MIT meta-analyses

Brenton M. Wiernik

05/22/2021

## Meta-analysis methods

### Outcome metric

To maximize comparability of effects across studies, we used change scores from pretest to posttest as the outcome variable, expressed in *z*-scores. For group-level randomized control trials, we standardized *z*-scores using pooled pretest standard deviation across control and treatment groups. For individual-patient data case series, 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 normalized 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 pretest POMP scores, with individual test scores nested within patients nested within studies. From these models, we used the population intercept as the estimated mean POMP score and the patient-level random effects standard deviation as the estimated POMP score *SD*. We used this mean and *SD* to standardize the pretest and posttest POMP scores.

### Group-level RCT

For group-level RCT studies, we computed group-level 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 multilevel 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 ability domain moderator nested within studies (Konstantopoulos, 2011). We used a homoskedastic 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., ) 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 (Knapp & Hartung, 2003) -distribution method 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 2.4-0) (Viechtbauer, 2010).[[3]](#footnote-3)

### Individual patient data (IPD) case series

For the IPD case series studies, we computed individual-level scores as the difference between pretest and posttest *z*-scores (the mean difference in this scores is the pretest-posttest Hedges’ *g*, ). We then pooled data across studies using a three-level random-effects IPD meta-analysis, with individual scores 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., 2015).

### Moderator Analyses

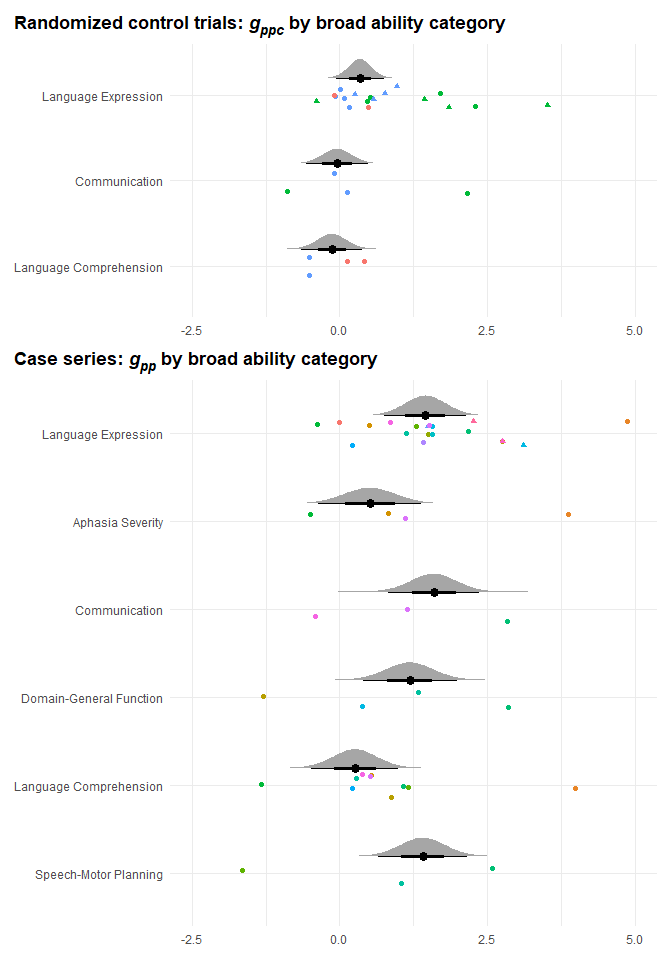
For the RCT meta-analyses, we fit a meta-regression model including (1) ability domain, (2) whether the study used a validated test or ad hoc measure, and (3) the domain × validated interaction as moderators. Next, we fit another model adding (1) the mean months post-onset (MPO) across treatment and control groups and (2) the difference in mean MPO between treatment and control groups as additional moderators.

For the IPD meta-analyses, we also fit a meta-regression model including (1) ability domain, (2) whether the study used a validated test or ad hoc measure, and (3) the domain × validated interaction as moderators. We then fit 3 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. Third, we fit a model including whether the tested stimuli for the outcome were also included in the training set (“trained”) or not (“untrained”). For this third analysis, all studies used ad hoc measures of language comprehension, so the domain and validation moderators were omitted from the model.

## Results

Study-level standardized mean difference scores and meta-analytic mean differences by ability domain are shown in Figure 2. Full meta-regression results tables are available in the online supplemental materials.

#### Figure 2. Results of meta-analyses



*Note.* Points are study-level standardized mean pretest-posttest difference scores, either adjusted for a control group (*gppc*) or not (*gpp*). Points with different colors drawn from different studies. = validated measure; = unvalidated measure. Large points are mean *gppc* with 66% (thick bar) and 95% confidence (thin bar) intervals and *t*-distribution confidence densities. For case series, one aphasia severity study with *gppc* = .88 omitted.

### RCT studies

Overall, RCT studies showed a small to moderate 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 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 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 =*.31 [95% CI .13, .93]).

#### Moderator analyses

Two studies included several ad hoc measures of language expression. These tended to show larger effects than validated language expression measures (∆*g̅* = .23 [95% CI −.04, .49]). This may plausibly indicate that ad hoc measures inflate the effect (e.g., validated measures may have better reliability and specificity to detect true treatment effects; cf. Ivanova & Hallowell, 2013).

When aphasia stage (months post-onset) terms were added to the RCT model, neither mean MPO across groups (∆*g̅* per month = −.006 [95% CI −.021, .010]) nor difference in mean MPO between MIT and control groups (∆*g̅* per month = −.005 [95% CI −.020, .010]) showed meaningful relationships with MIT treatment effects. Importantly, effect sizes for RCT analyses were drawn from only 3 studies, so these group-level MPO analyses have limited power to estimate impacts of MPO on MIT treatment effects.

### IPD Case Series studies

Compared to RCT studies, case series studies 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 language expression (repetition) tasks. Overall aphasia severity and language comprehension appeared to show somewhat smaller effects, but confidence intervals on these differences were wide. Effects were estimated to be highly heterogeneous across studies (*τ* [between-studies] = 1.15 [95% CI .70, 1.69]), to a degree that MIT was estimated to be harmful in some settings [e.g., 95% normal-theory prediction interval for language expression ranged −0.90 to 3.80; IntHout et al. (2016)].

#### Moderator analyses

Four studies included several ad hoc measures of language expression. As with RCT studies, these tended to show larger effects than validated language expression measures (∆*g̅* = 1.08 [95% CI −.32, 2.51]).

When aphasia stage (months post-onset) was added to the IPD 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 protocol appeared to show larger treatment effects, though the confidence interval on this difference was very wide (∆*g̅* = .79 [95% CI −.99, 2.57]).

For language expression studies that used ad hoc measures, when patients were tested using the same stimuli used in training, apparent treatment effects were much larger than when patients were tested using familiar (trained) stimuli (∆*g̅* = 1.56 [95% CI 1.08, 2.03]). Apparent improvement on novel (untrained) stimuli was similar in magnitude to that for validated language expression measures (*g̅* = 1.40 [95% CI .67, 2.31]). Thus, assessing patient improvement using trained stimuli appears to exaggerate MIT treatment effects to an even greater degree than case series designs alone.[^5]

[^5] In a supplemental analysis including the trained vs. untrained moderator, months post-onset (MPO), and their interaction, MPO had a negligible impact on the size of the trained vs. untrained effect (*β* = .004 [95% CI −.01, .02]; estimated effect for 12 months, .05 [95% CI −.13, .22]; compared to model with only trained vs. untrained moderator: likelihood-ratio *χ*(2) = 3.44, *p* = .179, ∆AIC (Akaike information criterion): 0.564).

## 

## Supplemental Tables

#### 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* (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., 2015, 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* (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., 2015, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | *SE* | *t* | 95% conf. int. |
| *g̅* (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., 2015, 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* (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., 2015, 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., 2015, pp. 34–35); inference should be based on the profile likelihood confidence intervals.

## References

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

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

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>

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>

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>

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*. R Foundation for Statistical Computing. <http://www.r-project.org/>

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>

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>

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 meaningful change if we excluded these studies from results. [↑](#footnote-ref-1)
2. For comparison, we also estimated models with 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 group-level 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 ability domain variable did not converge, likely due to the limited co-occurrence of specific pairs of ability categories within any one study. [↑](#footnote-ref-4)