medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Melodic Intonation Therapy for aphasia: A multi-level meta-analysis of randomised-controlled-trial and individual-participant data

Tudor Popescu[1,2*], Benjamin Stahl[3-6], Brenton M. Wiernik[7], Felix Haiduk[1], Michaela Zemanek[1], Hannah Helm[3], Theresa Matzinger[1,8], Roland Beisteiner[2], W. Tecumseh Fitch[1]

[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; benjamin.stahl@charite.de*

[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*

[8] *University of Vienna, Department of English, Vienna, Austria*

[*] corresponding author: Tudor Popescu, Department of Behavioural and Cognitive Biology; University of Vienna, A-1030 Vienna; tudor.popescu@univie.ac.at; https://orcid.org/0000-0001-5890-1520

Please note: We welcome questions, comments, citation, and constructive criticism, bearing in mind this is a draft preprint, subject to revision, and currently undergoing peer-review. Please direct any correspondence to the first author.

Recommended citation: Popescu, T., Stahl, B., Wiernik, B., Haiduk, F., Zemanek, M., Helm, H., Matzinger, T., Beisteiner, R., & Fitch, W. T. (forthcoming). *Melodic Intonation Therapy for aphasia: A multi-level meta-analysis of randomised-controlled-trial and individual-participant data*.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

ABSTRACT

Background and objectives: Melodic Intonation Therapy (MIT) is a prominent rehabilitation programme for individuals with post-stroke aphasia. Despite substantial progress in recent years, the efficacy of MIT remains not fully understood. Based on *a-priori* hypotheses, the present meta-analysis investigated the efficacy of MIT while considering quality of outcomes (psychometrically validated *vs.* unvalidated measures), experimental design (presence *vs.* absence of randomisation and control group), influence of spontaneous recovery (quantified as number of months post-stroke), MIT version applied (original *vs.* modified protocol), and level of generalisation (performance on trained *vs.* untrained items).

Methods: An extensive literature search in all major online databases and trials registers, including also solicitations for unpublished studies, identified 606 studies (years searched: 1973-2021). Inclusion criteria: randomised controlled trial (RCT) data or case reports on adults with aphasia; pre-post assessment of language performance. Exclusion criteria: substantial variation from original MIT protocol; unvalidated outcomes, unless both trained and untrained items were compared; essential information not indicated/retrievable. Following PRISMA guidelines, studies were double-coded. Multi-level mixed- and random-effects models were used to separately meta-analyse RCT and non-RCT data.

Main outcomes and measures: Measures of language performance focused on aphasia severity, everyday communication ability, domain-general function, language comprehension, non-communicative language expression, and speech-motor planning.

Results: The final sample consisted of 22 studies, comprising data from 129 patients overall. Best-quality evidence from RCTs with validated measures estimated a small-to-moderate standardised MIT treatment effect of g_{ppc} = .35 [-0.08, 0.78] for non-communicative language expression, with substantial uncertainty. Unvalidated outcomes appeared to attenuate MIT's effect size by 23% (non-RCT) to 43% (RCT) when compared to validated outcomes. Moreover, MIT's apparent effect size was 5.7 times larger for non-RCT data compared to RCT data. Effect size also decreased with number of months post-stroke, suggesting the non-RCT estimate is confounded with spontaneous recovery, primarily within the first year post-stroke. In contrast, variation from the original MIT protocol did not systematically alter benefit from treatment. Crucially, analyses demonstrated significantly improved performance on trained and untrained items. The latter finding arose mainly from gains in repetition tasks, rather than other domains of verbal expression including everyday communication ability.

Discussion: 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.

Keywords: melodic intonation therapy; meta-analysis; speech; singing; post-stroke aphasia; experimental design

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1. 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 often occurs in combination with aphasia. Although about a third of individuals with neurological communication disorders do not recover completely[1], rehabilitation programmes can improve language performance even in the chronic stage of symptoms[2].

Melodic Intonation Therapy (MIT) is a prominent rehabilitation programme originally developed for individuals with non-fluent aphasia[3]. Drawing on the observation that individuals with neurological communication disorders are often able to sing entire pieces of text fluently[4–6], 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[7]. Hypotheses on MIT's neural mechanisms have been discussed elsewhere[8].

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)[9], but not in the *chronic* stage of aphasia (i.e., more than 6–12 months after stroke)[10]. From a methodological point of view, influences of spontaneous recovery are generally lower in the *chronic* stage of aphasia, as suggested by RCT data[11] and meta-analyses[12]. 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., 13].

So far, there are several systematic reviews on MIT[e.g., 14,15] and two meta-analyses[16,17]. These meta-analyses reflect a relatively limited amount of RCT data[16] or dichotomise post-treatment improvement in a way that prevents specific estimates of effect size[17]. 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, namely how the effect size of MIT is systematically altered by:

1. **Psychometric quality of outcomes**: use of validated vs. unvalidated tests in outcome measures;
2. **Experimental design**: RCT vs. non-RCT studies;
3. **Confound by spontaneous recovery**, decreasing with number of months post-onset of stroke (MPO);
4. **Variance in protocol**: original vs. slightly modified MIT variants;
5. **Degree of generalisability**: performance on trained vs. untrained items.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2. METHODS

2.1. Eligibility criteria

We defined the following basic **inclusion criteria** for primary 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.

We chose to include case reports with individual patient data (IPD) to increase the pool of evidence. To determine the influence of experimental design on treatment outcome, we analysed RCT and non-RCT studies separately and comparatively.

After removal of duplicate items (see section 1 in the online Supplementary Materials), the following **exclusion criteria** were applied to remaining studies:

- substantial variation from original MIT protocol[3]. We accepted *minor* changes to the MIT protocol (and examined the effect of the categorical variable: original *vs.* 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 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;
- 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);
- publication in non-peer-reviewed or predatory journal.

Taken together, the included studies comprise 129 patients (59 in RCTs; 70 in IPDs) and 62 controls (all in RCTs). The full list of included and excluded studies can be found in eTable 1 and eTable 2 (Supplementary Materials).

2.2. Search strategy

To obtain high search sensitivity, we used both free-text and subject headings in databases for our search, not restricting by language or publication form[18]. Figure 1 shows the PRISMA statement chart (Preferred Reporting Items for Systematic Reviews and Meta-Analyses[19]), which summarises the study counts at all stages of the search. The full counts are given in section 1 of the Supplementary Materials, which also documents the full literature search procedure, including search terms, databases used, and attempts made to reach the "grey literature".

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

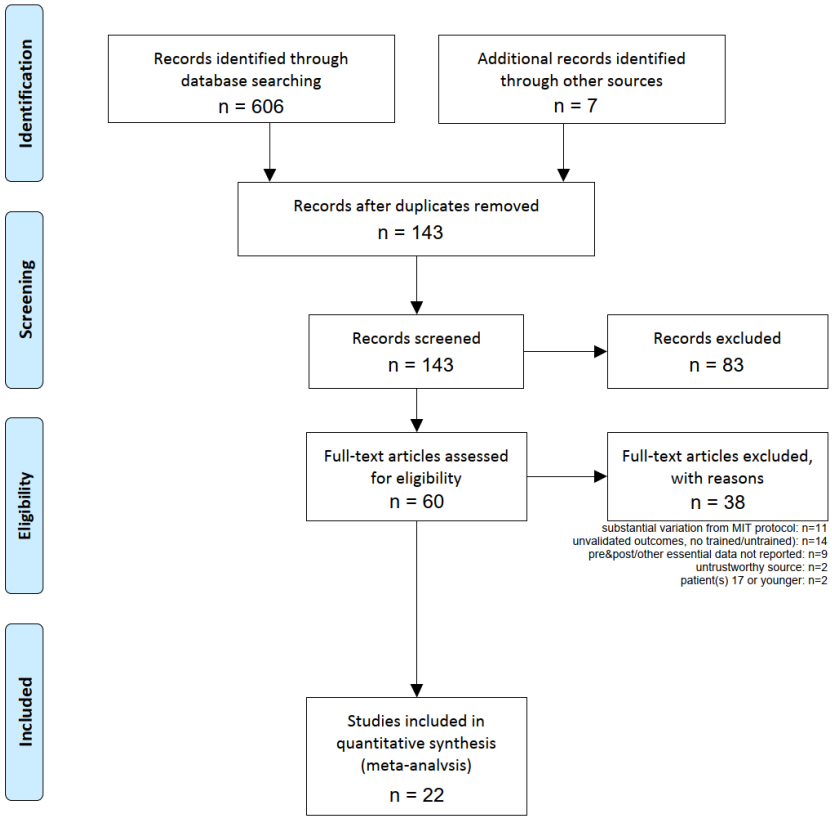


Figure 1: **Flow diagram from the PRISMA statement (**http://prisma-statement.org/prismastatement/Checklist.aspx). The lists of included and excluded studies are in eTables 1 and 2 of the Supplementary Materials, respectively.

Furthermore, we followed the guidelines and standards in the Methodological Expectations of Cochrane Intervention Reviews (MECIR) handbook, and those in the PRISMA checklist (see Supplementary Materials).

2.3. 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, which amounted to errors arisen from numerically estimating data reported in plot format only.

2.4. Tests and outcome measures in primary studies

All test batteries reported in the studies considered, and their validation status, are reported in eTable 3 of the Supplementary Materials. eTable 4 *ibid.* shows in detail which of the subtests of these test batteries are measuring which linguistic *Ability*; the associated *Target syndrome* (aphasia or apraxia of speech); and the hierarchical categorisation scheme that determined in each case the dependent variable meta-analysed (*Domain*).

2.5. MIT variants

Aside from the original version of MIT, eight variants were reported in the studies considered, of which we included five: "SIPARI", "Music therapy", "Singing therapy", "Speech–Music Therapy for Aphasia"

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(SMTA), and "Modified Melodic Intonation Therapy" (MMIT). The excluded MIT variants (cf exclusion criteria) were: choir therapy, metrical pacing technique (MPT), music therapy combined with SLT.

2.6. Meta-analysis methods

2.6.1 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[a] normalised raw scores to reflect the proportion of the maximum possible score, POMP[20]. 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 2). 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 2: **The nested multilevel model employed**. Standard deviations (τ) are shown at the measure level (📏), nested within patients (👤), nested within studies (📖).

For models specifically fitted to the RCT and the case report data respectively, see section 4 in the Supplementary Materials.

2.6.2 Moderator analyses

For the RCT meta-analyses, we fitted 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, to test the effect of time since stroke, 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 a similar meta-regression model with Domain, Validated, and Domain × Validated moderators. We tested additional moderators by fitting two 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.

[a] 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.7. Data availability

All data generated during the making of this work, including raw materials, analysis scripts, and supplementary materials, have been uploaded to the OSF repository https://osf.io/gcjqr/.

3. Results

Study-level standardised mean difference scores and meta-analytic mean differences by Domain are shown in Figure 3. Full meta-regression results tables are reported in the Supplementary Materials, eTables 5–10.

Figure 3: **Results of meta-analyses**. Points are study-level standardised mean pretest-posttest difference scores, either adjusted for a control group (g_{ppc}) or not (g_{pp}). Points of different colours are drawn from different studies. Large points are mean $g_{pp(c)}$ 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 $g_{ppc} = -4.88$ is not displayed.

3.1. RCT data

Overall, RCT data showed a small-to-moderate pretest-posttest effect of MIT on aphasia outcomes, after accounting for the control group ($\bar{g} = .31$ [95% CI −.01, .63]). These results were primarily based on

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

measures of Non-communicative Language Expression (i.e., focus on verbal utterances *per se*, such as in tasks requiring repetition of words and sentences). Other abilities were less commonly assessed, including Communication (i.e., verbal utterances used for social interaction in everyday situations) and Language Comprehension (i.e., understanding the meaning of verbal utterances). 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 3). Effects were estimated to be somewhat heterogeneous across studies (random effects standard deviation, τ = .33 [95% CI .15, 1.01]).

Two included RCTs included several unvalidated measures of Non-communicative Language Expression. For these measures, treatment effects for untrained items were somewhat smaller than those for validated measures, though the confidence interval for this difference was fairly wide ($\Delta\bar{g}$ = −.15 [95% CI −.46, .15]). As expected, estimated treatment effects were much larger when patients were tested using trained items ($\Delta\bar{g}$ = .99 [95% CI .60, 1.39]; trained vs. untrained items contrast: 1.15 [95% CI .74, 1.56]). Smaller effect sizes for unvalidated measures may be attributable to poorer reliability compared to validated measures; measurement error tends to attenuate effect sizes[21–23].

When aphasia stage (MPO) was added to the RCT model, neither mean MPO across groups ($\Delta\bar{g}$ per month = −.008 [95% CI −.024, .008]) nor difference in mean MPO between MIT and control groups ($\Delta\bar{g}$ 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.

3.2. Case report data

Compared to RCT studies, case reports with no control group estimated much larger effects of MIT ($\bar{g}$ = 1.72 [95% CI 1.00, 2.42]). 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] = 1.41 [95% CI .89, 2.05]), 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 −0.88 to +4.90[24].

Four case report studies included several unvalidated measures of Non-communicative Language Expression. As with RCT studies, treatment effects for untrained items on unvalidated measures appeared to be smaller than those for validated measures (with a wide confidence interval; $\Delta\bar{g}$ = −.47 [95% CI −2.40, 1.46]). Also similar to RCTs, apparent treatment effects were much larger for trained items ($\Delta\bar{g}$ = 2.37 [95% CI .44, 4.31]; trained vs. untrained items contrast: 2.84 [95% CI 1.21, 4.48]).

When aphasia stage (MPO) was added to the case reports model, MPO showed a moderate negative relationship with treatment effects ($\Delta\bar{g}$ per month = −.02 [95% CI −.03, −.01]; estimated effect for 12 months, −.18 [95% CI −.30, −.07]; estimated effect for 24 months, −.37 [95% CI −.61, −.14]).

Compared to studies that used the original MIT protocol, studies that used a modified variant of the protocol appeared to show somewhat larger treatment effects, though the confidence interval on this difference was very wide ($\Delta\bar{g}$ = .56 [95% CI −.92, 2.03]).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4. Discussion

The present meta-analysis aimed to investigate the efficacy of MIT while accounting for crucial methodological aspects of primary studies such as control comparisons, randomised group allocation, use of validated outcomes, and variance in MIT protocol. It also examined the confounding effect of spontaneous recovery, and the degree to which MIT's effect generalises to untrained items.

Overall, we found that MIT had a limited positive treatment effect in specific domains (mainly repetition tasks), in line with previous meta-analyses. However, our results reveal that poor methodology may introduce substantial bias into estimated treatment effects. Concerning RCT studies of Non-communicative Language Expression, using unvalidated outcomes for untrained items may attenuate MIT's effect size by about 43% when compared to validated outcomes ($\bar{g}_{unvalidated}$ = .20 vs. $\bar{g}_{validated}$ = .35). Holding language domain and outcome validity constant, MIT's effect size proved to be 5.7 times larger for non-RCT data compared to RCT data ($\bar{g}_{case\ report}$ = 2.01 vs. $\bar{g}_{RCT}$ = .35 for validated Non-communicative Language Expression measures). Implications and possible sources for each of these findings are discussed below.

4.1. Research implications

The current results indicate that appropriate study design can help reduce confound to obtain more realistic estimates. In particular, these results re-affirm the importance of setting up and adjusting for adequate control interventions. Otherwise, most of the changes observed in case reports—visible in inflated estimates of efficacy (the 5.7 factor reported)[b]—are inseparable from phenomena of spontaneous recovery, and ultimately regression to the mean, none of which are due to the treatment.

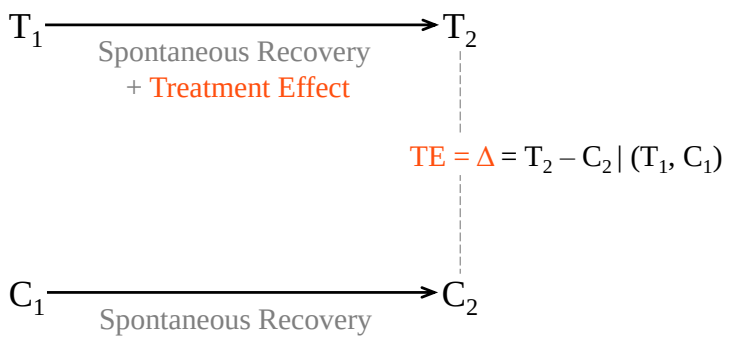
Effect sizes were found to decrease with number of months post-stroke (MPO) for IPD studies, indicating that progress in language performance reported in the late subacute or consolidation stage of aphasia may arise from influences of spontaneous recovery[c]. Currently available data do not allow conclusions about whether MIT's effect size increases or decreases with MPO, given the general lack of positive RCT evidence on speech-language therapy in subacute aphasia[16]. 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 of MIT in a reliable way.

Figure 4 schematically illustrates the need for a control group in order to estimate the treatment effect (TE), net of any effects due merely to the passage of time, such as (in disease) spontaneous recovery. Case series report T_2-T_1 and tend to interpret it as the TE. This however confounds TE with the spontaneous recovery effect (SR). To isolate TE, a control group is needed: from it, we compute T_2-C_2 (accounting for baseline differences at time 1) to estimate the actual TE. Figure e1 in the Supplementary Discussion additionally illustrates these relations in the form of causal diagrams (directed acyclic graphs).

[b] See also eTable 11 in the Supplementary Materials, which reports RCT meta-analyses only considering the change in control groups.

[c] See also eTable 12 in the Supplementary Materials, which reports IPD meta-analyses with MPO as a moderator for pretest scores only.

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



T_2 - $T_{1:}$ TE + SR

T_2 - C_2 | (T_1, C_1): (TE + SR) – SR = TE

Figure 4: **Treatment and spontaneous recovery effects in interventions**. T_1 and T_2 represent the treatment group, at time 1 and time 2 (pre and post treatment). C_1 and C_2 represent the control group, at the same two time points. TE, treatment effect. The | operator denotes "accounting for". The first equation shows how TE gets confounded with SR. The second equation shows how the confounding effect of SR can be removed. Also see Figure e1 (Supplementary Materials) for a causal diagram.

4.2. Clinical implications

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[25]. Our RCT results indicate negligible progress on validated outcomes of everyday communication ability with MIT. The number of non-repetition outcomes was comparatively small, regardless of experimental design, implying that benefits from MIT cannot be ruled out completely; nonetheless, current evidence does not support 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[2]. 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 entail a substantial increase in quality of life[15]. 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'[26], 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[27]. Conversely, it would be recommendable for future studies to address the impact of MIT on quality of life.

4.3. Limitations and future directions

As with any meta-analysis, the conclusiveness of the results strongly depends on the quantity and quality of the available sources. Our rigorous eligibility criteria left us with a low number of included studies. This small sample size in turn led to large confidence intervals, which necessarily limit the strength of our conclusions. As new clinical studies using MIT become available, future meta-analyses will be able to draw conclusions and issue recommendations with greater confidence – *insofar as these studies are less beset by methodological shortcomings of the sort we have pointed out*.

As it stands, such shortcomings in our included studies call for yet further caution in interpreting the results. Our meta-analysis considered various methodological aspects largely neglected in previous work. In particular, it carefully determined the psychometric quality of each outcome, relative to recently

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

defined standards in aphasia research[28]. In addition, our evaluation 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. Our results confirm the overall efficacy of MIT in repetition tasks, albeit to a smaller degree than previously reported.

Interestingly, deviations in the MIT protocol used relative to the original did not systematically alter the effect size. This finding casts doubt on the notion that the original composition and hierarchical structure of MIT are indispensable for improving language performance. However, few of the included studies employed an MIT variant, and their individual effects are heterogeneous. Therefore, our results can express no certainty about the impact of modifications to the original MIT protocol, and instead highlight the need for high-quality research exploring the influence of specific modifications.

Using unvalidated outcomes, cross-sectional and longitudinal multiple-case studies have examined the role of different MIT elements: melody and rhythm[e.g., 29], vocal expression in unison or alone[30], left-hand tapping[e.g., 31], and formulaicity of verbal utterances[e.g., 32]. Possible methodological reasons for seemingly contradictory data, as well as conjectured mechanisms of MIT, have been discussed[e.g., 33]. 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[34] and language formulaicity[35].

4.4. Conclusion

We here present the first meta-analysis on MIT that attempts to monitor the effects of various methodological caveats in interpreting the outcome of previous studies, such as lack of validated outcomes, control group or randomisation. 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.

5. Funding statement

This work was supported by a research-cluster grant from the Medical University of Vienna and University of Vienna (SO10300020), awarded to W.T.F. and R.B.

6. Acknowledgements

We gratefully acknowledge Ajay Halai and Yina Quique Buitrago for their helpful comments; Joost Hurkmans for providing unpublished work; and Sarah Wallace, Luisa Krein and Emily Braun for providing test-related information.

7. Author contributions

CRediT author statement:

- Conceptualisation: TP, BS, WTF
- Methodology: TP, BMW
- Software: TP, BMW
- Validation: TP
- Formal analysis: TP, BMW
- Investigation: TP
- Resources: TP, RB, WTF
- Literature search and curation: TP, HH, MZ
- Data curation: TP, BMW, FH, TM
- Writing - original draft: TP, BS
- Writing - review & editing: TP, BS, BMW, FH, WTF
- Visualisation: TP, BMW
- Supervision: TP, BS
- Project administration: TP, RB, WTF
- Funding acquisition: RB, WTF

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

8. Competing interests

The authors declare no competing interests.

9. Table of figures (figure legends)

Figure 1: Flow diagram from the PRISMA statement (http://prisma-statement.org/prismastatement/Checklist.aspx). The lists of included and excluded studies are in eTables 1 and 2 of the Supplementary Materials, respectively. 5
Figure 2: The nested multilevel model employed. Standard deviations (τ) are shown at the measure level (📏), nested within patients (👤), nested within studies (📖). 6
Figure 3: Results of meta-analyses. Points are study-level standardised mean pretest-posttest difference scores, either adjusted for a control group (g_{ppc}) or not (g_{pp}). Points of different colours are drawn from different studies. Large points are mean $g_{pp(c)}$ 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 $g_{ppc} = -4.88$ is not displayed. 7
Figure 4: Treatment and spontaneous recovery effects in interventions. T_1 and T_2 represent the treatment group, at time 1 and time 2 (pre and post treatment). C_1 and C_2 represent the control group, at the same two time points. TE, treatment effect. The | operator denotes "accounting for". The first equation shows how TE gets confounded with SR. The second equation shows how the confounding effect of SR can be removed. Also see Figure e1 (Supplementary Materials) for a causal diagram. 10

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

10. REFERENCES

1. Engelter, S. T. *et al.* Epidemiology of aphasia attributable to first ischemic stroke: incidence, severity, fluency, etiology, and thrombolysis. *Stroke* **37**, 1379–1384 (2006).
2. Breitenstein, C. *et al.* 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**, 1528–1538 (2017).
3. Albert, M. L., Sparks, R. W. & Helm, N. A. Melodic Intonation Therapy for Aphasia. *Arch. Neurol.* **29**, 130–131 (1973).
4. Gerstman, H. L. A case of aphasia. *J. Speech Hear. Disord.* **29**, 89–91 (1964).
5. Mills, C. K. Treatment of aphasia by training. *J. Am. Med. Assoc.* **43**, 1940–1949 (1904).
6. Yamadori, A., Osumi, Y., Masuhara, S. & Okubo, M. Preservation of singing in Broca's aphasia. *J. Neurol. Neurosurg. Psychiatry* **40**, 221–224 (1977).
7. Helm-Estabrooks, N., Morgan, A. R. & Nicholas, M. *Melodic intonation therapy*. (Pro-Ed, 1989).
8. Merrett, D. L., Peretz, I. & Wilson, S. J. Neurobiological, Cognitive, and Emotional Mechanisms in Melodic Intonation Therapy. *Front. Hum. Neurosci.* **8**, (2014).
9. van der Meulen, I., van de Sandt-Koenderman, W. Mieke. E., Heijenbrok-Kal, M. H., Visch-Brink, E. G. & Ribbers, G. M. The Efficacy and Timing of Melodic Intonation Therapy in Subacute Aphasia. *Neurorehabil. Neural Repair* **28**, 536–544 (2014).
10. Van Der Meulen, I., Van De Sandt-Koenderman, M. W. M. E., Heijenbrok, M. H., Visch-Brink, E. & Ribbers, G. M. Melodic Intonation Therapy in Chronic Aphasia: Evidence from a Pilot Randomized Controlled Trial. *Front. Hum. Neurosci.* **10**, (2016).
11. Doppelbauer, L. *et al.* Long-Term Stability of Short-Term Intensive Language-Action Therapy in Chronic Aphasia: A 1-2 year Follow-Up Study. *Neurorehabil. Neural Repair* 15459683211029236 (2021) doi:10.1177/15459683211029235.
12. RELEASE Collaborators. Predictors of Poststroke Aphasia Recovery: A Systematic Review-Informed Individual Participant Data Meta-Analysis. *Stroke* **52**, 1778–1787 (2021).
13. Blomert, L., Kean, M. L., Koster, C. & Schokker, J. Amsterdam—Nijmegen everyday language test: construction, reliability and validity. *Aphasiology* **8**, 381–407 (1994).
14. van der Meulen, I., van de Sandt-Koenderman, M. E. & Ribbers, G. M. Melodic Intonation Therapy: Present Controversies and Future Opportunities. *Arch. Phys. Med. Rehabil.* **93**, S46–S52 (2012).
15. Zumbansen, A., Peretz, I. & Hébert, S. Melodic Intonation Therapy: Back to Basics for Future Research. *Front. Neurol.* **5**, (2014).
16. Brady, M. C., Kelly, H., Godwin, J. & Enderby, P. Speech and language therapy for aphasia following stroke. *Cochrane Database Syst. Rev.* (2016) doi:10.1002/14651858.CD000425.pub3.
17. Zumbansen, A. & Tremblay, P. 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**, 466–497 (2018).
18. *Cochrane Handbook for Systematic Reviews of Interventions*. (Cochrane, 2021).
19. Moher, D. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann. Intern. Med.* **151**, 264 (2009).
20. Cohen, P., Cohen, J., Aiken, L. S. & West, S. G. The problem of units and the circumstance for POMP. *Multivar. Behav. Res.* **34**, 315–346 (1999).
21. Ivanova, M. V. & Hallowell, B. A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology* **27**, 891–920 (2013).
22. Wiernik, B. M. & Dahlke, J. A. Obtaining unbiased results in meta-analysis: the importance of correcting for statistical artifacts. *Adv. Methods Pract. Psychol. Sci.* **3**, 94–123 (2020).
23. van Smeden, M., Lash, T. L. & Groenwold, R. H. H. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int. J. Epidemiol.* **49**, 338–347 (2020).
24. IntHout, J., Ioannidis, J. P. A., Rovers, M. M. & Goeman, J. J. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* **6**, e010247 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2021.08.28.21262764; this version posted January 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

25. Stahl, B., Mohr, B., Dreyer, F. R., Lucchese, G. & Pulvermüller, F. Communicative-Pragmatic Assessment Is Sensitive and Time-Effective in Measuring the Outcome of Aphasia Therapy. *Front. Hum. Neurosci.* **11**, (2017).
26. Revicki, D., Hays, R. D., Cella, D. & Sloan, J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* **61**, 102–109 (2008).
27. Van Lancker Sidtis, D. *Foundations of Familiar Language: Formulaic Expressions, Lexical Bundles, and Collocations at Work and Play*. (Wiley-Blackwell, 2021).
28. Wallace, S. J. *et al.* A core outcome set for aphasia treatment research: The ROMA consensus statement. *Int. J. Stroke* **14**, 180–185 (2019).
29. Kershenbaum, A., Nicholas, M. L., Hunsaker, E. & Zipse, L. Speak along without the song: what promotes fluency in people with aphasia? *Aphasiology* **33**, 405–428 (2019).
30. Racette, A., Bard, C. & Peretz, I. Making non-fluent aphasics speak: sing along! *Brain* **129**, 2571–2584 (2006).
31. Zipse, L., Worek, A., Guarino, A. J. & Shattuck-Hufnagel, S. Tapped out: do people with aphasia have rhythm processing deficits? *J. Speech Lang. Hear. Res. JSLHR* **57**, 2234–2245 (2014).
32. Stahl, B., Kotz, S. A., Henseler, I., Turner, R. & Geyer, S. Rhythm in disguise: why singing may not hold the key to recovery from aphasia. *Brain* **134**, 3083–3093 (2011).
33. Stahl, B. & Kotz, S. A. Facing the music: three issues in current research on singing and aphasia. *Front. Psychol.* **5**, 1033 (2014).
34. Aichert, I., Späth, M. & Ziegler, W. The role of metrical information in apraxia of speech. Perceptual and acoustic analyses of word stress. *Neuropsychologia* **82**, 171–178 (2016).
35. Stahl, B., Gawron, B., Regenbrecht, F., Flöel, A. & Kotz, S. A. Formulaic Language Resources May Help Overcome Difficulties in Speech-Motor Planning after Stroke. *PloS One* **15**, e0233608 (2020).