Operationalising treatment success in aphasia rehabilitation

Caterina Breitenstein[1*], Katerina Hilari[2], Maya Menahemi-Falkov,[3-4], Miranda L. Rose[3-4], Sarah J. Wallace[4-5], Marian C Brady[6], Argye E. Hillis[7-9], Swathi Kiran[10], Jerzy P. Szaflarski[11], Donna C Tippett[7-8,12], Evy Visch-Brink[13], & Klaus Willmes[14]

[1] Department of Neurology with Institute of Translational Neurology, University of Muenster, Germany

[2] School of Health Sciences, Centre for Language and Communication Science Research, City, University of London, London, United Kingdom

[3] School of Allied Health, Human Services and Sport, La Trobe University, Melbourne, Australia

[4] Centre of Research Excellence in Aphasia Recovery and Rehabilitation, La Trobe University, Melbourne, Australia

[5] School of Health and Rehabilitation Sciences, Queensland Aphasia Research Centre, The University of Queensland, Australia

[6] Nursing, Midwifery and Allied Health Professions Research Unit, Glasgow Caledonian University, United Kingdom

[7] Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[8] Department of Physical Medicine and Rehabilitation Johns Hopkins University School of Medicine, Baltimore, MD, USA

[9] Department of Cognitive Science, Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, USA

[10] College of Health & Rehabilitation Sciences: Sargent College, Boston University, Boston, MA, USA

[11] Departments of Neurology, Neurosurgery, and Neurobiology, University of Alabama at Birmingham Heersink School of Medicine, Birmingham, AL, USA.

[12] Department of Otolaryngology-Head and Neck Surgery,
Johns Hopkins University School of Medicine, Baltimore, MD, USA

[13] Department of Neurology and Neurosurgery, Erasmus University Medical Centre, Rotterdam, The Netherlands

[14] Department of Neurology, University Hospital RWTH Aachen, Germany

*corresponding author

Keywords: aphasia, treatment effectiveness, patient-relevant outcome, benchmarking healthcare, individualised medicine

This paper is part of a special series of papers in *Aphasiology* on Methodological Issues in Aphasia Trials. The series comprises tutorial-type papers with core recommendations for aphasia intervention studies and randomised controlled trials. The series is guest edited by Professor Katerina Hilari, Dr Caterina Breitenstein, Dr Erin Godecke, Dr Helen Kelly and Professor Miranda Rose on behalf of the Trials for Aphasia Panel of the Collaboration of Aphasia Trialists https://www.aphasiatrials.org/

Abstract

Background

Treatment success is the desired outcome in aphasia rehabilitation. However, to date, there is a lack of consensus on what defines a 'successful' result on a given aphasia outcome measurement instrument (OMI).

Aim

In this methodological paper, we present strategies for how to define and measure treatment success on a given OMI at the group level, as well as for an individual person with aphasia. The latter is particularly important when research findings from group studies are clinically implemented for individuals in rehabilitation.

Scope

We start by presenting methods to calculate the average *statistically significant* change across several (group) studies (e.g., standardised mean difference, raw unstandardised mean difference) for a given OMI. Such metrics are useful to summarise an overall effect of the intervention of interest, particularly in meta-analyses. However, benchmarks based on group effects are not feasible for assessing an individual participant's treatment success and thus for determining the proportion of patients who had a beneficial response to therapy (overall response rate of an intervention). We therefore recommend a distribution-based approach to determine benchmarks of *statistically significant* treatment response at the individual level, i.e., the 'smallest detectable change' for a given OMI, whichrefers to the smallest change that can be detected by the OMI beyond measurement error. However, the statistical significance of an individual treatment effect does not necessarily correspond to its clinical impact. This requires an additional indicator. The benchmark to determine a *clinically relevant* improvement on a given OMI is the 'minimal important change'. The minimally important change is defined as the smallest OMI change scoreperceived as *important* by the relevant stakeholder group (i.e., people with aphasia, their relatives/caregivers, clinicians). It therefore requires relating the individual OMI

change scores to ‘anchors’, i.e., meaningful external criteria, preferably based on patient-perceived therapy success. Currently, there is no consensus regarding the optimal ‘anchors’ and their respective definition of clinically important change in aphasia outcome research.

Conclusions/Recommendations

Operationalising individual treatment success based on both statistically significant and (patient-reported) clinically meaningful benchmarks is a key priority in aphasia rehabilitation. Availability of such measures will (a) facilitate estimates of therapy response rate in intervention studies and thus optimise therapeutic decisions and (b) provide stakeholder groups (e.g., thesociety, the stroke team, people with aphasia, family, clinicians, healthcare professionals with objective, statistically reliable and meaningful feedback on individual treatment response in the clinical setting.

Background

The priorities of people with stroke and aphasia are of paramount importance in both research and rehabilitation. A James Lind Alliance priority-setting partnership in the UK asked stroke survivors, their family/carers, and health professionals what matters most to people affected by stroke (Stroke Association, 2021). Assessingand treating communication difficulties was ranked third among the top ten research priorities related to life post-stroke and thus was given an even higher priority than the recovery of abilities necessary for everyday life such as returning to work or driving (ranked 6th) and exercises to improve strength and fitness (ranked 9th).

Additionally, the inability to communicate was rated as 'equal to or worse than death' by the majority of acute/early subacute stroke patients (Everett et al., 2021). A similar priority setting partnership activity~~,~~ was limited to stroke survivors with aphasia, their family/carers and speech and language therapists and identified the top 10 research priorities relating to chronic aphasia (Franklin et al., 2018). This activity identified the *most effective aphasia treatment*(s) as the highest research priority. A requirement for a treatment to be effective is treatment success. However, to date there is no consensus of what constitutes treatment success in aphasia. It is unquestionable that the optimal outcome of aphasia rehabilitation is complete resolution of the deficits with return of language functions, activities, and social life participation to the pre-stroke state (a 'cure' so to speak). The reality, however, is that approximately 40% of people with stroke initially present with aphasia (Mitchell et al.,2021) and, about 65% of stroke survivors leave the hospital with a disability (National Institute for Health and Care Excellence, 2019). At three-months, about 25% of stroke survivors still have aphasia (Ali et al., 2015), and at one year, 20% suffer from persistent communication disability (Dijkerman et al., 1996; El Hachioui et al., 2013). Thus, complete language recovery may be an unrealistic outcome expectation for many stroke survivors with aphasia. Complete recovery may be ~~,~~particularly unlikely for people with aphasia in the later stages of stroke, when the brain's powerful spontaneous recovery processes have diminished (Bernhardt et al., 2017; Pedersen et al., 1995).

A suitable definition for treatment success in late sub-acute (3-6 months post-stroke) (Bernhardt et al., 2017) and chronic (>6 months post-stroke) aphasia is that the *a priori* defined intervention targets have been significantly modified in the desired direction after

the intervention and remain stable for an *a priori* defined duration after the treatment. In aphasia rehabilitation, stakeholder groups may differ in what they consider a successful treatment outcome (Rai et al., 2015). Even within any given stakeholder group, the definition may depend on factors such as aphasia type, severity or time since stroke.

As such, we will briefly describe various stakeholder perspectives on treatment success in post-stroke aphasia and will then suggest available standardised outcome measurement instruments (~~/~~OMIs) for each of these perspectives. We ordered the sectionsin terms of how far removed each of these perspectives is to the person with aphasia, starting with the 'socioeconomic/societal' and the 'stroke team' perspectives, followed by the 'aphasia rehabilitation' perspective.

STAKEHOLDER PERSPECTIVES AND MEASUREMENT OF THEIR RESPECTIVE INTERVENTION GOALS

Socioeconomic/societal perspective

Across the globe, stroke is a major cause of long-term disability, accounting for about one third of overall ~~(~~direct~~)~~ healthcare costs worldwide and being accompanied by enormous indirect costs such as loss of productivity (Rochmah et al., 2021). From a socioeconomic or societal perspective, stroke interventions should thus aim to significantly reduce the costs associated with acute stroke care and rehabilitation as well as (if applicable) increase the probability of a vocational rehabilitation.

Costs of rehabilitation services. For stroke survivors across age groups, the *reduction of rehabilitation service needs* is an important treatment outcome from a socioeconomic perspective (Rai et al., 2015). The annual cost for treating stroke totaled about 10% of the overall costs for the 19 major groups of neurological and mental disordersin Europe in 2010 (Olesen et al., 2012). Stroke with aphasia significantly contributes to the high costs, starting as early as the first week after the stroke, due to higher inpatientcomplication rates with longer hospital stays as compared to stroke without aphasia (Boehme et al., 2016; Lazar & Boehme, 2017). Overall, about 50% more rehabilitation services are required for those with aphasia after the initial stroke compared to strokes

without aphasia (Flowers et al., 2016). The attributable 1-year medical costs were approximately 8.5% higher in stroke with aphasia compared to without aphasia (Ellis et al., 2012).

Cost effectiveness of aphasia interventions. From a societal perspective, particularly from the healthcare funder's point of view, any treatment provided as part of routine clinical care should be cost effective, i.e., the costs associated with a treatment should be aligned with the expected gains in health outcome. A frequent approach to determine the cost effectiveness of stroke interventions (and of other diseases) is to relate the incremental costs of an intervention to the increment of *'Quality-adjusted Life Years (QALYs)'* compared to a control condition, usually standard care (Raftery et al., 2015). This approach is also known as '*cost-utility*' analysis.

QALYs refer to the expected number of years lived in 'perfect health' after a disease started (i.e., post the initial stroke). 'Perfect health' is assigned a utility value of 1, whereas death is given a utility value of 0. Living one year in 'perfect health' after a stroke would yield a QALY of one, surviving half a year in 'perfect health' would yield a QALY of 0.5 (formula: utility value x years of life). If the post-stroke condition is merely 'half of perfect health', a utility value of 0.5 would be assigned according to the formula, yielding a QALY of 0.5 for one year of survival (or 0.25 QALY for half a year of survival). The 'utility value' of a disease before and after an intervention is frequently estimated by scoring the generic health status of a patient using the European Quality of Life 5 Dimensions (EQ-5D) (Rabin & de Charro, 2001). The EQ-5D captures health status on five dimensions (mobility, self-care, usual activities, pain or discomfort, anxiety or depression), but does not assess communication health status. The questionnaire is available in more than 200 languages, two different response formats (3 or 5 levels of degree of 'problems') and for different stakeholders (self- report versus proxy rating) (https://euroqol.org/eq-5d-instruments; weblink last accessed on 20th October 2021). The psychometrically superior 5-level EQ-5D-5L (Janssen et al., 2013) has been widely used in general stroke outcome research , but only a few aphasia intervention study protocols to date include the EQ-5D (either 3-level~~L~~ or 5-level~~L~~ version) for estimating the cost effectiveness of an intervention (Hilari et al., 2019; Northcott et al., 2019; Rose et al., 2019; Stahl et al., 2019; Tarrant et al., 2018; van der Gaag & Brooks, 2008). For future aphasia trials, we recommend the assessment of an extended version of the EQ-5D-5L, which captures an additional cognitive dimension of stroke (EQ-5D-5L+C) (de Graaf et al.,

2020). Furthermore, an aphasia-friendly pictorial variant of the EQ-5D-3L/-5L is available (Whitehurst et al., 2018) and has been validated in a recent aphasia intervention study (Big CACTUS) (Palmer et al., 2019). Palmer and colleagues (2019) estimated that the incremental cost of self-managed computerised naming therapy as an adjunct to usual care was less than $1.000/patient for each QALYgained in comparison to usual care alone in a late subacute/chronic post-stroke aphasia sample. However, intervention-related improvements in health-related quality of life as assessed with the EQ-5D-5L were very small (< 0.04 QALYs, depending on anomia severity atbaseline); (Latimer et al., 2020).

An alternative cost-effectiveness analysis approach is to calculate incremental cost-effectiveness ratios, i.e., the costs incurred by routine clinical service to achieve 1% gain in the (primary) OMI. A recent retrospective cost-effectiveness analysis based on 19 single-subject experimental post-stroke aphasia interventions studies reported that the incremental cost of obtaining a 1% improvement on the primary outcome measure increased from $ 7 to $ 40 across the first 17 therapy sessions (based on clinical service salary levels in the year 2006), and no measurable improvements occurred after session 17 (Ellis et al., 2014). Several aphasia trial protocols (Godecke et al., 2018; Rose et al., 2021) describe the planned calculation of incremental cost-effectiveness ratios, but thesetrial results are not yet available.

A third strategy to align intervention costs with the expected health improvement is to determine the resources required fora particular intervention when delivered as part of routine clinical service. A recent examplefor such a '*costing-only*' study is a feasibility trial on the effects of virtual group social support in aphasia rehabilitation (Marshall et al., 2020), reporting average intervention cost(excluding hardware) of £1,364 ($2,000) per participant. However, the efficacy of the intervention probed in this study still needs to be determined in a properly powered randomised controlled trial.

Return to work. A successful intervention from a socioeconomic/societal perspective will ideally allow stroke survivors ofworking age to *return to pre-stroke employment*. This working-age subgroup forms about

20% of the stroke population (Heuschmann et al., 2010), with an increase in stroke incidence at younger ages over the past decades (Kissela et al., 2012).

In a recent Finnish registry-based study, the likelihood of *not* returning to work within a year after a stroke was about three times higher for working-age stroke survivors with moderate/severe aphasia as compared to those without aphasia (Aarnio et al., 2018). Within a working-age French aphasia sample, only about 15% had returned to work within 1.5 years post-stroke (Doucet et al., 2012). Additional reports in the literature are consistent in that less than one quarter of working-age PWA return to work post-stroke (Black-Schaffer & Osberg, 1990; Caporali & Basso, 2003; Graham et al., 2011; Hinckley, 1998; Parr et al., 1997).

Additionally, among the small fraction of stroke survivors who manage to return, few return to full time employment or to their pre-stroke occupational roles (Hinckley, 1998). Besides possibly stroke/aphasia severity (Ashley et al., 2019), factors such as workplace flexibility, social support, and personal motivation may contribute to successful return to work (Hinckley, 2002). Nevertheless, return to work is not a frequent treatment outcome to date in aphasia rehabilitation, even though the importance of speech and language therapy to facilitate return to work in those with mild aphasia was identified decades ago (Darley et al., 1980). Moreover, return to work was rated a top desired activity by PWA in the late sub-acute stage post-stroke (Haley et al., 2019).

To summarise, from a socioeconomic or societal perspective, any intervention provided as part of routine clinical care should not only be effective and (if applicable) aimed at vocational rehabilitation, but its costs need to be aligned with the expected health-related improvement. We therefore consider it pivotal for clinical trials in aphasia rehabilitation to report formal cost effectiveness analyses for the probed intervention(s). These analyses should adhere to established guidelines such as those laid out by the Task Force report of the International Society for Pharmacoeconomics and Outcomes Research Randomized Clinical Trials—Cost-Effectiveness Analysis (ISPOR RCT-CEA; Ramsey et al., 2015) or by the EuropeanStroke Organisation Health Economics working group (Cadilhac et al., 2020).

The perspective of the stroke team.

Neurological disorders including stroke remain the second leading cause of death worldwide (Global Burden of Disease Neurology Collaborators, 2019). Therefore, from a physician's perspective, the primary treatment goals early after stroke typically are to ensure and prolong survival (Ekker et al., 2019) and to limit the degree of damage to the brain. Rapid resolution of neurological deficits (including language) by restoring blood flow to dysfunctional tissue (Hillis, 2007; Hillis et al., 2003) and strategies to prevent a recurrent stroke (Kleindorfer et al., 2021) are additional critical goalsof the early stage after a stroke.

During the later stages post-stroke, *functional independence in activities of daily living* for improving the stroke survivors' health-related quality of life is considered the primary goal of both physicians and non-physician rehabilitation specialists. For example, in the UK the National Clinical Guideline for Stroke (developed inclose cooperation with patient representatives as has been recommended by Hinckley et al., 2014) highlights that "patients almost always interpret their illness in terms of its impact on their activities and social participation" (Intercollegiate Stroke Working Party, 2016, p. 6). Therefore, *activity and social participation* should be preferentially targeted in stroke rehabilitation in the UK. In the United States, the clinical practice guidelines for rehabilitation of aphasia suggest that speech and language therapy should focus on facilitating the recovery of communication, and assist in development of strategies to facilitate communication, decrease isolation, and provide a supportive environment (Winstein et al., 2016). In Germany, the major focus of intervention outcome is on gaining *maximum functional independence* in everyday activities (Gerdes et al., 2012).

A variety of standardised outcome measurement instruments (OMIs) are used globally to quantify the *amount of assistance required* to carry out various activities of daily living. According to a recent systematic review, the OMIs most frequently used in stroke trials are the Functional Independence Measure (FIM) and Functional Assessment Measure (FAM)

(Galeoto et al., 2019). The FIM and FAM predominantly refer to bodily functions (such as grooming, dressing, eating, physical mobility) rather than languageand communication. For example, two of the 18 FIM items assess communication, yet refer to basic daily needs while in hospital, such as communicating hunger/thirst and discomfort. The FIM does not assess more complex real-life communication needs, such as scheduling appointments by phone, negotiating home finances, or selecting and contractingan internet provider. Furthermore, interrater agreement and thus replicability of the cognitive FIM items is rather low (Ottenbacher et al., 1996). Patient and public involvement during construction of these scales was also limited (Granger et al., 1986).

Similar criticism applies to another widely used OMI to assess functional independence in activities of daily life, the Barthel Index (Mahoney & Barthel, 1965). The Barthel Index is also limited to items referring to basic physical needs and physical mobility, which explained merely 24% of the variance of aphasia recovery across the first year after stroke (El Hachioui et al., 2013). Another frequently used OMI in stroke outcomeresearch is the Modified Rankin Scale (mRS), which is a 7-point Likert rating scale for clinicians, ranging from 0 = 'no symptoms' across increasing levels of functional dependence to 6 = 'dead' (Broderick et al., 2017). A mRS score of 0 to 2 (no or minor symptoms, no assistance required) is considered a desirable outcome in stroke trials (Broderick et al., 2017), and a change score of one has been interpreted as clinically relevant (Harrison et al., 2013). The mRS has been recommended as a core measure for global disability in stroke trials on sensorimotor recovery (Kwakkel et al., 2017), but the scale is not suitable to assess cognitive and social functioning (de Haan et al., 1995). Its sensitivity to detect change instroke rehabilitation trials has been questioned (McGill et al., 2021).

In terms of assessing *functional independence in everyday communication* activities, a range of psychometrically sound aphasia measures exists (Wallace et al., 2020). For example, the following OMIs tap into communication activities or functional communication: Amsterdam Nijmegen Everyday Language Test/ANELT (Blomert et al., 1994); The Scenario Test (van der Meulen et al., 2010); Communicative Effectiveness Index/CETI (Lomas et al., 1989);, Communication Outcome after Stroke (COAST) scale (Long et al., 2008)]; Communication Activities of Daily Living-Third Edition (CADL-3) (Holland et al., 2018); and the Aphasia Communication Outcome Measure (ACOM) (Hula et al., 2015).

Many of the measures have sound psychometric properties (e.g., inter-/intra-rater and retest-reliability, measures of validity). However, most fail to meet all the psychometric standards recommended by the COnsensus-based Standards for the selectionof health Measurement INstruments (COSMIN) guidelines, particularly with respect to content and structural validity (Mokkink et al., 2016; Mokkink et al., 2010). Furthermore, even psychometrically top-ranking and widely used OMIs do not directly assess functional independence in 'real' everyday life communication. Rather, they use an indirect approach, such as role played communication scenarios (e.g., ANELT, The Scenario Test) or (proxy-rated) questionnaires (e.g., CETI).

An additional scale to assess communication independence is the single-item Aphasia Severity Rating Scale of the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001), which is a 6-point rating scale for clinicians (ranging from 0 = 'no usable speech or auditory comprehension' to 5 = 'minimal discernible speech handicap') and has been used to assess the *degree of assistance required for verbally communicating* about everyday life topics, e.g., family life, work, hobbies. The psychometric properties of this clinician-rated scale have not yet been evaluated, but the scale proved sensitive to the degree of recovery from aphasia across the first year post-stroke (El Hachioui et al., 2013).

The American Speech-Language-Hearing Association Functional Assessment of Communication Skills for adults (ASHA-FACS) (Frattali et al., 1995) also entails a scale 'Communicative Independence Scale' to grade the *degree of assistance required for performing verbal and nonverbal interactions* in four different life domains (social communication; communication of basic needs; reading/writing/number concepts; daily planning). However, the ASHA-FACS requires multiple observational sessions with the person with aphasia prior to the rating (Worrall L., 2000) and has thus limited clinical feasibility. Furthermore, analyses employing item response theory (IRT) modeling suggested that the social communication scale category labels may need to be reduced to four instead of seven; and responsiveness to changehas not yet been demonstrated (Meier et al., 2017).

Thus, the development of an easily administered and psychometrically sound OMI, which fulfills the criteria established by the COSMIN group to determine *functional independence in everyday life communication* scenarios is thus urgently required in aphasia rehabilitation.

The aphasia rehabilitation perspective

The views of multiple other stakeholder groups (most importantly people with aphasia (PWA), but also their family members, aphasia researchers, clinicians, and managers) have been systematically explored with a series of e-Delphi consensus and nominal group technique studies (structured small-group discussion to reach a consensus).These studies will be presented in turn before synthesising their findings.

People with aphasia

Despite the manifest language and communication challenges, PWA need to be involved in the discussion of intervention targets (Hinckley et al., 2014). Recently, PWA in the chronic stage post-stroke who were based in six different countries spanning five continents were asked, "What would you most like to change about your communication and the way aphasia affects your life?" (Wallace et al., 2017b, p. 1374). Content analysis of the group's responses revealed that top treatment outcomes spanned all components of the International Classification of Functioning, Disability and Health (WHO, 2001). Improved communication (including improved language functions) as well as increased life participation (e.g., ability to participate in conversations) and recovered normality (e.g., to be functionally independent) were the top patient-defined[1]outcomes of an aphasia intervention.

Family members of PWA

In the same study (Wallace et al., 2017b), the authors asked family members of PWAs which aphasia treatment outcomes for the PWA *they* would rank highest. Family members also ranked improved communication (beyond the communication of basic needs) as well as recovered normality as the two most important treatment outcomes.

[1] A patient-defined approach focusses on patients' treatment expectations in deciding what constitutes a successful treatment outcome (Zeppieri Jr. & George, 2017). A 'patient-defined' outcome may be a 'patient-reported' outcome, but self-reporting is not a critical requirement here.

Aphasia researchers

Following an international e-Delphi consensus approach, top outcome priorities of this stakeholder group were patient-reported impact of and satisfaction with treatment, improved communication-related quality of life, and improved language function in modalities relevant to the study's aims (Wallace et al., 2016).

Aphasia clinicians and managers

The ability to take part in conversations and to communicate/participate in various settings/roles were seen as the most important treatment outcomes for PWA~~s~~ from the perspective of aphasia clinicians and managers (Wallace et al., 2017a).

Synthesis of the findings for the four different stakeholder groups (PWA and their family members; researchers, clinicians, see Table 1) showed agreement for the following top treatment outcomes: Improving language function, communication, emotional well-being, quality of life (QoL) as well as patient-reported impact of and satisfaction with treatment (Wallace et al., 2019).

Table 1. Important aphasia treatment outcomes by the four different 'aphasia rehabilitation' stakeholder groups (based onWallace et al., 2016; Wallace et al., 2017a; Wallace et al., 2017b)

People with Aphasia	Family Members	Clinicians & Managers	Aphasia Researchers
1. Improved communication 2. Increased life participation 3. Changed attitudes through increased awareness and education about aphasia 4. Recovered normality 5. Improved physical and emotional well-being 6. Improved health services	*Outcomes for the person with aphasia:* 1. Improved communication 2. Recovered normality 3. Improved physical and emotional well-being 4. Increased life participation *Outcomes for family members:* 1. Improved communication 2. Increased life participation 3. Improved health and support services 4. Changed attitudes through increased awareness and education about aphasia 5. Improved emotional well-being 6. Recovered normality	*Outcomes for person with aphasia:* 1. Good psychosocial well-being 2. Able to participate in different roles and contexts 3. Positive feelings about communication 4. Satisfied and feels that they have improved 5. Able to communicate information of varying complexity 6. Improved communication 7. Able to participate in conversation 8. Able to communicate in different roles and contexts 9. Able to use multimodal communication/ strategies to support communication 10. The goals of the person with aphasia have been met 11. The communicative environment is enhanced	1. Impact of treatment from the perspective of the person with aphasia 2. Communication-related quality of life 3. Satisfaction with intervention from the perspective of the person with aphasia 4. Language functioning in modalities relevant to study aims 5. Satisfaction with ability to communicate from the perspective of the person with aphasia 6. Satisfaction with participation in activities from the perspective of the person with aphasia

People with Aphasia	Family Members	Clinicians & Managers	Aphasia Researchers
		12. Improved functioning, reduced disability, and able to be discharged *Outcomes for family/carers/significant others:* 1. Better communication partners 2. Good knowledge about aphasia and better attitudes towards people with aphasia 3. Less third-party disability 4. Engage in therapy (for the person with aphasia) *Outcomes relating to health services:* 1. Access to services and funding 2. Efficient use of resources and measurement of outcomes *Outcomes relating to health professionals:* 1. Greater awareness about aphasia and how to support communication	

After identification of the mutually agreed upon treatment outcomes across the four stakeholder groups, an additional consensus process was initiated to select the most appropriate psychometrically robust OMIs for each of these six top treatment outcomes. This consensus was based on a scoping review ofaphasia OMIs (Wallace et al., 2020) and an aphasia expert consensus meeting incorporatingthe most highly published aphasia treatment researchers in the Web of Science database. This consensus yielded, a Research Outcome Measurement in Aphasia (ROMA) core outcome set for assessing three of the six prioritised treatment outcomes (see Figure 1: language, emotional well-being, QoL; (Wallace et al., 2018). In a second aphasia expert meeting, additional consensus was reached on how to assess communication (Wallace et al., 2021). Psychometrically sound OMIs to assess patient-reported impact of and satisfaction with treatment, respectively, are not yet available in aphasia rehabilitation and thus not yet included in the ROMA core outcome set (cf., Figure 1). The development of treatment impact/satisfaction OMIs is a key priority in aphasia rehabilitation research, and their construction should be informed by the recommendations of the Core Outcome Measures in Effectiveness Trials/COMET (http://www.comet-initiative.org) and COSMIN (http://www.cosmin.nl) initiatives (both weblinks last accessed on 20th October 2021).

It is plausible though that top treatment outcome(s) vary for subgroups, depending on aphasia type and severity as well as stage post-stroke (Gallagher et al., 1993) and cultural group (Sanderson et al., 2012). This possibility also needs to be systematically examined in future studies to ensure that the determination of treatment success is based on criteria which are meaningful to the key stakeholder subgroup.

Research Outcome Measurement in Aphasia (ROMA) core outcome set

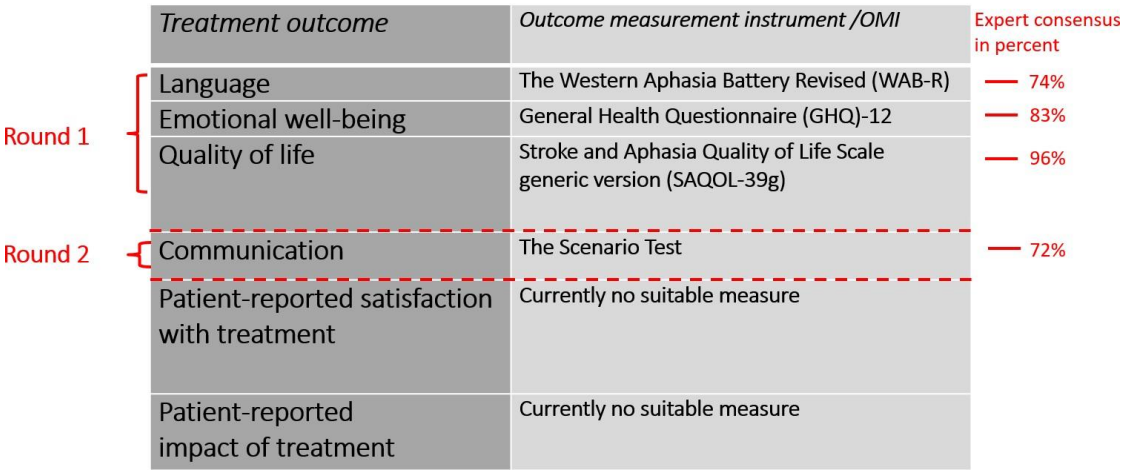

Figure 1: Top treatment outcomes across stakeholder groups and the respective OMI included in the ROMA core outcome set (Wallace et al., 2021; Wallace et al., 2018). Expert consensus varied for the various treatment outcomes: The strongest consensus was reached for a psychometrically sound OMI to assess health-related quality of life (SAQOL-39g) (Hilari et al., 2003; Hilari et al., 2009) followed by an OMI to measure emotional well-being (GHQ-12) (Goldberg et al., 1997). Consensus was lower for OMIs to assess language (WAB-R) (Kertesz, 2007) and functional communication (The Scenario Test) (van der Meulen et al., 2010), respectively.

In the preceding paragraphs, we summarized which treatment outcomes various stakeholder groups rated as top priorities. Where possible, we also presented information on the best available standardized OMI to assess the respective treatment outcome. In the following sections, we will discuss methodological approaches to determine whether a 'significant modification' of therapy targets (in the sense of treatment success) is indeed reflected by the selected OMIs for a given stakeholder group.

APPROACHES TO DETERMINE IF A 'SIGNIFICANT MODIFICATION' OF THERAPY TARGETS IS CAPTURED BY AN OMI

Aim

Knowing which treatment outcomes are of top priority for a given stakeholder group and observing numerical gains from pre to post intervention is not sufficient to determine the success of an intervention. The main aim of this methodological paper is to present strategies for both aphasia researchers and clinicians to separate 'random' score changes from 'true' score changes.

We differentiate between approaches that focus on the overall success of an intervention based on group-level data (mean changes from pre to post intervention, or between groups post intervention) versus approaches that allow inferences about an individual's treatment success (exceeding acritical change score for a given OMI).

Critical change scores at the individual level are required for providing evidence-based feedback on treatment outcome in routine clinical care. However, a critical change score is also needed to compute therapy *response rates* in treatment effectiveness trials. The response rate is the percentage of participants who were treated 'per protocol' (the on-treatment group) and showed beneficial effects from pre to post intervention that were equal to or larger than the critical change score ('treatment responders'). To date, large scale aphasia randomised controlled trials (RCTs) lack reports of such therapy response rates (Menahemi-Falkov et al., 2021). However, the treatment response rate in a given population should be reportedin *any* aphasia intervention study.

When discussing critical cut-off scores for determining individual treatment success, a further distinction will be made between the statistical versus the clinical significance of an individual change score. A *statistically* significant change score reflects a high probability (usually 95% or greater confidence, depending on the selected statistical threshold) that the observed change is real and did not occur by chance. A *clinically* significant change implies that the target stakeholder group considers the magnitude of theachieved change to be clinically meaningful.

Scope

The overall success of an intervention based on group-level data

Levels of scientific evidence for an intervention are based on the methodological quality of the study design on which the findings are based, particularly its validity. Treatment efficacy and effectiveness questions are most appropriately addressed by (systematic reviews with) meta-analyses of RCTs or n-of-1 trials (highest evidence level) or by sufficiently powered single (parallel or crossover) RCTs or observational studies with dramatic effect (https://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf; last accessed on 20th October 2021). To determine the scientific evidence-base of a specific intervention, one should select the highest level of evidence available in the hierarchy.

Evidence base derived from single trials

Traditionally, reporting overall treatment effects in RCTs involves determining through appropriate statistical analyses whether the mean change score from pre to post treatment or the post treatment score in the intervention group/arm differs *significantly* at a pre-specified type-I error level (usually $\alpha = 0.05$) from the mean change score or the post treatment score of a control group/arm with no (active ingredient of the) intervention of interest. The statistical comparisons are ideally conducted separately for each of the OMIs that have been selected as either primary (the most important outcome for the relevant stakeholder) or secondary outcomes for the study's purposes. For the OMI selectedas the study's primary outcome, a statistically significant group/arm difference in mean change/post intervention score is interpreted as a demonstration of a treatment effect for the *intervention*. This group effect does not imply a significant treatment effect for every *individual* participant. As pointed out above, a statistically significant group treatment effect with an acceptable effect size in a RCT may be driven by only a proportion of the study's sample, with the remaining sample not benefitting from the intervention (Menahemi-Falkov et al., 2021).

The *p-value* of the statistical group/arm test statistic is not only affected by the 'true' size of the treatment effect, but also by the study's sample size. The larger the sample size, the

more likely it is to find a significant group difference for a small intervention effect. Therefore, to quantify the *magnitude* of a treatment effect independently of sample size, estimates of the *effect size* are typically provided alongside the results of the statistical test. For example, Cohen's *d* is a widely used effect size for comparing the mean change scores of two independent groups. For equal sample sizes in both groups, and normally distributed continuous OMI scores, Cohen's d is calculated by simply subtracting the mean score of one group from that of the other (M1 – M2), dividing the result by the pooled standard deviation (SD) of the change score. If SDs of the two groups/arms differ substantially (violating the assumption of homogeneity of variance), the SD of the control group/condition is applied instead of the pooled SD (Glass et al., 1981). Variations of Cohen'sd are available for unequal sample sizes, single-group pre versus post treatment designs (taking the correlation between the two assessments into account) and other applications.
A useful resource for online calculation of effect sizes is:
https://www.psychometrica.de/effect_size.html (Lenhard & Lenhard, 2016); website link last accessed on 20th October 2021).

It is important to report the confidence interval of the effect size to see its potential range. If the *confidence interval includes zero*, the difference between the intervention and control groups is *not* statistically significant. With this in mind, effect size graphs (e.g., forest plots) typically provided in meta-analyses may be easier to comprehend.

In statistics, all applications which express the size of the treatment effect relative to the variability observed in the respective study samples are referred to as *'standardised mean difference (SMD)*' (Faraone, 2008). We will adopt this terminology (instead of referring to 'effect sizes') in the following sections.

SMD is equivalent to a 'z-score' of a standard normal distribution, i.e., a SMD = 1 indicates that the means of the two groups/arms differ by one SD, *SMD*= 2 refers to a difference of two standard deviations, and so forth. A SMD of 0 means that the distributions of the two

groups completely overlap, i.e., there is no difference in means between groups/arms; and thus there is no evidence that the two treatments differ in benefit conferred to the group. A SMD smaller than 0.2 implies that the size of the treatment effect is considered small (Cohen, 1988) even if the statistical comparison of the two means yielded asignificant difference. Cohen's qualitative operationalisation of treatment effect sizes appliesto both a single effect size from a given RCT and the average effect size estimate from ameta-analysis. Recommendations are that the SMD should be at least 0.5 (equal to half a standard deviation difference in means, considered an intermediate effect size) to indicate a medium treatment effect; a SMD of at least 0.8 is often considered a large treatment effect (Cohen, 1988). Interpretation of SMDs is not always straightforward. For example, the SMD is not directly affected by sample size. However, a small SMD may be induced by either a large shift in a "noisy" sample (with large inter-individual differences in change scores) or asmall shift in a very homogenous sample.

To illustrate the interpretation of a SMD, let us assume a SMD = 0.5, which is a realistic treatment effect expectation in aphasia RCTs comparing an intervention to a no-treatment/waiting list control condition (Breitenstein et al., 2017). Assuming a normal distribution of change scores, approximately two-thirds (69.1%) of the intervention group will show a difference in pre-post treatment performance above the mean difference of the groups, but one third will not; the intervention and control groups will overlap by 80% (Magnusson, 2020). Even with a larger SMD of 0.8, less than 80% of the participants in the intervention group will score above the mean of the groups, and the groups will overlap by 70%. To summarise, the SMD index (particularly in combination with the associated confidence interval; (Kelley, 2007) provides information on the degree of overlap in change scores, but is not suitable to determine the percentage of participants in the intervention group that showed a treatment effect at the *individual* level. To compute the intervention's response rate, individual participants need to be classified as responders versus or non-responders based on a cut-off score for a given OMI (see section below).

SMDs provide an estimate of the *overall* treatment effect for theprimary/secondary outcomes in a single study and can also be used to *contrast* treatment effects for different outcomes or different subgroups.

Evidence derived from meta-analysis

SMDs are also regularly applied to calculate the *average* magnitude of a treatment effect on a single outcome/construct across several (group) studies for the purpose of a secondary analysis such as meta-analysis (Baguley, 2009). However, the group studies compiled within a meta-analysis need to be comparable with respect to sample, outcome, and intervention characteristics to avoid (clinical and) statistical heterogeneity (Deeks et al., 2021). In other words, SMDs from diverse RCTs will produce meaningless results.

A SMD is not expressed in the original measurement unit of the OMI because the mean (change) score is adjusted to the variability within the sample(s). An alternate effect size measure, however, the raw unstandardised *mean difference* (Higgins et al., 2021) is expressed in the original units of analysis. The mean difference is the raw difference in mean (change) scores of two groups or arms (mean 1 minus mean 2), generally presented in conjunction with a confidence interval. The use of mean difference instead of SMD is recommended when the OMI is identical across studies and when the OMI is measured on an intuitively meaningful scale (such as the aphasia quotient [AQ] of the Western Aphasia Battery – Revised [WAB-R]) (Kertesz, 2007). However, when comparing the two point estimates (SMD versus mean difference) directly with regard to bias and efficiency using Monte Carlo simulations (a repeated random sampling method), the SMD outperformed the mean difference when datasets had small sample sizes or, large within-study variability and when the data were not normally distributed (B. T. Johnson & Huedo-Medina, 2013). These latter features frequently apply to aphasia intervention data sets (Brady et al., 2016; Lazar & Antoniello, 2008). Therefore, the use of SMD as an effect size measure maybe more appropriate unless samples sizes are sufficiently large. However, neither point estimate yielded an advantage in terms of bias and efficiency for datasets with high skewness and kurtosis.

Nevertheless, an application of the mean difference to delineate average treatment effects across aphasia rehabilitation studies for each of three different OMIs (one of which is incorporated in the ROMA core outcome set (see above: Wallace et al., 2018) was recently published (Gilmore et al., 2019).; Table 2 presents the mean differences ('summary intervention effect estimates' according to the Cochrane terminology for meta-analyses; https://handbook-5-1.cochrane.org/chapter_9/9_analysing_data_and_undertaking_meta_analyses.htm [weblink last accessed on 20th October 2021) across studies for the three OMIs, separated for within- and between-group comparisons (with large benchmark differences for one of the OMIs). These benchmarks present either the mean score difference from pre to post intervention (within-group designs) or the mean score difference at the post assessment between the intervention and the control groups (between-group designs), averaged across the analysed studies, respectively. The resulting benchmarks (e.g., a WAB-R AQ change score of at least 5.03, i.e., ≥6 points from pre to post intervention) represent the mean change/group difference scores across the analysed studies, but these benchmarks do *not* reflect cut-off scores to determine *individual* treatment success.

Additionally, SMD and mean difference are indicators of statistical importance, not of clinical relevance as they simply indicate the *average* numerical magnitude of a treatment effect observed across several studies. [2]

Table 2: Mean differences (unstandardised) across trials and the respective 95% confidence intervals for the WAB-AQ, Boston Naming Test (Kaplan et al., 2002) and Communicative Effectiveness Index (Lomas et al., 1989) reported by (Gilmore et al., 2019)

OMI	Within-group designs Mean Difference (CI)	Between-group designs Mean Difference (CI)
WAB-AQ	**5.03 points** (3.95-6.10)	**5.05 points** (1.64-8.46)
BNT	**3.30 points** (2.43-4.18)	**0.55 points** (-1.325-2.433, ns)
CETI	**10.37 points** (6.08-14.66)	Not available

OMI = Outcome measurement instrument; MD = mean difference (unstandardised);
CI = confidence interval, ns = no significant difference between treated and untreated groups across studies

In addition to the discussed effect sizes based on means, effect sizes can also be calculated based on *binary data* (e.g., depression rates as treatment outcome) or dichotomization of continuous measures. In the present paper, we do not focus on these effect sizes due to their current limited use in aphasia outcome research. The interested reader is referred to (Borenstein et al., 2009).

In summary, Figure 2 provides an overview of recommended procedures to determine treatment success in aphasia rehabilitation. The figure differentiates approaches to (i) examine the *overall* efficacy/effectiveness of an intervention ('Does it work or not?'), (ii) describe the magnitude of the overall treatment effect ('How well does it work?'), and (iii) classify individual participants into treatment responders and non-responders ('For whom does it work?').

With respect to the first approach (overall treatment effect), a *p*-value smaller than the pre-specified significance level α from a statistical test (t-test or ANOVA) is typically interpreted as suggesting that the intervention worked in the population and is notthe product of purely random variation. A tutorial on statistical tests and the interpretation of their results is provided by (Greenland et al., 2016). The second approach (treatment effect sizes) provides detail on how large the effect is, numerically, on the group level.

[2] An additional more recent method to evaluate average intervention effects, yet based on small-N designs (e.g., case-control series), is linear mixed-effects modeling, a type of multiple regression analysis. This method also yields a standardised treatment effect estimate for an OMI, but requires frequently repeated (e.g., daily) assessments of the OMI over the course of the treatment (Wiley & Rapp, 2019).

However, neither of these two group-based approaches is suitable to directly infer the clinical relevance of a score gain from pre to post intervention nor to determine treatment success for an individual with aphasia (or treatment responder rates within a study sample). Nevertheless, in some aphasia rehabilitation studies, large, statistically significant treatment effects at the group level have been interpreted as surrogates for clinically relevant ('meaningful') changes (Babbitt et al., 2016; Cherney, 2010; Elman & Bernstein-Ellis, 1999; Gilmore et al., 2019; Godecke et al., 2020; Katz & Wertz, 1997; Persad et al., 2013; Wenke et al., 2018). Large treatment effects are, of course, more meaningful than smaller treatment effects. For judgments of clinical relevance, however, it is essential that the relevant stakeholder group considers the outcome as meaningful. The stakeholder group not only has to *perceive* a treatment effect in everyday life performance after the intervention is terminated, but also needs to view the perceived change as *important* (Middel & van Sonderen, 2002). We will discuss this issue in more detail below in the section on determining individual treatment success.

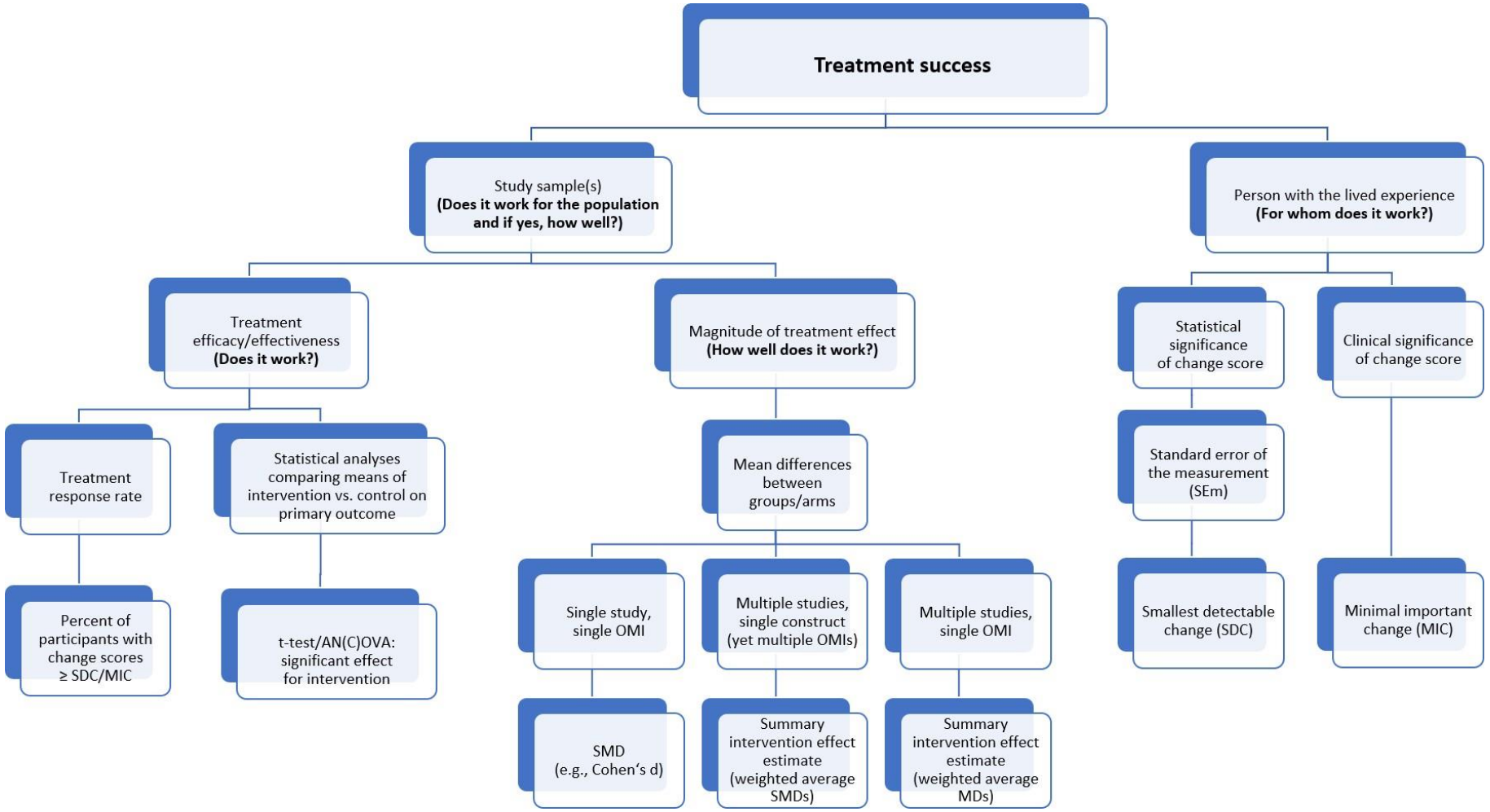

Figure 2: Overview of procedures to determine the success of an intervention (based on study samples) and individual treatment success (based predominantly on the patient perspective and clinician prescription needs). SDC = smallest detectable change; MIC = minimal important change;

$p \leq \alpha$ (α usually 0.05); AN(C)OVA = analysis of (co-)variance; OMI = outcome measurement instrument; SMD = standardised mean difference; MD = mean difference; SEm = standard error of the measurement.

Individual success of an intervention

Although summary SMDs and mean differences for OMIs based on group data are useful to quantify the overall effect of the intervention of interest, particularly when used inmeta-analyses, they are not suitable for assessing treatment success of individual participants. Determination of individual treatment success requires benchmarks to classify individuals with *positive* versus *no effect* or even *negative* treatment outcomes.

In the following section we therefore discuss best practice approaches in medical outcome research to determine treatment success for an individual person with aphasia (de Vet & Terwee, 2010).

a) *Statistically* significant change scores at the *individual* level

We start with distribution-based approaches to identify benchmarks of a *statistically significant* treatment response at the *individual* level. Improvement from pre to post intervention is commonly measured by the individual's change score (post-treatment score minus pre-treatment score) in the relevant OMI. However, scores of repeated administrations of the same measure frequently differ even when testing conditions match closely and no interim intervention period occurred. The observed test scores are subject to random score fluctuations due to the imprecision of the test, defined as *standard error of measurement (SEm)*. The SEm is not to be confused with the *standard error of the mean* (SEM) in statistics (an estimate based on the group's standard deviation divided by the square root of its sample size) which indicates how far *a group's* mean value is likely to vary from the 'true' (population) mean. The SEm, on the other hand, reflects the discrepancy between the observed and the 'true' score of a *single person*. The rationale for calculating SEm is based on the model of classical (psychometric) test theory (Lord & Novick, 1968) which postulates that an observed test score is the sum of an individual's latent 'true' score plus the OMI's measurement error, ~~and~~ neither of which ~~the latter two~~ is directly observable. The 'true' score is the score expected if an individual was assessed on an infinite number of occasions with a given test and thus reflects a constant 'trait' of the individual. The SEm isequal to the deviation of the observed from the 'true' score due to random measurementeffects (e.g., fatigue, inattention).

The SEm is a function of the *variability* (SD) of observed scores in the OMI's evaluation sample at the baseline assessment (based on the assumption that individual error variances are approximately equal in the population) (cf., Willmes, 1985) and the precision (reliability) of the OMI, expressed in the original units of the OMI:

$$SEm^{*} = SD_{OMI} \times -\sqrt{(1 - reliability_{OMI})}$$

* More precisely, the formula refers to the standard error of measurement = $SE_{meas,}$, the acronym suggested by (McManus, 2012)

The reliability of a test expresses the degree of agreement of the observed scores with the 'true' scores of the participants (i.e., precision). When the OMI is perfectly reliable (i.e., reliability = 1), the SEm is zero (SD_{OMI} x 0 = 0), indicating that any assessment with the OMI yields an identical score for a given individual and any observed score is identical with the individual's 'true' score. On the opposite, with a completely unreliable OMI, the SEm is close (or equal) to the *variability* of observed scores in the OMI's evaluation sample (SD_{OMI}).

SEm estimation requires stability of the measured construct (*parallel or tau-equivalent measures*)[3] across repeated assessments with the same OMI. If this stability assumption is violated (i.e., if the 'true' score has changed from the first to the second assessment), the formula does not yield an accurate reflection of the SEm due to overestimation of the measurement error. If the stability assumption applies, all available reliability estimates (parallel test versions, internal consistency, split-half, test-retest, structural equation modelling) may be applied to estimate the SEm (Danner, 2016). The different approaches may yield different reliability *estimates*, but a given OMI has only *one* reliability.

In principle, the intended application of the test score(s) determines which of the various reliability estimates is most appropriate. Test-retest reliability estimates are recommended

[3] Basic requirements for reliability estimation in classical test theory are *parallel* (or less strictly *tau-equivalent*) *measures*, i.e., (1) stability of subjects' 'true' scores (traits) across repeated assessments with the same test (or a constant 'true' score change for *all* individuals of the reference population) and (2) the test administrations have identical error variances and thus observed score variances in the reference population (for tau-equivalent measures the error variances may differ).

when “the measurement errors of primary concern are the fluctuations of an examinee's observed scores around the true score because of temporary changes in the examinee's state” (Crocker & Algina, 2008, p. 133).

Accordingly, the authors of the COSMIN manual (Mokkink, Boers, van der Vleuten, Patrick, et al., 2020) recently recommended a point estimate of reliability for calculation of the SEm (as part of the formula to determine the Smallest Detectable Change~~/SDC~~ across repeated assessments, see below) that is computed as the intraclass correlation coefficient (ICC: two-factor mixed model, single measure, absolute agreement variant) for test-retest data.[4] Test-retest reliability estimates such as the ICC (agreement/single measure variant)[5], which are based on ratios of variance components, are in principle suitable reliability estimates to detect minimal score changes between assessments because the variance between repeated assessments (without intermittent intervention) is accounted for. If the ‘true’ test score remains stable from the first assessment to the retest, any observed score difference will be attributable to (random) measurement error.

However, the requirement of stable ‘true’ scores may not hold when a test is repeated after weeks or months. Even in healthy subjects, it is unlikely that individual ‘true’ scores remain identical across repeated assessments spaced apart by weeks or months – except for highly stable ‘traits’ like intelligence or personality factors. In stroke samples, stable ‘true’ scores are even less likely because of spontaneous recovery and other plasticity processes of the brain, particularly in the early stages post-stroke, but which may last for years after the initial stroke for functions such as language (L. Johnson et al., 2019; Sachs et al., 2020; Smania et al., 2010), as well as cognitive and emotional fluctuations or concomitant neurodegenerative processes affecting task performance. Then, any score difference from the first assessment to the retest will not be attributable to random measurement error

[4] In case of rating scales with two or more raters the computation of generalizability coefficients or kappa coefficients was proposed by the COSMIN group (Mokkink, Boers, van der Vleuten, Patrick, et al., 2020).

[5] Other variants of the ICC (consistency/single measure variant or average variants) as well as the Pearson correlation coefficient are less suitable because systematic score level differences between assessments are of no concern (McGraw & Wong, 1996; also cf., Mokkink, Boers, van der Vleuten, Patrick, et al., 2020, p.37).

alone but may be affected by a large intra-individual variability regarding changes in 'true' scores. Thus, the assumption of stability of 'true' scores (or a constant score change for *all* individuals of the reference population) across repeated assessments may not hold in post-stroke aphasia samples. With few exceptions to date (see Table 3 below), test-retest intervals for common aphasia OMIs comprised several weeks (or in the case of the WAB evaluation even years) – or entirely lacked the assessment of test-retest reliability during the test evaluation process.

With *very short test-retest intervals* (e.g., two days), changes in 'true' scores may be negligible even in post-stroke aphasia samples, particularly in the chronic stage post-stroke. However, the repeated administration of a performance test requiring several hours of language stimulation (particularly for tests tapping ~~on~~ a specific language component like single word naming) may have the effect of a high intensity intervention, leading to changes in 'true' scores in this population (also see Ellis et al., 2014, who demonstrated that the largest treatment effects in aphasia rehabilitation typically occur within the first three sessions).

When the assumption of stability of 'true' scores is violated, a low test-retest reliability estimate can reflect either the low reliability of the test instrument or an instability of the measured construct. Test-retest reliability has thus been considered a "somewhat inaccurate estimate of the theoretical reliability coefficient" (Crocker & Algina, 2008, p. 134). The COSMIN author group also explicitly stated in their recommendations that "patients should be stable with regard to the construct to be measured between the repeated measurements" (Mokkink, Boers, van der Vleuten, Patrick, et al., 2020, p. 33; see also Rousson, 2011, for applying test-retest reliability estimates for the Smallest Detectable Change calculation).

Reliability estimates based on a *single assessment* (such as Cronbach's alpha)[6] have been recommended in situations in which 'true' score consistency cannot be expected across time (Willmes, 1985). The so-called 'internal consistency reliability estimates' examine how consistently the individual responds to similar items of the test instrument. It is also assumed that responses will be similar for all other possible items measuring the same construct. Cronbach's alpha, the standard psychometric index of a test's internal consistency, 'simulates' repeated test administrations by treating the individual items of a test as if they were separate assessments. This assumes that the various items of a test measure the same (unidimensional) construct, i.e., all items of the test measure the same 'true' score variable (with possibly different error variances and a constant difference in item difficulty). This assumption should be probed by exploratory/confirmatory factor analysis for any OMI prior to applying Cronbach's alpha to SEm/Smallest Detectable Change calculations. If the model of unidimensionality does not hold, but the error variables of the items are uncorrelated (i.e., there is no other latent variable than the one true-score variable that contributes to the covariation of items), then Cronbach's alpha is a *lower* bound to the reliability of the total score (Lord & Novick, 1968).

Furthermore, aphasia sample sizes for estimating an OMI's internal consistency (such as Cronbach's alpha) frequently exceeded 100 whereas sample sizes to estimate test-retest reliability were generally much smaller (most frequently $n \leq 20$). This difference in sample size numbers may be due to the additional burden of carrying out a second assessment for test-retest reliability estimation. Given the current considerable sample size differences, confidence intervals for the reliability point estimates based on repeated assessments will be much wider than confidence intervals based on a single assessment (cf., Table 3). A practical advantage of using Cronbach's alpha is that the coefficient is routinely reported in

[6] Alternatively, the split-half reliability coefficient with "upgrading" to the full test length via the Spearman-Brown formula may also be applied. Parallel versions of the same test are less frequent in aphasia rehabilitation, but the Pearson correlation between parallel versions would also be adequate. For either reliability estimate, the model of 'essentially tau parallel' measures needs to hold. If only the weaker model of 'tau-congeneric' measures holds, McDonald's omega is the reliability estimate of choice (Padilla, 2019).

psychometric evaluation studies of OMIs in aphasia rehabilitation, so the SEm/Smallest Detectable Change calculation should be already feasible for the most common aphasia outcome measures.

To summarise, when selecting a reliability estimate for SEm calculation, there is a trade-off between possibly (a) overestimating the measurement error of an OMI when using a test-retest reliability estimate (such as the ICC single measure/agreement variant) because changes in 'true' score are inseparable from measurement error of the test instrument versus (b) underestimating the SEm when using a consistency reliability estimate (such as Cronbach's alpha) because the unidimensionality assumption may not hold for the OMI. We illustrated this trade-off in Table 3 (adapted from Menahemi-Falkov et al., 2021) by comparing SEm and Smallest Detectable Change based on reliability estimates requiring single (Cronbach's alpha) versus repeated (test retest) test administrations for common aphasia OMIs, highlighting the (in some cases) discrepancies, particularly for an OMI assessing self-reported quality of life.

To probe the practical effects of applying different reliability estimates, we calculated SEm and Smallest Detectable Change based on both Cronbach's alpha and the ICC recommended by the COSMIN group (Mokkink, Boers, van der Vleuten, Patrick, et al., 2020; ICC variant: two way mixed, agreement, single measures) for two test instruments for which both reliability estimates were available (see Table 3). There were only minor differences in SEm/ Smallest Detectable Change for an overall language measure (the German Aachen Aphasia Test/AAT) and a standardised measure of verbal communication (German version ofthe Amsterdam Nijmegen Everyday Language Test/ANELT)~~.~~, yet with a narrower confidence interval for Cronbach's alpha compared to the ICC (confidence intervals were available for only one of the two measures) due to much larger sample sizes for estimating Cronbach's alpha (based on a single assessment). Overall, the two reliability estimates yielded highly comparable reliability coefficients (particularly with a very short test-retest interval of two days, as was the case for the AAT evaluation sample). Therefore, they should yield very similar treatment responder rates in intervention studies when calculating Smallest Detectable Changes.

As a future scenario, we recommend that test-retest indices (e.g., two-way mixed/agreement/single measure variant of the ICC) are calculated in evaluation studies of

aphasia OMIs that are based on *appropriate test-retest intervals* and *sufficiently large sample sizes (*preferably exceeding n=100). The appropriate test-retest interval depends on the stage post-stroke and the construct to be measured; it should be long enough to minimise memory or practice effects, yet short enough to avoid changes in 'true' score (Crocker & Algina, 2008)[7]. Statistical approaches to plan sample size calculation for achieving sufficiently high reliabilities in OMI evaluation studies are discussed elsewhere (Charter, 2008; Terry & Kelley, 2012). Samples also need to be *representative* for the targeted population (e.g., a clinically relevant population of post-stroke aphasia), and within these populations, benchmarks such as SEm and Smallest Detectable Change may need to be *calculated separately for various subgroups*, respectively. An additional future requirement for OMI evaluations is to use *a standard test evaluation setting* which is not part of a randomised control design because treatment expectations may induce additional systematic error between repeated (baseline) assessments.

With respect to the *interpretation* of the SEm, we noted some misconceptions in the literature. The SEm can be considered one standard deviation of the mean error of a measurement. As such, SEm may also be used in combination with an individual's observed score from a *single assessment* to construct the upper and lower confidence interval bounds of the actual scores likely to be obtained given an individual's 'true' score[8].

Assuming normally distributed scores, there is a 68% confidence that, given an individual with some fixed but unknown true score, the spread of actual scores will be between plus/minus one SEm around the observed score (or with a 96% confidence between plus/minus two SEm). Imagine an individual scored 50 on the WAB-R AQ. In combination with the OMI's SEm (3.56 points for the WAB-R AQ based on the test-retest reliability

[7] An alternate option is that *parallel* versions of a test may be administered for *repeated* assessments to disentangle changes in 'true' scores from measurement error (e.g., using bi-factor models; Mokkink, Boers, van der Vleuten, Bouter, et al., 2020, p.31).

[8] More precisely, this is the definition for the standard error of measurement = SE_{meas} according to (McManus, 2012), which needs to be differentiated from the other two types of SEm, which are the standard error of estimation = SE_{est} and the standard error of prediction = SE_{pred}.

reported in the original psychometric evaluation study by Shewan & Kertesz, 1980; also cf., Menahemi-Falkov et al., 2021), a confidence interval can be defined where this individual is likely to score if the OMI would be administered again under the same conditions (i.e., without a therapy-induced change of the 'true' score). This (68%) confidence interval will stretch from 46.44 (observed score of 50 minus one SEm) to 53.56 points (observed score of 50 plus one SEm). This is the amount of score variation that needs to be expected simply by random fluctuations (i.e., imprecision of the test). However, the SEm is *not* a benchmark of a 'statistically significant change' of the ('true') test score of an individual.

The general formula for calculating a 90 or 95% confidence interval of a SEm (assuming a standard normally distributed error variable) is:

$CI_{SEmeas} = Y_{OMI}$ +/- z-score * $SD_{OMI} * \sqrt{(1 - reliability_{OMI})}$

with Y_{OMI} representing the individual's score on the OMI, 'z-score' is the (1-alpha/2) – quantile of the standard normal distribution needed to compute the two-sided confidence interval boundary values (z = 1.96 for 95% confidence; z = 1.65 for 90% confidence) and SD_{OMI} is the OMI's standard deviation in the evaluation sample.

To summarize, the SEm is an estimation of the *expected random score variation* when *no 'true' change* has occurred between repeated assessments (Furlan & Sterr, 2018). Successful treatment outcome, however, implies that the 'true' score of an individual has *changed* in the desired direction. The most frequently used *change* index of an OMI is the Smallest Detectable Change (also referred to as Minimal Detectable Change/MDC in the literature). The Smallest Detectable Change is a cut-off value which indicates the minimum change score required to be considered a 'true' change that can be detected beyond measurement error with a certain confidence (de Vet & Terwee, 2010). If an individual's change score is smaller than the Smallest Detectable Change cut-off score, itis considered indistinguishable from measurement error due to the imperfect reliability of the OMI. The Smallest Detectable Change is based on the SEm (see above for formula) and is calculated using the following formula:

Smallest Detectable Change = SEm x z-score x √2

The standard (two-sided) confidence levels applied to Smallest Detectable Change cut-off scores are 90 and 95% (SDC_{90} and SDC_{95}), but (a two-sided) 90% confidence is regarded as sufficient for interventions unlikely to have serious adverse outcomes (Chen et al., 2012; Donoghue et al., 2009). There is an additional argument for choosing a more liberal type I error level of 10% (two-sided) in clinical populations like PWA: with a liberal type-I error level, the type-II error level, i.e., the error of 'overlooking' a 'true' difference, is reduced. In Table 3 we provide the cut-off scores for an individual treatment success for common aphasia OMIs based on a systematic review of intervention effects in post-stroke aphasia rehabilitation (Menahemi-Falkov et al., 2021).

Table 3: Comparisons of smallest detectable change cut-off scores (SDC_{90}) for individual treatment success based on different reliability estimates (internal consistency versus test-retest) for various aphasia outcome measurement instruments (adapted from Menahemi-Falkov et al., 2021)

	Language	Theoretical OMI score range	n for coefficient alpha (stage post-stroke)	n for ICC (stage post-stroke)	Mean ICC test-retest interval	Coefficient alpha (α)	ICC[1] agreement (test-retest)	CI for coefficient alpha	CI for ICC agreement	SEm[5] α / ICC	SDC_{90} α / ICC
BNT	English	0-60	75 (stage not reported)			0.98	-	Not reported	-	**2.69/-**	**6.22/-**
WAB AQ[3]	English	0-100	Not reported for WAB AQ (entire WAB: n=140, stage not reported)	38 (chronic)	median: 18.5 months (range: 6 months to 6.5 years)	Not reported for WAB AQ (0.91 for entire WAB)	0.97 (only *Pearson correlation* coefficient reported for WAB AQ)	Not reported	Not reported	**-/3.56**	**-/8.26** (reliability estimate: Pearson)
ANELT A-scale	German	10-50	150 (chronic)	78 (chronic)	21 days	0.95	0.93[2]	0.94-0.96	0.90-0.96	**2.44 / 2.88**	**5.65 / 6.68**
AAT profile height	German	20-80 (T-score range)	120 (mixed)	20 (chronic)	2 days	0.996[4]	0.99	[4]	Not reported	**0.63[4] / 0.74[4]**	**1.4 / 1.72**
CETI	English	0-100	22 (mixed)	11	~60 days	0.90	0.94 (ICC variant not reported)	Not reported	0.87-0.99	**4.96 / 3.84**	**11.53 / 8.91**
SAQOL-39g	English	1-5	71 (chronic)	18 (chronic)	7 days	0.95	0.91	0.93-0.96	0.76-0.97	**0.16 / 0.21**	**0.36 / 0.49**
	German	1-5	154/*52*[6] (chronic)	78/53[6] (chronic)	21 days	0.91/0.92[6]	0.73/0.80[6]	0.88-0.93 0.88-0.95[6]	0.60-0.82 0.68-0.88[6]	**0.17 (0.16[6]) / 0.29 (0.25[6])**	**0.39 (0.37[6])/ 0.67 (0.58[6])**

BNT = Boston Naming Test; WAB AQ = Western Aphasia Battery Aphasia; ANELT = Amsterdam Nijmegen Everyday Language Test; AAT = Aachen Aphasia Test; CETI = Communicative Effectiveness Index. SAQOL-39g: Stroke and Aphasia Quality of Life Scale 39 generic version; n = sample size; ICC = intraclass correlation coefficient, CI = confidence interval; SEm = standard error of measurement; SDC_{90} =smallest detectable change with 90% confidence;

[1]: ICC variant: 2-way random/mixed, single measures;

[2]: Parallel test versions used at the two assessments;

[3]Please note that the aphasia quotient/AQ of the revised version of the WAB (WAB-R) has not been psychometrically evaluated to date. An exception is the recent publication by (Dekhtyar et al., 2020) who compared in-person versus videoconference administration of the WAB-R in a small sample of n=20 chronic PWA and reported an AQ ICC (agreement variant; not reported whether single or average measures ICC type was applied) = 0.99, CI = 0.978-0.998 (test-retest interval: 7-14 days, SD = 22.17 for in-person/22.68 for videoconference administration).

[4] The AAT profile height estimate is computed differently as a reliability (Cronbach's alpha) weighted average of subtest T-scores. No confidence interval is reported.

[5]Estimates presented for SEm are based on the formula for the standard error of the measurement ($SE_{meas} = SD_{OMI} \times -\sqrt{(1 - reliability_{OMI})}$; formulas for calculation of the two other types of SEm ($SE_{est} = SD_{OMI} \times -\sqrt{reliability_{OMI}\ x\ (1 - reliability_{OMI})}$ and $SE_{pred} = = SD_{OMI} \times -\sqrt{(1 - reliability_{OMI})^2}$) are described in McManus (2012).

[6]results for subgroup with moderate to mild language comprehension impairment based on AAT (T ≥50; n=53) to match the standardisation sample of the original English SAQOL-39g.

So, when an individual's change score from pre to post treatment is equal to or larger than the OMI's SDC_{90}, the change can be considered *statistically significantly different from a change of zero in the 'true' score* with a 90% confidence (two-sided). Here is an example: The SDC_{90} for the WAB-R AQ in chronic post-stroke aphasia has been estimated at 8.26 points (see Table 3; the benchmark has been reported in Menahemi-Falkov et al., 2021)[9]. If an individual scored 50 on the WAB-R AQ at the baseline assessment, the post-treatment score would have to be at least (50 + 8.26 =) 58.26 points to be considered a statistically significant improvement in 'true' performance level with 90% confidence. A deterioration in scores from baseline to post-treatment would be considered statistically significant if the post-treatment score is equal to or smaller than (50 - 8.26 =) 41.74 points (again with 90% confidence). Any post-treatment score between 41.74 and 58.26 points thus potentially represents *random fluctuation* compared to the baseline score. This example also shows that the confidence interval for a given test score may stretch across traditional boundaries of aphasia severity classification levels (50 is the cut-off score between severe and moderate aphasia on the WAB-R AQ). To demonstrate a statistically significant individual treatment success, an individual change score from pre to post treatment needs to be equal to or exceed the SDC in the desired direction. It is noted that for the WAB-R AQ, the SDC_{90} cut-off score is about one third larger than the *mean* score difference observed across studies for the group level (≥6 points; see above). For other OMI's in aphasia rehabilitation, the discrepancy between group-level and individual benchmarks may be even larger.

It might seem surprising that the *individual's* benchmark for a statistically reliable change on the WAB-AQ (SDC= 8.26) exceeds the frequently reported 5-point benchmark of 'clinical significance' for WAB-AQ change scores (e.g., Babbitt et al., 2016; Cherney, 2010; Elman & Bernstein-Ellis, 1999; Eom & Sung, 2016; Falconer & Antonucci, 2012; Godecke et al., 2020; Katz & Wertz, 1997; Kempler & Goral, 2011; Maher et al., 2006; Mozeiko et al., 2018; Peach et al., 2019; Persad et al., 2013). The 5-point benchmark, however, has been introduced "as the amount of change clinicians might accept as indicating improvement" (Katz & Wertz,

[9] The WAB-R AQ benchmark for a statistically significant individual change score will presumably be higher in acute stroke samples because of greater sample variability and thus a larger SEm. This needs to be addressed in future studies.

1997, p. 501) and is not based on a formal clinical consensus process or an anchor-based approach involving relevant stakeholder groups (see section below on "*Clinically significant change scores*") and should be interpreted accordingly. This is particularly important given that the mean WAB-AQ score difference between two assessments *without* (reported) intermittent intervention yielded 5.32 points in the original WAB evaluation study for a sample of n=38 persons with chronic post-stroke aphasia (Shewan & Kertesz, 1980).

It is feasible, however, that the SDC cut-off scores reported in Table 3 will be lower in future aphasia OMI evaluation studies with sufficiently large sample sizes. As an example, in the recently completed Australian COMPARE trial [study protocol: Rose et al., 2019; results not yet published], the individual benchmark for the WAB-R AQ was recalculated based on n=152 participants with chronic post-stroke aphasia (who underwent two assessments with the WAB-R-AQ spaced >31 days apart prior to commencing the study intervention). Given a smaller standard deviation (18.28 WAB-R AQ points) in this sample compared to the original (smaller) WAB evaluation sample, the SDC_{90} equaled 7.09 and was thus about 1 point lower compared to the WAB-R-AQ benchmark calculation reported in Table 3 (which was based on the original WAB evaluation study with n = 38 PWA). The benchmark for a *clinically* meaningful change (see below for definition) on the WAB-R AQ still needs to be determined following a formal consensus protocol and involving people with aphasia.

In general, besides determining individual treatment success, the Smallest Detectable Change may also be applied in group studies to determine the intervention response rate, which is the proportion of study participants with a change score equal to or larger than the primary outcome's Smallest Detectable Change benchmark. To date, intervention response rates have rarely been reported in aphasia intervention trials or in systematic reviews on treatment outcome and maintenance (Menahemi-Falkov et al., 2021),despite being easy to calculate and easily understood by all stakeholder groups. Individual response rates in aphasia rehabilitation intervention studies – in addition to significant groupdifferences - complement the intervention's evidence base and support subgroup analyses of treatment effects and should thus be routinely reported in aphasia intervention studies.

In addition, reporting of the exact 95% (binomial) confidence interval (Clopper & Pearson, 1934) for the underlying response probability is also important. Determining this confidence interval requires the number of responders and the study sample size as entries in freely

accessible online confidence interval calculation software (such as https://sample-size.net/confidence-interval-proportion/; weblink last accessed 20th October 2021]. Interpretation of the data would indicate that an intervention has a responder probability between the lower and upper confidence interval bounds (with 95% confidence, depending on the selected confidence level).

Analysing individual responses can help to reveal profiles of treatment responders versus non-responders (Fridriksson & Hillis, 2021), so it is critical to have comprehensive participant descriptors in clinical trials. Recent consensus-based research has established international and multidisciplinary agreement on minimum reporting requirements for participant characteristics in aphasia research (Isaacs et al., 2021). The resulting profiles may be used to develop individually tailored therapies and to inform inclusion criteria for future clinical trials using a particular intervention. Last but not least, a dichotomous classification of treatment responses into success and failure allows determination of the intervention's 'odds ratio (OR)', which is an indicator of the association strength between two events (e.g., intervention and outcome) and which is a further common treatment effect measure used in the medical field.

In routine clinical care, reporting treatment response rates will contribute to raising realistic hopes for PWA starting a particular intervention and will also allow comparison of the response rates across treatment centres as part of a quality assurance strategy.

b) *Clinically significant* change scores

The statistical significance of an individual (or group) change score does not necessarily correspond to its clinical benefit: A statistically significant change may be numerically so small that the effect is considered clinically irrelevant. Vice versa, statistically non-significant changes may be clinically highly relevant. The benchmark to determine a *clinically significant* improvement on a given OMI is the *Minimal Important Change* (de Vet & Terwee, 2010). In the literature *Minimally Important Difference* and *Minimal Clinically Important Difference* are frequently used as synonyms. We will avoid the latter terms here to adhere to the COSMIN guidelines which discourage the term 'differences' when referring to longitudinal 'changes' in an individual person (de Vet & Terwee, 2010).

The Minimal Important Change for an OMI is defined as the *smallest change score* above which treatment outcome is experienced *as relevant* or *meaningful* by the relevant stakeholder group (PWA, their relatives/carers, clinicians, funders – depending on the study's aim). For computation of the Minimal Important Change, the OMI's change score from pre to post treatment is related to an independent external standard, the 'anchor'. The 'anchor' measure needs to be meaningful for the respective stakeholder group and can be bio- or physiological, performance-based (objective tests) or subjective (e.g., patient- or clinician-reported rating scales or questionnaires). Critically, the 'anchor' and the respective OMI should have a moderate to high correlation ($r \geq 0.5$) to demonstrate that they measure a comparable construct (Devji et al., 2020). An OMI can have one or multiple 'anchors', depending on the purpose.

For the key stakeholder group of PWA, the 'anchor(s)' should reflect "the extent to which changes in a patient's functioning or wellbeing meets the patient's needs or expectations" (Ware, 1992, p. 3). Frequently used patient-defined 'anchors' are (a) the patient's overall rating of perceived change ('anchor' question), (b) a patient's overall rating of satisfaction with the treatment and (c) a patient-reported score or item derived from a questionnaire (Devji et al., 2020). Patient-defined 'anchors' are predominantly *patient-reported outcome measures (PROMs)*, but this is not a requirement as long as the PWA stakeholder group considers the 'anchor' to be meaningful (e.g., a clinical endpoint).

To date, 'anchors' have rarely been defined in stroke outcome research (van Bloemendaal et al., 2012). In the recent randomized controlled 'Big CACTUS' trial, at least 10% improvement in a study-specific naming task was considered a 'clinically meaningful' change (Palmer et al., 2019). This criterion was derived from discussions with clinicians on the trial team and the aphasia patient-and-public-involvement (PPI) group. For standardised communication or language OMIs in aphasia rehabilitation, Minimal Important Changes have not been identified. Their development should be a key research endeavour for future studies inthe field.

The one exception is the estimation of the Minimal Important Change for the Singapore version of the Stroke and Aphasia Quality of Life scale -39 item generic version (SAQOL-39g) in stroke survivors with and without aphasia (Guo et al., 2015). This study used a clinician-reported measure of global disability, the modified Rankin scale (mRS) (Broderick et al.,

2017), as an ‘anchor’ for SAQOL-39g change scores. Change by at least one mRS level (total score range: 0-6) was considered clinically meaningful by the authors. Based on the mRS change scores from baseline (3 months post-stroke) to follow-up (12 months post-stroke), stroke survivors were grouped into (1) improvement of at least one level, (2) stable performance and (3) decline of at least one level. For each of the three subgroups, SAQOL-39g scores of the baseline and follow-up assessments were statistically compared. These analyses yielded a statistically significant SAQOL-39g change for the ‘improved’ mRS group only. Minimal Important Change was defined as the *average SAQOL-39g change score* from baseline to follow-up *inthe ‘improved’ group* (0.21 points). Therefore, an improvement of at least 0.21 points on theSAQOL-39g can be considered a clinically meaningful change in stroke survivors with and without aphasia in the first year post-stroke, but the statistical benchmark for a significant change is much higher (SDC_{90} = 0.39 for the English SAQOL-39g; cf., Table 3). The study requires replication in an appropriately powered aphasia sample to determine the validity ofthe mRS as an ‘anchor’ for aphasia treatment success.

On the face of it, the mRS seems inappropriate as an anchor for aphasia trials. Consider a person with Broca’s aphasia who was a professional radio show host, who is able to look after themselves independently, but cannot return to work because of speech and communication difficulties. The person would have to recover normal speech and communication to return to work in order to improve from a mRS score of 2 (“Slight disability: Able to look after own affairs without assistance, but unable to carry out all previous activities”) to 1 (“No significant disability. Able to carry out all usual activities, despite some symptoms”). Furthermore, given that the mRS utilises an ordinal scale, it has never been demonstrated to our knowledge that a score change from 2 to 1 indicates the same degree of improvementas, for example, a score improvement from 3 to 2.

Additionally, a patient-defined ‘anchor’ may be more clinically relevant and to date more sophisticated (logistic regression modelling) approaches for Minimal Important Change determination are available (Terluin et al., 2015; Terwee et al., 2021). Thestudy by Guo et al. (2015) did, however, show that the general application of an ‘anchor- based’ approach is feasible in a stroke sample including PWA.

Whereas psychometric properties of (primary) OMIs should adhere to stringent methodological standards (Mokkink et al., 2016), 'anchors' may be highly subjective measures. Until more refined 'anchor' measures are available in aphasia rehabilitation, the participant's overall rating of perceived treatment *impact* on their communication could be used. '*Acceptability'* of the treatment might be used as a separate 'anchor' question. We are aware that this 'anchor-question' may not be ideal because there may be circumstances outside the aphasia treatment itself leading to a perceived *non-impact* or *lack of acceptability* of treatment (e.g., concomitant other medical problems), but this may be an intrinsic problem of every patient-reported 'anchor'. It may be helpful to determine separate Minimal Important Change scores for each of the relevant stakeholder groups to illustrate the degree of agreement or disparity.

A recent guideline proposed that 'anchor' questions should include a specific time frame (Devji et al., 2020). For aphasia outcome research, the following 'anchor' question referring to treatment *impact* may be used: "How much has your[communication/language/quality of life/general well-being, depending on the OMI] changed since your last visit/since the treatment started?" . Responses can vary on a 6-point Likert scale ranging from -2= 'much worse', - 1 = 'slightly worse', 0= 'no change', +1 = 'slightlyimproved', +2= 'much improved' to +3 = 'completely recovered', as has been proposed by (Revicki et al., 2008). People with aphasia should be involved in determining the cut-off level for a perceived 'important' change ('slightly' or 'much' or 'completely' improved/recovered).This 'anchor' question may be used for estimating the Minimal Important Change for a study's primary outcome. Additionally, participants may be queried whether they feel that their treatment could have been changed in some way and how.

Conclusions/Recommendations

In this methodological ~~(~~tutorial~~)~~ paper we aimed to advance the application of best-practice approaches to determining treatment success in aphasia rehabilitation research studies (see Table 4 for a summary). We differentiated strategies focused on group level analysis (single trials and compiling across multiple trials) to demonstrate overall efficacy/ effectiveness of an intervention (standardised mean differences; raw unstandardised mean

differences) from those focused on the individual level. Group level effects are critical information for appropriately powering future clinical trials using the same OMI. ~~O~~On the other hand, individual therapy outcomes are vital for identifying treatment response rates in clinical trials and for accurate treatment evaluation of individuals in clinicalpractice. Specifically, we recommended methods to calculate benchmarks of statistically significant as well as clinically relevant score changes on an OMI for an individual.

Statistical benchmarks are dependent on the evaluation sample on which their calculation is based ~~on~~. Given the limited sample sizes reported in evaluation studies of aphasia OMIs to date (cf., Table 3), it seems pivotal to re-evaluate the psychometric properties for the majority of OMIs in aphasia rehabilitation research based on appropriate evaluation designs (e.g., appropriate test-retest intervals) and sufficiently large sample sizes (at least n=100 for estimating reliability coefficients) (Willmes, 1985). With larger sample sizes, the standard deviation of the sample will decrease and so will the SEm/ Smallest Detectable Change for an OMI (see formula and example from the recent COMPARE trial above).

Given the heterogeneity of aphasia samples with respect to stroke- and aphasia-related factors, it may be required to estimate the Smallest Detectable Change and Minimal Important Change for a given OMI separately for various subgroups, depending on age, sex, time post-stroke, aphasia severity and type as well as other factors. For example, the SEm for the WAB-R may vary with aphasia severity because the measurement error may not be distributed equally across the aphasia severity continuum (with greater measurement error for very low and very high WAB scores) (Hula et al., 2010). The field urgently needs to address these discrepancies so that accurate interpretation of results from previous and future trials can be made.

In summary, operationalising individual treatment success based on Smallest Detectable Change and Minimal Important Change (preferably based on a patient- reported 'anchor' of perceived treatment impact) is a key priority in aphasia rehabilitation.This operationalization will help to (a) identify the therapy response rate in intervention studies in order to optimise therapeutic decisions in routine clinical care and (b) provide all stakeholders (e.g., PWA, family, clinicians, health insurances) with objective, statistically reliable and meaningful feedback on the actual individual treatment response in the clinical setting.

Table 4: Summary of current best-practice approaches to determine treatment success in aphasia rehabilitation

What?	**How to?**	**Interpretation**
Define treatment success	Intervention targets *a priori* defined for key stakeholder group	Targets have been significantly modified in the desired direction after the intervention and remain stable?
Top treatment outcomes	Dependent on key stakeholder group: 1) Socioeconomic/societal perspective 2) Stroke team perspective 3) Aphasia rehabilitation perspective a) Aphasia researchers b) Aphasia clinicians and managers c) PWA d) Family members of PWA	Domains (informed ROMA COS development) 1) Return to employment (if applicable), rehabilitation service needs, QALYs, cost effectiveness 2) Survival, restoration of blood flow, prevention of recurrent stroke, functional independence in everyday life, health-related QoL a) Treatment impact/satisfaction, communication-related QoL, language functions b) Conversation participation, communication c) Communication, life participation, recovered normality, emotional well-being d) Communication, recovered normality, emotional well-being
Group-level benchmarks for treatment success	<u>Single RCT:</u> Mean difference between groups/condition (from pre to post intervention or post intervention adjusted for baseline performance)	Overall statistically significant effect of an intervention in one RCT

What?	How to?	Interpretation
	Standardised mean difference/Cohen's d (> 0.50)	Point estimate of respective treatment effect (magnitude), subgroup comparisons of treatment effect (if applicable)
	<u>Meta-analysis:</u> Standardised mean difference /Cohen's d (>0.50)	*Average* magnitude of a treatment effect across several (group) studies (mean difference between intervention/control relative to sample variability)
	Raw (unstandardised) mean difference	*Average* magnitude of a treatment effect across several (group) studies (mean difference between intervention/control)
Individual benchmarks for treatment success	Distribution-based	Minimal statistically significant score change from pre to post intervention on OMI for an individual (Smallest Detectable Change)
	Anchor-based (preferably using patient-reported ratings of perceived treatment impact and treatment satisfaction)	Minimal score change from pre to post intervention on OMI for an individual perceived as meaningful by key stakeholder group (Minimal Important Change)
Response rate for an intervention	Distribution-based	Percent of study participants exceeding minimal statistically significant score change from pre to post intervention on OMI (Smallest Detectable Change)
	Anchor-based (using a patient-reported anchor)	Percent of study participants exceeding minimal score change from pre to post intervention on OMI perceived as meaningful by key stakeholder group (Minimal Important Change)

OMI = outcome measurement instrument, QoL = quality of life, ROMA COS = Research Outcome Measurement in Aphasia Core Outcome Set (Wallace et al., 2021; Wallace et al., 2018), QALYs = Quality-adjusted Life Years

CLINICAL IMPLICATIONS

Well-designed and well-managed randomised controlled trials form the basis for translating research data into clinical practice. Understanding the distinction between group versus individual level outcomes as well as statistically significant versus clinically meaningful therapy-induced changes is vital in interpreting study results and determining their applicability to clinical practice. We therefore propose a set of recommendations for aphasia researchers and clinicians for determining the success of an intervention:

- Intervention goals and the treatment approach suitable for people with aphasia may vary depending on stakeholder perspectives (e.g., socioeconomic, ~~acute care~~ stroke team, aphasia rehabilitation) as well as by time after the stroke, aphasia severity and type.
- The ROMA 'core outcome set' (Wallace et al., 2018) is a globally applicable minimum set of standardised outcome measurement instruments to assess treatment success based on the aphasia rehabilitation stakeholder perspective (people with aphasia and their family members, aphasia researchers, clinicians and managers). Four standardised outcome measurement instruments have been selected through aphasia expert consensus to assess the prioritised treatment outcomes identified by this stakeholder group: language (WAB-R), communication (The Scenario Test), emotional well-being (GHQ-12), and quality of life (SAQOL-39g). Depending on the individual treatment goals, additional standardised outcome measurement instruments may be administered to determine treatment success.
- Results from group-level analysis cannot serve as benchmarks for therapy success for the individual client. This is also true for high-quality randomised controlled trials and systematic reviews, where results are averaged across all participants. Such group-level statistical analyses summarize the overall effect of an intervention for the entire population of post-stroke aphasia, but do not allow inferences about the treatment success for an individual.
- In assessing an individual's therapy success, benchmarks should be based on both the "smallest detectable change" (the smallest 'statistically significant' change score for an outcome measure) and the "minimal important change" (the smallest change score for an outcome measure that is considered important by the relevant stakeholder groups). Either benchmark may vary for different languages and aphasia

subgroups (depending e.g. on the individual's age, time post-stroke, aphasia severity, aphasia type).

Figure 3 summarises the key questions to address when operationalising aphasia treatment success for an individual.

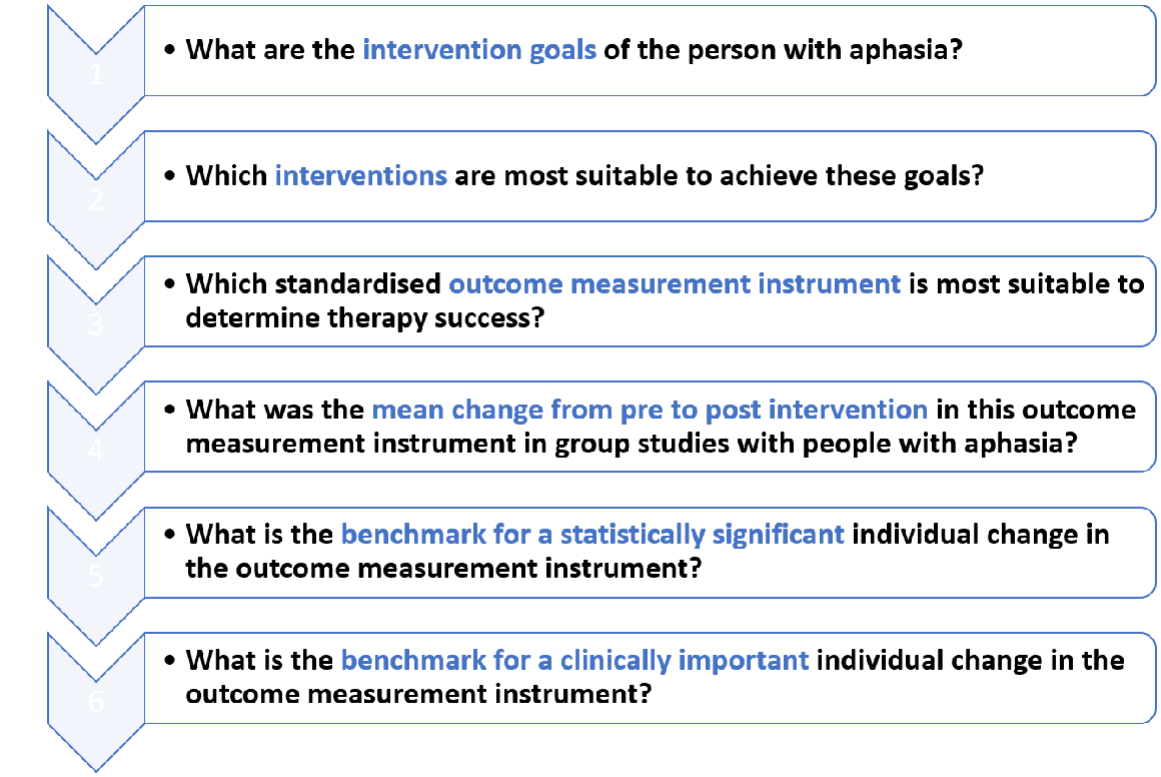

Figure 3: Key questions to address when operationalising aphasia treatment success for an individual

Disclosure of interest. The authors report no conflict of interest.

References

Aarnio, K., Rodriguez-Pardo, J., Siegerink, B., Hardt, J., Broman, J., Tulkki, L., Haapaniemi, E., Kaste, M., Tatlisumak, T., & Putaala, J. (2018). Return to work after ischemic stroke in young adults: A registry-based follow-up study. *Neurology*, *91*(20), e1909-e1917. https://doi.org/10.1212/WNL.0000000000006510

Ali, M., Lyden, P., Brady, M., & Collaboration, V. (2015). Aphasia and Dysarthria in Acute Stroke: Recovery and Functional Outcome. *Int J Stroke*, *10*(3), 400-406. https://doi.org/10.1111/ijs.12067

Ashley, K. D., Lee, L. T., & Heaton, K. (2019). Return to Work Among Stroke Survivors. *Workplace Health Saf*, *67*(2), 87-94. https://doi.org/10.1177/2165079918812483

Babbitt, E. M., Worrall, L., & Cherney, L. R. (2016). Who benefits from an intensive comprehensive aphasia program? *Topics in Language Disorders*, *36*(2), 168-184.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603-617.

Bernhardt, J., Hayward, K. S., Kwakkel, G., Ward, N. S., Wolf, S. L., Borschmann, K., Krakauer, J. W., Boyd, L. A., Carmichael, S. T., Corbett, D., & Cramer, S. C. (2017). Agreed definitions and a shared vision for new standards in stroke recovery research: The Stroke Recovery and Rehabilitation Roundtable taskforce. *Int J Stroke*, *12*(5), 444-450. https://doi.org/10.1177/1747493017711816

Black-Schaffer, R. M., & Osberg, J. S. (1990). Return to work after stroke: development of a predictive model. *Arch Phys Med Rehabil*, *71*(5), 285-290. (Not in File)

Blomert, L., Kean, M. L., Koster, C., & Schokker, J. (1994). Amsterdam Nijmegen Everyday Language Test: construction, reliability and validity. *Aphasiology*, *8*, 381-407.

Boehme, A. K., Martin-Schild, S., Marshall, R. S., & Lazar, R. M. (2016). Effect of aphasia on acute stroke outcomes. *Neurology*, *87*(22), 2348-2354. https://doi.org/10.1212/WNL.0000000000003297

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis. Chichester:* John Wiley & Sons, Ltd.

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database Syst Rev*, *6*, CD000425. http://www.ncbi.nlm.nih.gov/pubmed/27245310 (Not in File)

Breitenstein, C., Grewe, T., Floel, 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., Wigbers, F., Ruhmkorf, C., Hempen, I., List, J., Baumgaertner, A., & group, F. E. 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. *Lancet*, *389*(10078), 1528-1538. https://doi.org/10.1016/S0140-6736(17)30067-3

Broderick, J. P., Adeoye, O., & Elm, J. (2017). Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials. *Stroke*, *48*(7), 2007-2012. https://doi.org/10.1161/STROKEAHA.117.017866

Cadilhac, D. A., Kim, J., Wilson, A., Berge, E., Patel, A., Ali, M., Saver, J., Christensen, H., Cuche, M., Crews, S., Wu, O., Provoyeur, M., McMeekin, P., Durand-Zaleski, I., Ford, G. A., Muhlemann, N., Bath, P. M., Abdul-Rahim, A. H., Sunnerhagen, K., Meretoja, A., Thijs, V., Weimar, C., Massaro, A., Ranta, A., Lees, K. R., & group, E. S. O. H. E. W. (2020). Improving economic evaluations in stroke: A report from the ESO Health Economics Working Group. *Eur Stroke J*, *5*(2), 184-192. https://doi.org/10.1177/2396987319897466

Caporali, A., & Basso, A. (2003). A survey of long-term outcome of aphasia and of chances of gainful employment. *Aphasiology*, *17*(9), 815-834.

Charter, R. A. (2008). Statistical approaches to achieving sufficiently high test score reliabilities for research purposes. *J Gen Psychol*, *135*(3), 241-251. https://doi.org/10.3200/GENP.135.3.241-251

Chen, S., Wolf, S. L., Zhang, Q., Thompson, P. A., & Winstein, C. J. (2012). Minimal detectable change of the actual amount of use test and the motor activity log: the EXCITE Trial. *Neurorehabil Neural Repair*, *26*(5), 507-514. https://doi.org/10.1177/1545968311425048

Cherney, L. R. (2010). Oral reading for language in aphasia: impact of aphasia severity on cross-modal outcomes in chronic nonfluent aphasia. *Semin Speech Lang*, *31*(1), 42-51. https://doi.org/10.1055/s-0029-1244952

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, 404-413.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Global Burden of Disease Neurology Collaborators (2019). Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*,*18*(5), 459-480. https://doi.org/10.1016/S1474-4422(18)30499-X

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory.* Cengage Learning.

Danner, D. (2016). *Reliability – The precision of a measurement. GESIS Survey Guidelines.* GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_011

Darley, F., Helm, N., Holland, A., & Linebaugh, C. (1980, June 1-5). Techniques in treating mild or high-level aphasic impairment - Panel discussion. 10th Clinical AphasiologyConference, Bar Harbor, ME, USA.

de Graaf, J. A., Kuijpers, M., Visser-Meily, J., Kappelle, L. J., & Post, M. (2020). Validity of an enhanced EQ-5D-5L measure with an added cognitive dimension in patients with stroke. *Clin Rehabil*, *34*(4), 545-550. https://doi.org/10.1177/0269215520907990

de Haan, R., Limburg, M., Bossuyt, P., van der Meulen, J., & Aaronson, N. (1995). The clinical meaning of Rankin 'handicap' grades after stroke. *Stroke*, *26*(11), 2027-2030. https://doi.org/10.1161/01.str.26.11.2027

de Vet, H. C., & Terwee, C. B. (2010). The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol*, *63*(7), 804-805; author reply 806. https://doi.org/10.1016/j.jclinepi.2009.12.015

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2021). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (Ed.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).* Cochrane.

Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., & Kiran, S. (2020). Videoconference Administration of the Western Aphasia Battery-Revised: Feasibility and Validity. *Am J Speech Lang Pathol*, *29*(2), 673-687. https://doi.org/10.1044/2019_AJSLP-19-00023

Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Vernooij, R. W., Huang, H., Rizwan, Y., Siemieniuk, R., Lytvyn, L., Patrick, D. L., Ebrahim, S., Furukawa, T., Nesrallah, G., Schunemann, H. J., Bhandari, M., Thabane, L., & Guyatt, G. H. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ*, *369*, m1714. https://doi.org/10.1136/bmj.m1714

Dijkerman, H. C., Wood, V. A., & Hewer, R. L. (1996). Long-term outcome after discharge from a stroke rehabilitation unit. *J R Coll Physicians Lond*, *30*(6), 538-546.

Donoghue, D., Physiotherapy Research and Older People (PROP) group, & Stokes, E. K. (2009). How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *J Rehabil Med*, *41*(5), 343-346. https://doi.org/10.2340/16501977-0337

Doucet, T., Muller, F., Verdun-Esquer, C., Debelleix, X., & Brochard, P. (2012). Returning to work after a stroke: a retrospective study at the Physical and Rehabilitation Medicine Center La Tour de Gassies. *Ann Phys Rehabil Med*, *55*(2), 112-127.

Ekker, M. S., Verhoeven, J. I., Vaartjes, I., Jolink, W. M. T., Klijn, C. J. M., & de Leeuw, F. E. (2019). Association of Stroke Among Adults Aged 18 to 49 Years With Long-term Mortality. *JAMA*, *321*(21), 2113-2123. https://doi.org/10.1001/jama.2019.6560

El Hachioui, H., Lingsma, H. F., van de Sandt-Koenderman, M. W., Dippel, D. W., Koudstaal, P. J., & Visch-Brink, E. G. (2013). Long-term prognosis of aphasia after stroke. *J Neurol Neurosurg Psychiatry*, *84*(3), 310-315.

Ellis, C., Lindrooth, R. C., & Horner, J. (2014). Retrospective cost-effectiveness analysis of treatments for aphasia: an approach using experimental data. *Am J Speech Lang Pathol*, *23*(2), 186-195. https://doi.org/10.1044/2013_AJSLP-13-0037

Ellis, C., Simpson, A. N., Bonilha, H., Mauldin, P. D., & Simpson, K. N. (2012). The one-year attributable cost of poststroke aphasia. *Stroke*, *43*(5), 1429-1431.

Elman, R. J., & Bernstein-Ellis, E. (1999). The efficacy of group communication treatment in adults with chronic aphasia. *J Speech Lang Hear Res*, *42*(2), 411-419.

Eom, B., & Sung, J. E. (2016). The effects of sentence repetition–based working memory treatment on sentence comprehension abilities in individuals with aphasia. *American Journal of Speech-Language Pathology*, *25*(4S), S823-S838.

Everett, E. A., Everett, W., Brier, M. R., & White, P. (2021). Appraisal of Health States Worse Than Death in Patients With Acute Stroke. *Neurology Clinical Practice 11*(1), 43-48. https://doi.org/10.1212/CPJ.0000000000000856

Falconer, C., & Antonucci, S. M. (2012). Use of semantic feature analysis in group discourse treatment for aphasia: Extension and expansion. *Aphasiology*, *26*(1), 64-82.

Faraone, S. V. (2008). Interpreting estimates of treatment effects: implications for managed care. *P T*, *33*(12), 700-711.

Flowers, H. L., Skoretz, S. A., Silver, F. L., Rochon, E., Fang, J., Flamand-Roze, C., & Martino, R. (2016). Poststroke Aphasia Frequency, Recovery, and Outcomes: A Systematic Review and Meta-Analysis. *Arch Phys Med Rehabil*, *97*(12), 2188-2201 e2188. https://doi.org/10.1016/j.apmr.2016.03.006

Franklin, S., Hahrhen, D., Hayes, M., Mc Manus, S. D., & Pollock, A. (2018). Top 10 research priorities relating to aphasia following stroke. *Aphasiology*, *32*(11), 1388-1395.

Frattali, C. M., Thompson, C. M., Holland, A. L., Wohl, C. B., & Ferketic, M. M. (1995). The FACS of life ASHA facs--a functional outcome measure for adults. *ASHA*, *37*(4), 40-46.

Fridriksson, J., & Hillis, A. E. (2021). Current Approaches to the Treatment of Post-Stroke Aphasia. *J Stroke*, *23*(2), 183-201. https://doi.org/10.5853/jos.2020.05015

Furlan, L., & Sterr, A. (2018). The Applicability of Standard Error of Measurement and Minimal Detectable Change to Motor Learning Research-A Behavioral Study. *Front Hum Neurosci*, *12*, 95. https://doi.org/10.3389/fnhum.2018.00095

Galeoto, G., Iori, F., De Santis, R., Santilli, V., Mollica, R., Marquez, M. A., Sansoni, J., & Berardi, A. (2019). The outcome measures for loss of functionality in the activities of daily living of adults after stroke: a systematic review. *Top Stroke Rehabil*, *26*(3), 236-245. https://doi.org/10.1080/10749357.2019.1574060

Gallagher, M., Hares, T., Spencer, J., Bradshaw, C., & Webb, I. (1993). The nominal group technique: a research tool for general practice? *Fam Pract*, *10*(1), 76-81. https://doi.org/10.1093/fampra/10.1.76

Gerdes, N., Funke, U. N., Schuwer, U., Themann, P., Pfeiffer, G., & Meffert, C. (2012). Selbstaendigkeits-Index fur die Neurologische und Geriatrische Rehabilitation (SINGER)" - Entwicklung und Validierung eines neuen Assessment-Instruments. [Scores of Independence for Neurologic and Geriatric Rehabilitation (SINGER)" - development and validation of a new assessment instrument.]. *Rehabilitation (Stuttg)*, *51*(5), 289-299. https://doi.org/10.1055/s-0031-1287805

Gilmore, N., Dwyer, M., & Kiran, S. (2019). Benchmarks of Significant Change After Aphasia Rehabilitation. *Arch Phys Med Rehabil*, *100*(6), 1131-1139 e1187. https://doi.org/10.1016/j.apmr.2018.08.177

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research.* Sage.

Godecke, E., Armstrong, E., Rai, T., Ciccone, N., Rose, M. L., Middleton, S., Whitworth, A., Holland, A., Ellery, F., Hankey, G. J., Cadilhac, D. A., Bernhardt, J., & Group, V. C. (2020). A randomized control trial of intensive aphasia therapy after acute stroke: The Very Early Rehabilitation for SpEech (VERSE) study. *Int J Stroke*, 1747493020961926. https://doi.org/10.1177/1747493020961926

Godecke, E., Rai, T., Cadilhac, D. A., Armstrong, E., Middleton, S., Ciccone, N., Whitworth, A., Rose, M. L., Holland, A., Ellery, F., Hankey, G. J., Bernhardt, J., & Collaboration, V. (2018). Statistical analysis plan (SAP) for the Very Early Rehabilitation in Speech (VERSE) after stroke trial: an international 3-arm clinical trial to determine the effectiveness of early, intensive, prescribed, direct aphasia therapy. *Int J Stroke*, *13*(8), 863-880. https://doi.org/10.1177/1747493018790055

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol Med*, *27*(1), 191-197. https://doi.org/10.1017/s0033291796004242

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston Diagnostic Aphasia Examination (BDAE-3)* (3rd ed.). Pro-Ed.

Graham, J. R., Pereira, S., & Teasell, R. (2011). Aphasia and return to work in younger stroke survivors. *Aphasiology*, *25*(8), 952-960.

Granger, C. V., Hamilton, B. B., Keith, R. A., Zielezny, M., & Sherwin, F. S. (1986). Advances in functional assessment for medical rehabilitation. *Topics in Geriatric Rehabilitation*, *1*(3), 59-74.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*, *31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Guo, Y. E., Togher, L., Power, E., Heard, R., Luo, N., Yap, P., & Koh, G. C. H. (2015). Sensitivity to change and responsiveness of the Stroke and Aphasia Quality-of-Life Scale (SAQOL) in a Singapore stroke population. *Aphasiology*, *31*(4), 427.

Haley, K. L., Womack, J. L., Harmon, T. G., McCulloch, K. L., & Faldowski, R. A. (2019). Life activity choices by people with aphasia: repeated interviews and proxy agreement. *Aphasiology*, *33*(6), 710-730. https://doi.org/10.1080/02687038.2018.1506087

Harrison, J. K., McArthur, K. S., & Quinn, T. J. (2013). Assessment scales in stroke: clinimetric and clinical considerations. *Clin Interv Aging*, *8*, 201-211. https://doi.org/10.2147/CIA.S32405

Heuschmann, P. U., Busse, O., Wagner, M., Endres, M., Villringer, A., Röther, J., Kolominsky-Rabas, P. L., & Berger, K. (2010). Schlaganfallhäufigkeit und Versorgung von Schlaganfallpatienten in Deutschland [Stroke frequency and routine care for stroke patients in Germany]. *Akt Neurologie*, *37*(7), 333-340.

Higgins, J. P. T., Li, T., & Deeks, J. J. (2021). Chapter 6: Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021).* . Cochrane. www.training.cochrane.org/handbook

Hilari, K., Behn, N., Marshall, J., Simpson, A., Thomas, S., Northcott, S., Flood, C., McVicker, S., Jofre-Bonet, M., Moss, B., James, K., & Goldsmith, K. (2019). Adjustment with aphasia after stroke: study protocol for a pilot feasibility randomised controlled trial for SUpporting wellbeing through PEeR Befriending (SUPERB). *Pilot Feasibility Stud*, *5*, 14. https://doi.org/10.1186/s40814-019-0397-6

Hilari, K., Byng, S., Lamping, D. L., & Smith, S. C. (2003). Stroke and Aphasia Quality of Life Scale-39 (SAQOL-39): evaluation of acceptability, reliability, and validity. *Stroke*, *34*(8), 1944-1950.

Hilari, K., Lamping, D. L., Smith, S. C., Northcott, S., Lamb, A., & Marshall, J. (2009). Psychometric properties of the Stroke and Aphasia Quality of Life Scale (SAQOL-39) in a generic stroke population. *Clin Rehabil*, *23*(6), 544-557. https://doi.org/10.1177/0269215508101729

Hillis, A. E. (2007). Magnetic resonance perfusion imaging in the study of language. *Brain Lang*, *102*(2), 165-175. https://doi.org/10.1016/j.bandl.2006.04.016

Hillis, A. E., Ulatowski, J. A., Barker, P. B., Torbey, M., Ziai, W., Beauchamp, N. J., Oh, S., & Wityk, R. J. (2003). A pilot randomized trial of induced blood pressure elevation: effects on function and focal perfusion in acute and subacute stroke. *Cerebrovasc Dis*, *16*(3), 236-246. https://doi.org/10.1159/000071122

Hinckley, J. J., Boyle, E., Lombard, D., & Bartels-Tobin, L. (2014). Towards a consumer-informed research agenda for aphasia: preliminary work. *Disabil Rehabil*, *36*(12), 1042-1050. https://doi.org/10.3109/09638288.2013.829528

Hinckley, J. J. (1998). Investigating the predictors of lifestyle satisfaction among younger adults with chronic aphasia. *Aphasiology*, *12*, 509-518.

Hinckley, J. J. (2002). Vocational and social outcomes of adults with chronic aphasia. *J Commun Disord*, *35*(6), 543-560. https://doi.org/10.1016/s0021-9924(02)00119-3

Holland, A. L., Wozniak, L., & Fromm, D. (2018). *CADL-3 : communication activities of daily living.* Pro-Ed.

Hula, W. D., Donovan, N. J., Kendall, D. L., & Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology*, *24*(11), 1326-1341.

Hula, W. D., Doyle, P. J., Stone, C. A., Austermann Hula, S. N., Kellough, S., Wambaugh, J. L., Ross, K. B., Schumacher, J. G., & St Jacque, A. (2015). The Aphasia Communication Outcome Measure (ACOM): Dimensionality, Item Bank Calibration, and Initial Validation. *J Speech Lang Hear Res*, *58*(3), 906-919. https://doi.org/10.1044/2015_JSLHR-L-14-0235

Intercollegiate Stroke Working Group Party (2016). *National [UK] clinical guideline for stroke (5th ed.).* London: Royal College of Physicians.

Isaacs, M., Ali, M., Brady, M. C., & Wallace, S. J. (2021). *Establishing standards for reporting participant characteristics in post-stroke aphasia research: An international e-Delphi exercise and consensus meeting.* [Manuscript in preparation]. School of Health and Rehabilitation Sciences, The University of Queensland, Australia.

Janssen, M. F., Pickard, A. S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P., & Busschbach, J. (2013). Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res*, *22*(7), 1717-1727. https://doi.org/10.1007/s11136-012-0322-4

Johnson, B. T., & Huedo-Medina, T. B. (2013). *Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions* (www.effectivehealthcare.ahrq.gov/reports/final.cfm

Johnson, L., Basilakos, A., Yourganov, G., Cai, B., Bonilha, L., Rorden, C., & Fridriksson, J. (2019). Progression of Aphasia Severity in the Chronic Stages of Stroke. *Am J Speech Lang Pathol*, *28*(2), 639-649. https://doi.org/10.1044/2018_AJSLP-18-0123

Kaplan, E. F., Goodglas, H., & Weinstraub, S. (2002). *The Boston Naming Test* ((2nd Editon) ed.). Lea Febiger.

Katz, R. C., & Wertz, R. T. (1997). The efficacy of computer-provided reading treatment for chronic aphasic adults. *J Speech Lang Hear Res*, *40*(3), 493-507. https://doi.org/10.1044/jslhr.4003.493

Kelley, K. (2007). Confidence Intervals for Standardized Effect Sizes:Theory, Application, and Implementation. *Journal of Statistical Software*, *20*(8), 1-24. https://doi.org/10.18637/jss.v020.i08

Kempler, D., & Goral, M. (2011). A comparison of drill-and communication-based treatment for aphasia. *Aphasiology*, *25*(11), 1327-1346.

Kertesz, A. (2007). *The Western Aphasia Battery-Revised*. Grune & Stratton.

Kissela, B. M., Khoury, J. C., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., Khatri, P., Ferioli, S., De Los Rios La Rosa, F., Broderick, J. P., & Kleindorfer, D. O. (2012). Age at stroke: temporal trends in stroke incidence in a large, biracial population. *Neurology*, *79*(17), 1781-1787. https://doi.org/10.1212/WNL.0b013e318270401d

Kleindorfer, D. O., Towfighi, A., Chaturvedi, S., Cockroft, K. M., Gutierrez, J., Lombardi-Hill, D., Kamel, H., Kernan, W. N., Kittner, S. J., Leira, E. C., Lennon, O., Meschia, J. F., Nguyen, T. N., Pollak, P. M., Santangeli, P., Sharrief, A. Z., Smith, S. C., Jr., Turan, T. N., & Williams, L. S. (2021). 2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: A Guideline From the American Heart Association/American Stroke Association. *Stroke*,*52*(7), e364-e467. https://doi.org/10.1161/STR.0000000000000375

Kwakkel, G., Lannin, N. A., Borschmann, K., English, C., Ali, M., Churilov, L., Saposnik, G., Winstein, C., van Wegen, E. E. H., Wolf, S. L., Krakauer, J. W., & Bernhardt, J. (2017). Standardized Measurement of Sensorimotor Recovery in Stroke Trials: Consensus-Based Core Recommendations from the Stroke Recovery and Rehabilitation Roundtable. *Neurorehabil Neural Repair*, *31*(9), 784-792. https://doi.org/10.1177/1545968317732662

Latimer, N. R., Bhadhuri, A., Alshreef, A. O., Palmer, R., Cross, E., Dimairo, M., Julious, S., Cooper, C., Enderby, P., Brady, M. C., Bowen, A., Bradley, E., & Harrison, M. (2020). Self-managed, computerised word finding therapy as an add-on to usual care for chronic aphasia post-stroke: An economic evaluation. *Clin Rehabil*, 269215520975348. https://doi.org/10.1177/0269215520975348

Lazar, R. M., & Antoniello, D. (2008). Variability in recovery from aphasia. *Curr Neurol Neurosci Rep*, *8*(6), 497-502. https://doi.org/10.1007/s11910-008-0079-x

Lazar, R. M., & Boehme, A. K. (2017). Aphasia As a Predictor of Stroke Outcome. *Curr Neurol Neurosci Rep*, *17*(11), 83. https://doi.org/10.1007/s11910-017-0797-z

Lenhard, W., & Lenhard, A. (2016). *Calculation of Effect Sizes*. Psychometrica https://www.psychometrica.de/effect_size.html.

Lomas, J., Pickard, L., Bester, S., Elbard, H., Finlayson, A., & Zoghaib, C. (1989). The communicative effectiveness index: development and psychometric evaluation of a functional communication measure for adult aphasia. *J Speech Hear Disord*, *54*(1), 113-124.

Long, A., Hesketh, A., Paszek, G., Booth, M., & Bowen, A. (2008). Development of a reliable self-report outcome measure for pragmatic trials of communication therapy following stroke: the Communication Outcome after Stroke (COAST) scale. *Clin Rehabil*, *22*(12), 1083-1094. https://doi.org/10.1177/0269215508090091

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Magnusson, K. (2020). *Interpreting Cohen's d Effect Size.* https://rpsychologist.com/d3/cohend/

Maher, L. M., Kendall, D., Swearengin, J. A., Rodriguez, A., Leon, S. A., Pingel, K., Holland, A., & Rothi, L. J. (2006). A pilot study of use-dependent learning in the context of Constraint Induced Language Therapy. *Journal of the International Neuropsychological Society*, *12*(6), 843-852.

Mahoney, F. I., & Barthel, D. W. (1965). Functional Evaluation: The Barthel Index. *Md State Med J*, *14*, 61-65.

Marshall, J., Devane, N., Talbot, R., Caute, A., Cruice, M., Hilari, K., MacKenzie, G., Maguire, K., Patel, A., Roper, A., & Wilson, S. (2020). A randomised trial of social support group intervention for people with aphasia: A Novel application of virtual reality. *PLoS One*, *15*(9), e0239715. https://doi.org/10.1371/journal.pone.0239715

McGill, K., Sackley, C., Godwin, J., McGarry, J., & Brady, M. C. (2021, in press). Using the Barthel Index and modified Rankin Scale as outcome measures for stroke rehabilitation trials; an examinationof appropriateness and minimum sample size requirements. *Journal of Stroke and Cerebrovascular Diseases. ~~[Manuscript in preparation.] Nursing, Midwifery and Allied Health Professions Research Unit, Glasgow Caledonian University, United Kingdom.~~*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30-46.

McManus, I. C. (2012). The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Med Teach*, *34*(7), 569-576. https://doi.org/10.3109/0142159X.2012.670318

Meier, E. L., Johnson, J. P., Villard, S., & Kiran, S. (2017). Does Naming Therapy Make Ordering in a Restaurant Easier? Dynamics of Co-Occurring Change in Cognitive-Linguistic and Functional Communication Skills in Aphasia. *Am J Speech Lang Pathol*, *26*(2), 266-280. https://doi.org/10.1044/2016_AJSLP-16-0028

Menahemi-Falkov, M., Breitenstein, C., Pierce, J. E., Hill, A. J., O'Halloran, R., & Rose, M. L. (2021). A systematic review of maintenance following intensive therapy programs in chronic post-stroke aphasia: Importance of individual response analysis. *Disability and Rehabilitation*, 1-16. https://doi.org/10.1080/09638288.2021.1955303

Middel, B., & van Sonderen, E. (2002). Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *Int J Integr Care*, *2*, e15. https://doi.org/10.5334/ijic.65

Mitchell, C., Gittins, M., Tyson, S., Vail, A., Conroy, P., Paley, L., & Bowen, A. (2021). Prevalence of aphasia and dysarthria among inpatient strokesurvivors: describing the population, therapy provision and outcomes on discharge. *Aphasiology*, *35*(7), 950-960. https://doi.org/10.1080/02687038.2020.1759772

Mokkink, L. B., Boers, M., van der Vleuten, C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument. User manual. *BMC Med Res Methodol*, *20*(1), 293. https://www.cosmin.nl/wp-content/uploads/user-manual-COSMIN-Risk-of-Bias-tool_v4_JAN_final.pdf

Mokkink, L. B., Boers, M., van der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol*, *20*(1), 293. https://doi.org/10.1186/s12874-020-01179-5

Mokkink, L. B., Prinsen, C. A., Bouter, L. M., Vet, H. C., & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*, *20*(2), 105-113. https://doi.org/10.1590/bjpt-rbf.2014.0143

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*, *19*(4), 539-549. https://doi.org/10.1007/s11136-010-9606-8

Mozeiko, J., Myers, E. B., & Coelho, C. A. (2018). Treatment Response to a Double Administration of Constraint-Induced Language Therapy in Chronic Aphasia. *Journal of Speech, Language, and Hearing Research*, *61*, 1664-1690.

National Institute for Health and Care Excellence, N. I. C. E. (2019). *NICEimpact Stroke.* (https://www.nice.org.uk/Media/Default/About/what-we-do/Into-practice/measuring-uptake/NICE-Impact-stroke.pdf

Northcott, S., Simpson, A., Thomas, S. A., Hirani, S. P., Flood, C., & Hilari, K. (2019). Solution Focused brief therapy In post-stroke Aphasia (SOFIA Trial): protocol for a feasibility randomised controlled trial. *AMRC Open Research*, *1*(11). https://doi.org/https://doi.org/10.12688/amrcopenres.12873.2

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jonsson, B., group, C. s., & European Brain, C. (2012). The economic cost of brain disorders in Europe. *Eur J Neurol*, *19*(1), 155-162. https://doi.org/10.1111/j.1468-1331.2011.03590.x

Ottenbacher, K. J., Hsu, Y., Granger, C. V., & Fiedler, R. C. (1996). The reliability of the functional independence measure: a quantitative review. *Arch Phys Med Rehabil*, *77*(12), 1226-1232. https://doi.org/10.1016/s0003-9993(96)90184-7

Padilla, M. (2019). A primer on reliability via coefficient alpha and omega. . *Archives of Psychology*, *3*(8), 1-15.

Palmer, R., Dimairo, M., Cooper, C., Enderby, P., Brady, M., Bowen, A., Latimer, N., Julious, S., Cross, E., Alshreef, A., Harrison, M., Bradley, E., Witts, H., & Chater, T. (2019). Self-managed, computerised speech and language therapy for patients with chronic aphasia post-stroke compared with usual care or attention control (Big CACTUS): a multicentre, single-blinded, randomised controlled trial. *Lancet Neurol*, *18*(9), 821-833. https://doi.org/10.1016/S1474-4422(19)30192-9

Parr, S., Byng, S., Gilpin, S., & Ireland, C. (1997). *Talking about aphasia: Living with Loss of Language After Stroke.* Open University Press.

Peach, R. K., Beck, K. M., Gorman, M., & Fisher, C. (2019). Clinical outcomes following language-specific attention treatment versus direct attention training for aphasia: a comparative effectiveness study. *Journal of Speech, Language, and Hearing Research*, *62*(8), 2785-2811.

Pedersen, P. M., Jorgensen, H. S., Nakayama, H., Raaschou, H. O., & Olsen, T. S. (1995). Aphasia in acute stroke: incidence, determinants, and recovery. *Ann Neurol*, *38*(4), 659-666. (In File)

Persad, C., Wozniak, L., & Kostopoulos, E. (2013). Retrospective analysis of outcomes from two intensive comprehensive aphasia programs. *Top Stroke Rehabil*, *20*(5), 388-397. (Not in File)

Rabin, R., & de Charro, F. (2001). EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*, *33*(5), 337-343. https://doi.org/10.3109/07853890109002087

Raftery, J., Young, A., Stanton, L., Milne, R., Cook, A., Turner, D., & Davidson, P. (2015). Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme. *Health Technol Assess*, *19*(11), 1-138. https://doi.org/10.3310/hta19110

Rai, S. K., Yazdany, J., Fortin, P. R., & Avina-Zubieta, J. A. (2015). Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis Res Ther*, *17*, 143. https://doi.org/10.1186/s13075-015-0658-6

Ramsey, S. D., Willke, R. J., Glick, H., Reed, S. D., Augustovski, F., Jonsson, B., Briggs, A., & Sullivan, S. D. (2015). Cost-effectiveness analysis alongside clinical trials II-An ISPOR Good Research Practices Task Force report. *Value Health*, *18*(2), 161-172. https://doi.org/10.1016/j.jval.2015.02.001

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

Rochmah, T. N., Rahmawati, I. T., Dahlui, M., Budiarto, W., & Bilqis, N. (2021). Economic Burden of Stroke Disease: A Systematic Review. *Int J Environ Res Public Health*, *18*(14). https://doi.org/10.3390/ijerph18147552

Rose, M. L., Copland, D., Nickels, L., Togher, L., Meinzer, M., Rai, T., Cadilhac, D. A., Kim, J., Foster, A., Carragher, M., Hurley, M., & Godecke, E. (2019). Constraint-induced or multi-modal personalized aphasia rehabilitation (COMPARE): A randomized controlled trial for stroke-related chronic aphasia. *Int J Stroke*, *14*(9), 972-976. https://doi.org/10.1177/1747493019870401

Rose, M. L., Rai, T., Copland, D., Nickels, L., Togher, L., Meinzer, M., Godecke, E., Kim, J., Cadilhac, D. A., Hurley, M., Wilcox, C., & Carragher, M. (2021). Statistical analysis plan for the COMPARE trial: a 3-arm randomised controlled trial comparing the effectiveness of Constraint-induced Aphasia Therapy Plus and Multi-modality Aphasia Therapy to usual care in chronic post-stroke aphasia (COMPARE). *Trials*, *22*(1), 303. https://doi.org/10.1186/s13063-021-05238-0

Rousson, V. (2011). Assessing inter-rater reliability when the raters are fixed: Two concepts and two estimates. *Biometrical Journal*, *53*(3), 477–490.

Sachs, A., Rising, K., & Beeson, P. M. (2020). A Retrospective Study of Long-Term Improvement on the Boston Naming Test. *Am J Speech Lang Pathol*, *29*(1S), 425-436. https://doi.org/10.1044/2019_AJSLP-CAC48-18-0224

Sanderson, T., Hewlett, S., Calnan, M., Morris, M., Raza, K., & Kumar, K. (2012). Exploring the cultural validity of rheumatology outcomes. *Br J Nurs*, *21*(17), 1015-1020, 1522-1523. https://doi.org/10.12968/bjon.2012.21.17.1015

Shewan, C. M., & Kertesz, A. (1980). Reliability and validity characteristics of the Western Aphasia Battery (WAB). *J Speech Hear Disord*, *45*(3), 308-324.

Smania, N., Gandolfi, M., Aglioti, S. M., Girardi, P., Fiaschi, A., & Girardi, F. (2010). How long is the recovery of global aphasia? Twenty-five years of follow-up in a patient with left hemisphere stroke. *Neurorehabil Neural Repair*, *24*(9), 871-875. https://doi.org/10.1177/1545968310368962

Stahl, B., Darkow, R., von Podewils, V., Meinzer, M., Grittner, U., Reinhold, T., Grewe, T., Breitenstein, C., & Floel, A. (2019). Transcranial Direct Current Stimulation to Enhance Training Effectiveness in Chronic Post-Stroke Aphasia: A Randomized Controlled Trial Protocol. *Front Neurol*, *10*, 1089. https://doi.org/10.3389/fneur.2019.01089

Stroke Association, U. (2021). *Shaping Stroke Research to Rebuild Lives: The Stroke Priority Setting Partnership results for investment.* (https://www.stroke.org.uk/sites/default/files/research/stroke_priority_setting_partnership_full_report.pdf

Tarrant, M., Carter, M., Dean, S. G., Taylor, R. S., Warren, F. C., Spencer, A., Adamson, J., Landa, P., Code, C., & Calitri, R. (2018). Singing for people with aphasia (SPA): a protocol for a pilot randomised controlled trial of a group singing intervention to improve well-being. *BMJ Open*, *8*(9), e025167. https://doi.org/10.1136/bmjopen-2018-025167

Terluin, B., Eekhout, I., Terwee, C. B., & de Vet, H. C. (2015). Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *J Clin Epidemiol*, *68*(12), 1388-1396. https://doi.org/10.1016/j.jclinepi.2015.03.015

Terry, L., & Kelley, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology*, *65*, 371-401.

Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J. S., Terluin, B., Cella, D., Griffith, P., & Mokkink, L. B. (2021). Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Qual Life Res*, *30*(10), 2729-2754. https://doi.org/10.1007/s11136-021-02925-y

van Bloemendaal, M., van de Water, A. T., & van de Port, I. G. (2012). Walking tests for stroke survivors: a systematic review of their measurement properties. *Disabil Rehabil*, *34*(26), 2207-2221. https://doi.org/10.3109/09638288.2012.680649

van der Gaag, A., & Brooks, R. (2008). Economic aspects of a therapy and support service for people with long-term stroke and aphasia. *Int J Lang Commun Disord*, *43*(3), 233-244. https://doi.org/10.1080/13682820701560376

van der Meulen, I., W.M., v. d. S.-K., Duivenvoorden, H. J., & Ribbers, G. M. (2010). Measuring verbal and non-verbal communication in aphasia: reliability, validity, and sensitivity to change of the Scenario Test. *Int J Lang Commun Disord*, *45*(4), 424-435.

Wallace, S. J., Worrall, L., Le Dorze, G., Brandenburg, C., Foulkes, J., & Rose, T. A. (2020). Many ways of measuring: a scoping review of measurement instruments for use with people with aphasia. *Aphasiology*. https://doi.org/10.1080/02687038.2020.1836318

Wallace, S. J., Worrall, L., Rose, T., Alyahya, R. S. W., Babbitt, E., Beeke, S., de Beer, C., Bose, A., Bowen, A., Brady, M., Breitenstein, C., Bruehl, S., Bryant, L., Cherney, L., Conroy, P., Copland, D., Croteau, C., Cruice, M., Dipper, L., Hilari, K., Howe, T., Kelly, H., Kiran, S., Laska, A., Marshall, J., Murray, L., Patterson, J., Quinting, J., Rochon, E., Rose, M., Rubi-Fessen, I., Sage, K., Simmons-Mackie, N., Visch-Brink, E., Volkmer, A., Webster, J., Whitworth, A., & Le Dorze, G. (2021). *Measuring communication as a core outcome in aphasia trials: Results of the ROMA-2 international core outcome set development meeting.* [Manuscript in preparation]. School of Health and Rehabilitation Sciences, The University of Queensland, Australia.

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2016). Core Outcomes in Aphasia Treatment Research: An e-Delphi Consensus Study of International Aphasia Researchers. *Am J Speech Lang Pathol*, *25*(4S), S729-S742. https://doi.org/10.1044/2016_AJSLP-15-0150

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2017a). Which treatment outcomes are most important to aphasia clinicians and managers? An international e-Delphi consensus study. *Aphasiology*, *31*(6), 643-673. https://doi.org/10.1080/02687038.2016.1186265

Wallace, S. J., Worrall, L., Rose, T., & Le Dorze, G. (2019). Using the International Classification of Functioning, Disability, and Health to identify outcome domains for a core outcome set for aphasia: a comparison of stakeholder perspectives. *Disabil Rehabil*, *41*(5), 564-573. https://doi.org/10.1080/09638288.2017.1400593

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., Babbitt, E., Bose, A., Brady, M., Cherney, L. R., Copland, D., Cruice, M., Enderby, P., Hersh, D., Howe, T., Kelly, H., Kiran, S., Laska, A. C., Marshall, J., Nicholas, M., Patterson, J., Pearl, G., Rochon, E., Rose, M., Sage, K., Small, S., & Webster, J. (2018). A core outcome set for aphasia treatment research: The ROMA consensus statement. *Int J Stroke*, 1747493018806200. https://doi.org/10.1177/1747493018806200

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Cruice, M., Isaksen, J., Kong, A. P. H., Simmons-Mackie, N., Scarinci, N., & Gauvreau, C. A. (2017b). Which outcomes are most important to people with aphasia and their families? an international nominal group technique study framed within the ICF. *Disabil Rehabil*, *39*(14), 1364-1379. https://doi.org/10.1080/09638288.2016.1194899

Ware, J. E., Jr. (1992). Measures for a new era of health assessments. In A. L. Stewart & J. E. Ware (Eds.), *Measuring Functioning and Well-being: The Medical Outcomes Study Approach.* (pp. 3-11). Duke University Press.

Wenke, R., Cardell, E., Lawrie, M., & Gunning, D. (2018). Communication and well-being outcomes of a hybrid service delivery model of intensive impairment-based treatment for aphasia in the hospital setting: a pilot study. *Disabil Rehabil*, *40*(13), 1532-1541. https://doi.org/10.1080/09638288.2017.1300949

Whitehurst, D. G. T., Latimer, N. R., Kagan, A., Palmer, R., Simmons-Mackie, N., Victor, J. C., & Hoch, J. S. (2018). Developing Accessible, Pictorial Versions of Health-Related Quality-of-Life Instruments Suitable for Economic Evaluation: A Report of Preliminary Studies Conducted in Canada and the United Kingdom. *Pharmacoecon Open*, *2*(3), 225-231. https://doi.org/10.1007/s41669-018-0083-2

World Health Organisation/WHO (2001). *International Classification of Functioning, Disability and Health.* (https://www.who.int/standards/classifications/international-classification-of-functioning- disability-and-health

Wiley, R. W., & Rapp, B. (2019). Statistical analysis in Small-N Designs: using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology*, *33*(1), 1-30. https://doi.org/10.1080/02687038.2018.1454884

Willmes, K. (1985). An approach to analyzing a single subject's scores obtained in a standardized test with application to the Aachen Aphasia Test (AAT). *J Clin Exp Neuropsychol*, *7*(4), 331-352. https://doi.org/10.1080/01688638508401268

Winstein, C. J., Stein, J., Arena, R., Bates, B., Cherney, L. R., Cramer, S. C., Deruyter, F., Eng, J. J., Fisher, B., Harvey, R. L., Lang, C. E., MacKay-Lyons, M., Ottenbacher, K. J., Pugh, S., Reeves, M. J., Richards, L. G., Stiers, W., Zorowitz, R. D., American Heart Association Stroke Council, C. o. C., Stroke Nursing, C. o. C. C., Council on Quality of, C., & Outcomes, R. (2016). Guidelines for Adult Stroke Rehabilitation and Recovery: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke*, *47*(6), e98-e169. https://doi.org/10.1161/STR.0000000000000098

Worrall L., Y. E. (2000). Effectiveness of functional communication therapy by volunteers for people with aphasia following stroke. *Aphasiology*, *14*, 911-924.

Zeppieri Jr., G., & George, S. Z. (2017). Patient-defined desired outcome, success criteria, and expectation in outpatient physical therapy: a longitudinal assessment. *Health and Quality of Life Outcomes*, *15*(29), 1-10.