**Melodic Intonation Therapy: A multi-level meta-analysis**

Tudor Popescu1-2\*, Benjamin Stahl3-6, Brenton M. Wiernik7, Hannah Helm3, Michaela Zemanek1, Felix Haiduk1, Theresa Matzinger1, Roland Beisteiner2, Tecumseh W. Fitch1

1 University of Vienna, Department of Behavioural and Cognitive Biology, Vienna, Austria

2 Medical University of Vienna, Department of Neurology, Vienna, Austria

3 Medical School Berlin, Department of Psychology, Berlin, Germany

4 Charité Universitätsmedizin, Department of Neurology, Berlin, Germany

5 Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neurophysics, Leipzig, Germany

6 Universitätsmedizin Greifswald, Department of Neurology, Greifswald, Germany

7 University of South Florida, Department of Psychology, Tampa, Florida, USA

\*: tudor.popescu@univie.ac.at

# Abstract

INTRODUCTION

Melodic Intonation Therapy (MIT) is a prominent rehabilitation programme for individuals with post-stroke aphasia. Despite substantial progress in recent years, the overall efficacy and clinical relevance of MIT remain not fully understood: available studies are often constrained by lack of standardised outcomes, randomisation or control group. These limitations challenge the interpretation of the data.

AIMS

The present meta-analysis seeks to determine the efficacy and clinical relevance of MIT while considering quality of outcomes (psychometrically validated *versus* unvalidated measures), experimental design (presence *versus* absence of randomisation and control group), influence of spontaneous recovery (quantified as number of months post-stroke), MIT version applied (original *versus* modified protocol), and level of generalisation (testing performance on trained *versus* untrained items).

METHODS

An extensive literature search led to X studies being selected, X of which met the inclusion criteria. Multi-level mixed- and random-effects models were used to separately meta-analyse randomised controlled trial (RCT) and non-RCT data.

REULTS

Unvalidated outcomes inflated the effect size of MIT by a factor of X when compared to validated outcomes. Similarly, MIT effect size was 5.X times larger for non-RCT data compared to RCT data. Effect size also decreased with number of months post-stroke, suggesting confounds related to spontaneous recovery, primarily within the first year post-stroke. In contrast, variation of the original MIT protocol did not systematically alter the benefit from treatment. Crucially, analyses demonstrate significantly improved language performance on trained and untrained items; the latter finding mainly arose from gains in repetition tasks, but not in other domains of verbal expression, including everyday communication ability (*g̅* = −0.04 [95% CI −0.57, 0.48]).

CONCLUSIONS

Accounting for various methodological aspects, the current results confirm the promising role of MIT in improving language performance on trained items and in repetition tasks, while highlighting possible limitations in promoting everyday communication ability.

# Introduction

Stroke survivors often experience a profound loss of communication skills, among them a syndrome known as aphasia. This syndrome may manifest as severe difficulty in verbal expression, referred to as ‘non-fluent aphasia.’ In addition, stroke survivors frequently suffer from impaired speech-motor planning. Known as ‘apraxia of speech,’ this syndrome typically occurs in combination with aphasia. Although about a third of individuals with neurological communication disorders do not recover completely (Engelter et al., 2006), rehabilitation programmes can improve language performance even in the chronic stage of symptoms (Breitenstein et al., 2017).

Melodic Intonation Therapy (MIT) is a prominent rehabilitation programme originally developed for individuals with non-fluent aphasia (Albert et al., 1973). Drawing on the observation that individuals with neurological communication disorders are often able to sing entire pieces of text fluently (Gerstman, 1964; Mills, 1904; Yamadori et al., 1977), MIT uses melody, rhythm, vocal expression (in unison and alone), left-hand tapping, formulaic and non-formulaic verbal utterances, as well as other therapeutic elements, in a hierarchically structured protocol (Helm-Estabrooks et al., 1989). To date, randomised controlled trial (RCT) data have confirmed the efficacy of MIT on validated outcomes in the *late subacute* or *consolidation* stage of aphasia (i.e., up to 12 months after stroke; van der Meulen et al., 2014); but not in the *chronic* stage of aphasia (i.e. more than 6–12 months after stroke; Van Der Meulen et al., 2016).

From a methodological point of view, influences of spontaneous recovery are generally lower in the *chronic* stage of aphasia, as suggested by RCT data (Doppelbauer et al., 2021) and meta-analyses (RELEASE Collaborators, 2021). Therefore, it is important to consider stage of symptoms post-stroke. Moreover, speech-language therapy seeks to promote performance on untrained items. Consistent with this goal, the present work distinguishes progress on trained items—learning resulting from using the same set of utterances both during treatment and subsequent assessment—from the more desirable goal of attaining generalisation to untrained items, ideally in the context of everyday communication to ensure ecological validity (e.g., Blomert et al., 1994).

So far, to our knowledge there are systematic reviews on MIT (e.g., van der Meulen et al., 2012; Zumbansen et al., 2014) and two meta-analyses (Brady et al., 2016; Zumbansen & Tremblay, 2018). These meta-analyses reflect a relatively limited amount of RCT data (Brady et al., 2016) or dichotomise post-treatment improvement in a way that prevents specific estimates of effect size (Zumbansen & Tremblay, 2018). Given the substantial burden of disease associated with aphasia, the present meta-analysis attempts to provide a deeper understanding about the potential and limitations of MIT. To achieve this goal, the current analyses synthesise available studies on MIT to address five research questions:

1. **Psychometric quality of outcomes.** Does the use of validated *versus* unvalidated outcomes systematically alter the resulting effect size of MIT?
2. **Experimental design.** Do RCT and non-RCT results in the context of MIT differ systematically in terms of effect size?
3. **Aphasia stage.** Do influences related to spontaneous recovery, quantified as number of months post-stroke-onset (MPO), affect the effect size of MIT?
4. **Variants of MIT protocol.** Do variations of the original MIT protocol alter the resulting effect size?
5. **Generalisability.** Apart from trained items, does MIT enhance performance on untrained items and, if so, does the resulting effect size demonstrate gains on measures of everyday communication ability?

# Methods

## Eligibility criteria

### Inclusion criteria

We defined the following basic inclusion criteria for studies to be considered for the present meta-analysis:

* empirical study that administered MIT to adult individuals (age 18 or over) with aphasia, with or without a control group;
* language-related outcomes in pre-post assessment;
* publication in peer-reviewed journal.

Despite the limited quality of evidence in non-RCT designs (case reports), we chose to include and analyse those separately and comparatively.

### Exclusion criteria

After the removal of duplicate items (see *Appendix 2: Dealing with duplicate findings*), the following exclusion criteria were applied to remaining studies:

* substantial variation from original MIT protocol (Albert et al., 1973). We accepted *minor* changes to the MIT protocol (and examined the effect of the categorical variable: original *versus* modified MIT), as long as the protocol met all of the following features:
  + - melody-based vocal expression;
    - some form of rhythmic pacing (e.g., left-hand tapping);
    - use of verbal utterances known from everyday communicative interaction;
* unvalidated (ad hoc) outcome measures; no published or otherwise accessible validation study for the particular test battery. Exception: if a study included both trained and untrained items for an unvalidated measure, we included it to determine the degree of generalisation by comparing performance on trained and untrained items (see Results);
* other essential data not reported and / or not retrievable, even after contacting the authors (e.g., no sample size or standard error, insufficient information to compute an effect size).

### List of included studies (after application of exclusion criteria)

*(to paste here the* ***@includedStudies*** *tab from main sheet, possibly with certain essential study features added as columns from STUDIES:ALL, such as MIT type etc)*

For a list of the excluded studies, see Appendix 4: List of excluded studies.

## Search strategy

The literature search strategy was devised and implemented by authors TP and MZ. This was designed to obtain high search sensitivity, using both free-text and subject headings in databases, which were not restricted by language or publication form (Higgins et al., 2021).

### Electronic searches

We searched the following databases: *Cochrane Library* (last searched 25.02.2021), *CINAHL EBSCOhost* (last searched 27.02.2021), *PsycINFO OVID* (1806 to August 2020), *Web of Science* (last searched 25.02.2021), *PubMed* (last searched 26.02.2021), *Scopus* (last searched 26.02.2021), *Medline* *OVID* (1946 to February 2021) *PSYNDEX OVID* (1977 to February 2021), *Music Periodicals Database* (last searched 26.02.2021), and *ProQuest Dissertations & Theses Global* (last searched 27.02.2021).

We also searched trials registers, including *International Clinical Trials Registry Platform* (ICTRP,<https://www.isrctn.com/>; last searched 11.03.2021), *National Research Register* (UK),<http://www.nihr.ac.uk/>; last searched 11.03.2021), *Clinical Trials.gov* ([www.clinicaltrials.gov](http://www.clinicaltrials.gov); last searched 11.03.2021), *Netherlands Trials Register*[www.trialregister.nl](http://www.trialregister.nl); last searched 11.03.2021), and the *German clinical trials Register*<https://www.drks.de/drks_web/>; last searched 11.03.2021)

Additionally, we performed searches in Google Scholar (last searched 11.03.2021) and in the grey literature database OpenGrey.eu (<http://www.opengrey.eu/>; last searched 11.03.2021). Messages soliciting any unpublished data were additionally sent to:

* aphasia associations
  + National Aphasia Association, NAA, https://www.aphasia.org
  + Australian Aphasia Association (AAA), https://aphasia.org.au
  + Fédération Nationale des Aphasiques de France (FNAF), http://aphasie.fr/
* music therapy associations
  + American Music Therapy Association (AMTA), https://www.musictherapy.org/
  + British Association for Music Therapy (BAMT), https://www.bamt.org
* mailing lists
  + AUDITORY
  + Musicology-all
* authors of the included studies

Finally, to ensure no studies were omitted we consulted the list of studies in published systematic reviews and meta-analyses concerning MIT (Brady et al., 2016; Hurkmans et al., 2012; van der Meulen et al., 2012; Zumbansen et al., 2014; Zumbansen & Tremblay, 2018).

Search histories are provided in *Appendix 1: Search terms*, and the number of search hits and duplicate removal in *Appendix 2: Dealing with duplicate findings*

### Extracting empirical studies and excluding not relevant studies

Since no filters relating to methods used or publication type were applied to our searches, we manually separated and kept the empirical studies from the overall results.

## Study coding and double-coding

All studies were coded by the first author (TP). Two of the authors (FH, TM) re-coded all studies, verifying the cross-coder consistency. Agreement among the three coders occurred in a majority of cases, and any discrepancies found between coding sheets were solved by consensus. The ICCs (intraclass correlations) were >0.9 in the remaining cases, such as errors arisen from numerically estimating data reported in plot format only.

## Tests and outcome measures in primary studies

The table below shows all the tests reported in the primary studies considered, and the reason for excluding some of them.

*((to paste here the @****testBatteries*** *tab from main sheet))*

The table below shows how the different measures (subtests) from the batteries of validated tests that we considered, contributed towards the relevant linguistic *Abilities*, and towards the meta-analysed dependent variables (which we deemed *Domains*).

*((to paste here the @****categScheme*** *tab from main sheet))*

## MIT variants

The table below shows all the MIT variants that we considered, across both included and excluded studies.

*(to paste here, or perhaps in the SUpplMat, the @****MIT variants*** *tab from main sheet)*

## Meta-analysis methods

### Computed outcome metric

To maximise comparability of effects across studies, we used change scores from pre-test to post-test as the outcome variable, expressed in *z*-scores. For group-level studies (the RCTs in the current analyses), we standardised *z*-scores using pooled pre-test standard deviation across control and treatment groups. For individual patient data studies (the case reports in the current analyses), we computed *z*-scores in one of three ways. For studies that reported results as *z*-scores (e.g., based on test norms), we used the *z*-scores directly. For studies that reported results as percentile scores (e.g., based on test norms), we converted these to *z*-scores using the quantiles of the standard Normal distribution. For other studies, we estimated *z*-scores using the following procedure. We first converted normalised raw scores to reflect the proportion of the maximum possible score (POMP; Cohen et al., 1999).[[1]](#footnote-1) Next, we estimated a three-level random-intercept model for the pre-test POMP scores, with individual test scores nested within patients nested within studies (see Figure 1). From these models, we used the population intercept as the estimated POMP score *mean*, and the patient-level random effects standard deviation as the estimated POMP score *SD* (τ). We then used this *mean* and *SD* to standardise the pre-test and post-test POMP scores.



Figure 1: Three-level model employed, with SDs (τ) at study-, patient- and measure-level.

### RCTs

All RCTs were reported at the group-level. We computed effect sizes as the pretest-posttest-control group Hedges’ *g*:  (Morris, 2008). We computed the variance for each *g* using the method of Morris (2008). We estimated multi-level mixed effects meta-regression models to account for effect size dependency, with random intercepts for each study. We first fit an overall meta-analysis combining all effect sizes. Second, we fit additional meta-regression models including potential moderator variables. For these meta-regression models, we included random slopes for the *Domain* moderator, nested within studies (Konstantopoulos, 2011). We used a homoscedastic compound symmetric structure for the random effects, estimating a single random effects variance and correlation for all abilities.[[2]](#footnote-2) We estimated the amount of heterogeneity (i.e., *τ*) using the restricted maximum-likelihood estimator (Viechtbauer, 2005). We computed confidence intervals for meta-regression coefficients and mean treatment effects using the Knapp and Hartung *t*-distribution method (Knapp & Hartung, 2003), and for the random effects components using profile likelihood. We estimated models using *R* (version 4.1.0; R Core Team, 2021) and the *metafor* package (version 3.-01 Viechtbauer, 2010).[[3]](#footnote-3)

### Case reports

All case reports reported results as individual-level data, so we analyzed these studies using individual patient data (IPD) meta-analysis. We computed individual-level scores as the difference between pre-test and post-test *z*-scores (the mean difference in these scores is the pretest-posttest Hedges’ *g*, *gpp*). We then pooled data across studies using a three-level random-effects IPD meta-analysis, with individual scores again (cf. Figure 1) nested within patients nested within studies (Riley et al., 2010). Similar to the group-level RCT meta-analyses, we first fit an overall model including all data points with no moderators, then fit additional models including potential moderator variables as predictors. For these models, we included random intercepts for patients and studies.[[4]](#footnote-4) We estimated random effects components using REML and computed confidence intervals using profile likelihood. We estimated models using *R* (version 4.1.0; R Core Team, 2021) and the *lme4* package (version 1.1-27; Bates et al., 2015a).

### Moderator analyses

For the RCT meta-analyses, we fit a meta-regression model with the moderators (1) Domain (cf. section 2.4); (2) whether the study used validated tests as its outcome measures, or unvalidated ones (for unvalidated measures, we treated trained and untrained items as separate groups to avoid confounding measure validation and training effects); and (3) the Domain × Validated interaction. Next, we fit another model adding the additional moderators of (1) mean MPO across treatment and control groups; and (2) the difference in mean MPO between treatment and control groups.

For the case report meta-analyses, we initially fitted the same meta-regression model with three moderators as for RCTs. We then fit 2 additional models adding one moderator at a time to this baseline model. First, we fit a model adding individual-level MPO. Second, we fit a model adding whether a study used the original MIT protocol or a modified protocol.

# Results

Study-level standardised mean difference scores and meta-analytic mean differences by Domain are shown in Figure 2. Full meta-regression results tables are reported in the Supplementary Materials.



Figure 2: Results of meta-analyses. Points are study-level standardised mean pretest-posttest difference scores, either adjusted for a control group (*gppc*) or not (*gpp*). Points of different colours are drawn from different studies. Large points are mean *gpp(c)* for validated measures with 66% (thick bar) and 95% (thin bar) confidence intervals and *t*-distribution confidence densities. For case reports, one aphasia severity study with *gppc* = .88 not displayed.

## RCT data

Overall, RCT data showed a small to moderate the pretest-posttest effect of MIT on aphasia outcomes, after accounting for the control group (*g̅* = .31 [95% CI −.01, .63]). These results were primarily based on Non-Communicative Language Expression (repetition) tasks. Other abilities were less commonly assessed. In moderator analyses, effects appeared to be much weaker for Communication and Language Comprehension tasks than for Non-Communicative Language Expression, but confidence intervals for these differences were wide (see Figure 2). Effects were estimated to be somewhat heterogeneous across studies (random effects standard deviation, *τ =*.33 [95% CI .15, 1.01]).

Two studies included several unvalidated measures of Non-Communicative Language Expression. For these measures, treatment effects for untrained items were similar to those for validated measures (∆*g̅* = −.15 [95% CI −.46, .15]). As expected, estimated treatment effects were much larger when patients were tested using trained items (∆*g̅* = .99 [95% CI .60, 1.39]; trained vs. untrained items contrast: 1.15 [95% CI .74, 1.56]). The similarity of effects for validated measures and untrained stimuli on unvalidated measures suggests that ad hoc measures may perform similarly in these contexts, so long as assessment stimuli are untrained (but cf. Ivanova & Hallowell, 2013, for a discussion of other benefits of measure validation).

When aphasia stage (MPO) was added to the RCT model, neither mean MPO across groups (∆*g̅* per month = −.008 [95% CI −.024, .008]) nor difference in mean MPO between MIT and control groups (∆*g̅* per month = −.004 [95% CI −.020, .011]) showed meaningful relationships with MIT treatment effects. Importantly, effect sizes for RCT analyses were drawn from only three studies, so these group-level MPO analyses have limited power to estimate the impact of MPO on MIT treatment effects.

## Case report data

Compared to RCT studies, case reports without a control group estimated much larger effects of MIT (*g̅* = 1.43 [95% CI .82, 2.02]). As with RCT studies, these results were primarily based on Non-Communicative Language Expression (repetition) tasks. Overall aphasia severity and language comprehension appeared to show somewhat smaller effects, but confidence intervals on these differences were very wide. Effects were estimated to be highly heterogeneous across studies (*τ* [between-studies] = .93 [95% CI .73, 1.19]), to the degree that MIT was even estimated to be harmful in a small proportion of settings (for instance, the 95% normal-theory prediction interval for Non-Communicative Language Expression ranged −0.49 to +3.40; IntHout et al., 2016).

Four studies included several unvalidatedmeasures of Non-Communicative Language Expression. As with RCT studies, treatment effects for untrained items were similar to those for validated measures (with a wide confidence interval; ∆*g̅* = .31 [95% CI −1.16, 1.80]), but apparent treatment effects were much larger for trained items (∆*g̅* = 1.86 [95% CI .40, 3.35]; trained vs. untrained items contrast: 1.56 [95% CI .67, 2.44]).

When aphasia stage (MPO) was added to the case reports model, MPO showed a moderate negative relationship with treatment effects (∆*g̅* per month = −.02 [95% CI −.02, −.01]; estimated effect for 12 months, −.19 [95% CI −.28, −.10]; estimated effect for 24 months, −.38 [95% CI −.55, −.20]).

Compared to studies that used the original MIT protocol, studies that used a modified variant of the protocol appeared to show larger treatment effects, though the confidence interval on this difference was very wide (∆*g̅* = .79 [95% CI −.99, 2.57]).

# Discussion

The present meta-analysis aimed to investigate the efficacy and clinical relevance of MIT while accounting for crucial methodological aspects of primary studies, such as validated outcomes, use of control comparison, and randomised allocation to treatment groups. Our results reveal that previously-trained items on unvalidated outcome measures inflate the effect size of MIT by a factor of ≈ 3.8 when compared to validated outcome measures. Similarly, the apparent effect size of MIT proved to be ≈ 4.2 times larger for non-RCT data compared to RCT data, pointing out that rigorous experimental design can help reduce confounds, to obtain more realistic estimates. In particular, this results reaffirms that adjusting for the control group is critical. Without it, most of the change observed in case reports, visible in inflated estimates of the efficacy of MIT, are inseparable from the phenomena of spontaneous recovery, as well as, ultimately, regression to the mean, none of which are due to the treatment.

Effect size also decreased with number of months post-stroke for IPD studies, indicating that progress in language performance reported in the acute stage of aphasia may confound MIT effects with influences of spontaneous recovery. In the late subacute or consolidation stages of aphasia, much of the spontaneous recovery is likely to have already occurred. This finding is also consistent with the smaller results for RCTs, which also account for spontaneous recovery effects by adjustment against a control group.

Taken together, these results suggest that validated outcomes, randomised-controlled designs and inclusion of individuals with chronic aphasia are essential prerequisites to determine the efficacy and clinical relevance of MIT (that is to say, statistical significance *and* substantial effect size) in a reliable way. Interestingly, variation of the original MIT protocol did not systematically alter the effect size, challenging the idea that modification of the treatment necessarily diminishes its outcome. Although MIT appears to benefit language performance on trained items, gains on untrained items emerged mostly from progress in repetition tasks. In contrast, the current results failed to demonstrate comparable progress in everyday communication ability on validated outcomes. These findings have implications in clinical practice.

As with any meta-analysis, the conclusiveness of the results strongly depends on the quality of the source material. As always, methodological shortcomings of primary studies emphasise the need for caution in interpreting the results. Nonetheless, the present meta-analysis considered various methodological aspects that tended to be neglected in previous work. In particular, our meta-analyses carefully determined the psychometric quality of each outcome, relative to recently defined standards in aphasia research (Wallace et al., 2019). In addition, the analyses accounted for quality of the research design, in terms of using control interventions and group randomisation to address unspecific influences, including bias due to placebo effects. In light of heterogeneity in treatment administration across studies, the analyses also reflect changes to the original MIT protocol. Our results demonstrate that the overall efficacy of MIT in repetition tasks appears to persist, albeit to a smaller degree than previously reported.

The current results indicate that changes to the original MIT protocol do not systematically alter the benefit from treatment. This finding casts doubt on the notion that the original composition and hierarchical structure of MIT are indispensable for improving language performance. Using unvalidated outcomes, cross-sectional and longitudinal multiple-case studies have examined the role of different MIT elements: melody and rhythm (e.g., Kershenbaum et al., 2019), vocal expression in unison or alone (Racette et al., 2006), left-hand tapping (e.g., Zipse et al., 2014), and formulaicity of verbal utterances (e.g., Stahl et al., 2011). Possible methodological reasons for seemingly contradictory data, as well as conjectured mechanism of MIT, have been discussed (Stahl & Kotz, 2014). Obviously, the present results do not offer insight into any of these mechanisms. If indeed adherence to the original MIT protocol does not manifest in significantly elevated language performance, our results encourage future research to optimise the composition and structure of the treatment, to increase its efficacy in the rehabilitation of neurological communication disorders. For example, individuals with apraxia of speech may benefit from several elements of MIT, such as rhythmic pacing (Aichert et al., 2016) and language formulaicity (Stahl et al., 2020).

According to the present meta-analysis, MIT leads to gains mainly in repetition tasks that reflect the ability to reproduce prior utterances in exactly the same form. Although this ability may facilitate the acquisition of novel words, it is not entirely clear to what extent it ultimately affects verbal behaviour in everyday communicative situations (Stahl et al., 2017). Our results indicate negligible progress on validated outcomes of everyday communication ability with MIT (*g̅* = −0.04 [95% CI −0.57, 0.48]). However, the number of studies reporting non-repetition outcomes was very comparatively very small. This produced wide confidence intervals and therefore uncertainty as to MIT’s effectiveness for these outcomes. Thus, although benefits from MIT cannot be ruled out, present evidence neither clearly supports them. In contrast, large-scale RCT data demonstrate that combining selected non-MIT methods *can* lead to moderate gains on validated outcomes of communication ability (Breitenstein et al., 2017). This finding suggests that individuals with aphasia should not rely exclusively on MIT if the primary goal is to improve everyday communication. Still, our meta-analysis should not undermine the importance of MIT-mediated progress on trained items. In individuals with severe forms of aphasia, this ‘palliative’ use of MIT may nonetheless entail a substantial increase in quality of life (Zumbansen et al., 2014). Critically, individuals with aphasia may perceive notable progress in language performance irrespective of statistically significant gains on validated outcomes. Known as ‘minimal clinically-important difference’ (Revicki et al., 2008), this diagnostic approach may be especially valuable for individuals where MIT can help establish a repertoire of trained phrases to convey basic needs in daily life (Van Lancker Sidtis, 2021).

We here present the first meta-analysis on MIT that attempts to rule out various methodological caveats in interpreting the outcome of the treatment, such as lack of validated outcomes, control group or randomisation in previous studies. Accounting for each of these issues in a rigorous way, the results of our meta-analyses confirm the promising role of MIT in improving language performance on trained items and in repetition tasks, while highlighting possible limitations in promoting everyday communication ability. We hope that the current work will be helpful for clinicians, patients and families to make informed decisions about their treatment options to support recovery from post-stroke aphasia.

# Acknowledgements

We gratefully acknowledge:

* Joost Hurkmans for providing unpublished work.
* Sarah Wallace for providing test-related information.
* Luisa Krein and Emily Braun for information regarding test scores.
* Ajay Halai and Yina Quique Buitrago for very useful comments.

# Author contributions (CRediT author statement)

Conceptualization: TP, BS, WTF

Methodology: TP, BMW

Software: TP, BMW

Validation: TP

Formal analysis: TP, BMW

Investigation: TP

Resources: TP, RB, WTF

Data Curation: TP, BMW, FH, TM

Writing - Original Draft: TP

Writing - Review & Editing: TP, BS, BMW, FH, TM, RB, WTF

Visualization: TP

Supervision: TP, BS

Project administration: TP

Funding acquisition: RB, WTF

additional ones not included in the CRediT author statement

literature search and curation: TP, HH, MZ

# References

Aichert, I., Späth, M., & Ziegler, W. (2016). The role of metrical information in apraxia of speech. Perceptual and acoustic analyses of word stress. *Neuropsychologia*, *82*, 171–178. https://doi.org/10.1016/j.neuropsychologia.2016.01.009

Albert, M. L., Sparks, R. W., & Helm, N. A. (1973). Melodic Intonation Therapy for Aphasia. *Archives of Neurology*, *29*(2), 130–131. https://doi.org/10.1001/archneur.1973.00490260074018

Bates, D. (2006, May 19). *[R] lmer, p-values and all that*. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015a). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015b). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blomert, L., Kean, M. L., Koster, C., & Schokker, J. (1994). Amsterdam—Nijmegen everyday language test: Construction, reliability and validity. *Aphasiology*, *8*(4), 381–407. https://doi.org/10.1080/02687039408248666

Brady, M. C., Kelly, H., Godwin, J., & Enderby, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, *5*. https://doi.org/10.1002/14651858.CD000425.pub3

Breitenstein, C., Grewe, T., Flöel, A., Ziegler, W., Springer, L., Martus, P., Huber, W., Willmes, K., Ringelstein, E. B., Haeusler, K. G., Abel, S., Glindemann, R., Domahs, F., Regenbrecht, F., Schlenck, K.-J., Thomas, M., Obrig, H., de Langen, E., Rocker, R., … Bamborschke, S. (2017). Intensive speech and language therapy in patients with chronic aphasia after stroke: A randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *The Lancet*, *389*(10078), 1528–1538. https://doi.org/10.1016/S0140-6736(17)30067-3

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. https://doi.org/10.1207/S15327906MBR3403\_2

Doppelbauer, L., Mohr, B., Dreyer, F. R., Stahl, B., Büscher, V., & Pulvermüller, F. (2021). Long-Term Stability of Short-Term Intensive Language-Action Therapy in Chronic Aphasia: A 1-2 year Follow-Up Study. *Neurorehabilitation and Neural Repair*, 15459683211029236. https://doi.org/10.1177/15459683211029235

Engelter, S. T., Gostynski, M., Papa, S., Frei, M., Born, C., Ajdacic-Gross, V., Gutzwiller, F., & Lyrer, P. A. (2006). Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke*, *37*(6), 1379–1384. https://doi.org/10.1161/01.STR.0000221815.64093.8c

Gerstman, H. L. (1964). A case of aphasia. *The Journal of Speech and Hearing Disorders*, *29*, 89–91. https://doi.org/10.1044/jshd.2901.89

Helm-Estabrooks, N., Morgan, A. R., & Nicholas, M. (1989). *Melodic intonation therapy*. Pro-Ed.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (Eds.). (2021). *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane. www.training.cochrane.org/handbook

Hurkmans, J., Bruijn, M. de, Boonstra, A. M., Jonkers, R., Bastiaanse, R., Arendzen, H., & Reinders-Messelink, H. A. (2012). Music in the treatment of neurological language and speech disorders: A systematic review. *Aphasiology*, *26*(1), 1–19. https://doi.org/10.1080/02687038.2011.602514

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Ivanova, M. V., & Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology*, *27*(8), 891–920. https://doi.org/10.1080/02687038.2013.805728

Kershenbaum, A., Nicholas, M. L., Hunsaker, E., & Zipse, L. (2019). Speak along without the song: What promotes fluency in people with aphasia? *Aphasiology*, *33*(4), 405–428. https://doi.org/10.1080/02687038.2017.1413487

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis: Three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Mills, C. K. (1904). Treatment of aphasia by training. *Journal of the American Medical Association*, *43*(26), 1940–1949.

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

R Core Team. (2021). *R: A language and environment for statistical computing* (4.1.0) [Computer software]. R Foundation for Statistical Computing. http://www.r-project.org/

Racette, A., Bard, C., & Peretz, I. (2006). Making non-fluent aphasics speak: Sing along! *Brain*, *129*(10), 2571–2584. https://doi.org/10.1093/brain/awl250

RELEASE Collaborators. (2021). Predictors of Poststroke Aphasia Recovery: A Systematic Review-Informed Individual Participant Data Meta-Analysis. *Stroke*, *52*(5), 1778–1787. https://doi.org/10.1161/STROKEAHA.120.031162

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, *61*(2), 102–109. https://doi.org/10.1016/j.jclinepi.2007.03.012

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*, c221. https://doi.org/10.1136/bmj.c221

Stahl, B., Gawron, B., Regenbrecht, F., Flöel, A., & Kotz, S. A. (2020). Formulaic Language Resources May Help Overcome Difficulties in Speech-Motor Planning after Stroke. *PloS One*, *15*(6), e0233608. https://doi.org/10.1371/journal.pone.0233608

Stahl, B., & Kotz, S. A. (2014). Facing the music: Three issues in current research on singing and aphasia. *Frontiers in Psychology*, *5*, 1033. https://doi.org/10.3389/fpsyg.2014.01033

Stahl, B., Kotz, S. A., Henseler, I., Turner, R., & Geyer, S. (2011). Rhythm in disguise: Why singing may not hold the key to recovery from aphasia. *Brain*, *134*(10), 3083–3093. https://doi.org/10.1093/brain/awr240

Stahl, B., Mohr, B., Dreyer, F. R., Lucchese, G., & Pulvermüller, F. (2017). Communicative-Pragmatic Assessment Is Sensitive and Time-Effective in Measuring the Outcome of Aphasia Therapy. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00223

van der Meulen, I., van de Sandt-Koenderman, M. E., & Ribbers, G. M. (2012). Melodic Intonation Therapy: Present Controversies and Future Opportunities. *Archives of Physical Medicine and Rehabilitation*, *93*(1, Supplement), S46–S52. https://doi.org/10.1016/j.apmr.2011.05.029

Van Der Meulen, I., Van De Sandt-Koenderman, M. W. M. E., Heijenbrok, M. H., Visch-Brink, E., & Ribbers, G. M. (2016). Melodic Intonation Therapy in Chronic Aphasia: Evidence from a Pilot Randomized Controlled Trial. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00533

van der Meulen, I., van de Sandt-Koenderman, W. Mieke. E., Heijenbrok-Kal, M. H., Visch-Brink, E. G., & Ribbers, G. M. (2014). The Efficacy and Timing of Melodic Intonation Therapy in Subacute Aphasia. *Neurorehabilitation and Neural Repair*, *28*(6), 536–544. https://doi.org/10.1177/1545968313517753

Van Lancker Sidtis, D. (2021). *Foundations of Familiar Language: Formulaic Expressions, Lexical Bundles, and Collocations at Work and Play* (1st edition). Wiley-Blackwell.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

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

Yamadori, A., Osumi, Y., Masuhara, S., & Okubo, M. (1977). Preservation of singing in Broca’s aphasia. *Journal of Neurology, Neurosurgery, and Psychiatry*, *40*(3), 221–224. https://doi.org/10.1136/jnnp.40.3.221

Zipse, L., Worek, A., Guarino, A. J., & Shattuck-Hufnagel, S. (2014). Tapped out: Do people with aphasia have rhythm processing deficits? *Journal of Speech, Language, and Hearing Research: JSLHR*, *57*(6), 2234–2245. https://doi.org/10.1044/2014\_JSLHR-L-13-0309

Zumbansen, A., Peretz, I., & Hébert, S. (2014). Melodic Intonation Therapy: Back to Basics for Future Research. *Frontiers in Neurology*, *5*. https://doi.org/10.3389/fneur.2014.00007

Zumbansen, A., & Tremblay, P. (2018). Music-based interventions for aphasia could act through a motor-speech mechanism: A systematic review and case–control analysis of published individual participant data. *Aphasiology*, *33*(4), 466–497. https://doi.org/10.1080/02687038.2018.1506089

# Appendix 1: Search terms

*CINAHL:*

((MM "Aphasia+") OR "aphasia" OR (MH "Aphasia, Broca") OR (MH "Aphasia, Transcortical Sensory") OR (MH "Aphasia, Wernicke")) AND (((MH singing or singing ) AND ("speech therapy" or (MH "Speech Therapy+")) OR "melodic intonation therapy")))

ClinicalTrials.gov

"melodic intonation therapy"

*Cochrane Library*

"melodic intonation therapy" AND aphasia

("speech therapy“ in Ti Abstr Key OR MeSH descriptor [speech therapy] explode all trees) AND singing\* in Ti Abstr Key

*Deutsches Register klinischer Studien (DRKS)/German clinical trials register*<https://www.drks.de/drks_web/>

melodic intonation therapy

melodische intonationstherapie

*Google Scholar*

„melodic intonation therapy“

„melodische intonationstherapie“

*ICTRP (International Clinical Trials Registry Platform)*

"melodic intonation therapy"

*Medline OVID*

("melodic intonation therapy".ab. or "melodic intonation therapy".ti. or "melodic intonation therapy".id. or ((singing.ab. or singing.id. or singing.ti. or singing/) AND (speech therapy/ or "speech therapy".ab. or "speech therapy".id. or "speech therapy".ti.))) AND (aphasia.ab. or aphasia.id. or aphasia.ti. or exp aphasia)

*Music Periodicals Database*

"melodic intonation therapy" and aphasia

*National Research Register (UK):* [*http://www.nihr.ac.uk/*](http://www.nihr.ac.uk/)

"melodic intonation therapy"

*Netherlands Trials Register www.trialregister.nl*

Melodic intonation

*OpenGrey.eu http://www.opengrey.eu/*

„melodic intonation therapy“

*ProQuest Dissertations & Theses Global*

„melodic intonation therapy“

*PsycINFO OVID*

("melodic intonation therapy".ab. or "melodic intonation therapy".ti. or "melodic intonation therapy".id. or ((singing.ab. or singing.id. or singing.ti. or singing/) AND (speech therapy/ or "speech therapy".ab. or "speech therapy".id. or "speech therapy".ti.))) AND (aphasia.ab. or aphasia.id. or aphasia.ti. or exp aphasia)

*PSYNDEX OVID*

("melodic intonation therapy".ab. or "melodic intonation therapy".ti. or "melodic intonation therapy".id. or ((singing.ab. or singing.id. or singing.ti. or singing/) AND (speech therapy/ or "speech therapy".ab. or "speech therapy".id. or "speech therapy".ti.))) AND (aphasia.ab. or aphasia.id. or aphasia.ti. or exp aphasia)

*PubMed*

(„melodic intonation therapy“ OR ((singing [MeSH Terms] OR singing) AND ("speech therapy"[MeSH Terms] OR "speech therapy")) AND (aphasia[MeSH Terms] OR aphasia)

*Scopus*

((( TITLE-ABS-KEY ( singing ) AND (TITLE-ABS-KEY ( "speech therapy" OR "language therapy" ))) OR ( TITLE-ABS-KEY ( "melodic intonation therapy" ))) AND ( TITLE-ABS-KEY ( aphasia))

*Web of Science*

TOPIC: (("melodic intonation therapy") OR TOPIC: ("singing" AND ("speech therapy" OR "language therapy")) AND TOPIC: aphasia

# Appendix 2: Dealing with duplicate findings

The systematic literature search in literature databases yielded 606 hits; through searching the trials registers 7 additional trials were found. These 613 items have been subjected to a duplicate check for identical publications found through the different search tools combined. They were checked with the “duplicate detection” feature of the reference management software Zotero, with the following procedure: When all of the duplicate items had the same publication form and all of the duplicates had an abstract, we kept the first one in the list (by order of importing into Zotero); if not, we kept the first item of the duplicates having an abstract. With different publication entries from the same study we followed the preference hierarchy: journal article > book chapter > conference proceedings. When a journal article was based on a dissertation, we kept the journal article and referred to the dissertation in the notes. After this procedure, 357 items remained. More duplicates were deleted manually (313 remaining).

## Appendix 2a: Detecting further duplicates and initial checking of the exclusion criteria

To continue our search of duplicates we manually checked the 156 studies remaining at this stage. These were sorted by title and compared in terms of dates, authors and types of article in order to detect duplicates that had trivial differences in those fields that prevented automated duplicate detection. From this step, the 143 abstracts were read and checked against the exclusion criteria.

After this step, 99 articles remained that checked all eligibility criteria. As a final step, the full texts of all remaining articles were reviewed, in order to make sure the studies contained all the information needed to be included in this meta-analysis (cf. exclusion criterion #N). During a first round, 13 articles were excluded and in the second round further 32 articles. This left 54 articles that fit our inclusion criteria.

See also the Prisma flow chart in Appendix N, which summarises the study counts above.

## Appendix 2b: Post hoc examination of articles and further exclusion

For those 54 articles found suitable we introduced an added post hoc scan using our exclusion criteria. 33 articles were excluded, leaving a total of 21 articles finally selected for inclusion in the meta-analysis.

# Appendix 3: PRISMA statement (checklist/flow diagram)

We used the PRISMA statement as found at http://prisma-statement.org/prismastatement/Checklist.aspx.

[[ bring latest version of that diagram below, from the local PDF ]]

# Appendix 4: List of excluded studies

*((paste here the @****excludedStudies*** *tab))*

# Supplementary Materials [[to be made into separate doc for submission]]

## 

## Supplementary Results

### Table 1. Overall RCT meta-analyses

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | Statistic | *df* | *p* | 95% conf. int. |
| *g̅* | 0.31 | 0.16 | 2.00 | 25 | 0.057 | [−0.01, 0.63] |
| τ | 0.25 |  | 212.08 | 25 | < 0.001 | [ 0.10, 1.11] |

*Note.* *g̅* = mean pretest-posttest difference (*gppc*; accounting for control group); τ = estimated random effects standard deviation across studies; Statistic = *t* value for *g̅* and *QE* value for τ; confidence intervals computed using *t* distributions for *g̅* and profile likelihood for τ and ρ.

### Table 2. RCT meta-analyses of broad ability categories

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | Statistic | *df* | *p* | 95% conf. int. |
| *g̅* (Non-communicative Language Expression) | 0.34 | 0.19 | 1.78 | 22 | 0.089 | [−0.06, 0.75] |
| *g̅* (Communication) | −0.04 | 0.25 | −0.16 | 22 | 0.872 | [−0.57, 0.48] |
| *g̅* (Language Comprehension) | −0.13 | 0.25 | −0.52 | 22 | 0.611 | [−0.64, 0.39] |
| ∆*g̅* (ad hoc test) | 0.23 | 0.13 | 1.76 | 22 | 0.092 | [−0.04, 0.49] |
| τ | 0.31 |  | 183.68 | 22 | < 0.001 | [ 0.13, 0.93] |
| ρ | −0.09 |  |  | 22 |  | [−0.45, 0.92] |

*Note.* *g̅* = mean pretest-posttest difference (*gppc*; accounting for control group); τ = estimated random effects standard deviation across studies; ρ = estimated correlation among *g* treatment effects between measures of different ability domains across studies; Statistic = *t* value for *g̅* and *QE* value for τ; confidence intervals computed using *t* distributions for *g̅* and profile likelihood for τ.

### Table 3. Overall IPD meta-analyses

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* | 1.43 | 0.30 | 4.84 | [0.82, 2.02] |
| τ | 1.00 |  |  | [0.59, 1.55] |
| σ (person) | 0.93 |  |  | [0.73, 1.20] |
| σ (measure) | 1.45 |  |  | [1.38, 1.52] |

*Note.* *g̅* = mean pretest-posttest difference (*gpc*; not accounting for any control group); τ = estimated random effects standard deviation across studies; σ (person) = estimated random effects standard deviation across persons (within study); σ (measure) = estimated random effects standard deviation across measures (within person); confidence intervals computed using profile likelihood; *p* values omitted as the appropriate denominator degrees of freedom for linear mixed effects models is ill-defined (Bates, 2006; Bates et al., 2015b, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

### Table 4. IPD meta-analyses of broad ability categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* (Non-communicative Language Expression) | 1.45 | 0.35 | 4.12 | [ 0.74, 2.14] |
| *g̅* (Aphasia Severity) | 0.52 | 0.44 | 1.16 | [−0.37, 1.38] |
| *g̅* (Communication) | 1.59 | 0.39 | 4.08 | [ 0.81, 2.35] |
| *g̅* (Domain-General Function) | 1.19 | 0.40 | 2.94 | [ 0.38, 1.97] |
| *g̅* (Language Comprehension) | 0.26 | 0.38 | 0.70 | [−0.50, 0.99] |
| *g̅* (Speech-Motor Planning) | 1.41 | 0.38 | 3.68 | [ 0.64, 2.15] |
| ∆*g̅* (ad hoc test) | 1.08 | 0.71 | 1.52 | [−0.32, 2.51] |
| τ | 1.15 |  |  | [ 0.70, 1.69] |
| σ (person) | 0.93 |  |  | [ 0.73, 1.19] |
| σ (measure) | 1.40 |  |  | [ 1.33, 1.47] |

*Note.* *g̅* = mean pretest-posttest difference (*gpc*; not accounting for any control group); ∆*g̅* = estimated difference in *g̅* between validated and ad hoc measures; note that only language expression included ad hoc measures; τ = estimated random effects standard deviation across studies; σ (person) = estimated random effects standard deviation across persons (within study); σ (measure) = estimated random effects standard deviation across measures (within person); confidence intervals computed using profile likelihood; *p* values omitted as the appropriate denominator degrees of freedom for linear mixed effects models is ill-defined (Bates, 2006; Bates et al., 2015b, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

### Table 5. IPD meta-analyses with aphasia stage (months post-onset, MPO) as a moderator

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* (Non-communicative Language Expression) | 1.57 | 0.21 | 7.33 | [ 1.16, 1.99] |
| *g̅* (Aphasia Severity) | 0.87 | 0.43 | 2.05 | [ 0.04, 1.70] |
| *g̅* (Communication) | 2.08 | 0.26 | 8.04 | [ 1.57, 2.58] |
| *g̅* (Domain-General Function) | 2.07 | 0.29 | 7.09 | [ 1.50, 2.64] |
| *g̅* (Language Comprehension) | 0.59 | 0.26 | 2.25 | [ 0.08, 1.11] |
| *g̅* (Speech-Motor Planning) | 1.86 | 0.25 | 7.49 | [ 1.38, 2.35] |
| ∆*g̅* (ad hoc test) | 1.09 | 0.42 | 2.59 | [ 0.27, 1.92] |
| ∆*g̅* (per month post-onset) | −0.02 | 0.0038 | -4.16 | [−0.02, −0.01] |
| τ | 0.00 |  |  | [ 0.00, 0.51] |
| σ (person) | 0.98 |  |  | [ 0.75, 1.24] |
| σ (measure) | 1.31 |  |  | [ 1.24, 1.38] |

*Note.* *g̅* = mean pretest-posttest difference (*gpc*; not accounting for any control group); ∆*g̅* = estimated difference in *g̅*; note that only language expression included ad hoc measures; τ = estimated random effects standard deviation across studies; σ (person) = estimated random effects standard deviation across persons (within study); σ (measure) = estimated random effects standard deviation across measures (within person); confidence intervals computed using profile likelihood; *p* values omitted as the appropriate denominator degrees of freedom for linear mixed effects models is ill-defined (Bates, 2006; Bates et al., 2015b, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

### Table 6. IPD meta-analyses with MIT protocol as a moderator

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* (Non-communicative Language Expression) | 1.35 | 0.82 | 1.64 | [−0.17, 2.86] |
| *g̅* (Aphasia Severity) | 0.21 | 0.92 | 0.23 | [−1.48, 1.90] |
| *g̅* (Communication) | 1.57 | 0.85 | 1.86 | [ 0.01, 3.13] |
| *g̅* (Domain-General Function) | 1.14 | 0.85 | 1.34 | [−0.43, 2.71] |
| *g̅* (Language Comprehension) | 0.05 | 0.84 | 0.05 | [−1.50, 1.59] |
| *g̅* (Speech-Motor Planning) | 1.34 | 0.84 | 1.60 | [−0.21, 2.89] |
| ∆*g̅* (ad hoc test) | 0.25 | 1.08 | 0.23 | [−1.75, 2.24] |
| ∆*g̅* (modified MIT protocol) | 0.79 | 0.96 | 0.82 | [−0.99, 2.57] |
| τ | 1.38 |  |  | [ 0.70, 2.01] |
| σ (person) | 0.98 |  |  | [ 0.75, 1.30] |
| σ (measure) | 1.42 |  |  | [ 1.35, 1.49] |

*Note.* *g̅* = mean pretest-posttest difference (*gpc*; not accounting for any control group); ∆*g̅* = estimated difference in *g̅*; note that only language expression included ad hoc measures; τ = estimated random effects standard deviation across studies; σ (person) = estimated random effects standard deviation across persons (within study); σ (measure) = estimated random effects standard deviation across measures (within person); confidence intervals computed using profile likelihood; *p* values omitted as the appropriate denominator degrees of freedom for linear mixed effects models is ill-defined (Bates, 2006; Bates et al., 2015b, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

### Table 7. IPD meta-analyses of studies using ad hoc language expression measures: Comparing trained vs untrained stimuli

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* (untrained stimuli) | 1.40 | 0.37 | 3.76 | [0.67, 2.31] |
| ∆*g̅* (trained stimuli) | 1.56 | 0.24 | 6.58 | [1.08, 2.03] |
| τ | 0.47 |  |  | [0.00, 1.40] |
| σ (person) | 0.61 |  |  | [0.27, 1.23] |
| σ (measure) | 0.73 |  |  | [0.56, 0.95] |

*Note.* *g̅* = mean pretest-posttest difference (*gpc*; not accounting for any control group); ∆*g̅* = estimated difference in *g̅*; note that only language expression included ad hoc measures; τ = estimated random effects standard deviation across studies; σ (person) = estimated random effects standard deviation across persons (within study); σ (measure) = estimated random effects standard deviation across measures (within person); confidence intervals computed using profile likelihood; *p* values omitted as the appropriate denominator degrees of freedom for linear mixed effects models is ill-defined (Bates, 2006; Bates et al., 2015b, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

## Supplementary References

# [[To do for this manuscript]]

* check that reporting conforms to the guidelines in the
  + PRISMA statement (Moher, 2009)
  + cochrane handbook for systematic reviews of interventions, https://training.cochrane.org/handbook/current

1. For a small number of studies, it was not possible to determine the maximum or minimum possible scores. For these studies, we computed POMP scores using the maximum and minimum *observed* scores in the sample. Results did not change meaningfully if we excluded these studies from results. [↑](#footnote-ref-1)
2. For comparison, we also estimated models with unequal random effects variances across dependent variables. This did not improve model fit based on AICc comparison or likelihood ratio tests. [↑](#footnote-ref-2)
3. As only 3 RCT studies were identified, it was not possible to apply methods to detect publication-bias or other small-sample effects (e.g., tests of funnel plot asymmetry). [↑](#footnote-ref-3)
4. Models with random slopes for the *Domain* variable did not converge, likely due to the limited co-occurrence of specific pairs of those within any one study. [↑](#footnote-ref-4)