effect size with CI). Peer-reviewed articles, dissertations, conference papers; any language; no date limit.

3. Information sources and search strategy

Databases searched 14–15 October 2026: PubMed (web interface), ERIC (API), Crossref (bibliographic query, screened to the first 200 records per query), Scopus (run by K.A.T. through institutional access with the string below; exported as CSV). Web of Science Core Collection was planned with the same string but could not be searched, for lack of institutional access (see the table below). Supplementary: backward citation chasing from included studies and from prior reviews (Degé et al., 2022; Rodriguez-Gomez & Talero-Gutiérrez, 2022; Jouin et al., 2026; Hua, 2026; Sio & Lortie-Forgues, 2024), and forward citation chasing of the five seed studies (Koutsoupidou & Hargreaves, 2009; Lewis & Lovatt, 2013; Sowden et al., 2015; Hainselin et al., 2018; Norgaard et al., 2019).

Search strings:

- PubMed: `improvisation AND (creativity OR "divergent thinking" OR "executive function" OR "working memory")` [all fields], no limits.
- ERIC: `improvisation AND (creativity OR "creative thinking" OR "divergent thinking" OR "executive function" OR "working memory" OR inhibition OR "cognitive flexibility").`
- Scopus (TITLE-ABS-KEY): `(improvis* OR improv) AND ("executive function*" OR "working memory" OR inhibit* OR "cognitive flexibility" OR "set shifting" OR "divergent thinking" OR "creative thinking" OR creativity OR originality) AND (intervention OR training OR instruction OR program* OR course OR workshop OR class* OR lesson* OR experiment* OR random* OR "control group" OR pretest OR "pre-test" OR posttest).`
- Web of Science (TS=): same string as Scopus.

4. Study selection and data extraction

Titles/abstracts screened against criteria; full texts retrieved for all potentially eligible records. Screening and coding performed by K.A.T. with the assistance of a generative AI tool (disclosed); S.S.N. independently screens a random 25% of full-text records and independently codes all included studies on design, comparison type, outcome category and extracted statistics; agreement reported (κ for categorical codes; ICC for continuous data); disagreements resolved by discussion.

Coded variables: study ID; publication type/year; country; sample age group (preschool 3–6; primary 7–12; adolescent 13–18; university/young adult 19–30; adult 31–59; older adult ≥60); n per group; design (RCT / cluster-RCT / quasi-experimental / within-subject counterbalanced); comparison type (passive: no-treatment/wait-list/business-as-usual; active: matched activity); improvisation domain (music; rhythm-based music; theatre/drama; dance/movement; verbal/storytelling; mixed arts); dose (sessions, weeks, minutes per session, total minutes); acute (1 session) vs training (≥2 sessions); outcome domain (EF: inhibition, working memory, flexibility, attention, composite; creativity: DT fluency, flexibility, originality, elaboration, composite; convergent; domain-specific product/performance); measure name; timing of outcome (immediate ≤1 week post; delayed); statistics extracted.

5. Effect sizes

Positive values favour the improvisation condition. For pre–post–control designs: Morris (2008) $d_{ppc2} = c_p \times [(M_{post,T} - M_{pre,T}) - (M_{post,C} - M_{pre,C})] / SD_{pre,pooled}$, with the sampling variance of Morris (2008) assuming pre–post correlation $\varrho = .50$ (sensitivity: .30 and .70). For post-only randomized designs: Hedges' g on post-test means. For within-subject acute designs: standardized difference between conditions divided by the pooled condition SD, with variance for correlated designs ($\varrho = .50$ assumed). Reaction-time and error measures reverse-scored. Where only test statistics are reported, d is derived from t or F(1, df); where only "non-significant" is reported, $d = 0$ is imputed in a sensitivity analysis.

6. Synthesis

Three-level random-effects model (effect sizes nested in studies; REML; Knapp–Hartung adjustment for tests), fitted separately for EF and creativity outcomes and overall. Heterogeneity: τ^2 at levels 2 and 3, I^2 decomposition. Moderator analyses by mixed-effects meta-regression for the variables in §4 (minimum $k = 4$ per level). Robustness: cluster-robust (CR2) standard errors; leave-one-study-out; exclusion of non-randomized studies; exclusion of effect sizes derived from imputation. Publication bias: contour-enhanced funnel plot, multilevel Egger regression (standard error as moderator), PET-PEESE, and trim-and-fill on study-averaged effects. Risk of bias: randomization, allocation concealment/blinding of outcome assessors, attrition, baseline equivalence, selective reporting (adapted from RoB 2/ROBINS-I), summarized as low/some concerns/high and tested as a moderator.

Software: Python 3 (numpy, scipy, pandas, statsmodels); custom REML implementation of the three-level model, validated against published worked examples. Analysis code and coding sheet to be deposited as supplementary material.

Part B. Search documentation (as run)

Source	Interface	Date	String / query	Records
Scopus	Documents search; field "Article title, Abstract, Keywords"	14 Oct 2026	`(improvis* OR improv) AND ("executive function*" OR "working memory" OR inhibit* OR "cognitive flexibility" OR "set shifting" OR "divergent thinking" OR "creative thinking" OR creativity OR originality) AND (intervention OR training OR instruction OR program* OR course OR workshop OR class* OR lesson* OR experiment* OR random* OR "control group" OR pretest OR "pre-test" OR posttest)`	1129 (CSV export, citation information)
PubMed	Web interface, all fields, no limits	14 Oct 2026 (record list re-exported 15 Oct 2026)	`improvisation AND (creativity OR "divergent thinking" OR "executive function" OR "working memory")`	336
ERIC	Public API (api.ies.ed.gov/eric)	14 Oct 2026 (counts re-run 15 Oct 2026)	(1) `improvisation AND (creativity OR "executive function")` — 241; (2) `improvisation AND "divergent thinking"` — 8; (3) `improvisation AND "executive function"` — 4; (4) `improv AND creativity AND (experiment OR "control group" OR pretest)` — 1	254 (246 unique within ERIC)
Crossref	REST API, `query.bibliographic`, first 60 records per query, relevance order	15 Oct 2026	(1) improvisation creativity training experiment control group; (2) improvisation divergent thinking intervention pretest posttest; (3) improvisation executive function training randomized; (4) improvisation working memory inhibition cognitive flexibility; (5) improvisational theatre creative thinking students experimental; (6) music improvisation creativity children program effects; (7) dance improvisation creativity intervention control; (8) improv training cognitive flexibility divergent thinking adults	480
Web of Science	—	—	Not searched (no institutional access)	0
Other methods	Backward citation chasing (reference lists of included studies and of Biasutti, 2017; Chandler, 2018; Degé & Frischen, 2022; Rodriguez-Gomez & Talero-Gutiérrez, 2022; Phipps, 2025); targeted web searches in English, Spanish, Portuguese, German, Turkish and Russian; one suggestion from a colleague	14–15 Oct 2026	—	reports sought for retrieval that were not already in the database set: see Figure 2

Deduplication was performed by DOI and by normalized title (lower-case, diacritics and punctuation removed, stop words removed) with the script `prisma_tally.py` (Supplementary Material S3); the record lists for PubMed, ERIC and Crossref and the Scopus export are included in S3.

Departures from the protocol (summary; see Section 2.7 of the article)

1. Forward citation chasing of the five seed studies through Scopus was not performed systematically, because Scopus access became available only at the end of the search period.
2. Crossref was queried with eight bibliographic strings and screened to the first 60 records per query rather than the 200 stated in the protocol; Crossref's bibliographic query has no field restriction and weak relevance ranking, and once the Scopus search became available Crossref served only as a supplementary source.
3. The planned independent screening of a 25% random sample of full-text records and the independent coding of all included studies by the second author were replaced by the procedure described in Section 2.4 of the article: a second check of every coding-sheet value against the source (20 values by the first author, the remainder in documented AI-assisted passes); a consistency test of the extracted means against the test statistics reported in the sources; and a complete independent second extraction of the 15 studies with available full text by the second author, made with the assistance of a separate instance of the same AI tool without access to the first coding sheet and checked by her against the PDFs, compared field by field with the first and resolved by consensus. The independent screening of a sample of full-text records was not performed, and the second extraction is not blind double coding by two human coders.
4. The planned moderator test of executive-function subtypes could not be carried out (fewer than three studies per subtype) and was replaced by a descriptive account.
5. For the one study in which outcomes were measured at the level of dyads (Harinck & Dooren, 2023), the dyad was accepted as the unit of analysis: effect sizes were computed from the dyad-level test statistics with the number of dyads as the sample size, so that the dependence of individuals within dyads is already absorbed in the reported statistics and no further adjustment was applied.
6. Risk of bias was rated with five items (random allocation, baseline equivalence, blinding of outcome assessment or objective scoring, completeness of outcome data, completeness of reporting); allocation concealment was not rated separately because no study reported it, and the planned moderator test of risk of bias was not run because no study was rated low overall.

Supplementary Material S2. Effect-size derivations

Table S2. Effect sizes and their derivation (ϱ = .50).

#	Study / sample	Outcome	Measure	Dir.	n_T / n_C	Statistics used	Method	g	SE	Set
1	Hainselin et al. (2018)	creativity: DT fluency	AUT fluency (raw)	+	18 / 12	T: 4.111 (1.632) → 6.222 (3.878); C: 4.167 (1.249) → 4.458 (1.437)	ppc2	1.19	0.41	main
2	Hainselin et al. (2018)	creativity: DT flexibility	AUT flexibility ratio	+	18 / 12	T: 0.478 (0.273) → 0.786 (0.248); C: 0.682 (0.249) → 0.683 (0.236)	ppc2	1.13	0.41	main
3	Hainselin et al. (2018)	creativity: DT originality	AUT originality ratio	+	18 / 12	T: 0.691 (0.498) → 1.514 (0.426); C: 0.822 (0.434) → 0.812 (0.368)	ppc2	1.71	0.45	main
4	Hainselin et al. (2018)	creativity: DT elaboration	AUT elaboration ratio	+	18 / 12	T: 0.29 (0.267) → 0.466 (0.907); C: 0.262 (0.404) → 0.277 (0.336)	ppc2	0.48	0.38	main
5	Norgaard et al. (2019) Grade 8	EF: flexibility	WCST % perseverative errors	−	34 / 19	T: 15.57 (7.57) → 11.63 (3.23); C: 14.09 (5.29) → 12.95 (5.81)	ppc2	0.40	0.29	main
6	Norgaard et al. (2019) Grade 7	EF: flexibility	WCST % perseverative errors	−	31 / 38	t_T = 3.48, t_C = 3.36	within_t	−0.08	0.24	main
7	Norgaard et al. (2019) Grade 7	EF: inhibition	Stroop inhibition RT (ms)	−	35 / 41	T: 92.64 (66.23) → 64.41 (65.24); C: 77.59 (81.26) → 92.38 (66.25)	ppc2	0.57	0.24	main
8	Norgaard et al. (2019) Grade 8	EF: inhibition	Stroop inhibition RT (ms)	−	36 / 21	—	imputed0	0.00	0.28	imputed-zero
9	Norgaard et al. (2019) Grade 7	EF: working memory	Forward digit span	+	35 / 41	—	imputed0	0.00	0.23	imputed-zero
10	Norgaard et al. (2019) Grade 8	EF: working memory	Forward digit span	+	36 / 21	—	imputed0	0.00	0.28	imputed-zero
11	Hansen et al. (2020)	EF: EF composite	D-KEFS Design Fluency composite (conditions 1-3)	+	14 / 13	t = 2.06	change_t	0.77	0.41	main
12	Sowden et al. (2015) Exp. 1	creativity: DT originality	Instances task originality	+	14 / 13	F = 7.93	F_between	1.05	0.40	main
13	Sowden et al. (2015) Exp. 1	creativity: DT fluency	Instances task fluency	+	14 / 13	—	imputed0	0.00	0.39	imputed-zero
14	Sowden et al. (2015) Exp. 1	creativity: domain product	Instances task CAT creativity rating	+	14 / 13	F = 7.83	F_between	1.05	0.40	main
15	Sowden et al. (2015) Exp. 1	creativity: domain product	Toy design CAT originality	+	14 / 13	F = 15.93	F_between	1.49	0.43	main
16	Sowden et al. (2015) Exp. 1	creativity: domain product	Toy design CAT creativity	+	14 / 13	F = 9.78	F_between	1.17	0.41	main
17	Sowden et al. (2015) Exp. 2	creativity: DT originality	TTCT incomplete figures originality	+	17 / 17†	F = 7.09	F_ancova	0.77	0.35	main
18	Sowden et al. (2015) Exp. 2	creativity: DT elaboration	TTCT incomplete figures elaboration	+	17 / 17†	F = 6.24	F_ancova	0.72	0.35	main
19	Sowden et al. (2015) Exp. 2	creativity: DT composite	TTCT abstractness of titles	+	17 / 17	—	imputed0	0.00	0.35	imputed-zero
20	Yanicheva & Nikitina (2020)	creativity: DT flexibility	Torrance Unusual Uses flexibility (post, mean ranks)	+	20 / 18	mean ranks 25.6 vs 12.8	ranks	1.37	0.35	main
21	Yanicheva & Nikitina (2020)	creativity: DT originality	Mednick RAT originality index (post, mean ranks)	+	20 / 18	mean ranks 24.3 vs 14.2	ranks	1.01	0.34	main
22	Lewis & Lovatt (2013) Exp. 1	creativity: DT fluency	AUT fluency	+	21 / 20	T: 5.33 (2.16) → 7.46 (2.42); C: 5.12 (2.72) → 5.63 (2.41)	ppc2	0.65	0.32	main
23	Lewis & Lovatt (2013) Exp. 1	creativity: DT originality	AUT originality	+	21 / 20	T: 1.33 (1.24) → 3.43 (1.66); C: 1.95 (1.54) → 2.3 (2.21)	ppc2	1.23	0.35	main
24	Lewis & Lovatt (2013) Exp. 1	creativity: DT elaboration	AUT elaboration	+	21 / 20	T: 2.14 (1.24) → 2.57 (1.94); C: 2.11 (1.88) → 2.05 (1.43)	ppc2	0.30	0.32	main
25	Lewis & Lovatt (2013) Exp. 1	creativity: DT flexibility	AUT flexibility	+	21 / 20	T: 3.49 (1.06) → 4.87 (1.57); C: 3.57 (1.55) → 4.02 (1.51)	ppc2	0.69	0.32	main
26	Lewis & Lovatt (2013) Exp. 2	creativity: DT fluency	AUT fluency	+	24 / 12	T: 8 (4.37) → 9.56 (3.71); C: 8.5 (4.25) → 7.75 (3.85)	ppc2	0.52	0.36	main
27	Lewis & Lovatt (2013) Exp. 2	creativity: DT originality	AUT originality	+	24 / 12	T: 2.58 (2.48) → 4.5 (3.78); C: 2.67 (2.61) → 2 (1.48)	ppc2	1.00	0.38	main
28	Lewis & Lovatt (2013) Exp. 2	creativity: DT elaboration	AUT elaboration	+	24 / 12	T: 3.46 (2.41) → 3.5 (2.06); C: 4.17 (2.95) → 3.17 (2.21)	ppc2	0.39	0.36	main
29	Lewis & Lovatt (2013) Exp. 2	creativity: DT flexibility	AUT flexibility	+	24 / 12	T: 7.21 (3.21) → 5.96 (2.29); C: 7.08 (3.26) → 5.54 (2.75)	ppc2	0.09	0.36	main

#	Study / sample	Outcome	Measure	Dir.	n_T / n_C	Statistics used	Method	g	SE	Set
30	Felsman et al. (2020) Exp. 1	creativity: DT fluency	AUT fluency	+	37 / 37	T: 6.76 (3.18) → 7.23 (2.67); C: 5.73 (2.92) → 5.76 (2.81)	ppc2	0.14	0.23	main
31	Felsman et al. (2020) Exp. 1	creativity: DT flexibility	AUT flexibility	+	37 / 37	T: 6.36 (2.96) → 6.5 (2.4); C: 5.26 (2.48) → 5.37 (2.65)	ppc2	0.01	0.23	main
32	Felsman et al. (2020) Exp. 1	creativity: DT elaboration	AUT elaboration	+	37 / 37	T: 2.99 (2.1) → 3.18 (2.35); C: 2.76 (1.56) → 2.77 (2.48)	ppc2	0.10	0.23	main
33	Felsman et al. (2020) Exp. 1	creativity: DT originality	AUT originality	+	37 / 37	T: 2.32 (1.78) → 3.03 (1.91); C: 2.14 (2.24) → 1.81 (1.68)	ppc2	0.51	0.24	main
34	Felsman et al. (2020) Exp. 2	creativity: DT fluency	AUT fluency	+	67 / 64	T: 6.47 (2.76) → 7.13 (3.12); C: 5.92 (2.96) → 5.99 (3.09)	ppc2	0.21	0.18	main
35	Felsman et al. (2020) Exp. 2	creativity: DT flexibility	AUT flexibility	+	67 / 64	T: 6.21 (2.55) → 6.81 (2.83); C: 5.59 (2.6) → 5.69 (2.83)	ppc2	0.19	0.18	main
36	Felsman et al. (2020) Exp. 2	creativity: DT elaboration	AUT elaboration	+	67 / 64	T: 2.33 (2) → 2.54 (1.79); C: 2.54 (2) → 2.07 (1.94)	ppc2	0.34	0.18	main
37	Felsman et al. (2020) Exp. 2	creativity: DT originality	AUT originality	+	67 / 64	T: 2.12 (1.54) → 2.3 (2.22); C: 1.55 (1.55) → 1.89 (1.77)	ppc2	−0.10	0.18	main
38	Koutsoupidou & Hargreaves (2009)	creativity: domain product	MCTM-II musical flexibility	+	12 / 13	t = 3.29	change_t	1.28	0.45	main
39	Koutsoupidou & Hargreaves (2009)	creativity: domain product	MCTM-II musical originality	+	12 / 13	t = 5.76	change_t	2.23	0.53	main
40	Koutsoupidou & Hargreaves (2009)	creativity: domain product	MCTM-II musical syntax	+	12 / 13	t = 6.78	change_t	2.62	0.57	main
41	Koutsoupidou & Hargreaves (2009)	creativity: domain product	MCTM-II musical extensiveness	+	12 / 13	—	imputed0	0.00	0.41	imputed-zero
42	Harinck & Dooren (2023) Study 1	creativity: DT fluency	AUT fluency (coffee mug)	+	23 / 22†	F = 13.36	F_between	1.07	0.31	main
43	Harinck & Dooren (2023) Study 1	creativity: DT flexibility	AUT flexibility	+	23 / 22†	F = 14.85	F_between	1.13	0.32	main
44	Harinck & Dooren (2023) Study 2	creativity: DT fluency	AUT fluency (paper clip)	+	19 / 24	post T: 15.31 (4.14); C: 12.25 (4.12)	post_only	0.73	0.31	main
45	Harinck & Dooren (2023) Study 2	creativity: DT flexibility	AUT flexibility	+	19 / 24	post T: 13.16 (4); C: 11.54 (3.48)	post_only	0.43	0.30	main
46	Schwenke et al. (2021)	creativity: DT composite	Verbaler Kreativitäts-Test (VKT) total	+	30 / 28	t = 3.37	change_t	0.87	0.28	main
47	Santos et al. (2021)	EF: planning	Clock Drawing Test	+	15 / 12	T: 5.47 (1.85) → 6.8 (2.24); C: 6.77 (1.3) → 6.08 (1.89)	ppc2	1.20	0.43	main
48	Santos et al. (2021)	EF: attention	Trail Making Test A (s)	−	15 / 12	T: 66.44 (29.58) → 60.31 (20.94); C: 54.46 (14.7) → 50.36 (16.16)	ppc2	0.08	0.39	main
49	Santos et al. (2021)	EF: verbal fluency	Verbal fluency, phonemic	+	14 / 12	T: 11.93 (3.47) → 12 (4); C: 12.77 (4.3) → 11.83 (3.38)	ppc2	0.25	0.40	main
50	Santos et al. (2021)	EF: verbal fluency	Verbal fluency, semantic	+	14 / 12	T: 17.67 (5.29) → 17.07 (4.85); C: 15 (4.94) → 15.83 (4.86)	ppc2	−0.27	0.40	main
51	Santos et al. (2021)	EF: inhibition	Five Digits Test inhibition (z)	−	15 / 12	—	imputed0	0.00	0.39	imputed-zero
52	Santos et al. (2021)	EF: flexibility	Five Digits Test flexibility (z)	−	15 / 12	—	imputed0	0.00	0.39	imputed-zero
53	Richard et al. (2021)	creativity: domain product	Motor creativity fluency (agility ladder)	+	26 / 30	d = 0.23	d_reported	0.23	0.27	main
54	Richard et al. (2021)	creativity: domain product	Motor creativity flexibility	+	26 / 30	d = 0.58	d_reported	0.57	0.28	main
55	Richard et al. (2021)	creativity: domain product	Motor creativity originality	+	26 / 30	d = 0.18	d_reported	0.18	0.27	main
56	Richard et al. (2021)	creativity: DT fluency	Verbal DT fluency (Unusual Uses)	+	26 / 30	d = 0.06	d_reported	0.06	0.27	main
57	Richard et al. (2021)	creativity: DT flexibility	Verbal DT flexibility	+	26 / 30	d = −0.07	d_reported	−0.07	0.27	main
58	Richard et al. (2021)	creativity: DT originality	Verbal DT originality	+	26 / 30	d = 0.03	d_reported	0.03	0.27	main
59	Richard et al. (2021)	creativity: DT fluency	Figural DT fluency (TTCT incomplete figures)	+	26 / 30	d = 0.13	d_reported	0.13	0.27	main
60	Richard et al. (2021)	creativity: DT flexibility	Figural DT flexibility	+	26 / 30	d = 0.15	d_reported	0.15	0.27	main
61	Richard et al. (2021)	creativity: DT originality	Figural DT originality	+	26 / 30	d = 0.5	d_reported	0.49	0.27	main
62	Bugos et al. (2024)	EF: working memory	PASAT correct	+	17 / 20	T: 57.39 (17.94) → 67.33 (14.75); C: 57.58 (16.12) → 59.21 (14.12)	ppc2	0.48	0.34	broad
63	Bugos et al. (2024)	EF: flexibility	Cued Stroop errors, mixed condition	−	17 / 20	T: 11.5 (8.15) → 5.44 (3.78); C: 9.47 (6.56) → 8.37 (7.13)	ppc2	0.66	0.34	broad
64	Bugos et al. (2024)	EF: flexibility	Cued Stroop RT, mixed condition (ms)	−	17 / 20	T: 1272.17 (175.66) → 1234.47 (199.65); C: 1190.53 (133.69) → 1159.29 (207.47)	ppc2	0.04	0.33	broad
65	Bugos et al. (2024)	EF: verbal fluency	D-KEFS letter fluency	+	17 / 20	T: 41.28 (14.7) → 43.72 (14.43); C: 43.53 (8.59) → 42 (10.48)	ppc2	0.33	0.33	broad

#	Study / sample	Outcome	Measure	Dir.	n_T / n_C	Statistics used	Method	g	SE	Set
66	Bugos et al. (2024)	EF: verbal fluency	D-KEFS category fluency	+	17 / 20	T: 37.67 (9.12) → 41.39 (9.81); C: 39.89 (8.85) → 39.47 (8.44)	ppc2	0.45	0.34	broad
67	Bugos et al. (2024)	EF: flexibility	D-KEFS category switching	+	17 / 20	T: 14.39 (2.55) → 14.33 (3.93); C: 14.26 (2.49) → 14.58 (3.27)	ppc2	−0.15	0.33	broad
68	West et al. (2017)	creativity: DT composite	TCT-DP individual creativity (figural production test)	+	45 / 38	T: 17.9 (7.6) → 27.7 (12.1); C: 20.7 (8.6) → 21.7 (10)	ppc2	1.08	0.24	main
69	West et al. (2017)	creativity: domain product	DOG group creativity (CAT-rated group product)	+	14 / 11	F = 4.67	F_ancova	0.73	0.40	main
70	Méndez Martínez & Fernandez-Rio (2021)	creativity: domain product	Motor creativity fluency (object-transformation task)	+	79 / 84	T: 4.96 (2.81) → 8.79 (3.42); C: 5.11 (2.53) → 7.69 (2.64)	ppc2	0.47	0.16	main
71	Méndez Martínez & Fernandez-Rio (2021)	creativity: domain product	Motor creativity flexibility	+	79 / 84	T: 1.31 (0.59) → 2.51 (1.33); C: 1.41 (0.67) → 2.06 (1.08)	ppc2	0.87	0.16	main
72	Méndez Martínez & Fernandez-Rio (2021)	creativity: domain product	Motor creativity imagination	+	79 / 84	T: 1.99 (0.56) → 2.58 (0.84); C: 2.06 (0.58) → 2.37 (0.66)	ppc2	0.49	0.16	main
73	Méndez Martínez & Fernandez-Rio (2021)	creativity: domain product	Motor creativity originality	+	79 / 84	T: 1.89 (0.67) → 2.2 (0.68); C: 1.85 (0.59) → 1.78 (0.5)	ppc2	0.60	0.16	main
74	Coubard et al. (2011)	EF: flexibility	Rule Shift Cards profile score (/4)	+	16 / 94	T: 2.19 (1.12) → 2.69 (1.16); C: 3.0645 (1.0674) → 3.0533 (1.1466)	ppc2	0.47	0.27	main
75	Coubard et al. (2011)	EF: flexibility	Rule Shift Cards switch error rate (%)	−	16 / 94	T: 43.7 (30) → 27.1 (32.04); C: 18.4234 (29.9019) → 19.5915 (31.9495)	ppc2	0.59	0.27	main
76	Coubard et al. (2011)	EF: inhibition	Stroop interference ratio (correct cond. 3 / cond. 1)	+	16 / 94	T: 0.46 (0.12) → 0.48 (0.12); C: 0.5057 (0.0886) → 0.4957 (0.0886)	ppc2	0.32	0.27	main
77	Coubard et al. (2011)	EF: inhibition	Stroop interference error rate (%)	−	16 / 94	T: 3.29 (7.16) → 6.65 (10.28); C: 2.1689 (7.156) → 2.1035 (10.1944)	ppc2	−0.48	0.27	main
78	Coubard et al. (2011)	EF: attention	Arithmetic word problems, low planning demand, accuracy	+	16 / 94	T: 0.87 (0.36) → 0.69 (0.4); C: 0.8401 (0.3378) → 0.8087 (0.4114)	ppc2	−0.43	0.27	main
79	Coubard et al. (2011)	EF: attention	Arithmetic word problems, high planning demand, accuracy	+	16 / 94	T: 0.37 (0.48) → 0.31 (0.48); C: 0.3527 (0.4856) → 0.4385 (0.497)	ppc2	−0.30	0.27	main
80	Jin et al. (2026) Exp. 2	creativity: DT fluency	AUT fluency	+	32 / 35	T: 2.86 (1.68) → 4.05 (2.24); C: 2.76 (1.44) → 2.65 (1.55)	ppc2	0.82	0.26	main
81	Jin et al. (2026) Exp. 2	creativity: DT flexibility	AUT flexibility	+	32 / 35	T: 2.35 (1.35) → 3.49 (1.66); C: 2.09 (1.07) → 2.19 (1.06)	ppc2	0.85	0.26	main
82	Jin et al. (2026) Exp. 2	creativity: DT originality	AUT originality (rarity-based, 1–5)	+	32 / 35	T: 1.47 (0.81) → 2.21 (0.85); C: 1.45 (0.55) → 1.41 (0.51)	ppc2	1.12	0.27	main
83	Karakelle (2009)	creativity: DT fluency	Total fluency (circle drawing + alternate uses)	+	15 / 15	T: 32.27 (15.03) → 40.93 (14.9); C: 30.67 (9.24) → 27.4 (9.94)	ppc2	0.93	0.39	broad
84	Karakelle (2009)	creativity: DT flexibility	Total flexibility (circle drawing + alternate uses)	+	15 / 15	T: 28.73 (13.45) → 36.4 (13.89); C: 26.07 (7.07) → 25.27 (8.37)	ppc2	0.77	0.38	broad

Note. Dir. = scoring direction (− = lower raw score is better; effect reversed so that positive favours improvisation). Methods: ppc2 = Morris d_ppc2 from pre/post M and SD of both conditions; post_only = Hedges' g from post-test M and SD; F_between = $d = \sqrt{F} \cdot \sqrt{(1/n_t + 1/n_c)}$ (between-groups F with 1 df, post-only); F_ancova = d from ANCOVA F with error variance divided by $(1 - \varrho^2)$; change_t = d from independent-samples t on change scores (pre-test SD metric via $\sqrt{(2(1-\varrho))}$); within_t = d from paired t of each condition (within-condition change, pre-test SD metric); ranks = d from Mann–Whitney mean ranks via probability of superiority (Ruscio, 2008); d_reported = reported Cohen's d on change scores, small-sample corrected; imputed0 = reported as non-significant without statistics; d = 0 (sensitivity analysis only). Set: main = strict definition, reported statistics; imputed-zero = included only in the imputed-zero sensitivity analysis; broad = included only in the broad-definition sensitivity analysis. Notes on individual derivations are in the coding sheet (Supplementary S3). † Group sizes not reported by condition; the split shown is assumed (Harinck & Van Dooren, 2023, Study 1: 45 dyads, 23/22; Sowden et al., 2015, Exp. 2: 34 children, 17/17); the sensitivity of the estimates to the split is reported in Section 3.5.

Supplementary Material S3. Coding sheet, record lists, analysis code, and verification log

Provided as a separate archive (S3_code_and_data_v3.zip), also openly available at Zenodo (version 3.0, https://doi.org/10.5281/zenodo.24157009; all versions, https://doi.org/10.5281/zenodo.23422047). The archive contains the coding sheet (`coding_sheet.csv`, one row per effect size with the source location of every value), the computed effect sizes, the full model output for ϱ = .30, .50 and .70, the record lists returned by each database search, all analysis and figure scripts (Python 3.10 or later) with a single script, `code/reproduce.sh`, that regenerates every data file, table, and figure from the coding sheet and the record lists, the screening and extraction logs, and the verification log. A README in the archive describes each file and records a clean-directory reproduction check in which every regenerated table and figure matched the manuscript.

Supplementary Material S4. Moderator analyses for creative-thinking outcomes

Table S4. Moderator analyses restricted to creative-thinking outcomes (k = 54, m = 12).

Moderator	Level	k	m	g [95% CI]	Omnibus test
Outcome subtype	DT elaboration	6	4	0.34 [0.14, 0.53]	F(4, 47) = 1.52, p = .21
	DT flexibility	11	7	0.60 [0.17, 1.03]	
	DT fluency	10	6	0.56 [0.26, 0.86]	
	DT originality	11	7	0.81 [0.39, 1.23]	
	Domain-specific product	14	5	0.93 [0.32, 1.54]	
Comparison condition	Active	49	9	0.77 [0.45, 1.09]	F(1, 52) = 0.70, p = .41
	Passive	5	3	1.02 [0.76, 1.28]	
Acute vs training	Training (≥ 2 sessions)	25	7	0.94 [0.52, 1.35]	F(1, 52) = 0.79, p = .38
	Acute (1 session)	29	5	0.69 [0.38, 0.99]	
Domain of improvisation	Dance	16	3	0.75 [0.27, 1.22]	F(2, 47) = 0.48, p = .62
	Music	9	3	0.97 [−0.15, 2.10]	
	Theatre	25	8	0.80 [0.54, 1.07]	
Age group	Adult	8	3	0.86 [0.39, 1.34]	F(2, 47) = 2.47, p = .10
	Primary	16	4	1.18 [0.76, 1.60]	
	University	26	5	0.51 [0.20, 0.81]	
Allocation	Non-randomized	25	7	0.94 [0.52, 1.35]	F(1, 52) = 0.79, p = .38
	Randomized	29	5	0.69 [0.38, 0.99]	
Total minutes (log)	slope	49	11	0.010 [−0.101, 0.122]	p = .86
Number of sessions (log)	slope	54	12	0.090 [−0.119, 0.300]	p = .39
Mean age (years)	slope	54	12	−0.015 [−0.038, 0.007]	p = .18

Supplementary Material S5. Reports excluded at full-text assessment, reports not retrieved, and risk-of-bias ratings

Table S5a. Reports assessed at full text and excluded, with the primary reason for exclusion (a secondary reason, where one applies, is given in parentheses), and reports that could not be retrieved.

Report	Primary reason	Note
Biasutti & Mangiacotti (2018). Assessing a cognitive music training for older participants: A randomised controlled trial. *International Journal of Geriatric Psychiatry*. https://doi.org/10.1002/gps.4721	clinical sample	cognitive decline
Richard, Lebeau, Becker, Boiangin & Tenenbaum (2018). Developing cognitive and motor creativity in children through an exercise program using nonlinear pedagogy principles. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2018.1530913	improvisation not the defining component	nonlinear pedagogy programme
Bugos, DeMarie, Stokes & Power (2022). Multimodal music training enhances executive functions in children: Results of a randomized controlled trial. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.14857	improvisation not the defining component	multimodal programme
Cordero, Amaro-Fuenzalida, Cid & Aguirre (2026). Impact of a play-based intervention on creativity and cognitive flexibility measures. *Acta Psychologica*. https://doi.org/10.1016/j.actpsy.2026.106709	improvisation not the defining component	play-based executive-function workshop
Li (2022). Creativity level of students in the course of collaborative flute jamming. *Música Hodie*. https://doi.org/10.5216/mh.v22.70327	insufficient or inconsistent statistics	
Richard, Halliwell & Tenenbaum (2017). Effects of an improvisation intervention on elite figure skaters' performance, self esteem, creativity, and mindfulness skills. *The Sport Psychologist*. https://doi.org/10.1123/tsp.2016-0059	no comparison condition (self-report outcomes)	n = 9
Daugėlienė & Strakšienė (2017). Development of students' musical creative thinking by applying musical improvisation. *Pedagogika*. https://doi.org/10.15823/p.2017.07	no comparison condition	n = 7
Navarro Ramón & Chacón-López (2021). The impact of musical improvisation on children's creative thinking. *Thinking Skills and Creativity*. https://doi.org/10.1016/j.tsc.2021.100839	no comparison condition	
Lindquist, Liggett, Muhammad, Seltzer, Kim, Barish & Smith (2021). Effects of improv training on older adults in a long term care facility. *Gerontology and Geriatric Medicine*. https://doi.org/10.1177/23337214211016111	no comparison condition (self-report outcomes)	
Bassis, Rybko & Maor (2023). It's never too late to improvise: The impact of theatre improvisation on elderly people. *Experimental Aging Research*. https://doi.org/10.1080/0361073X.2022.2059208	no comparison condition (no eligible outcome)	
Liang, Walgenbach & Schweinberger (2026). From discipline to improvisation: A practice-based model for training creativity. *Education Sciences*. https://doi.org/10.3390/educsci16081311	no comparison condition	single-case introspective account
Hu, Li, Jiang & Chen (2024). The change in aesthetic experience and empathic concern predicts theory of mind ability: Evidence from drama improvisation. *The Arts in Psychotherapy*. https://doi.org/10.1016/j.aip.2024.102167	no eligible outcome	
Dou, Jia & Ge (2021). Improvisational dance-based psychological training of college students' dance improvement. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.663223	no instructional intervention (self-report outcomes)	experience-group comparison
Schwenke, Bleichner, Fjaellingsdal, Klarhölter, Dshemuchadse & Scherbaum (2024). Improving by improvising: The impact of improvisational theatre on handling expectation violation during social interactions. *Journal of Creativity*. https://doi.org/10.1016/j.yjoc.2024.100082	no instructional intervention	expert–novice comparison and within-subject task study
Keisari, Krueger, Ben-David & Hainselin (2024). New horizon in improving ageing with improvisational theatre. *Age and Ageing*. https://doi.org/10.1093/ageing/afae087	no new controlled data	commentary
Berkemeyer, Binnewies, Hilbert & Maunz (2025). Creative and healthy through improv: Effects of training improvisational theatre on creative and work-related outcomes. *Thinking Skills and Creativity*. https://doi.org/10.1016/j.tsc.2025.101757	self-report outcomes only	

Report	Primary reason	Note
Pan, He, Chen, Jiang & Jamaludin (2025). The effect of improvisation and performance on engagement and creativity among music performance students: A quasi-experimental study. *Humanities and Social Sciences Letters*. https://doi.org/10.18488/73.v13i3.4316	self-report outcomes only	

Reports not retrieved. The following reports could not be obtained through the channels available to the authors (no institutional access to SAGE or Taylor & Francis journals; no response to requests). Eligibility could therefore not be assessed; the abstract (as indexed in OpenAlex) is summarized for each.

Report	Abstract (indexed)
Hansen & Oxoby (2017). An earned presence: Studying the effect of multi-task improvisation systems on cognitive and learning capacity. *Connection Science*. https://doi.org/10.1080/09540091.2016.1277692	two pilot studies of performance-generating systems and executive functions; potentially eligible or precursor of Hansen et al. (2020)
Mourey (2020). Improv comedy and modern marketing education: Exploring consequences for divergent thinking, self-efficacy, and collaboration. *Journal of Marketing Education*. https://doi.org/10.1177/0273475318822087	10-week improvisation course with divergent-thinking outcomes; potentially eligible
Neville & Makopoulou (2021). Effect of a six-week dance-based physical education intervention on primary school children's creativity: A pilot study. *European Physical Education Review*. https://doi.org/10.1177/1356336X20939586	controlled pilot (n = 40) with Alternative Uses outcome (d = 0.36); improvisation framing not determinable
Zhang (2023). The role of musical improvisation in the development of creative thinking in children. *Cultura y Educación*. https://doi.org/10.1080/11356405.2023.2200528	experimental and control groups with pre- and post-tests of musical creativity; potentially eligible
Yao & Qin (2024). Musical improvisation for college students: Personal creativity and thinking skills. *Cambridge Journal of Education*. https://doi.org/10.1080/0305764X.2024.2426475	pre–post t-tests; comparison condition not described
Chang (2024). The effectiveness of improvisation in the development of creative thinking in dance education. *Cultura y Educación*. https://doi.org/10.1177/11356405241280165	pre–post design with 561 students; comparison condition not described

Note. Records excluded at the title/abstract stage are not listed individually; the screened record lists with screening flags are provided in Supplementary Material S3.

Table S5b. Risk-of-bias ratings for the included studies (L = low, S = some concerns, H = high).

Study	Random allocation	Baseline equivalence	Blinded / objective scoring	Attrition	Complete reporting	Overall	Note
Koutsoupidou & Hargreaves (2009)	H	L	S	L	S	H	intact classes (authors: random assignment of groups not feasible); pre-test means stated as almost equal; MCTM-II judge scoring, rater blinding not stated; all 25 children analysed; F reported for all four subscales, extensiveness as non-significant without means
Coubard et al. (2011)	H	H	L	S	L	H	allocation by district of residence; Rule Shift pre-test difference to the disadvantage of the improvisation group (profile 2.19 vs 3.06, d = −0.81; switch errors d = −0.84), reported as non-significant; objective paper-and-pencil scoring; attrition not reported; means, SEs and F reported for all outcomes
Lewis & Lovatt (2013)	S	L	L	L	L	S	randomized (group-level in Exp. 1), method not stated; pre-test differences non-significant; raters blind to condition and time (p. 50); no attrition (df match n)

Study	Random allocation	Baseline equivalence	Blinded / objective scoring	Attrition	Complete reporting	Overall	Note
Sowden et al. (2015)	S	H	S	L	S	H	randomized (stratified by sex in Exp. 1), method not stated; Exp. 1 post-only, Exp. 2 elaboration differed at pre-test, t(32) = 2.66, p = .012; CAT and TTCT rater blinding not stated; fluency (Exp. 1) and abstractness (Exp. 2) reported without means
West et al. (2017)	H	S	S	L	L	H	intact work teams; creativity pre-tests not compared (control higher on TCT-DP); CAT judges, blinding not stated; attrition 5/50 and 5/43
Hainselin et al. (2018)	S	H	S	S	L	H	allocation to groups not stated; AUT pre-tests not compared; control group higher on the flexibility ratio at pre-test (d = −0.77); scorer and blinding not stated; 5 of 35 excluded, distribution across groups not given; all subscores with means, SDs and F; data on OSF
Norgaard et al. (2019)	H	S	L	S	S	H	intact band classes; improvisation and comparison cohorts not compared at pre-test; computerized tasks (PEBL); 10 absent at post-test plus 13 excluded from WCST, distribution not given; Grade 7 WCST means, Grade 8 Stroop and digit span reported without statistics
Felsman et al. (2020)	S	S	L	L	L	S	groups of 5-8 randomized to condition, method not stated; pre-test comparison not reported (ANCOVA-adjusted; Exp. 2 originality 2.12 vs 1.55); raters blind to condition and time; no attrition; full reporting
Hansen et al. (2020)	H	S	L	L	S	H	course enrolment vs recruited volunteers; pre-test equivalence asserted, not reported; standardized D-KEFS scoring, tester independent of the instructor; 1 of 15 in the intervention group did not complete the quantitative tests (reason not reported), none lost in the control group (14 and 13 analysed); no means reported, change-score t only
Yanicheva & Nikitina (2020)	S	S	S	S	H	H	allocation not stated (groups 'equalized' on parameters); no pre-test data; researcher-scored tests, blinding not stated; attrition not reported; between-group results given only for the two significant measures (selective reporting)
Méndez Martínez & Fernandez-Rio (2021)	H	L	L	S	L	H	intact classes, four of eight allocated to the improvisation unit by the school timetable (authors: quasi-experimental; p. 5); pre-test equivalence shown (MANOVA F(4, 155) = 0.56); six expert raters blind to condition, ICC .97-.99; analysed n not reported, the post-test MANCOVA F(4, 140) implying about 146-149 of 163 (about 10%), distribution across conditions unknown; means, SDs and ANCOVA F for all four outcomes
Richard et al. (2021)	S	L	S	S	L	S	allocation method not stated; rater blinding not stated; attrition not reported
Santos et al. (2021)	S	H	L	H	S	H	randomized by a person external to the study, sequence generation not stated; outcome pre-tests not compared; improvisation group lower on clock drawing at pre-test (d = −0.80); assessors blind; control 16 to 12 (25%, unbalanced); FDT z-scores without SDs, p values only

Study	Random allocation	Baseline equivalence	Blinded / objective scoring	Attrition	Complete reporting	Overall	Note
Schwenke et al. (2021)	H	S	S	H	S	H	allocation by order of enrolment; outcome pre-tests not compared; VKT scoring, scorer blinding not stated; attrition 4/34 vs 12/40 (30%, unbalanced); F values without means
Harinck & Dooren (2023)	S	S	S	H	L	H	dyads randomized, method not stated; post-only design; research-assistant scoring, blinding not stated; Study 1: 5 of 50 dyads without divergent-thinking data (10%), distribution across conditions not reported; Study 2: 1 dyad excluded from all analyses and 6 set aside from the outcome analyses for inconsistent agreement reports, so that the analysed set (19/24) lacks 6 of 25 improvisation and 1 of 25 control dyads (24% vs 4%); fluency effect also reported for all 49 dyads ($F(1, 46) = 5.88$ vs $F(1, 41) = 5.85$), flexibility for the 43 only; full reporting
Jin et al. (2026)	S	L	L	L	S	S	randomized, method and unit not stated; baselines parallel; blind raters; no attrition; statistics of the Exp. 2 passive control group identical to those of the Exp. 1 control (apparent duplication, so the reporting for that group is incomplete; the active comparison is used)

Note. Each item was rated independently by both authors (the first author from the full reports and the extraction notes, with the assistance described in Section 2.4; the second author in the independent second extraction described there) and the ratings were compared and resolved by consensus against the rules of Section 2.5; item-level agreement before resolution was 55% (Cohen's $\kappa = .26$) over the 15 studies rated by both. Richard et al. (2021) carries the first author's rating only. The overall judgement is the worst item rating. Rating rules: (1) random allocation: L, random allocation with the method of sequence generation stated; S, randomization stated without its method, or allocation not described; H, intact groups or self-selection. (2) baseline equivalence: L, groups shown equivalent on the outcome at pre-test; S, equivalence not reported or post-only design, with no pre-test difference of half a pooled standard deviation or more in the reported data; H, a pre-test difference on an outcome reported as significant, or of at least half a pooled standard deviation ($|d| \geq 0.5$) computed from the reported pre-test means and SDs. (3) blinded or objective scoring: L, scorers blind to condition or objective scoring; S, blinding not stated; H, subjective scoring by the trainer. (4) attrition (incomplete outcome data): units allocated to a condition but absent from the analysis that yields the coded effect, whether lost or excluded, counted per condition; L, none (stated, or implied by the degrees of freedom), or below 20% in every condition with a between-condition difference below 10 percentage points; S, not reported, or a loss below 20% overall whose distribution across conditions is not reported; H, 20% or more in any condition, or a between-condition difference of 10 percentage points or more; in reports with several studies the item takes the worst rating among the included studies. (5) complete reporting: L, statistics reported for every pre-specified outcome; S, some outcomes reported only as non-significant or with incomplete statistics; H, outcomes omitted or selectively reported. The definitions of rule (4) (excluded units counted with lost ones, thresholds per condition, the 10-point criterion for imbalance, losses of unknown distribution, reports with several studies) were written out after the consensus rating; applied to all 16 studies, they reproduce every consensus rating. The $|d| \geq 0.5$ criterion of rule (2) was added at revision, in response to peer review; applied to the pre-test data of all studies, it changed item 2 from S to H for Coubard et al. (2011), Hainselin et al. (2018), and Santos et al. (2021) (largest pre-test differences −0.81 to −0.84, −0.77, and −0.80), whereas no other study reached $|d| = 0.45$.

Supplementary Material S6. Sensitivity to the assumed pre-test–post-test correlation

Table S6. Sensitivity of the main estimates to the assumed pre-test–post-test correlation ϱ.

Estimate	ϱ = .30	ϱ = .50 (main)	ϱ = .70
All outcomes	0.73 [0.48, 0.98]	0.70 [0.45, 0.94]	0.65 [0.43, 0.86]
Executive functions	0.20 [−0.08, 0.48]	0.19 [−0.08, 0.47]	0.19 [−0.08, 0.46]
Creative thinking	0.86 [0.60, 1.11]	0.82 [0.57, 1.07]	0.77 [0.54, 1.00]
All, imputed zeros included	0.65 [0.41, 0.89]	0.62 [0.39, 0.84]	0.57 [0.37, 0.78]
Randomized studies only	0.64 [0.37, 0.91]	0.63 [0.33, 0.93]	0.61 [0.31, 0.92]
Effects from reported means only	0.52 [0.31, 0.73]	0.52 [0.29, 0.75]	0.52 [0.29, 0.74]
Broad definition	0.71 [0.48, 0.93]	0.68 [0.46, 0.89]	0.63 [0.44, 0.83]
Egger slope (SE), p	2.83 (0.91), p = .003	3.11 (0.95), p = .002	3.34 (0.91), p < .001
PET estimate	−0.27 [−0.94, 0.40]	−0.27 [−0.89, 0.34]	−0.23 [−0.72, 0.27]
PEESE estimate	0.15 [−0.22, 0.51]	0.12 [−0.22, 0.46]	0.15 [−0.13, 0.43]
Trim-and-fill (k0; adjusted)	0; 0.71	5; 0.48	4; 0.49

Supplementary Material S7. PRISMA 2020 checklist

Locations refer to sections of the article (Page et al., 2021, BMJ 372:n71). "S" refers to the Supplementary Materials of this article.

Section and topic	Item	Checklist item	Location
Title	1	Identify the report as a systematic review.	Title ("A Systematic Review and Meta-Analysis")
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract (objective, eligibility, information sources and search date, risk-of-bias method, synthesis method, number of studies and participants, pooled effects with 95% CI, limitations, interpretation); registration: not registered, stated in 2.1
Introduction	3	Describe the rationale for the review in the context of existing knowledge.	1.1–1.3
	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	1.3 (hypotheses H1–H6), 1.4
Methods	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	2.2; grouping by outcome domain in 2.7
	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted. Specify the date when each source was last searched or consulted.	2.3; S1 Part B
	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	2.3; S1 Part B (full strings)
	8	Specify the methods used to decide whether a study met the inclusion criteria, including how many reviewers screened each record and each report retrieved, whether they worked independently, and details of automation tools used.	2.4 (including the re-screen of excluded records); Acknowledgments
	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and details of automation tools used.	2.4, 2.5; S3 (verification log)
	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought, and if not, the methods used to decide which results to collect.	2.2 (criterion 4), 2.5
	10b	List and define all other variables for which data were sought. Describe any assumptions made about any missing or unclear information.	2.5; S2 (assumptions per effect size)
	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and details of automation tools used.	2.5 (second paragraph, including the baseline criterion added at revision); S5 note
	12	Specify for each outcome the effect measure(s) used in the synthesis or presentation of results.	2.6
	13a	Describe the processes used to decide which studies were eligible for each synthesis.	2.7 (first sentences); 2.2
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics or data conversions.	2.6 (derivations, imputation of non-significant outcomes); S2
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	2.7; Figures 3–4; Table 2

Section and topic	Item	Checklist item	Location
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	2.7
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g., subgroup analysis, meta-regression).	2.7 (moderators; Holm and Benjamini–Hochberg adjustment)
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	2.7 (robustness, ϱ sensitivity, imputed zeros, broad definition; effect-size metric, randomized studies only, executive-function estimate with each study omitted, samples with experienced performers excluded)
	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	2.7 (funnel plot, Egger, PET–PEESE, trim-and-fill, small-study relation by comparison type); 2.6 (outcomes reported only as non-significant)
	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	2.7; Supplementary Material S8 (GRADE-style appraisal across the five domains, with the rules applied)
Results	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	3.1 (including the re-screen result); Figure 2
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	3.1; S5 Table S5a
	17	Cite each included study and present its characteristics.	3.2; Table 2
	18	Present assessments of risk of bias for each included study.	3.2; S5 Table S5b
	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision, ideally using structured tables or plots.	Figure 3; S2 (Table S2); S3 (coding sheet)
	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	3.2, 3.3
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	3.3–3.4; Tables 3–5; S4
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	3.4; Tables 3–5; S4
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	3.5; S6
	21	Present assessments of risk of bias due to missing results for each synthesis assessed.	3.5 (small-study effects, by comparison type); Figure 4
	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	3.5; 4.5; Supplementary Table S8
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	4.1–4.4
	23b	Discuss any limitations of the evidence included in the review.	4.5
	23c	Discuss any limitations of the review processes used.	4.5
	23d	Discuss implications of the results for practice, policy, and future research.	4.4, 4.6

Section and topic	Item	Checklist item	Location
Other information	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	2.1 (not registered; no registration number)
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	2.1; S1
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	2.7; S1 Part B
	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding statement
	26	Declare any competing interests of review authors.	Conflicts of Interest
	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data Availability Statement; S3 (Zenodo, version 3.0, https://doi.org/10.5281/zenodo.24157009; all versions, https://doi.org/10.5281/zenodo.23422047)

Supplementary Material S8. GRADE-style appraisal of the certainty of the evidence

Table S8. Certainty of the evidence for the two outcome families, appraised across the five GRADE domains (ϱ = .50; main analysis set of 16 studies).

Outcome family	Studies (samples; participants)	Pooled g [95% CI]; 95% prediction interval	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Certainty
Creative thinking (divergent-thinking tests, convergent tasks, rated products)	12 (16; 951)	0.82 [0.57, 1.07]; −0.03 to 1.67	Serious (−1): 4 of the 12 studies allocated intact groups or volunteers and no randomized study described its allocation sequence; 8 of the 12 rated high on at least one item and none low on every item (Table S5b); passive comparisons gave the largest effects	Not downgraded: heterogeneity is in magnitude, not direction (I^2 = 70%; all 12 study-level effects positive, 0.17 to 2.04); the prediction interval touches zero	Not downgraded: outcomes are performance-based; 14 of 54 effects are near-transfer product ratings, noted in Section 4.2	Not downgraded: 951 participants; the confidence interval excludes zero and the five randomized studies give 0.69 [0.38, 0.99]	Strongly suspected (−1): funnel asymmetry, Egger p = .002, conditional PET estimate −0.27 [−0.89, 0.34], trim-and-fill 0.48	**Low**
Executive functions (inhibition, flexibility, working memory, attention, composites)	4 (5; 293)	0.19 [−0.08, 0.47]; −0.62 to 1.01	Serious (−1): 3 of the 4 studies allocated intact groups or volunteers and all 4 rated high on at least one item; five outcomes in two studies reported only as non-significant (entered as zero only in sensitivity analysis)	Serious (−1): mixed within-study profiles (opposite-signed Stroop indices in one study); different task batteries in every study; I^2 = 58% within studies	Not downgraded: performance-based tasks; the profile is descriptive (Section 4.2)	Very serious (−2): four studies and 293 participants; the confidence interval includes both no effect and a moderate benefit	Not assessable with four studies; the small-study relation found for the set as a whole applies	**Very low**

Rules applied. The appraisal follows the five GRADE domains (Guyatt et al., 2008) but is informal: GRADE was designed for randomized evidence, and the present set mixes six randomized with ten non-randomized controlled studies. The body of evidence for each outcome family starts at *high* as controlled intervention evidence, and the non-randomized allocation of most studies is charged to the risk-of-bias domain rather than to the starting level; charging it to the starting level instead (non-randomized evidence starting at *low*) would leave the executive-function rating unchanged and lower the creative-thinking rating by one level. Downgrades: one level for a serious concern, two for a very serious concern. Risk of bias follows the item ratings of Table S5b; inconsistency is judged on the direction and heterogeneity of study-level effects and the prediction interval (Section 3.3); indirectness on the match between the measured outcomes and the outcomes the questions of Section 1.4 concern; imprecision on the width of the confidence interval relative to a smallest effect of interest of 0.2 and on the number of participants; publication bias on the small-study analyses of Section 3.5. The certainty ratings qualify the pooled estimates and the conclusions drawn from them in Sections 4.1 and 5: the creative-thinking estimate is a low-certainty estimate whose true value is probably smaller than 0.82, and the executive-function estimate is a very-low-certainty estimate that does not distinguish no effect from a moderate benefit.