PATIENT REPORTED OUTCOME MEASURES IN STUTTERING INTERVENTION: A SYSTEMATIC REVIEW

ADILA SHARIN R J
Register No.: P01II24S123002

A Dissertation is submitted as a part fulfilment for the degree of Master of

Science (Speech-Language Pathology)

University of Mysore,

Mysuru

ALL INDIA INSTITUTE OF SPEECH AND HEARING MANASAGANGOTHRI,

MYSURU-570006

June, 2026

CERTIFICATE

This is to certify that this dissertation entitled **"Patient-reported Outcome Measures in Stuttering Intervention: A Systematic Review"** is a bonafide work submitted as part of fulfilment of the degree of Master of Science (Speech Language Pathology) of the student with Registration Number P01II24S123002. This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any other university for the award of any other Diploma or Degree.

Mysuru
8th June, 2026

M. Pushpavathi

Dr. M. Pushpavathi

Director
All India Institute of Speech and Hearing
Manasagangothri, Mysuru-570006

CERTIFICATE

This is to certify that this dissertation entitled "Patient Reported Outcome Measures in Stuttering Intervention: A Systematic Review" has been prepared under my supervision and guidance. It is also being certified that this dissertation has not been submitted earlier to any other university for the award of any other Diploma or Degree.

Mysuru
June, 2026

Dr. Divya Seth
Guide
Assistant Professor in Speech Pathology
All India Institute of Speech and Hearing
Manasagangothri
Mysuru-570006

DECLARATION

This is to certify that this dissertation entitled "Patient Reported Outcome Measures in Stuttering Intervention: A Systematic Review" is the result of my own study under the guidance of Dr. Divya Seth, Assistant Professor of Speech Language Pathology, All India Institute of Speech and Hearing, Mysuru and has not been submitted earlier to any other university for the award of any other Diploma or Degree.

Mysuru
June, 2026

Registration no.
P01II24S123002

Acknowledgement

Alhamdulillahh*…Writing this dissertation has been an incredible journey, and I certainly didn't reach the finish line alone. There are so many wonderful people who have supported me along the way.*

First and foremost, my deepest gratitude goes to my guide, Dr. Divya Seth. Thank you ma'am for your unwavering guidance from day one. Your mentorship, insights, and continuous support shaped not only this dissertation but also my growth throughout this entire process. I am incredibly lucky to have had you as my guide.

My sincere gratitude extends to Dr. M. Pushpavathi, Director, AIISH, Mysuru, for providing the facilities and academic environment necessary to carry out this research.

I am also incredibly grateful to Dr. Santhosh sir, Dr. Mahesh sir, Ms. Audri ma'am, Ms. Vasupradha ma'am, and Ms.Sneha ma'am. Thank you all for stepping in and helping me so much, especially with the crucial keyword validations. Your time and insights made a world of difference to my work.

Most importantly, my deepest gratitude goes to my true foundation. To my "Ummachi" and "Uppachi", thank you for your endless love, your countless sacrifices, and all the prayers that made this entire milestone possible. This achievement is as much yours as it is mine. Thank you "Umma" and 'Uppa" for your constant support during this journey. I also want to take a moment to thank Rasal, Milu, ponnu. Thank you for always believing in me, keeping my spirits high, and being such a constant source of support throughout this journey.

To my dissertation partners, Akki and Neba, thank you for sharing this journey with me through every deadline, challenge, confusion, and achievement. Having both of you by my side made the dissertation process far more manageable and memorable.

To my charming girls—Neeraja, Agraja, Akki the makki, Paruu, and Nihalaa—who have shared my days, my classrooms, and our hostel corridors for the past seven years. We have grown together through the shifting seasons of this long journey. You have been my anchors when the academic tides were high, and my sanctuary at the end of exhausting days. For the echoes of our shared laughter, endless funny reels, the late-night conversations that chased the stress away, and for constantly keeping my feet on the ground—thank you for being the most beautiful support system a student could ever hope for.

A very special thank you goes to Shyam and Vismaya chechi. Thank you for always being there to cheer me on, and for so generously providing your notes and guidance at every single step of this journey.

I would also like to extend a special thank you to Swalih sir. Your guidance and support during my entrance preparations laid the very foundation for this journey, and I am so grateful for the role you played in getting me here.

I would also like to thank Vaishnavi, Aswin, Amal (the Kannan maman), 'Mad heads' for making this journey lighter, happier, and more memorable. The daily chaos, laughter, food breaks, and shared experiences became some of the best parts of postgraduate life.

My journey would be incomplete without thanking the wonderful people of Section A and Section B; thank you for your radiant positivity and the unforgettable memories we crafted together. Our shared struggles, vibrant celebrations, and endless conversations became the light that made every heavy challenge so much easier to bear.

I also reserve a very special place in my heart for our entire batch, the "Celestials," and our beloved "ECHO." Thank you for writing one of the most remarkable chapters of my life alongside me. The friendships we forged and the moments we lived will forever remain safely tucked in the corners of my heart.

To my dear juniors, especially the "Fresh Makkals,", "*DAPANS", Arshad, Ihsan, Fidha, Abdu, Nandhu, Shahul, Nihala, Anly, Azra, Nandhana, Lavanya and everyone else. Thank you for bringing so much fun, laughter, and cheerful energy into my life during these years.*

I also sincerely thank all my seniors, juniors, faculty members, classmates, friends, and bond staff who, in one way or another, contributed to making this journey smoother, more meaningful, and memorable.

My deepest gratitude belongs to my husband, Ajmal, who has been my absolute rock throughout this journey. Thank you for practically holding our world together and seamlessly taking care of everything behind the scenes so I could focus on my research. Whether it was sacrificing your own sleep to wait up patiently during my endless late-night writing marathons, or offering constant, unwavering encouragement when the dissertation stress peaked, you were my ultimate support system. You were the quiet strength behind this entire project, and I truly could not have crossed this finish line without you. This milestone is as much yours as it is mine.

Before I close this chapter, I want to pause and express a quiet word of gratitude to myself. Thank you for standing strong through the storms, for holding fast to my dreams when the shadows of self-doubt crept in, and for pressing on through the quiet, sleepless nights, the looming deadlines, and every mountain climbed.

Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

According to Yairi and Ambrose (2013), stuttering is a complicated neurodevelopmental communication issue that typically first manifests in early infancy and persists into adolescence and adulthood for a few. Although obvious speech disruptions like blocks, prolongations, and repetitions are frequently used to identify stuttering, its effects extend well beyond speech fluency. According to Bloodstein and Ratner (2008) and Yaruss and Quesal (2004), these consequences include emotional distress, avoidance behaviours, negative communication attitudes, and decreased engagement in social, professional, and academic contexts. Additionally, stuttering triggers improper reactions from others, including disapproval from fluent individuals, or negative internal responses to talking, like the feelings an individual experiences about their stuttering (ASHA, 2025). Therefore, it not only influences the speech fluency but also interferes with an individual's daily activities, social participation, and emotional experiences; hence, affecting several domains.

As the World Health Organization's International Classification of Functioning, Disability, and Health (WHO ICF, 2001) has grown in prominence, this guiding framework has been adopted by the field of fluency disorders, especially concerning stuttering (Eadie, 2003). It is suggested that the clinicians align their assessment and management practices for stuttering alongside ICF. The above framework allows measurement of several dimensions: (a) impairment-level indicators, such as stuttering frequency and severity; (b) activity and participation measures, which reflect how successfully and comfortably individuals communicate in real-world settings; and (c) personal and environmental factors, such as attitudes, anxiety, avoidance behaviours, and the availability of social support (Mokkink et al., 2018, 2024; Yaruss, 2010).

It is crucial to assess all components or domains that are impacted by stuttering, to ensure a comprehensive evaluation and effective intervention. Such an evaluation helps the clinicians to demonstrate the effectiveness of therapy, compare the impact of different interventions, and involve clients and families in shared decision-making. Most importantly, following an ICF perspective while planning and delivering intervention, which would include measuring outcomes across domains, ensures benefits that extend beyond speech fluency improving overall communication, social participation, and quality of life for people who stutter. (Mokkink et al., 2018, 2024).

Outcome measures are crucial instruments for assessing therapy efficacy and directing clinical judgement. These measurements are generally separated into two categories: Patient-Reported Outcome Measures (PROMs) and the Clinician-Reported Outcome Measures (CROMs). PROMs capture a person's subjective experience of their condition and its influence on their everyday activities, which evaluate constructs like perceived severity and impact, communication attitudes, participation limitations, treatment satisfaction, and health-influenced quality of life—which are all inherently subjective patient experiences (Churruca et al., 2021). On the other hand, CROMs are observations of a clinician about the patients' behaviours and have been used conventionally to measure treatment outcomes in intervention trials. In the recent years, PROMs have gained popularity in the healthcare industry as a result of the increased focus on person-centred treatment. They offer a methodical way to gather crucial evidence that might otherwise be difficult to gather formally. De Riesthal and Ross (2015) state that "the patient is the exclusive source of data for certain therapeutic effects" and emphasise the need of actively gathering the data that reflects the experiences and personal preferences of the clients. They also highlighted that PROMs provide a systematic and measurable method to monitor the critical evidence-based practice (EBP) elements.

There are several essential differences that exist between PROMs and CROMs. PROMs can capture the personal experiences such as emotions, avoidance behaviours and day-to-day communication, unlike CROMs that focus primarily on measuring speech behaviours. As a result, PROMs are better aligned with the real concerns of stutterers and their everyday situations (Koedoot et al., 2011; Yaruss, 2010). Further, PROMs are generally more accessible and cost-effective. Patients or parents can complete the short rating scales for measuring the severity at home, online, and on a regular basis. In fact, recent research findings indicate that PROMs are quite advantageous for large-scale studies because self-reported assessments often correlate well with the ratings given by the clinician or other professionals (Horton et al., 2024). Additionally, the measurement properties of the PROMs can vary, as some PROM tools have excellent or strong psychometric properties like validity and reliability, while others have poor measures such as floor and ceiling effects, and reduced sensitivity for specific groups. Similarly, in a few intervention trials in the field of speech language pathology, it was noted that some specific PROMs become less consistent or unresponsive over time (Koedoot et al., 2011). Lastly, PROMs are reported to be sensitive to contextual and personal variables, such as mood, expectations, and social context which enables the clinicians to record the changes in lived experiences of the individual that the CROMs failed to capture. Nevertheless, PROMs still require systematic and strong evaluation and standardisation according to the guidelines provided by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) before being served as a primary outcome measure (Mokkink et al., 2018, 2024).

In accordance with Yaruss (2010), it is essential to measure quality of life (QoL) as a patient-centered outcome when researching stuttering treatments. Emphasizing the limitations of traditional fluency tests, Yaruss reasoned that PROMs are the best approach to record the valid, real-world influences of stuttering on a patient's emotional health, social participation,

and communication confidence. He also explained about validated instruments like the Overall Assessment of the Speaker's Experience of Stuttering (OASES), which offers a methodical approach to evaluate the functional and personal outcomes of stuttering in a variety of contexts. Based on this study, using PROMs in stuttering research ensures that treatment results show actual gains in quality of life rather than only lower stuttering rate. This perspective closely aligns with modern patient-centered healthcare frameworks that rely heavily on lived experiences of the individual to assess treatment outcomes (Yaruss, 2010).

In stuttering intervention, PROMs capture features of stuttering that are not visible during a brief clinical examination, but are crucial to stuttering research and treatment. According to Sheehan's "stuttering as an iceberg analogy", in addition to the overt features of speech disfluencies and associated physical struggle, there are several covert features of stuttering (feelings & emotions) that constitute a major proportion of the problem. These are often difficult to assess, particularly through CROMs. In contrast, PROMs provide the finest insight into these hidden elements as well as the emotional impact that stuttering takes. Researchers and clinicians can more accurately assess whether therapy significantly improves daily communication and general quality of life by taking these viewpoints into account. Because of this, self-reports are a valuable addition to and, in many situations, a more accurate representation of what actually changes for a stutterer during intervention (Guntupalli et al., 2006). Some of the patient-reported measures are condition-specific tools (like OASES, WASSP, UTBAS-6, KiddyCAT, and Self-Stigma of Stuttering Scale) and proxy reports (like Palin Parent Rating Scales) by parents/caregivers (Morris et al., 2021).

Since traditional assessments primarily focused on reducing observable speech behaviours like repetitions and prolongations, Guntupalli et al. (2006) emphasized the importance of self-report measures to determine the true efficacy of a stuttering intervention. The authors aimed to reframe stuttering as a multi-faceted syndrome driven by experiential and

covert elements. Through a narrative review of studies investigating stuttering treatment and neuroimaging data, they critically evaluated the standard assessment procedures. They reported that while the conventional fluency-shaping intervention reduce the observable stuttering behaviours in clinical environment, generalization of the same and fading of treatment benefits over time are often a challenge. Additionally, measuring only overt behaviors may ignore the internal struggles, emotions, excessive physical effort, and avoidance experienced by the individual. The study concluded that self-report measures should be the primary outcomes assessing long-term treatment effectiveness, along with measurement of overt speech behaviours as secondary outcomes.

The accuracy of self-reported stuttering severity was also examined by Horton et al. (2024) as a possible outcome measure for extensive studies. Participants with self-reported stuttering of a wide age range from 5 to 84 years with 195 participants were included in the study, and their evaluations were contrasted with those of clinicians. The validity of self-reported severity scores was supported by the findings, which revealed a strong correlation between participant and clinician ratings. Crucially, the study showed that stutterers can accurately rate the severity of their stuttering using straightforward rating scales during clinical evaluations. The results thus support the use of PROMs together with clinician-reported outcomes in clinical practice and research.

The use of PROMs has not been restricted to only adults who stutter, but has been extended to treatment studies for children with stuttering. In case of young children, who may not be able to self-report the perceived outcomes of a treatment, the parents provide the information as proxy-reported outcome measures. According O'Brian et al. (2018), both the parent-reported stuttering severity ratings (SRs) and the conventional percentage of syllables stuttered (%SS) yielded significant effect sizes in a randomised controlled trial of early preschool stuttering intervention. Although the correlation between the two measures was only

moderate, the correlation between the two measures improved as stuttering frequency and severity declined. The study found no statistical reason to favour %SS over parent-reported SRs, noting that parent metrics are both highly reliable and potentially more practical for clinical trials involving young children. Similar findings were reported by Onslow et al. (2018) wherein no significant difference was seen between %SS and parent severity ratings. They reiterated that parent-reported severity rating can be served as a robust, valid and highly useful primary outcome measure for early childhood stuttering research.

Need for the study

Understanding a patient's outlook regarding their treatment and its impact on overall quality of life is an essential component of both person-centred care and evidence-based practice. PROMs provide the most direct and reliable means of assessing how patients themselves perceive their health, functioning, and impact of intervention. Use of PROMs also aligns with the WHO ICF framework and ensures a holistic assessment. A systematic evaluation and synthesis of the available PROMs is essential to guide clinical practice. Over the last decade, PROMs have gained popularity even in the field of speech-language pathology. These include PROMs in dysphagia (Patel et al., 2017, Moloney et al., 2023, Manduchi et al., 2024), speech sound disorder (Harding et al., 2024, Kearney et al., 2015), cleft lip and palate (Salinero et al., 2025, Nolan et al., 2025, Bennett et al., 2018, Ranganathan et al., 2015), dysphonia (Jeyaraman et al., 2025, Demirci et al., 2024, Bohlender(2023), Francis et al., 2017), motor speech disorders (Young et al., 2023, Carlozzi et al., 2018, Carlozzi et al 2017), cognitive communication disorders (Cohen et al., 2025, 2021), and aphasia (Dunne et al., 2023, Quique et al., 2023, Armour., 2019). However, to our knowledge, no studies to date have systematically reviewed and analyzed the patient-reported outcome measures specifically available for stuttering.

Stuttering is a complex communication disorder that impacts a person's emotional, social, and psychological health in addition to their speech fluency. Although traditional clinician-reported measures capture the observable aspects of stuttering, the literature review emphasises that these measures fall short in capturing the lived experiences of stutterers (Yaruss, 2010; Manning, 2010). Contributing insights in to quality of life, self-perceived severity, and involvement in daily communication, Patient-Reported Outcome Measures (PROMs) help close this gap (Koedoot et al., 2011; Horton et al., 2024; Van Ewijk et al., 2025).

A scoping review by Gardner et al. in 2024 evaluated the content validity of the PROMs for adults in speech-language pathology. They identified eight PROMs for adults who stutter. These included the following measures - Overall Assessment of the Speaker's Experience of Stuttering (OASES; Yaruss & Quesal, 2010), Self-stigma of Stuttering Scale (4S; Boyle, 2015), Satisfaction with Communication in Everyday Speaking Situations (SCESS; Karimi et al., 2018), Speech Situation Checklist-Revised (SSC-R; Vanryckeghem et al., 2017), Subjective Screening of Stuttering Severity (SSS; Riley et al., 2004), Inventory of Communication Attitudes (Watson et al., 1987), Self-Efficacy Scale for Adult Stutterers (SESAS; Ornstein & Manning, 1985) and BigCAT (Vanryckeghem & Brutten, 2011). Key constructs measured by these tools included confidence/self-stigma, communication experiences and attitudes associated with stuttering. Among the eight PROMs, OASES and 4S were found to meet most criteria for quality assessment as outlined in Francis et al. (2016) checklist.

Another recent review investigated the content validity of PROMs used for adults with communication disorders (Ewijk et al., 2025). PROMs specific to stuttering identified in this review included - Overall Assessment of the Speaker's Experience of Stuttering (OASES; Yaruss et al., 2006), Assessment of Language Use in Social Contexts for Adults (ALUSCA; Valente et al., 2019), Satisfaction with Communication in Everyday Speaking Situations

(SCESS; Karimi et al., 2018), and Stuttering Generalization Self-Measure (SGSM; Alameer et al., 2017). The review pointed out that although tools such as the OASES are still frequently used, more recent PROMs are appearing that have robust frameworks for examining the psychosocial effects of stuttering and communication participation. Further, the review also pointed out the existing gaps in regard to lack of appropriate PROMs across age groups and suitable to various cultural environments.

Overall, there is a shift in the trend for measuring treatment outcomes in stuttering intervention. Unlike the previous decades where CROMs served as the primary, and often the only measures to gauge treatment effectiveness, PROMs have become popular in the recent years. While PROMs for adults are relatively well-established, tools for children who stutter lack rigorous evaluation (Yaruss, 2010; Horton et al., 2024). The necessity for a systematic appraisal of these measures is further amplified by challenges related to their psychometric validation, cultural adaptation, and translation (Koedoot et al., 2011; Van Ewijk et al., 2025). Consequently, this study systematically reviews the PROMs currently utilized in both child and adult stuttering interventions. By rigorously evaluating these measures, this work will promote the use of culturally and developmentally appropriate tools, yield evidence-based clinical recommendations, and ultimately elevate the quality of stuttering intervention research.

PICO- based Research Question

What is the quality of patient-reported outcome measures (O) used to evaluate the stuttering intervention (I) and its efficacy among adults and children who stutter (P)?

Aim of the study

The study aimed to identify and systematically review the PROMs, designed specifically to assess the treatment outcomes in children and adults who stutter.

Objectives of the study

The specific objectives of the study were to identify and critically appraise the:

a. proxy-reported outcome measures or patient-reported outcome measures for children who stutter.
b. Patient-reported outcome measures for adults who stutter.

CHAPTER 2

METHOD

The study aimed to collect, compile, extract, and analyse the currently available literature for the patient-reported outcome measures (PROMs) in stuttering. The study followed the procedures of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis – COnsensus-based Standards for the selection of health Measurement INstruments for Outcome Measurement Instruments (PRISMA-COSMIN for OMIs, 2024) guidelines given by Mokkink et al. (2024) to answer the research question "What is the quality of patient-reported outcome measures (O) used to evaluate the stuttering intervention (I) and its efficacy among adults and children who stutter (P)?"

Search strategy

The prime objective of the search strategy was to obtain all the relevant published literature on measurement properties of PROMs in stuttering intervention. This was implemented through the following stages:

Generation and validation of keywords for searching the databases

The researchers identified the keywords from the research question and generated a list of 25 words initially that were synonyms, derivations, definitions or components of keywords. Following are a few examples of the keywords: "PROM's", "stuttering", "children", "preschoolers", "school-going children", "children who stutter", "adolescents and adults who stutter", **"quality of life", "functional communication", "self-perception", "**psychosocial impact of stuttering". These keywords were validated using Delphi method wherein the list was shared with five expert speech-language pathologists (the Delphi panel). The panel members were qualified speech-language pathologists (SLPs) who were nationally

accredited, and actively involved in the clinical- and research-based assessment and intervention of persons with fluency disorders.

The investigator explained the research question, aim, and objectives of the study individually to each of the member of the Delphi panel. Following this, they were asked to rate each generated keyword independently as either 'appropriate' or 'inappropriate'. They were also requested to provide remarks or suggestions, if any. Further, there was a provision for the addition of any new keywords to the given list by the members of the Delphi panel. Search words that met 90% agreement as 'relevant' were selected for the search syntax. Other words, those that did not meet the 90% agreement criterion were re-circulated among the panel members and investigators for review and re-validation. The final search string comprised of 17 words that were approved by 90% of the Delphi panel members as 'appropriate' during their independent reviews, after resolving the queries/concerns related to specific keywords and eliminating the overlapping terminologies.

Generation and Validation of Search Syntax

The 17 keywords finalised from the previous phase were incorporated into the format of search syntax using the Boolean search operators ('OR' & 'AND'). The search syntax was validated by the investigators using the probe articles. The list of articles which were retrieved was cross-referenced against the probe articles. However, it was noted that the inclusion of all 17 validated keywords failed to capture those probe articles. Hence, some of the keywords were refined by the investigators to ensure that all probe articles were retrieved using the search syntax. Also, to minimise the identification of irrelevant articles, the 'NOT' operator was introduced in the syntax. The final search syntax generated is given below:

("Reliability" OR "Psychometric" OR "Measurement" OR "Development" OR "Outcome measures" OR "Self-report measure") AND ("Quality of life" OR "Psychosocial" OR

"Speech" OR "Behaviour" OR "Attitude" OR "Severity") AND ("Stuttering" OR "Stammering" OR "Fluency disorder" OR "Stutter" OR "Communication") NOT ("Autism" OR "ADHD" OR "Mental illness" OR "Stroke" OR "Aphasia" OR "Dysarthria" OR "Motor Delay" OR "Cancer" OR "Review")

Execution of Search Syntax on Scientific Databases

A systematic literature search was conducted by using the finalized search syntax across four major databases - the NCBI-PubMed, Science Direct, ProQuest and CINAHL Ultimate. The reference list of relevant studies or articles were also hand-searched for relevance of inclusion and also to ensure no other eligible studies were overlooked.

Data selection

Rayyan software was utilised to streamline the initial screening and selection process. Articles collected from each of the electronic databases were exported to files in the form of either .CSV, Web of Science/CIW, BibTeX, or PubMed XML and compiled together in Rayyan QCRI (Ouzzani et al., 2016). Rayyan is a web-based free software platform and a user-friendly tool which is designed to conduct systematic reviews, scoping reviews, and other knowledge syntheses. It helps researchers to speed up screening and organising the relevant studies for a systematic analysis.

Regardless of the study design, all types of studies investigating the measurement properties of PROMs in stuttering intervention in children and adults were included in the review. A two-stage selection process was implemented to narrow down to the final set of studies. The study selection was done based on a pre-defined inclusion and exclusion criteria (see Table 1). The entire retrieved list of studies from each source was imported into the Rayyan software and duplicate records were eliminated. The remaining articles were retained for further data identification procedures as detailed below.

Table 2.1

Inclusion and exclusion criteria for study selection

Inclusion Criteria	Exclusion criteria
1. Study included human participants including children and adults and any gender with a primary, confirmed diagnosis of stuttering. 2. Studies that explain the development, evaluation, or measurement of the psychometric properties of Patient-Reported Outcome Measures (PROMs) used in stuttering interventions. 3. Primary research studies published in peer-reviewed academic journals. 4. Full text report should be available in English.	1. Study included participants presenting with stuttering alongside comorbid speech, language, or cognitive disorders. 2. Studies that do not focus on PROMs or fail to name or define the specific standardized outcome measure used. 3. Single-subject experimental designs. 4. Review articles, systematic reviews on different topics, practice guidelines, policy papers, non-peer-reviewed literature and conference papers. 5. Articles that are reported in languages other than English.

Data Identification

Level 1 and 2 (Title and abstract level screening)

Both the investigators of the current research were involved in Levels 1 and 2 screening. Each of them independently reviewed the titles and abstracts of the downloaded list of articles using the Rayyan platform. Screening was performed in view of the research

question, aim of the study, and the inclusion-exclusion criteria. The articles were categorized as 'included', 'excluded', or 'may be'. The titles that did not explain the PROMs were excluded. Further, the titles that described the efficacy/effectiveness of an intervention technique by using a PROM were not included for further screening. Upon the completion of Level 1 and 2 screening, independently by the two investigators, conflict resolution was performed. In this process, the articles that were categorized under the 'maybe' label, were further re-visited by the two investigators along with a third reviewer (subject expert), to make a decision (include/exclude).

Level 3 (Full-text Screening)

The studies that passed the levels 1 and 2 screening were included for full-text content screening. Here, the studies that met the inclusion criteria mentioned above were further reviewed for full-length screening to answer the aim and objectives of the present study.

Data Extraction

The investigator developed a data extraction template to optimize the process of evaluation of studies. The information included in the template were - author with year of study, tool developed, country/region where the study was conducted, type of study, sample population, no. of items in the scale including the subscale, response system, content validity, structural validity, internal consistency, other measurement properties and quality of the study of each PROM. For the quality assessment, the reviewer evaluated the methodological quality and the measurement properties of each study. Methodological weakness that potentially affected the evaluation of content validity for studies were also noted.

Data Analyses

Risk of Bias

The risk of bias for the included studies were conducted using the COSMIN Risk of Bias checklist (Mokkink et al., 2018). It is a checklist developed specifically to assess the methodological quality of studies assessing measurement properties of PROMs. It consists of 10 domains with a total of 116 items evaluating aspects such as PROM development, content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity and responsiveness. Each item on the checklist is assessed individually using a 5-point rating ranging from “very good” to “not applicable”.

The assessment and synthesis of the final tools followed the COSMIN 2024 guidelines that was executed in two phases:

Phase 1: In this phase, the investigator focused on understanding the *development of tool/instrument and its content validity*. First, the reviewer assessed the methodological quality of each study using the COSMIN Box 1 (Mokkink et al., 2024), which lists parameters to evaluate the relevance and comprehensiveness of the item, and pilot study conducted. Following this, the reviewer assessed the content validity of the PROM using COSMIN Box 2, which considers parameters such as the relevance, comprehensibility, and comprehensiveness of the PROM from clinician and patient perspectives, along with the integrated scores reflecting the overall relevance, comprehensibility and content validity.

Phase 2: This phase involved assessing remaining measurement properties and assigning the final recommendations. Then an overall quality of evidence was graded using the modified GRADE (Grading of Recommendations Assessment, Development and Evaluation)

approach. Lastly, based on COSMIN 2024 each PROM was categorised into 3 recommendation classes - *recommended for use, not recommended for use* and *no recommendation.*

CHAPTER 3

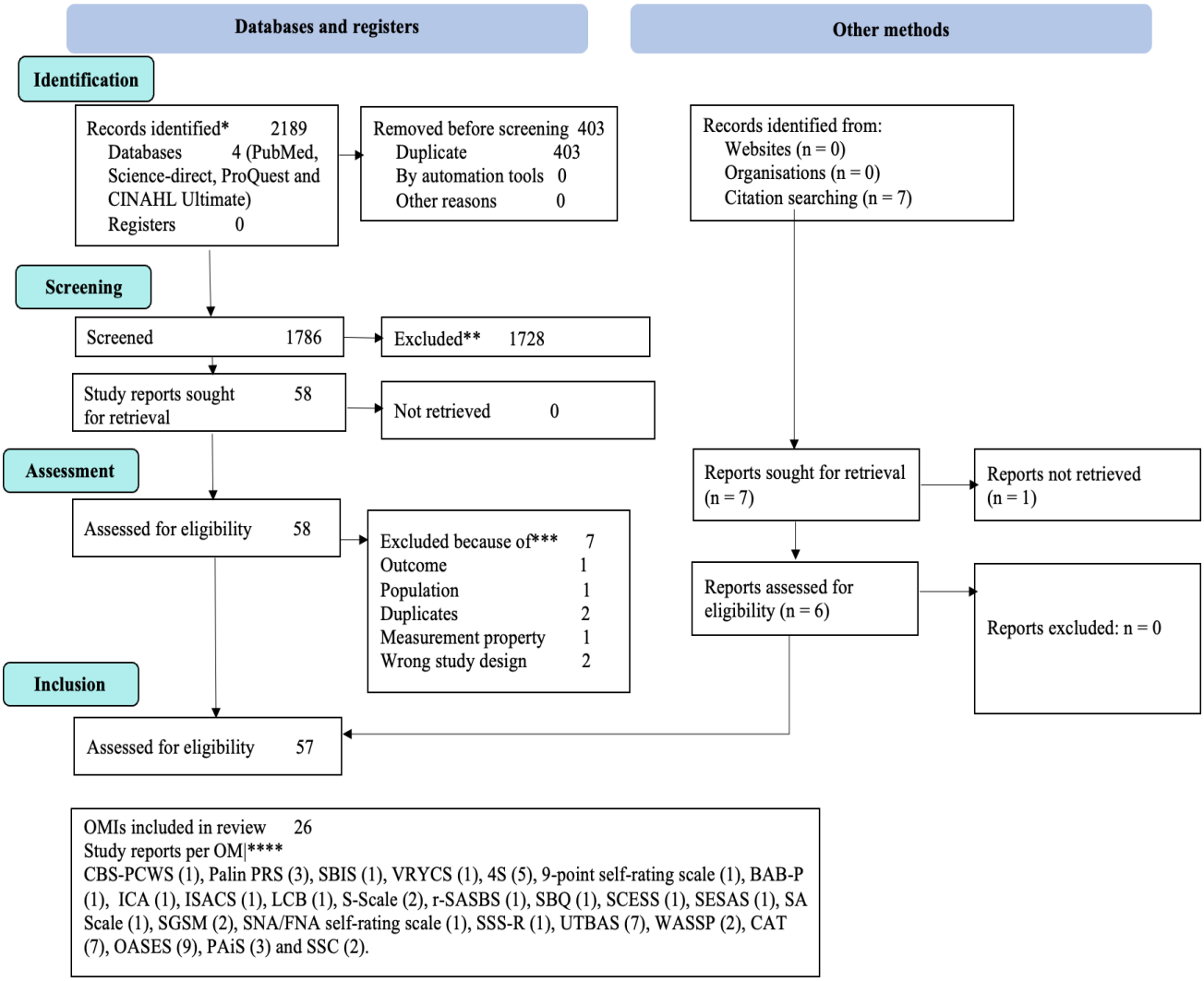
RESULTS

Result of Search Strategy

Through a thorough systematic search from the database including NCBI-PubMed, Science Direct, ProQuest and CINAHL Ultimate, a total of 2189 articles were identified. After removing 403 duplicates, 1786 articles were selected for Level 1 and 2 screening i.e., abstract and title level of screening. Based on the inclusion and exclusion criteria which were pre-defined, 1728 studies were excluded from the review and following these, 57 full-length articles were included for level 3 screening. During the full-length article screening, 6 articles were excluded. However, through the backward reference searching i.e., manual search of the reference lists provided in each study, 6 more relevant articles were identified. Hence, the review was done for a total of 57 articles that resulted in identification of 25 PROMs (See figure 3.1). These included both the patient reported as well as proxy/parent-reported measures (in case of CWS) to answer the generated question for the study: “What is the quality of patient-reported outcome measures (O) used to evaluate the stuttering intervention (I) and its efficacy among adults and children who stutter (P)?”

Figure 3.1

PRISMA-COSMIN for OMI (2024) flowchart



Note. *Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Study Characteristics

The 57 studies (Supplementary material S1 and S3) included in the review range from studies explaining regarding the development, validation, adaptation, or any psychometric evaluation of the outcome measures. Of these, 5 studies focused on development, 19 assessed

validity, reliability, or other psychometric properties, 14 explained the development and validation or psychometric evaluation, and 19 of them were reports of cross-cultural translation, adaptation, and validation. A total of 25 different instruments were identified based on the included studies (Table 3.1).

Table 3.1

List of different PROMs identified in the review

Sl. No.	**Patient Reported Outcome Measures (PROMs)**
1	Caregiver Burden Scale for Parents of Children with Stuttering (CBS-PCWS; Mehdizadeh Behtash et al., 2022) *
2	Palin PRS (PPRS; Millard & Davis, 2016) *
3	Short Behaviour Inhibition Scale (SBIS;Ntourou et al., 2020)*
4	Vanderbilt Responses to Your Child's Speech (VRYCS; Singer et al., 2022) *
5	Self-Stigma of Stuttering (4S) *
6	9-point self-rating scale (O'Brian et al., 2004)
7	Persian version of Behavioral Assessment Battery (BAB)
8	Inventory of Communication Attitudes (ICA) (Watson et al., 1987)
9	Impact Scale for Assessment of Cluttering and Stuttering (ISACS; Kelkar et al., 2024) *
10	Locus of Control of Behaviours (LCB) Scale (Craig et al., 1984)
11	S-Scale (Erickson, 1969)
12	revised-Scale of Avoidance and Struggle Behaviours in Stuttering (r-SASBS; Hancer & Tokgoz-Yilmaz, 2025) *
13	Safety Behaviours Questionnaire (SBQ; Azarinfar et al., 2022) *
14	Satisfaction with Communication in Everyday Speaking Situations Scale (SCESS; Karimi et al., 2018) *
15	Self-efficacy scale for Adults who Stutter SESAS (Ornstein & Manning, 1985)
16	Self -Assessment Scale (SA Scale)

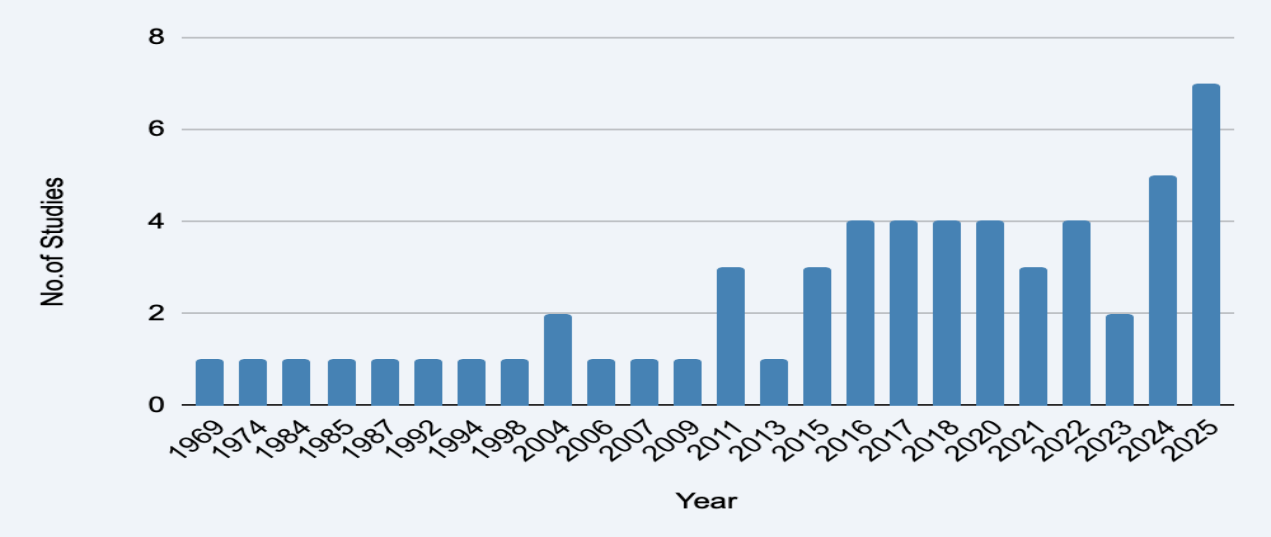
17	Stuttering Generalization Self-Measure (SGSM; Alameer et al., 2017) *
18	Speech Naturalness (SNA) and Feel Naturalness (FNA) self-rating scales (Finn & Ingham, 1994)
19	Subjective Stuttering Severity-Research Edition (SSS; Riley et al., 2004)
20	Unhelpful Thoughts and Beliefs About Stuttering (UTBAS;St Clare et al., 2009)*
21	The Wright and Ayre Stuttering Self-rating Profile (WASSP;Wright et al., 1998)*
22	Communication Attitude Test (CAT; Vanryckeghem & Brutten, 1992) *
23	Overall Assessment of Speakers Experience of Stuttering (OASES;Yaruss & Quesal, 2006)*
24	Premonitory Awareness in Stuttering (PAiS:Cholin et al., 2016)*
25	SSC-R (Vanryckeghem et al., 2017) *

Note. *Studies that evaluated content validity

The identified articles were published in various journals between 1969 and 2025 (see Figure 3.2). A significant increase was noted in the research for PROMs in the context of stuttering intervention since 2016. This change in trend reflects the increased focus on patient-centred outcomes and wider acceptance of the WHO ICF framework in the field of speech-language pathology, particularly in stuttering. Geographically, majority of the studies were done in the United States (n = 16), followed by Australia (n = 7), Turkey (n = 6), Iran (n = 6), India (n = 5) and also some notable contribution from Netherlands, Japan, Poland, Canada, Germany, and China. The geographical concentration of research in the United States, highlights the limited evidence in the Asian context and need for the same (see Figure 3.3).

Figure 3.2

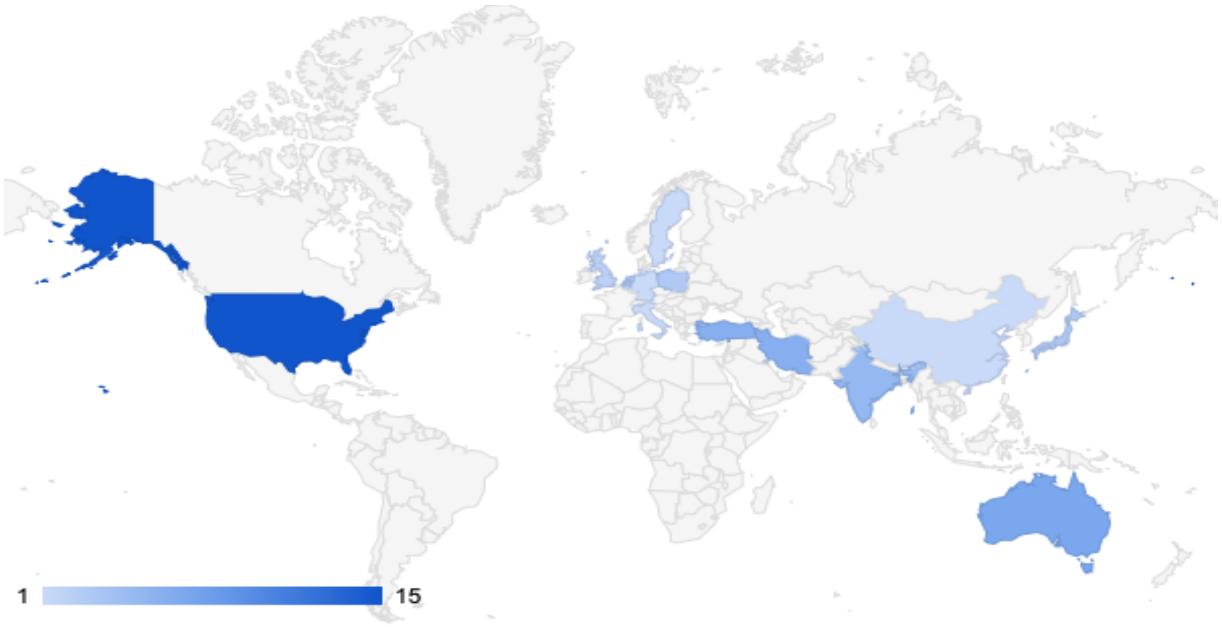
Occurrence of studies by year



Note. The figure represents the number of publications through the observed period. Each bar represents the total number of occurrences recorded in the respective year. Data reflect observation from 1969 through 2025, with a notable increase in frequency during recent years

Figure 3.3

Geographical distribution



Note. The map displays the geographical distribution of publications across countries. Colour intensity represents the frequency of occurrences, ranging from 1 (highest) to 15 (lowest). Countries with no recorded occurrences are shown in light grey. The United States recorded the highest number of occurrences.

With respect to the target population, the majority of the studies i.e., 36 out of 57 focused on adults who stutter (AWS). Further, 10 studies focused on children who stutter (CWS) and 11 included both the age groups, or targeted specifically the parents/caregivers of CWS (Supplementary Material Table S1). The proxy-reported outcome measures identified in the review included Palin Parent Rating Scale (Palin PRS;Millard&Davis, 2016), Caregiver Burden Scale for Parents of Children who Stutter (CBS-PCWS;Mehdizadeh Behtash et al., 2022) and the Vanderbilt Responses to Your Child's Speech (VRYCS;Singer et al., 2022) Scale. The sample size varied considerably across the studies, ranging from 10 participants (9-point self-rating scale;O'Brian et al., 2004) in a reliability study to 468 children in one of the most comprehensive validation study(SBIS;Ntourou et al., 2020).

Study Quality and Content Validity

The methodological quality of the included studies and their content validity were analysed according to PRISMA-COSMIN guidelines (Mokkink et al., 2024). The overall content validity of a tool as per the COSMIN guidelines, is rated on 3 parameters – *relevance, comprehensiveness* and *comprehensibility*. It was found that overall evidence for quality of studies was variable, with only a few studies that conducted a systematic evaluation of the content validity of PROM; however, there were methodological caveats that existed. Out of the 25 PROMs identified in the review, evidence for content validity was available only for 19 as explained below.

PROMs for Children who Stutter

The current review identified four PROMs for CWS, that were proxy-reported outcomes (rated by parents/caregivers). Caregiver Burden Scale for Parents of Children with Stuttering (CBS-PCWS; Mehdizadeh Behtash et al., 2022) followed a rigorous two-phase development process involving 15 parents of CWS for qualitative item generation, along with 12 expert SLPs for content validation. The study estimated both the Content Validity Ratio (CVR) and Content Validity Index (CVI), representing high content validity. Another popular proxy-reported measures, the Palin PRS (PPRS; Millard & Davis, 2016) involved an initial Delphi study in which 22 to 32 parents of CWS contributed 129 statements, subsequently refined for importance and consensus. The 2016 study revealed sufficient scores with high evidence (+/H). The Persian adaptation (Barzegar Bafrooei et al., 2025) employed a quantitative CVI and CVR analyses with expert SLPs providing explicit evidence for high content validity.

Content validity for the Short Behaviour Inhibition Scale (SBIS) was established by Ntourou et al. (2020) and was observed to be moderate. The study involved only parents of typically fluent children and no quantitative measures such as CVI were included. Similarly, the fourth proxy-reported outcome, Vanderbilt Responses to Your Child's Speech (VRYCS; Singer et al., 2022) rating scale was noted to have only moderate evidence. The content validation did not involve parents of CWS.

PROMs for Adults who Stutter

The original 4S (Boyle, 2013, 2015) was developed using semi-structured qualitative interviews with 15 AWS. The appraisal revealed sufficient high evidence (+/H) for all the three content validity measures. Similarly, sufficient scores with high evidence was observed for the Turkish adaptation of this tool (Turkish 4S-TR; Tiryaki et al., 2023). The Polish 4S (Wesierska

et al., 2025) underwent a multi-step translation process verified by double experts (SLPs who themselves stutter), receiving sufficient high evidence (+/H) across all domains. In a recent study, the Japanese short form 4S-J-16 (Limura et al., 2022) received sufficient high evidence for relevance and comprehensibility but only sufficient moderate evidence (?/M) for comprehensiveness, illustrating the inherent compromise in content validity that occurs during item reduction. Another cross-cultural study, the Polish adaptation of Behavior Assessment Battery (BAB; Węsierska et al., 2018) also showed sufficient high evidence (+/H) across all parameters of content validity.

A recent addition to the impact assessment has been the Impact Scale for Assessment of Cluttering and Stuttering (ISACS; Kelkar et al., 2024) received sufficient high evidence (+/H) for all three parameters of content validity. Another PROM for adults, the revised-Scale of Avoidance and Struggle Behaviours in Stuttering (r-SASBS; Hancer & Tokgoz-Yilmaz, 2025) was developed based on the compositions written by 15 AWS and supplemented by a literature review. The content was further evaluated by 10 expert reviewers yielding a CVR of 0.96 and CVI of 0.93. The quality assessment of this study revealed moderate level of content validity as the checks for the relevance, comprehensiveness, and comprehensibility were performed only by the experts, but not by the AWS. Similarly, the Safety Behaviours Questionnaire (SBQ; Azarinfar et al., 2022) applied a Delphi method involving 17 expert SLPs and 5 AWS who rated the relevance of items. Overall, the study exhibited moderate level of evidence.

The Satisfaction with Communication in Everyday Speaking Situations Scale (SCESS; Karimi et al., 2018) is a single-item tool for AWS to rate their communication satisfaction. The development and validation involved methodological rigor and content validity was established by 14 expert clinicians and based on a pilot study with 15 AWS. Overall, the tool is rated to have sufficient scores for high evidence. The original tool, Stuttering Generalization Self-

Measure (SGSM; Alameer et al., 2017), did not specify measures of content validity. However, content validity was established for Persian SGSM (Hozeili et al., 2024) by involving ten AWS and ten expert SLPs to rate the clarity, relevance, and comprehensibility of the test items. Both CVR and CVI were calculated for the SLP ratings. Though the AWS were asked to rate the relevance and clarity of test items, there was no scope to add any important relevant situations. Also, the sample size in both groups (AWS & SLPs) was low in the context of COSMIN-OMI guidelines. Based on these factors the content validity of Persian SGSM was rated to be moderate.

The content validity for the original UTBAS (St Clare et al., 2009) and the expanded UTBAS (Iverach et al., 2011) was found to be moderate (+/M) as the development of items was primarily based on a retrospective clinical record audit, rather than direct cognitive interviews with people who stutter or SLPs. Likewise, the brief UTBAS-6 (Iverach et al., 2016) was noted to have sufficient moderate evidence (+/M) for relevance and comprehensibility but indeterminate low evidence (?/L) for comprehensiveness, as the reduction to six items inherently limits the breadth of constructs covered. Cross-cultural adaptations of UTBAS in Turkish (Aydın Uysal & Ege, 2020), Japanese (Chua et al., 2017), and Italian (Bernardini et al., 2024) all lacked formal cognitive interviews with the target population, receiving indeterminate or low evidence ratings for content validity, particularly for comprehensiveness. In contrast, the Persian version (UTBAS-P; Farpour et al., 2025) involved ten AWS to evaluate the face and content validity of the tool. In the process, one culturally irrelevant item was removed, and six new culturally-specific items were added. All three parameters of content validity received sufficient scores of high evidence (+/H).

The original study on development of WASSP (Wright et al., 1998) involved a informal qualitative feedback from subject experts regarding the relevance of items; however, no formal methods were followed, thereby making its content validity low. In a recent investigation, the

Turkish version (WASSP-TR; Uysal & Köse, 2021) provided a substantially more comprehensive cross-cultural adaptation with high content validity (+/H).

PROMs for Children and Adults who Stutter

A widely used tool is the Communication Attitude Test (CAT; Vanryckeghem & Brutten, 1992). Though the tool is available in several languages, content validity is established only in a few contexts. The content validity for the Persian BigCAT (Valinejad et al., 2018) was evaluated by eight Persian-speaking SLPs with CVR > 0.75 indicating sufficient high evidence (+/H) for relevance, comprehensiveness, and comprehensibility. However, it lacked formal cognitive interviews with Persian-speaking AWS, which depicts a gap between expert and patient perspectives. The CAT-Kannada (Veerabhadrappa et al., 2020) and BigCAT-Kannada (Veerabhadrappa et al., 2021) relied on forward-backward translation by native speakers and item-check for cultural relevance by experienced SLPs. However, no qualitative interviews were conducted with children or adults who stutter. Hence, the content validity for the Kannada versions of CAT was rated as sufficient scores of moderate evidence.

One of the most popular tools in stuttering assessment has been the Overall Assessment of Speakers Experience in Stuttering (OASES; Yaruss & Quesal, 2006). The study explaining the development of OASES was found to have sufficient scores with moderate evidence for relevance, comprehensiveness, and comprehensibility. Originally developed for AWS, later versions for teenagers and school-age CWS were also made available. However, these were not available as published journal articles. Several cross-cultural adaptations of OASES are available - Dutch (Koedoot et al., 2011; Lankman et al., 2015), Japanese (Sakai et al., 2017), Swedish (Lindstrom et al., 2020), Polish (Wesierska et al., 2023), Kannada (Mahesh et al., 2024), and simplified version in Chinese (Ma et al., 2025). Though most of these adaptations followed the standard guidelines for reverse translation, none of them conducted any formal

quantitative measures for content validity such as the CVR or CVI. Nevertheless, psychometric properties have been validated as explained in the following section.

The Premonitory Awareness in Stuttering Scale (PAiS; Cholin et al., 2016) was observed to have sufficient scores with moderate evidence (+/M). A major limitation of the study, observed and acknowledged by the authors, was lack of differentiation between immediate and prospective anticipation. In a recent study, the 12-item PAiS was revised to an 8-item list in the English language (PAiS-R; Bies et al., 2025) that demonstrated sufficient scores with high evidence (+/H). Another version, the Turkish PAiS-TR (Uysar & Yasar, 2018) was adapted from the existing scale for CWS; however, the content validity was assessed to be low (+/L), owing to no involvement of experts or CWS in validating the tool. Also, the study provided limited information on the item added in the version for CWS as well as response forms. This further weakens the quality of the study.

The Speech Situation Checklist – Revised (SSC-R; Vanryckeghem et al., 2017) eliminated 13 irrelevant items that were reported to have poor sensitivity in discriminating AWS from AWNS. Also, the wording of the items were revised to facilitate better understanding. However, the content validity for the revised version was not established with the clinical group. Hence, the study was rated as indeterminate scores with low evidence (?/L). Another version of this tool, for CWS, the Speech Situation Checklist – Emotional Reactions in Kannada (SSC-ER-K; Veerabhadrappa et al., 2021) was translated via a forward-backward process without cognitive interviews with Kannada-speaking CWS, indicating a low evidence for relevance and comprehensiveness. Also, it was noted that the pilot study of the Kannada adaptation recruited only 5 CWS, which is considered insufficient sample size to establish content validity as per COSMIN-OMI guidelines.

Overall, most of the PROMs identified made initial efforts to establish content validity, though the methodological robustness varied significantly. A primary limitation across RPOMs was the absence of formal cognitive interviews with the target population, lack of formal documentation of concept saturation, and lack of re-establishing content validity for cross-cultural translations/adaptations.

Tools, Psychometric Properties and Strength of Evidence

This section summarizes the analysis for remaining psychometric properties (test-retest reliability, internal consistency, cross-cultural validity/measurement invariance, construct validity, concurrent validity, criterion validity, and responsiveness) of the 25 PROMs identified in the review. In addition, the strength of evidence related to each study for the tools identified were graded based on the modified Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Approach. Using this approach, recommendations on use of PROMs were made as: *recommended for use, potential to be recommended,* and *no recommendation*. Specific details on the psychometric properties may be found in the Appendix 1.

PROMs for Children who Stutter

The structural validity of CBS-PCWS (Mehdizadeh Behtash et al., 2022) was evaluated using exploratory factor analysis (EFA) on 205 parents of CWS that extracted five factors explaining 63.89% of total variance. Further, the internal consistency was excellent and the test-retest reliability was good. Measurement error was also evaluated. However, confirmatory factor analysis was not conducted for evaluating the structural validity and the test-retest sample was slightly sub-optimal. The tool was classified as *recommended for use* based on the strong and comprehensive psychometric evidence.

The Palin PRS (PPRS, Millard & Davis, 2016) evaluated structural validity using EFA with adequate sample, extracting three factors explaining 57% of variance with all item loadings above 0.40. Internal consistency was very good with α values ranging from 0.838 to 0.865. Normative data including percentile ranks and stanine scores were established which enhances the clinical utility. Test-retest reliability was not reported in this specific study. The tool was classified as recommended for use. Further, both the Turkish (PPRS-TR; Cangi et al., 2025) and Persian adaptations (PPRS-P; Barzegar Bafrooei et al., 2025) were observed to have good structural validity with adequate sample size along with high internal consistency. In addition, a remarkable test-retest reliability was also noted for both adaptations. Further, no floor or ceiling effects were detected in the Persian version. Nevertheless, both Turkish and Persian versions were classified as *potential to be recommended for use*.

The SBIS (Ntourou et al., 2020) employed the most robust structural validity methodology in this review utilising both EFA (n = 234) and independent CFA (n = 234) with excellent fit (RMSEA = 0.00, CFI = 1.00). The Internal consistency was strong, test-retest reliability was good, and the construct validity was excellent. Lastly, the predictive validity demonstrated that a higher behavioural inhibition significantly predicted severity of stuttering, frequency of stuttering, and KiddyCAT scores. The tool was classified as *recommended for use* based on COSMIN Guidelines. The VRYCS (Singer et al., 2022) rating scale used principal component analysis (PCA) to extract five factors explaining 59.3% of variance with adequate sample (n = 214) and revealed adequate structural validity. The internal consistency measures were acceptable. However, CFA, test-retest reliability, and external construct validity were not evaluated. Hence, the study was classified as *potential to be recommended for use with future research*.

PROMs for Adults who Stutter

The Self-Stigma of Stuttering Scale (4S) Original (Boyle, 2013) demonstrated a strong test-retest reliability, excellent internal consistency, and a good convergent validity against self-esteem, self-efficacy, and life satisfaction measures. The replication study (Boyle, 2015) replicated the three-factor EFA structure (though explaining only 39.5% of variance) for structural validity and demonstrated extensive construct validity against seven psychometric scales. Both studies were classified as *highly recommended for use*. The Turkish version of 4S-TR (Tiryaki et al., 2023) was the only adaptation study of 4S to conduct both independent EFA and CFA reporting adequate structural validity. With an excellent internal consistency and excellent test-retest reliability the PROM was classified as *highly recommended for use*. Likewise, the Polish 4S (Wesierska et al., 2025) demonstrated excellent internal consistency. However, structural validity and test-retest reliability were not evaluated, thereby classification being *potential to be recommended*. Furthermore, the Japanese short 4S-J-16 (Limura et al., 2022) was classified as *potential to be recommended* with good reliability and strong construct validity.

O'Brian et al. (2004) estimated only the inter-rater agreement for the 9-point self-rating scale that demonstrated a high inter-rater agreement between AWS and SLPs. Overall, the study was rated as low due to the small sample size and weak study design. Based on the study, the tool was classified as *potential to be recommended for use with further research*. In another investigation, Persian version of BAB in Persian (BAB-P; Węsierska et al. 2018) demonstrated an excellent internal consistency across all four subdomains for both AWS and AWNS and significantly differentiated between the two groups. Overall, the tool was rated as *recommended for use*. Similarly, the Inventory of Communication Attitudes (ICA) (Watson et al., 1987) showed strong psychometric properties - excellent internal consistency across all four response scales and a strong test-retest reliability. The discriminant validity between AWS and AWNS was highly significant across all 13 situations and four response scales. The tool

was rated as *recommended for use.* Another tool with good psychometric assessment was ISACS (Kelkar et al., 2024) which demonstrated strong overall internal consistency across subdomains, a good discriminant validity against normal fluency, and an expected correlations with SSI-4. However, owing to inadequate sample size for factor analysis and lack of test-retest reliability measures, the tool was classified as *potential to be recommended for use with further research*.

The Locus of Control of Behaviours (LCB) Scale (Craig et al., 1984) was observed to exhibit robust psychometric properties. The scale demonstrates acceptable structural validity explaining only 36.5% of variance; however, excellent test-retest reliability and a strong predictive validity was reported. Hence, the tool was classified as *recommended for use*. Unlike LCB, the r-SASBS (Hancer & Tokgoz-Yilmaz, 2025) demonstrated the most comprehensive and adequate structural validity by using the powered independent EFA (n = 165) that extracted two factors explaining 77.52% of variance, confirmed by independent CFA (n = 200) with excellent fit (CFI = 0.98, RMSEA = 0.075). The internal consistency as well as the test-retest reliability was excellent, while the discriminant validity was strong (Cohen's d = 3.67). Hence, the tool was classified as *recommended for use*.

The S-Scale (Erickson, 1969) demonstrated excellent internal consistency with strong concurrent and discriminant validity, and the scale was classified as *recommended for use*. Similarly, S-24 (Andrews & Cutler, 1974) demonstrated adequate internal consistency and high sensitivity to outcome in intervention i.e., reduction from mean 19.2 to 9.1 that significantly differentiated AWS from AWNS. Based on this, the tool was classified as *recommended for use*. The psychometric assessment of SA Scale (Huinck & Rietveld, 2007) involved only estimation of the low criterion validity that was found to be low. Also, the sample size was inadequate. Thus, the tool was classified as *potential to be recommended for use with further research.*

The Persian version of Safety Behaviours Questionnaire (SBQ; Azarinfar et al., 2022) conducted EFA on an adequate sample of 167 AWS that extracted four factors explaining 50% of variance indicating acceptable structural validity with good internal consistency ($\alpha = 0.89$). Assessing communication satisfaction, the SCESS scale (Karimi et al., 2018) was found to have good construct validity. However, the test-retest reliability and external construct validity were not evaluated for both these tools, thereby they were classified as *potential to be recommended for use* with further research. Another tool found to have similar properties was the SESAS (Ornstein & Manning, 1985) that demonstrated strong criterion validity, significant discriminant validity, and acceptable test-retest reliability. However, the sample size was significantly less for COSMIN standards, and structural validity and internal consistency were not evaluated. The overall rating was doubtful, and the classification was *potential to be recommended for use with further research*.

The SGSM (Alameer et al., 2017) demonstrated a very high test-retest reliability, strong intra-judge and inter-judge reliability, and significant discriminant validity. However, the study exhibited lack of statistical power due to inadequate sample size ($n = 43$) for assessing reliability and structural validity. The classification was *potential to be recommended for use with further research*. The Persian version of SGSM (Hozeili et al., 2024) showed excellent test-retest reliability, excellent inter-judge reliability, high internal consistency, and responsiveness. Owing to inadequate sample size (as per COSMIN recommendations), the structural validity was not evaluated, and the tool was classified as *potential to be recommended for use*.

The Speech Naturalness (SNA) and Feel Naturalness (FNA) self-rating scales (Finn & Ingham, 1994) demonstrated high intra-judge agreement and adequate concurrent validity against listener ratings, but were limited by a smaller sample size ($n = 12$) and single-item scope. The classification was *potential to be recommended for use with further research*.

Likewise, the Subjective Stuttering Severity-Research Edition (SSS; Riley et al., 2004) demonstrated good test-retest reliability and adequate criterion validity against %SS and PSI, but it was critically limited by an inadequate sample size and lack of assessment of structural validity. The classification was *potential to be recommended for use with future research*.

The original UTBAS (St Clare et al., 2009) demonstrated excellent internal consistency, strong convergent validity against Social Phobia Anxiety Inventory(SPAI) and Social Avoidance and Distress Scale (SADS), large known-groups differentiation, and significant responsiveness to CBT treatment change. The test-retest reliability was acceptable; however the sample size was inadequate according to COSMIN guidelines and structural validity could not be evaluated. The resulting classification was *potential to be recommended for use with further research*. Iverach et al. (2011) significantly strengthened the tool by establishing adequate structural validity, excellent internal consistency, and a strong test-retest reliability. Construct validity was assessed by establishing significant correlations with State-trait Anxiety Inventory – Trait (STAI-T) and the Fear of Negative Emotional Scale (FNE). Based on this study, UTBAS was classified as *recommended for use*. Further, the brief UTBAS-6 (Iverach et al., 2016) yielded an acceptable internal consistency and excellent diagnostic accuracy (sensitivity 93.4%, specificity 89.8%). However, test-retest reliability and external construct validity were not evaluated, resulting in a classification of *potential to be recommended for use with further research*.

Among the cross-cultural adaptations of UTBAS, the Japanese UTBAS-J (Chu et al., 2017) demonstrated exceptional internal consistency, adequate test-retest reliability, and a strong convergent validity against-24 and SA scale. The Turkish UTBAS-TR (Aydın Uysal & Ege, 2020) reported high internal consistency and good test-retest reliability. The Persian UTBAS-P (Farpour et al., 2025) showed high internal consistency and excellent convergent

validity against anxiety measures, but test-retest reliability was critically underpowered (n = 10). Further, the Italian UTBAS-ITA (Bernardini et al., 2024) demonstrated excellent test-retest reliability and a strong discriminant validity, with poor convergent validity. All of these cross-cultural adaptations failed to estimate the structural validity, primarily owing to inadequate sample size. Hence, they were all classified as *potential to be recommended for use with further research*.

The WASSP (Wright et al., 1998) is an introductory paper presented no formal psychometric validation data and was classified as *not recommended* in its 1998 form, but to address this gap comprehensively, the Turkish version of WASSP i.e., WASSP- TR (Uysal & Köse, 2021) assessed the psychometric properties. It demonstrated excellent CFA model fit (CFI = 0.99, RMSEA = 0.059) for structural validity, excellent internal consistency (α = 0.949 total), and strong test-retest reliability (r = 0.972). Hence, classified as *recommended for use*.

PROMs for Children and Adults who Stutter

Across all the studies evaluating psychometric properties of CAT, structural validity was not evaluated inherently lowering the quality of the study. The original CAT (Vanryckeghem & Brutten, 1992) assessed only the test-retest reliability using Pearson product-moment correlations and reported good reliability. The BigCAT (Vanryckeghem & Brutten, 2011) demonstrated excellent internal consistency for both AWS and AWNS, along with excellent discriminant validity. However, structural validity and content validity were not evaluated. The Persian BigCAT (Valinejad et al., 2018) was found to have good internal consistency, strong test-retest reliability, excellent criterion validity against the Erickson S-24 scale, and strong discriminant validity. Similarly, the English BigCAT (Vanryckeghem & Muir, 2016) and the Dutch KiddyCAT (Vanryckeghem et al., 2015) demonstrated good test-retest reliability for individuals with stuttering. The CAT-K (Veerabhadrappa et al., 2020)

showed good internal consistency for CWS but poor reliability for controls, with adequate construct validity (discriminant analysis accuracy 99.5%). The tool was limited by an underpowered test-retest sample (n = 10). In a subsequent investigation, the BigCAT-K (Veerabhadrappa et al., 2021) showed high internal consistency for AWS and AWNS, with good construct validity and an acceptable test-retest reliability. None of the studies evaluated structural validity. Overall, across studies, the tool was classified as *potential to be recommended for use with further research*.

OASES is one of the most popular tools with several cross-cultural adaptations available. The original OASES (Yaruss & Quesal, 2006) demonstrated an excellent internal consistency across all domains and satisfactory construct validity. The test-retest reliability was low in the original version. Among the cross-cultural adaptations, the simplified Chinese version, OASES-A-SC (Ma et al., 2025) was found to be psychometrically most robust, with excellent total internal consistency and good test-retest reliability. The Polish OASES-PL (Wesierska et al., 2023) was a translation and validation study for teenagers that showed good internal consistency across sub-scales and suboptimal structural validity. The Japanese OASES-A-J (Sakai et al., 2017) resulted in good internal consistency and good construct validity evaluated against the modified Erikson scale. Further, the Dutch OASES-A-D (Koedoot et al., 2011) reported acceptable subscale internal consistency and adequate concurrent validity, but no structural validity or test-retest data were evaluated. The Dutch OASES-S-D (Lankman et al., 2015) for school-age children showed low internal consistency, particularly for Section I. The Swedish OASES (Lindstrom et al., 2020) was validated among 139 people who stutter including adults, teenagers and CWS. The findings revealed severely problematic internal consistency for Section I among the teenagers and school-age CWS.

The Kannada OASES-A-K (Mahesh et al., 2024) was evaluated in 51 AWS limited by a small sample and assessment of test-retest reliability with smaller sample size (n = 10). The

revised OASES-A-R (Tichenor & Yaruss, 2024) employed IRT-based item selection and demonstrated acceptable subscale internal consistency and high concurrent validity against the original OASES-A, but lacked test-retest reliability data. Based on the above findings, OASES and its variants were classified as *potential to be recommended for use with further research*, except the Swedish version that was classified as *not recommended*.

The original PAiS (Cholin et al., 2016) demonstrated internal consistency and acceptable discriminant validity. Likewise, the Turkish version of PAiS-TR (Uysar & Yasar, 2018), was found to have remarkable test-retest reliability and strong construct validity against SSI-4. However, the structural validity was not evaluated. Both these versions were classified as *potential to be recommended for use with further research.* The revised English version of PAiS i.e., PAiS-R (Bies et al., 2025) showed clear improvement across all psychometric dimensions. PCA confirmed a single-factor solution explaining 54.5% of variance, internal consistency was excellent, and test-retest reliability was strong. Construct validity was assessed against subjective severity and rated as sufficient moderate evidence. The PAiS-R was classified as *recommended for use*.

The SSC-R (Vanryckeghem et al., 2017) evaluated structural validity using PCA, which resulted in seven factors for SSC-ER and six factors for SSC-SD explaining 70.46% and 70.50% of variance, respectively. This reflected good structural validity, though with inadequate sample as per COSMIN. Further, the internal consistency was excellent for SSC-ER and SSC-SD in PWS. Discriminant validity was also reported to be strong. Similarly, the SSC-ER-K (Veerabhadrappa et al., 2021) demonstrated excellent internal consistency for CWS and good internal consistency for CWNS, with very high test-retest reliability for CWS. Both the studies led to classification of the tool as *potential to be recommended for use with further research.*

Across all the PROMs evaluated, three of the PROMs were classified as *recommended for use*. However, majority of the tools were classified as *potential to be recommended for use with further research* owing to factors such as sample size inadequacy and lack of evaluation of structural validity or other psychometric properties. A *not recommended for use* classification was given only for the Swedish OASES (for teenager & school-age children) and the SA Scale due to inadequate and significantly poor psychometric measures.

CHAPTER 4

DISCUSSION

This systematic review focused on identifying various tools that can serve as patient-reported outcome measures (PROMs) in stuttering intervention. The review followed the PRISMA-COSMIN-2024 guidelines. Studies reviewed explained the development of the tool and evaluation of its psychometric or measurement properties. A total of 57 studies, published between January 1969 and November 2025 were included which led to the identification of 25 PROMs that may be potentially used to measure the treatment efficacy among adults and children who stutter. The various psychometric properties assessed across studies include - content validity, internal consistency, structural validity, and reliability measures. There was a notable increment in publications on PROMs in stuttering from 2016 reflecting the growing acceptance and due acknowledgment of the ICF framework in the field of speech language pathology, specifically in the area of stuttering. This shift in trend is attributed to an increase in global emphasis on person-centered care in the healthcare sector. The results of this review were discussed based on geographical distribution and population coverage, ICF framework, and effect of psychometric properties of PROMs on the clinical utility.

Geographical distribution and population coverage

The geographical distribution of the studies revealed a dominance from the United States followed by Australia, Turkey, Iran, and India. The dominance of more tools developed in these regions reflects a well-established research infrastructure, adequate research funding, and the availability of experts. In contrast, limited studies were identified from other geographical locations across the world unveiling inadequate research infrastructure and funding as well as barriers associated with language access and professional training in these regions. Several of the studies identified from Europe, Middle-East, and India were cross-

cultural adaptations of the tools developed in the West. In a scoping review of validated PROMs in adult speech-language pathology, Gardner et al. (2024) argued that there is a lack of developmentally and culturally diverse PROMs, and we must close this gap by developing some culturally appropriate tools for different age groups.

Similarly, a notable imbalance was found among the PROMs available for adults who stutter (AWS) versus those available for children who stutter (CWS). Majority of the studies focused on AWS (N = 36), while only 10 targeted CWS. The remaining studies either included both age groups or recruited parents/caregivers of CWS. Among the PROMs identified for CWS, the Palin PRS was found to be the most comprehensively developed tool, derived from a Delphi panel study involving parents of CWS and validated across three cultural contexts. Lesser PROMs for the paediatric group, in the field of speech-language pathology, were also reported by Gardner et al. (2024). Further, most existing PROMs/proxy-outcomes for CWS are adaptations of PROMs for adults. This reflects several practical and methodological challenges associated with the development and validation of PROMs among children. Development of a PROM is highly dependent on the inputs from the target group for whom the PROM is being developed. CWS, particularly the younger preschool children, may not be able to understand and communicate their perceptions of stuttering, treatment, and its impact. Hence, developing and validating PROMs for CWS is an important domain that warrants further research.

Mapping PROMs to the ICF Framework: Domain Classification

To understand the usefulness of PROMs identified in this review, all PROMs were systematically mapped onto the WHO ICF framework. Hence, the tools were categorized across the four domains: body structure and functions, activities and participation, personal factors and environmental factors. Table 4.1 below summarizes the categorization of PROMs across the domains of WHO ICF framework.

Table 4.1

Mapping of Patient-Reported Outcome Measures to the ICF Framework Domains and Sub-Domains

Sub-Domain	Patient-Reported Outcome Measures (PROMs)
Body Functions and Structures	1. 9-Point Self-Rating Scale (Stuttering Severity)[a,P]
	2. SA Scale[a,P]
	3. SNA & FNA Self-Rating Scales[a,P]
	4. SSS): Research Edition [Stuttering Severity subscale][a,P]
	5. PAiS & adaptations (PAiS-R; PAiS-TR)[a,c,P]
	6. SGSM & Persian version (P-SGSM)[a,P]
Activities and Participation	1.OASES/OASES-A, S, T & adaptations (OASES-A-D; OASES-A-J; OASES-A-K; OASES-A-R; OASES-PL; OASES-S-D; OASES-A-SC; Swedish OASES) [Section III: Communication in Daily Situations; Section IV: Quality of Life][a,c,d,P]
	2. ICA[a,P]
	3. SSC-ER & SD & Kannada version (SSC-ER-K)[a,c,P]
	4. BAB & BAB-Polish [SSC-ER and SSC-SD subtests][a,P]
	5. SCESS[a,P]
	6. SSS [Avoidance subscale][a,P]
	7. ISACS [Subscale III: Activities and Environmental Factors; Subscale IV: Participation/QoL][a,P]

Sub-Domain	Patient-Reported Outcome Measures (PROMs)
Personal Factors	1.BigCAT/CAT & adaptations (BigCAT-K Kannada; CAT-K; BigCAT Persian; Dutch KiddyCAT; English BigCAT)[a,c,P]
	2. S-Scale (Communication Attitudes Scale)[a,P]
	3. Modified Erickson Scale (S-24)[a,P]
	4. BAB [BigCAT subtest & Behavior Checklist subtest][a,P]
	5. UTBAS I, II, III) & adaptations (UTBAS-6; UTBAS-ITA; UTBAS-P; UTBAS-J; UTBAS-TR)[a,P]
	6. OASES [Section II: Reactions to Stuttering][a,c,d,P]
	7. WASSP & WASSP-TR [Avoidance, Behaviours, Thoughts and Feelings subscales][a,P]
	8. SBIS[b,c,R]
	9. r-SASBS[a,P]
	10. Safety Behaviors Questionnaire (Persian version)[a,P]
	11. SSS [Avoidance subscale & Locus of Control subscale][a,P]
	12. SESAS[a,P]
	13. LCB[a,P]
	14. 4S & adaptations (4S-Polish; 4S-TR Turkish; 4S-J-16 Japanese)[a,P]

Sub-Domain	Patient-Reported Outcome Measures (PROMs)
Environmental Factors	1. Palin PRS & adaptations (Palin PRS-TR Turkish; Palin PRS-P Persian)[b,c,R]
	2. CBS-PCWS[b,c,R]
	3. VRYCS[b,R]
Multidimensional	1. OASES & all versions – explicitly structured across all ICF components[a,c,d,P]
	2. ISACS & subscales directly mapped to ICF domains[a,P]
	3. BAB & BAB-Polish – covers behaviours, emotional reactions, speech disruption, and communication attitudes[a,P]
	4. WASSP & WASSP-TR – covers behaviours, thoughts, feelings, avoidance, and disadvantage[a,P]

Note. Classification is based on the International Classification of Functioning, Disability and Health (ICF; WHO, 2001). A total of 25 patient-reported outcome measures (PROMs) identified across 57 studies included in the systematic review are mapped to ICF. Tools classified under Personal Factors are organised by primary psychological construct. Tools listed under Multidimensional/Comprehensive (ICF-Based) span two or more ICF domains simultaneously and therefore also appear under their respective primary domain rows. Superscripts denote the target population age group and the reporter type for each measure.

Age group: [a] Adults who stutter (AWS; ≥18 years); [b] Preschool-age children who stutter (PS-CWS; 2–5 years); [c] School-age children who stutter (SA-CWS; 6–12 years); [d] Adolescents who stutter (AdWS; 13–17 years)

PROM type: [P] Patient/self-report; [R] Parent/caregiver-report

Abbreviations: OASES = Overall Assessment of the Speaker's Experience of Stuttering; UTBAS = Unhelpful Thoughts and Beliefs About Stuttering; BAB = Behavior Assessment Battery; WASSP = Wright and Ayre Stuttering Self-Rating Profile; ISACS = Impact Scale for Assessment of Cluttering and Stuttering; 4S = Self-Stigma of Stuttering Scale; CAT/BigCAT = Communication Attitude Test; SSC = Speech Situation Checklist; ICA = Inventory of Communication Attitudes; SCESS = Satisfaction with Communication in Everyday Speaking Situations; PAiS = Premonitory Awareness in Stuttering Scale; SGSM = Stuttering Generalization Self-Measure; r-SASBS = Revised Scale of Avoidance and Struggle Behaviours in Stuttering; SBIS = Short Behavioral Inhibition Scale; SESAS = Self-Efficacy Scale for Adult Stutterers; LCB = Locus of Control of Behaviour Scale; SSS = Subjective Screening of Stuttering; Palin PRS = Palin Parent Rating Scales; CBS-PCWS = Caregiver Burden Scale for Parents of Children Who Stutter; VRYCS = Vanderbilt Responses to Your Child's Speech; SA = Self-Assessment Scale; SNA/FNA = Speech/Feel Naturalness Assessment Scales.

Body functions and structures

This domain focuses on the speech behaviour and the underlying physiological systems. Conventionally, this domain of ICF has been chiefly assessed based on clinician-related outcome measures. A few PROMs identified through this review, that document the speech fluency and overall speech production as perceived by those who stutter, include - the nine-point self-rating scale (O'Brian et al., 2004), SA Scale (Huinck & Rietveld, 2007), SNA/FNA scales (Finn & Ingham, 1994), the stuttering severity subscale of the SSS-Research edition (Riley et al., 2004), and the PAiS (Cholin et al., 2016). These tools gauge aspects such as internal experience of stuttering, speech naturalness, perceived severity and the premonitory urge before stuttering. However, the evaluation of the psychometric properties of these tools indicate scope for further research. A major gap was identified in this domain of ICF, and it was recommended that PROMs measuring self-perception of physical and psychological functioning are needed for both children and adults who stutter.

Activities and Participation

This domain focuses of ICF focuses on how well an individual can perform the daily communication tasks and engage in daily activities. Several PROMs were identified under this domain as follows: Avoidance subscale of SSS-Research edition (Riley et al., 2004), ICA (Watson et al., 1987), BAB (Węsierska et al., 2018), SCESS (Karimi et al., 2018), SGSM (Alameer et al., 2017), SSC-R and SSC-R-K (Vanryckeghem et al., 2017; Veerabhadrappa et al., 2021), ISACS (subscales iii & iv), and the OASES (Yaruss & Quesal, 2006). and its various adaptations. These tools align with the framework of ICF by recognising that the stuttering affects person's ability to participate in job opportunities, education, and social participation significantly that measures the outcome than just speech fluency (Yaruss, 2010; Mokkink et al., 2024).

Personal factors

Personal contextual factors, the third domain in the ICF framework, includes the psychological and attitudinal relationship of individuals with their own condition. This includes their self-perception, emotional responses, coping patterns, and sense of control. This domain was identified to have some of the most clinically significant PROMs in stuttering, reflecting a well-established link between stuttering and the cognitive-emotional profile of the individual. The PROMs identified in the review were: CAT and BigCAT (Vanryckeghem & Brutten, 1992, 2011), BAB (Węsierska et al., 2018; BigCAT & Behavior checklist subtest), S-Scale (Erickson, 1969), the modified Erickson Scale or S-24 (Andrews & Cutler, 1974), UTBAS (St Clare et al., 2009; Iverach et al., 2011, 2016), WASSP (Wright et al., 1998; Uysal & Köse, 2021), r-SASBS (Hancer & Tokgoz-Yilmaz, 2025), Safety Behaviours Questionnaire (Azarinfar et al., 2022), SESAS (Ornstein & Manning, 1985), LCB Scale (Craig et al., 1984), SBIS (Ntourou et al., 2020), and the 4S (Boyle, 2013, 2015; Tiryaki et al., 2023). The higher number of PROMs in this domain reflects the extent and complexity of the psychological impacts of stuttering. It also underscores the importance of assessing personal factors through the course of stuttering intervention and using appropriate PROMs.

Environmental factors

The fourth domain of the ICF represents the external influences on stuttering – reactions and attitudes of others, social-communicative situations, and societal attitudes, and support system for those who stutter. Based on the review, a few PROMs for CWS were identified that gauge some of these aspects. These include: Caregiver Burden Scale for Parents of Children who Stutter (CBS-PCWS; Mehdizadeh Behtash et al., 2022), Palin Parent Rating Scales (Millard & Davis, 2016; Cangi et al., 2025; Barzegar Bafrooei et al., 2025), and Vanderbilt Responses to Your Child's Speech rating scale (VRYCS; Singer et al., 2022). In case of AWS, no PROM was identified that specifically evaluates the environmental factors. OASES was one

tool found that partially assesses some of the environmental influences. Though the environmental factors are not evaluated as in independent domain in OASES, items assessing this factor are embedded within broader sections of "communication in daily situations". Despites the critical impact of environmental factors like family support, peer attitudes, and societal reactions, there are limited number of tools to measure this domain in both children and adults who stutter. This is a key gap identified in the literature. It may be noted that no PROM assesses all the domains of ICF. Hence, encouraging clinicians to follow an ICF-informed assessment and intervention practices, authors suggest selecting suitable PROMs across domains for individual clients.

Effect of Psychometric Property on Clinical Utility of PROMs

The quality of psychometric evidence varied across the 25 PROMs, reflecting a wide range of study designs, sample sizes, and validation approaches. According to the COSMIN guidelines for OMIs 2024 grading approach, none of the instruments achieved the highest classification of a strong recommendation for clinical use. However, some tools, demonstrated strong psychometric properties and were classified as *recommended for use* or a *potential to be recommended for use with further research*. According to the COSMIN guidelines, the tools such as the 4S (Boyle, 2013, 2015; Tiryaki et al., 2023), CBS-PCWS (Mehdizadeh Behtash et al., 2022), ICA (Watson et al., 1987), Palin PRS (Millard & Davis, 2016), BAB (Węsierska et al., 2018), r-SASBS (Hancer & Tokgoz-Yilmaz, 2025), SBIS (Ntourou et al., 2020), the LCB Scale (Craig et al., 1984), S-Scale (Erickson, 1969), WASSP-TR (Uysal & Köse, 2021), and the UTBAS (Iverach et al., 2011) were classified as *recommended or highly recommended for clinical use.* These tools were observed to display strong psychometric properties such as adequately powered structural validity analyses using EFA/CFA/PCA, excellent internal consistency, acceptable to excellent test-retest reliability with adequate sample size, and evidence of construct validity through convergent and discriminant testing. The r-SASBS

demonstrated adequate EFA & CFA on adequately samples along with exceptional test-retest reliability data that made this scale to be the one of the most methodologically complete recently developed tools in the field.

On the other hand, some tools such as OASES and CAT exhibited some significant methodological weaknesses that limited their clinical utility and were classified as *potential to be recommended with further research on quality*. However, these are tool that are popular among clinicians and have well established test-retest reliability measures. They were found to be rated low on quality as per the COSMIN guidelines, primarily due to the sample size criteria. Apart from the inadequate sample size, other limitations observed across studies for measuring psychometric properties included absence of CFA/EFA or insufficient evaluation of structural validity. It should be noted that these observations do not imply that these tools are inefficient; rather, they reflect the challenge of obtaining large sample size in a condition like stuttering that has relatively low prevalence and assessment through the stringent lens of COSMIN guidelines.

Some of the cross-cultural adaptations of the PROMs were rated as low or moderate on quality. This may be attributed to the fact that most of these adaptations did not re-establish all psychometric properties, as they were based on previously standardized tools. Further, during the review several other versions of existing PROMs were identified that lack any psychometric data. These include data for PPRS from Norway, US, Korea, Italian, Swedish, Greek, Spanish, Czech and Polish (available on Palin PCI website), OASES normative for Brazilian (Osborn et al., 2012) and Australian (Tran et al., 2012) populations, and CAT translation to languages such as Slovenian (Gacnik &Vanryckeghem, 2014), Italian (Bernardini et al., 2009), Pakistan (Vanryckeghem & Mukati, 2006). Nevertheless, this review emphasized the importance of re-

evaluating psychometric properties during cross-cultural adaptations to ensure the both the constructs measured and the items remain relevant to the target population.

CHAPTER 5

SUMMARY AND CONCLUSION

This systematic review aimed to identify the patient reported outcome measures (PROMs) in stuttering intervention among children and adults who stutter. The current review followed the PRISMA-COSMIN for OMIs 2024 guidelines, a standard method to evaluate the quality of PROMs. After conducting a thorough search in four databases - PubMed, Science-direct, ProQuest and CINAHL Ultimate, a total of 2189 articles were selected. After removing the duplicates, title and abstract screening, 1728 articles were excluded. 57 studies published between January 1969 and November 2025 were included in the final review. Based on these, 25 PROMs were identified available for children and adults who stutter. The geographical synthesis indicated PROMs developed and validated across United States (n = 16), Australia (n = 7), Turkey (n = 6), Iran (n = 6), India (n = 5), and other locations (n = 17).

The findings of the study showed that PROMs for children are significantly lesser when compared to adults. The review identified 4 PROMs for CWS (Appendix A), among which three were proxy/parent-reported outcome measures. The CBS-PCWS and Palin PRS were found to have sufficient and high content validity compared to other PROMs for CWS. The study also identified 17 PROMs exclusively for AWS. Additionally, the study identified some PROMs that are available for both children and adults such as OASES and CAT. The psychometric analysis of the PROMs revealed that most of the tools were categorised as *potential to be recommended for use with further research*. The primary limitations/weakness recognized was inadequate sample size and lack of assessment of structural validity.

The PROMs were also mapped onto the ICF framework (See Table 4.1). It was observed that majority of the PROMs were related to two domains of the ICF: *activity and participation* and *personal factors*. Relatively lesser PROMs were noted for the *body structure*

and function component. Further, one domain identified to have scarcity of PROMs, particularly for AWS, was the *environmental factors.* Another important observation was that most PROMs were developed in the western context, and cross-cultural adaptations have been done in other parts of the world. Additionally, these adapted versions primarily involved the reverse translation, informal content validation, and reliability testing, but did not undergo a rigorous psychometric analysis.

Implications of the Study

The findings of the study provide insights into available PROMs for children and adults who stutter. The study attempts to map the PROMs onto the ICF framework that may help the clinicians to choose suitable measures capturing both the overt and covert behaviours of stuttering. Further, the study provides a comprehensive overview of the quality of existing PROMs and also highlights the PROMs along with specific psychometric properties that require further investigation. The review identified the lacunae with limited PROMs for CWS and emphasizes on the need to develop the same. Findings of the current review may facilitate clinically-informed practice among clinicians and researchers. Lastly, the gaps identified in existing PROMs, may help to design and adapt culturally and developmentally relevant tools for usage in a wide range of communities.

Limitations and Future Directions

Some of the potential limitations and future directions include:

- The search was restricted to peer-reviewed English articles. Hence, there is a potential that some of the PROMs developed in other languages were missed. Future research can include databases from other languages and publications to ensure better global coverage.

- The quality rating given for each study were based on the COSMIN checklist by the reviewers. Although every effort was made to be consistent and objective, some degree of subjectivity in ratings cannot be fully ruled out. Future studies could strengthen this process by involving multiple independent raters and reporting inter-rater reliability scores alongside COSMIN ratings.
- The search strategy was restricted to only four databases — PubMed, Science Direct, ProQuest, and CINAHL. Including additional databases such as the Scopus, EMBASE, and PsycINFO, could further strengthen the search and capture broader range of relevant studies.
- Future studies could focus on intervention trails that used PROM as a primary or secondary outcome to assess the quality of responsiveness in relation to specific treatment techniques.

REFERENCES

Alameer, M., Meteyard, L., & Ward, D. (2017). Stuttering generalization self-measure: Preliminary development of a self-measuring tool. *Journal of Fluency Disorders*, *53*, 41-51. 10.1016/j.jfludis.2017.04.001

Armour, M., Brady, S., Sayyad, A., & Krieger, R. (2019). Self-Reported Quality of Life Outcomes in Aphasia Using Life Participation Approach Values: 1-Year Outcomes. Archives of rehabilitation research and clinical translation, 1(3-4), 100025. https://doi.org/10.1016/j.arrct.2019.100025

American Speech-Language-Hearing Association (ASHA). (2025). *Stuttering, cluttering, and fluency disorders.* ASHA. Retrieved from ASHA website

Andrews, G., & Cutler, J. (1974). Stuttering therapy: The relation between changes in symptom level and attitudes. *Journal of Speech and Hearing Disorders*, *39*(3), 312–319. https://doi.org/10.1044/jshd.3903.312

Aydın Uysal, A., & Ege, P. (2020). Reliability and validity of the UTBAS-TR (The Unhelpful Thoughts and Beliefs Scale-the Turkish version) in the Turkish population. *International Journal of Speech-Language Pathology*, *22*(1), 24–29. https://doi.org/10.1080/17549507.2019.1568572

Azarinfar, M., Karimi, H., Jowkar, F., & Shafiei, B. (2022). Validity and reliability of safety behaviors questionnaire for Persian adults who stutter: A cultural perspective. *Journal of Communication Disorders*, *100*, 106251. 10.1016/j.jcomdis.2022.106251

Barzegar Bafrooei, E., Darouie, A., Maroufizadeh, S., & Farazi, M. (2025). Validation of the Persian version of the Palin Parent Rating Scales. *Folia Phoniatrica et Logopaedica*, *77*(1), 35–43. https://doi.org/10.1159/000539119

Bennett, K. G., Ranganathan, K., Patterson, A. K., Baker, M. K., Vercler, C. J., Kasten, S. J., ... & Waljee, J. F. (2018). Caregiver-reported outcomes and barriers to care among patients with cleft lip and palate. Plastic and reconstructive surgery, 142(6), 884e-891e.

Bernardini, S., Onnivello, S., & Lanfranchi, S. (2024). Italian normative data for the Unhelpful Thoughts and Beliefs about Stuttering (UTBAS) Scales for adults who stutter. *Journal of Fluency Disorders*, *81*, 106074. 10.1016/j.jfludis.2024.106074

Bernardini S, Vanryckeghem M, Brutten GJ,Cocco L, Zmarich C. Communication attitude of Italian children who do and do not stutter. J Commun Disord. 2009 Mar-Apr;42(2): 155–61. 10.1016/j.jcomdis.2008.10.003

Bloodstein, O., & Ratner, N. B. (2008). *A handbook on stuttering* (6th ed.). Delmar.

Blumgart, E., Tran, Y., Yaruss, J. S., & Craig, A. (2012). Australian normative data for the Overall Assessment of the Speaker's Experience of Stuttering. *Journal of Fluency Disorders*, *37*(2), 83-90. 10.1016/j.jfludis.2011.12.002

Bohlender J. E. (2023). Patient-reported outcome measures for assessing health-related quality of life in patients with voice and swallowing disorders. HNO, 71(9), 549–555. https://doi.org/10.1007/s00106-023-01346-2.

Boyle, M. P. (2013). Assessment of stigma associated with stuttering: Development and evaluation of the Self-Stigma of Stuttering Scale (4S). *Journal of Speech, Language, and Hearing Research*, *56*(5), 1517–1529. https://doi.org/10.1044/1092-4388(2013/12-0280

Boyle, M. P. (2015). Identifying correlates of self-stigma in adults who stutter: Further establishing the construct validity of the Self-Stigma of Stuttering Scale (4S). Journal of Fluency Disorders,43, 17–27. https://doi.org/10.1016/j.jfludis.2014.12.002

Bragatto, E. L., Osborn, E., Yaruss, J. S., Quesal, R., Schiefer, A. M., & Chiari, B. M. (2012). Brazilian version of the Overall Assessment of the Speaker's Experience of Stuttering-Adults protocol (OASES-A). *Jornal da Sociedade Brasileira de Fonoaudiologia*, *24*(2), 145-151. 10.1590/S2179-64912012000200010

Cangi, M. E., Aydin, N., & Millard, S. K. (2025). Turkish adaptation of the Palin Parent Rating Scales: A validity and reliability assessment. *Journal of Fluency Disorders*, *84*, 106118. 10.1016/j.jfludis.2025.106118

Carlozzi, N. E., Ready, R. E., Frank, S., Cella, D., Hahn, E. A., Goodnight, S. M., Schilling, S. G., Boileau, N. R., & Dayalu, P. (2017). Patient-reported outcomes in Huntington's disease: Quality of life in neurological disorders (Neuro-QoL) and Huntington's disease health-related quality of life (HDQLIFE) physical function measures. Movement disorders : official journal of the Movement Disorder Society, 32(7), 1096–1102. https://doi.org/10.1002/mds.27046

Carlozzi, N. E., Schilling, S., Kratz, A. L., Paulsen, J. S., Frank, S., & Stout, J. C. (2018). Understanding patient-reported outcome measures in Huntington disease: at what point is cognitive impairment related to poor measurement reliability?. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 27(10), 2541–2555. https://doi.org/10.1007/s11136-018-1912-6.

Cohen, M. L., Lanzi, A. M., & Boulton, A. J. (2021). Clinical Use of PROMIS, Neuro-QoL, TBI-QoL, and Other Patient-Reported Outcome Measures for Individual Adult Clients with Cognitive and Language Disorders. Seminars in speech and language, 42(3), 192–210. https://doi.org/10.1055/s-0041-1731365

Cholin, J., Heiler, S., Whillier, A., & Sommer, M. (2016). Premonitory Awareness in Stuttering Scale (PAiS). *Journal of Fluency Disorders*, *49*, 40–50. 10.1016/j.jfludis.2016.07.001

Chu, S. Y., Sakai, N., Mori, K., & Iverach, L. (2017). Japanese normative data for the Unhelpful Thoughts and Beliefs about Stuttering (UTBAS) Scales for adults who stutter. *Journal of Fluency Disorders*, *51*, 1–7. 10.1016/j.jfludis.2016.09.006

Churruca, K., Pomare, C., Ellis, L. A., Long, J. C., Henderson, S. B., Murphy, L. E., ... & Braithwaite, J. (2021). Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. *Health Expectations*, *24*(4), 1015-1024. 10.1111/hex.13254

Craig, A. R., Franklin, J. A., & Andrews, G. (1984). A scale to measure locus of control of behaviour. *British Journal of Medical Psychology*, *57*, 173–180. https://doi.org/10.1111/j.2044-8341.1984.tb01597.x

Demirci, A. N., İncebay, Ö., & Köse, A. (2024). Readability of Patient-Reported Outcome Measures Used in Voice Disorders. Journal of voice : official journal of the Voice Foundation, S0892-1997(24)00305-9. Advance online publication. https://doi.org/10.1016/j.jvoice.2024.09.014

de Riesthal, M., & Ross, K. B. (2015). Patient reported outcome measures in neurologic communication disorders: An update. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, *25*(3), 114-120.

Dunaway Young, S., Pasternak, A., Duong, T., McGrattan, K. E., Stranberg, S., Maczek, E., Dias, C., Tang, W., Parker, D., Levine, A., Rohan, A., Wolford, C., Martens, W., McDermott, M. P., Darras, B. T., & Day, J. W. (2023). Assessing Bulbar Function in

Spinal Muscular Atrophy Using Patient-Reported Outcomes. Journal of neuromuscular diseases, 10(2), 199–209. https://doi.org/10.3233/JND-221573

Dunne, M., Hoover, E., & DeDe, G. (2023). Efficacy of Aphasia Group Conversation Treatment via Telepractice on Language and Patient-Reported Outcome Measures. American journal of speech-language pathology, 32(5S), 2565–2579. https://doi.org/10.1044/2023_AJSLP-22-00306

Eadie, T. L. (2003). The ICF.

Eckstein, D. A., Wu, R. L., Akinbiyi, T., Silver, L., & Taub, P. J. (2011). Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. Plastic and reconstructive surgery, 128(5), 518e-526e.

Elsman, E. B., Mokkink, L. B., Terwee, C. B., Beaton, D., Gagnier, J. J., Tricco, A. C., ... & Offringa, M. (2024). Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. *Journal of Clinical Epidemiology*, *173*, 111422 https://doi.org/10.1007/s11136-024-03634-y.

Erickson, R. L. (1969). Assessing communication attitudes among stutterers. *Journal of Speech and Hearing Research*, *12*(4), 711–724. https://doi.org/10.1044/jshr.1204.711

Farpour, S., Shafie, B., Menzies, R., & Karimi, H. (2025). Reliability and validity of the unhelpful thoughts and beliefs scale for Persian-speaking adults who stutter (UTBAS-P): A cross-cultural examination of social anxiety in people who stutter. *Journal of Fluency Disorders*, *83*, 106099. 10.1016/j.jfludis.2025.106099

Finn, P., & Ingham, R. J. (1994). Stutterers' self-ratings of how natural speech sounds and feels. *Journal of Speech and Hearing Research*, *37*(2), 326–340. https://doi.org/10.1044/jshr.3702.326

Fontana, I., Piscopo, K., Mink van der Molen, A., Rezzonico, A. A., & Ombashi, S. (2025). A Validated Cultural Adaptation and Translation of the CLEFT-Q Patient Reported Outcome Measure (PROM) from English into Italian. Face, 6(1), 120-12.

Francis, D. O., McPheeters, M. L., Noud, M., Penson, D. F., & Feurer, I. D. (2016). Checklist to operationalize measurement characteristics of patient-reported outcome measures. *Systematic reviews*, *5*(1), 129.

Francis, D. O., Daniero, J. J., Hovis, K. L., Sathe, N., Jacobson, B., Penson, D. F., Feurer, I. D., & McPheeters, M. L. (2017). Voice-Related Patient-Reported Outcome Measures:

A Systematic Review of Instrument Development and Validation. Journal of speech, language, and hearing research : JSLHR, 60(1), 62–88. https://doi.org/10.1044/2016_JSLHR-S-16-0022

Gačnik M, Vanryckeghem M. Study of the Communication Attitude of Slovenian Children Who Do and Do Not Stutter. Cross-CultCommun. 2014; 10(5): 85–91. http://dx.doi.org/10.3968/5160

Gardner, D., Mitchell, V., Frakking, T., Weir, K. A., Canning, A., & Wenke, R. J. (2024). Validated patient reported outcome measures in speech-language pathology: A scoping review of adult practice. *International journal of speech-language pathology*, 1-30.

Gochman, G. E., Dwyer, C. D., Young, V. N., & Rosen, C. A. (2023). Exploring Patient's Preference of Patient-Reported Outcome Measures in Laryngeal Movement Disorders. The Laryngoscope, 133(6), 1448–1454. https://doi.org/10.1002/lary.30376

Guntupalli, V. K., Kalinowski, J., & Saltuklaroglu, T. (2006). The need for self-report data in the assessment of stuttering therapy efficacy: Repetitions and prolongations of speech. *International Journal of Language & Communication Disorders, 41*(1), 1–18. https://doi.org/10.1080/13682820500172212

Hancer, H., & Tokgoz-Yilmaz, S. (2024). Validity and reliability of the revised scale of avoidance and struggle behaviours in stuttering (r-SASBS). *International Journal of Language & Communication Disorders*, e13135. https://doi.org/10.1111/1460-6984.13135

Harding, S., Burr, S., Cleland, J., Stringer, H., & Wren, Y. (2024). Outcome measures for children with speech sound disorder: an umbrella review. BMJ open, 14(4), e081446.

Horton, S., Jackson, V., Boyce, J., Franken, M. C., Siemers, S., John, M. S., ... & Morgan, A. (2024). Self-reported stuttering severity is accurate: Informing methods for large-scale data collection in stuttering. *Journal of Speech, Language, and Hearing Research, 67*(10S), 4015–4024. https://doi.org/10.1044/2023_JSLHR-23-00081

Hozeili, E., Azimi, T., Ahmadi, A., Khoramshahi, H., Tahmasebi, N., & Dastoorpoor, M. (2024). Psychometric properties of the Persian version of the stuttering generalization

self-measure tool in adults who stutter. *Journal of Fluency Disorders*, *80,* 106056. 10.1016/j.jfludis.2024.106056

Huinck, W., & Rietveld, T. (2007). The validity of a simple outcome measure to assess stuttering therapy. *Folia Phoniatrica et Logopaedica*, *59*(2), 91–99. https://doi.org/10.1159/000098342

Iimura, D., Koyama, Y., Kondo, H., Toyomura, A., & Boyle, M. P. (2022). Development of a short Japanese version of the Self-Stigma of Stuttering Scale (4S-J-16): Translation and evaluation of validity and reliability. *Journal of Fluency Disorders*, *73*, 105917. 10.1016/j.jfludis.2022.105917

Iverach, L., Heard, R., Menzies, R., Lowe, R., O'Brian, S., Packman, A., & Onslow, M. (2016). A brief version of the Unhelpful Thoughts and Beliefs About Stuttering Scales: The UTBAS-6. *Journal of Speech, Language, and Hearing Research*, *59*(5), 964–972. https://doi.org/10.1044/2016_JSLHR-S-15-0167

Iverach, L., Menzies, R., Jones, M., O'Brian, S., Packman, A., & Onslow, M. (2010). Further development and validation of the Unhelpful Thoughts and Beliefs About Stuttering (UTBAS) scales: relationship to anxiety and social phobia among adults who stutter. *International Journal of Language & Communication Disorders*, 1–14. https://doi.org/10.3109/13682822.2010.495369

James, S., Brumfitt, S., & Cowell, P. (2009). The influence of communication situation on self-report in people who stutter. *International Journal of Speech-Language Pathology*, *11*(1), 34–44. https://doi.org/10.1080/17549500802588179

Johannisson, T. B., Wennerfeldt, S., Havstam, C., Naeslund, M., Jacobson, K., & Lohmander, A. (2009). The Communication Attitude Test (CAT-S): normative values for 220 Swedish children. *International Journal of Language & Communication Disorders*, *44*(6), 813–825. https://doi.org/10.1080/13682820903201084

Karimi, H., Onslow, M., Jones, M., O'Brian, S., Packman, A., Menzies, R., ... & Jelčić-Jakšić, S. (2018). The Satisfaction with Communication in Everyday Speaking Situations (SCESS) scale: An overarching outcome measure of treatment effect. *Journal of Fluency Disorders*, *58*, 77-85.

Kearney, E., Granata, F., Yunusova, Y., Van Lieshout, P., Hayden, D., & Namasivayam, A. (2015). Outcome measures in developmental speech sound disorders with a motor basis. Current Developmental Disorders Reports, 2(3), 253-272.

Klassen, A. F., Tsangaris, E., Forrest, C. R., Wong, K. W., Pusic, A. L., Cano, S. J., ... & Goodacre, T. (2012). Quality of life of children treated for cleft lip and/or palate: a systematic review. Journal of Plastic, Reconstructive & Aesthetic Surgery, 65(5), 547-557.

Koedoot, C., Versteegh, M., & Yaruss, J. S. (2011). Psychometric evaluation of the Dutch translation of the Overall Assessment of the Speaker's Experience of Stuttering for adults (OASES-AD). *Journal of Fluency Disorders, 36*(3), 222–230. https://doi.org/10.1016/j.jfludis.2011.02.002

Lankman, R. S., Yaruss, J. S., & Franken, M.-C. (2015). Validation and evaluation of the Dutch translation of the Overall Assessment of the Speaker's Experience of Stuttering for school-age children (OASES-S-D). *Journal of Fluency Disorders*, *45*, 27–37. 10.1016/j.jfludis.2015.05.003

Lindström, E., Nilsson, E., Nilsson, J., Schödin, I., Strömberg, N., Österberg, S., … Samson, I. (2020). Swedish outcomes of the Overall Assessment of the Speaker's Experience of Stuttering in an international perspective. *Logopedics Phoniatrics Vocology*, *45*(4), 181–189. https://doi.org/10.1080/14015439.2019.1695930

Ma, Y., An, R., Bin, J., Liu, C., Tsao, Y.-C., & Yaruss, J. S. (2025). Psychometric evaluation of the Simplified Chinese translation of the Overall Assessment of the Speaker's Experience of Stuttering-Adult. *Journal of Fluency Disorders*, *86*, 106170. 10.1016/j.jfludis.2025.106170

Mahesh, S., Pushpavathi, M., Seth, D., Saravanan, S., & Yaruss, J. S. (2024). Adaptation and validation of Overall Assessment of the Speaker's Experience of Stuttering for Adults in Kannada (OASES-A-K). *Folia Phoniatrica et Logopaedica*, *76*(1), 30–38. https://doi.org/10.1159/000531048

Mehdizadeh Behtash, M., Mansuri, B., Salmani, M., Tohidast, S. A., Zarjini, R., & Scherer, R. C. (2022). Development and evaluation of the psychometric properties of the caregiver

burden scale for parents of children who stutter (CBS-PCWS). *Journal of Fluency Disorders*, *73*, 105921. 10.1016/j.jfludis.2022.105921

Millard, S. K., & Davis, S. (2016). The Palin Parent Rating Scales: Parents' perspectives of childhood stuttering and its impact. *Journal of Speech, Language, and Hearing Research*, *59*(5), 950–963. https://doi.org/10.1044/2016_JSLHR-S-14-0137

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., et al. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11 (Placeholder1)136-018-1798-3 (SpringerLink)

Moloney, J., Regan, J., & Walshe, M. (2023). Patient reported outcome measures in dysphagia research following stroke: a scoping review and qualitative analysis. Dysphagia, 38(1), 181-190. 10.1007/s00455-022-10448-y

Morris, M. A., Perez, D., & McCarthy, M. (2021). The value of patient-reported outcome measures in clinical practice. *Health Expectations, 24*(3), 805–812. https://doi.org/10.1111/hex.13215

Nolan, K., MacAuley, Y., Byrne, S., & De Blacam, C. (2025). Patient-reported outcomes in Irish adolescents who were born with cleft lip and palate. The Surgeon, 23(1), 33-37.

Ntourou, K., Oyler DeFranco, E., Conture, E. G., Walden, T. A., & Mushtaq, N. (2020). A parent-report scale of behavioral inhibition: Validation and application to preschool-age children who do and do not stutter. *Journal of Fluency Disorders*, *63*, 105748. 10.1016/j.jfludis.2020.105748

O'Brian, S., Packman, A., & Onslow, M. (2004). Self-rating of stuttering severity as a clinical tool. *American Journal of Speech-Language Pathology*, *13*(3), 219–226. https://doi.org/10.1044/1058-0360(2004/023

O'Brian, S., Packman, A., Onslow, M., Jones, M., & Hearne, A. (2018). Parent-reported severity ratings of stuttering compared with percentage of syllables stuttered in randomized controlled trials of early stuttering treatment. *American Journal of Speech-Language Pathology, 27*(2S), 793–802. https://doi.org/10.1044/2017_AJSLP-17-0126

Ornstein, A. F., & Manning, W. H. (1985). Self-efficacy scaling by adult stutterers. *Journal of Communication Disorders*, *18*(4), 313-320.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Patel, D. A., Sharda, R., Hovis, K. L., Nichols, E. E., Sathe, N., Penson, D. F., ... & Francis, D. O. (2017). Patient-reported outcome measures in dysphagia: a systematic review of instrument development and validation. Diseases of the Esophagus, 30(5).

Quique, Y. M., Ashaie, S. A., Babbitt, E. M., Hurwitz, R., & Cherney, L. R. (2023). Fatigue Influences Social Participation in Aphasia: A Cross-sectional and Retrospective Study Using Patient-Reported Measures. Archives of physical medicine and rehabilitation, 104(8), 1282–1288. https://doi.org/10.1016/j.apmr.2023.02.013

Ranganathan, K., Vercler, C. J., Warschausky, S. A., MacEachern, M. P., Buchman, S. R., & Waljee, J. F. (2015). Comparative effectiveness studies examining patient-reported outcomes among children with cleft lip and/or palate: a systematic review. Plastic and reconstructive surgery, 135(1), 198-211.

Riley, J., Riley, G., & Maguire, G. (2004). Subjective screening of stuttering severity, locus of control and avoidance: Research edition. *Journal of Fluency Disorders*, *29*(1), 51–62. 10.1016/j.jfludis.2003.12.001

Robert, T. B., Robb, M. P., & Choo, A. L. (2025). Measuring anticipation of stuttering: Validation and revision of the English Premonitory Awareness in Stuttering Scale (PAiS-R). *Journal of Fluency Disorders*, *85*, 10.1016/j.jfludis.2025.106142

Salinero, L. K., Ahluwalia, V. S., Barrero, C. E., Wagner, C. S., Pontell, M. E., Magee, L., ... & Taylor, J. A. (2025). Long-term patient-reported outcomes in internationally adopted children with cleft lip and palate. The Cleft Palate Craniofacial Journal, 62(1), 44-50.

Sakai, N., Chu, S. Y., Mori, K., & Yaruss, J. S. (2017). The Japanese version of the overall assessment of the speaker's experience of stuttering for adults (OASES-A-J):

Translation and psychometric evaluation. *Journal of Fluency Disorders*, *51*, 50–59. 10.1016/j.jfludis.2016.11.002

Sitzman, T. J., Allori, A. C., & Thorburn, G. (2014). Measuring outcomes in cleft lip and palate treatment. Clinics in plastic surgery, 41(2), 311-319.

Singer, C. M., Kelly, E. M., White, A. Z., Zengin-Bolatkale, H., & Jones, R. M. (2022). Validation of the Vanderbilt Responses to Your Child's Speech rating scale for parents of young children who stutter. *Journal of Speech, Language, and Hearing Research*, *65*(12), 4652–4666. https://doi.org/10.1044/2022_JSLHR-22-00182

St Clare, T., Menzies, R. G., Onslow, M., Packman, A., Thompson, R., & Block, S. (2009). Unhelpful thoughts and beliefs linked to social anxiety in stuttering: development of a measure. *International Journal of Language & Communication Disorders*, *44*(3), 338–351. https://doi.org/10.1080/13682820802067529

Stefu, J., Slavych, B. K., & Zraick, R. I. (2023). Patient-reported outcome measures in voice: an updated readability analysis. Journal of Voice, 37(3), 465-e27

Tichenor, S. E., & Yaruss, J. S. (2024). Development and validation of a research version of the Overall Assessment of the Speaker's Experience of Stuttering- Adult (OASES-A-R). *Journal of Fluency Disorders*, *80*, 106060. 10.1016/j.jfludis.2024.106060

Tiryaki, N., Özdemir, R. S., Karsan, Ç., & Boyle, M. P. (2023). Turkish adaptation of the self-stigma of stuttering scale (4S): Study of validity and reliability (4S-TR). *Journal of Fluency Disorders*, *78*, 106020. 10.1016/j.jfludis.2023.106020

Tsangaris, E., Riff, K. W. W., Goodacre, T., Forrest, C. R., Dreise, M., Sykes, J., ... & Klassen, A. F. (2017). Establishing content validity of the CLEFT-Q: a new patient-reported outcome instrument for cleft lip/palate. Plastic and Reconstructive Surgery–Global Open, 5(4), e1305.

Ullrich, S., MacIntyre, T., & Leeson, L. (2022). Patient-reported outcome measures in speech and language therapy: Trends and challenges. *International Journal of Language & Communication Disorders, 57*(5), 1010–1025. https://doi.org/10.1111/1460-6984.12709

Uysal, A. A., & Yaşar, Ö. C. (2018). Premonitory awareness scale in children who stutter: PAIS-TR. *Journal of Experimental and Clinical Medicine*, *35*(2), 31–34. https://doi.org/10.5835/jecm.omu.35.02.001

Uysal, H. T., & Köse, A. (2021). The investigation of the validity and reliability of the Turkish version of the Wright and Ayre Stuttering Self-Rating Profile (WASSP). *International Journal of Language & Communication Disorders*, *56*(3), 653–661. https://doi.org/10.1111/1460-6984.12621

Van Ewijk, L., Hilari, K., Pais, A., & Volkmer, A. (2025). Communication patient reported outcome measures for adults with communication disorders: A systematic review of content validity. *International Journal of Language & Communication Disorders, 60*(3), e70050. https://doi.org/10.1111/ijlcd.70050.

Valente, A. R. S., Hall, A., Alvelos, H., Leahy, M., & Jesus, L. M. (2019). Reliability and validity evidence of the Assessment of Language Use in Social Contexts for Adults (ALUSCA). *Logopedics Phoniatrics Vocology*, *44*(4), 166-177.

Valinejad, V., Yadegari, F., & Vanryckeghem, M. (2020). Reliability, validity, and normative investigation of the Persian version of the Communication Attitude Test for Adults Who Stutter (BigCAT). *Applied Neuropsychology: Adult*, *27*(1), 44–48. https://doi.org/10.1080/23279095.2018.1480482

Vanryckeghem, M., & Brutten, G. J. (1992). The Communication Attitude Test: A test-retest reliability investigation. *Journal of Fluency Disorders*, *17*(3), 177–190. 10.1016/0094-730X(92)90010-N

Vanryckeghem, M., & Brutten, G. J. (2011). The BigCAT: A normative and comparative investigation of the communication attitude of nonstuttering and stuttering adults. *Journal of Communication Disorders*, *44*(2), 200–206. 10.1016/j.jcomdis.2010.09.005

Vanryckeghem, M., & Muir, M. (2016). The Communication Attitude Test for Adults who Stutter (BigCAT): A test-retest reliability investigation among adults who do and do not stutter. *Journal of Speech Pathology & Therapy*, *1*(2), 1000110. http://dx.doi.org/10.4172/jspt. 1000110

Vanryckeghem, M., De Niels, T., & Vanrobaeys, S. (2015). The KiddyCAT: A test-retest reliability investigation. *Cross-Cultural Communication*, *11*(4), 10–16. 10.3968/6778

Vanryckeghem, M., Matthews, M., & Xu, P. (2017). Speech Situation Checklist-Revised: Investigation with adults who do not stutter and treatment-seeking adults who stutter. *American Journal of Speech-Language Pathology*, *26*(4), 1129–1140. https://doi.org/10.1044/2017_AJSLP-16-0170

Vanryckeghem M, Mukati SA. The Behavior Assessment Battery: a preliminary study of non-stuttering Pakistani grade-school children. Int J Lang Commun Disord. 2006 Sep-Oct; 41(5): 583–9. 10.1080/13682820500402457

van Tellingen M, Hurkmans J, Terband H, Jonkers R, Maassen B. Music and musical elements in the treatment of childhood speech sound disorders: A systematic review of the literature. Int J Speech Lang Pathol. 2023 Aug;25(4):549-565. doi: 10.1080/17549507.2022.2097310. Epub 2022 Jul 28. PMID: 35900281.

Veerabhadrappa, R. C., Krishnakumar, J., Vanryckeghem, M., & Maruthy, S. (2021). Communication attitude of Kannada-speaking adults who do and do not stutter. *Journal of Fluency Disorders*, *70*, 105866. 10.1016/j.jfludis.2021.105866

Veerabhadrappa, R. C., Vanryckeghem, M., & Maruthy, S. (2021). Communication attitude of Kannada-speaking school-age children who do and do not stutter. *Folia Phoniatrica et Logopaedica*, *73*(2), 126–133. https://doi.org/10.1159/000505423

Veerabhadrappa, R. C., Vanryckeghem, M., & Maruthy, S. (2021). The Speech Situation Checklist—Emotional Reaction: Normative and comparative study of Kannada-speaking children who do and do not stutter. *International Journal of Speech-Language Pathology*, *23*(5), 559–568. https://doi.org/10.1080/17549507.2020.1862301

Watson, J. B., Gregory, H. H., & Kistler, D. J. (1987). Development and evaluation of an inventory to assess adult stutterers' communication attitudes. *Journal of fluency disorders*, *12*(6), 429-450.

Wright, L., Ayre, A., & Grogan, S. (1998). Outcome measurement in adult stuttering therapy: A self-rating profile. *International Journal of Language & Communication Disorders*, *33*(S1), 378–383. https://doi.org/10.3109/13682829809179455

Węsierska, K., Vanryckeghem, M., Krawczyk, A., Danielowska, M., Faściszewska, M., & Tuchowska, J. (2018). Behavior Assessment Battery: Normative and psychometric investigation among Polish adults who do and do not stutter. *American Journal of*

Speech-Language Pathology, *27*(3S), 1224–1234. https://doi.org/10.1044/2018_AJSLP-ODC11-17-0187

Węsierska, K., Yaruss, J. S., Kosacka, K., Kowalczyk, Ł., & Boroń, A. (2023). The experience of Polish individuals who stutter based on the OASES outcomes. *Journal of Fluency Disorders*, *77*, 105991. 10.1016/j.jfludis.2023.105991

Węsierska, M., Świsłocka, N., Węsierska, K., & Boyle, M. P. (2025). Polish adaptation of the Self-Stigma of Stuttering Scale: Scale development and analysis. *Journal of Fluency Disorders*, *85*, 106138. 10.1016/j.jfludis.2025.106138

Yairi, E., & Ambrose, N. G. (2013). *Epidemiology of stuttering: 21st century advances*. *Journal of Fluency Disorders, 38*(2), 66–87.

Yaruss, J. S., & Quesal, R. W. (2004). Stuttering and the International Classification of Functioning, Disability and Health (ICF): An update. *Journal of Communication Disorders, 37*(1), 35–52.

Yaruss, J. S., & Quesal, R. W. (2006). Overall Assessment of the Speaker's Experience of Stuttering (OASES): Documenting multiple outcomes in stuttering treatment. *Journal of Fluency Disorders*, *31*(2), 90–115. 10.1016/j.jfludis.2006.02.002

Yaruss, J. S. (2010). Assessing quality of life in stuttering treatment outcomes research. *Journal of Fluency Disorders, 35*(3), 190–202. https://doi.org/10.1016/j.jfludis.2010.05.010

Yorkston, K. M., Baylor, C. R., Dietz, J., Dudgeon, B. J., Eadie, T., Miller, R. M., & Amtmann, D. (2008). Developing a scale of communicative participation: A cognitive interviewing study. *Disability and rehabilitation*, *30*(6), 425-433.

APPENDIX A

Table S1

Studies included in review and psychometrics properties of the included tools

PROM Group	Tool, Author, Country, Type of Study	Items, Subscales & Response System	Sample Size & Population (e.g., Adults who stutter)	Overall Content Validity	Structural Validity	Internal Consistency	Other Measurement Properties (e.g., Reliability, Construct Validity)	Overall Quality of Study
4S (Self-Stigma)	Self-Stigma of Stuttering Scale (4S) - Original Development Boyle (2013) USA Scale development, structural validation, and psychometric	33 items (reduced from an initial pool of 73). 3 Subscales: 1. Stigma Awareness (14 items) 2. Stereotype Agreement (7 items) 3. Stigma Application/Self-	Total Participants (Phase 2): 291 AWS (after excluding missing data and no non-stuttering control group used). Gender: 181 males and 110 females. Age Range: 18 to 80 years. Mean Age: 41.6 years (SD = 15.1).	Phase 1: generated items from qualitative interviews with 15 PWS and literature review. Items were reviewed by an expert panel of 4 researchers for clarity, relevance, and redundancy and	Principal Axis Factoring with Varimax rotation on 291 AWS. Eliminated 20 cross-loaded items. Extracted 3 factors explaining 40.5% of total variance (Loadings: .38	Good overall internal consistency (Cronbach's α 0.87) Subscales: Awareness (α) = 0.84 Agreement (α) = 0.70 Application	Test-Retest Reliability: Total r = 0.80 over a 2-3 week interval (n = 58). Construct Validity: The 4S correlated negatively with the Generalized Self-Efficacy scale (r = -0.41), Rosenberg Self-Esteem Scale (r =	Very Good (Robust Phase 1 qualitative generation, adequately powered EFA, strong reliability and theoretical construct testing; lacks CFA).

evaluation.	Concurrence (12 items). 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree) Contains 11 reverse-scored items. Higher scores denote higher internalized stigma.	Subsamples: Item generation interviews (n = 15 PWS) Test-retest reliability (n = 58 AWS).	reducing the pool from 73 to 53 prior to EFA.	to .84)	(α) = 0.89	-0.47), and Satisfaction with Life Scale (r = -0.28).
Self-Stigma of Stuttering Scale (4S) Boyle (2015) USA Construct validity and scale validation study.	33 items. 3 Subscales: 1. Stigma Awareness (14 items) 2. Stereotype Agreement (7 items) 3. Stigma Application (12 items).	Total Participants: 354 AWS (after filtering out incomplete/non-stuttering and no control group used). Gender: 200 males, 95 females, 59 unspecified. Age Range: 18 to 84 years. Mean Age: 40.67	Built on Corrigan's established 4-stage psychological model of self-stigma; tool development originally validated through rigorous literature reviews and interviews with PWS in a 2013	PCA with varimax rotation was used. Replicated the original 3-factor structure, which accounted for 39.5% of the overall variance (primary factor loadings ranged from	overall internal consistency = Good (ranged from acceptable to excellent) Overall Cronbach's α = 0.89 Subscales: Awareness (α) = 0.81	Construct Validity: 4S scores showed moderate-to-strong negative correlations with hope (r = -0.51), empowerment (r = -0.52), quality of life (r = -0.40) and social support (r = 0.35) and positive correlations with anxiety (r = 0.42), depression (r = 0.44), and self-	Good (Adequately powered EFA and highly robust correlation testing against 7 separate validated psychological instruments, limited by convenience sampling and factors

	5-point Likert scale (1-Strongly disagree and 5-Strongly agree) Higher scores denote higher internalized stigma.	years (SD = 15.97).	study by the same author.	.39 to .77).	Agreement (α) = 0.75 Application (α) = 0.90	rated speech disruption (r = 0.45). Test-Retest Reliability: NE	explaining <50% variance).
Self-Stigma of Stuttering Scale, Short Japanese Version (4S-J-16) Limura et al. (2022) Japan Translation, validation and reliability study.	16 items (reduced from the original 33) 3 Subscales: 1. Stigma Awareness (5 items) 2. Stereotype Agreement (5 items) 3. Stigma Application (6 items). 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree)	Total Participants: 123 adults who stutter (AWS) and no non-stuttering control group used. Gender: 102 males (82.9%), 21 females (17.1%). Age Range: 18 to 79 years. Mean Age: 36.4 years (SD=13.0). Test-Retest Subsample: n=40 AWS.	The original 33 items underwent forward-backward translation verified by bilingual speech-language pathologists to ensure conceptual equivalence, followed by factor-based item reduction.	Exploratory Factor Analysis (EFA) using principal axis factoring and Promax rotation on the 33 items extracted 3 factors explaining 49.38% of the variance. 16 items with loadings >/= to 0.40 were retained to form the short version.	Good overall internal consistency (Cronbach's α = 0.85). Subscales: Awareness (α) = 0.81 Agreement (α) = 0.85 Application (α) = 0.61	Test-Retest Reliability: Total r = 0.82 over a 1-month interval (n=40) Construct Validity: The 4S-J-16 significantly correlated in expected directions with the Rosenberg Self-Esteem Scale (r=-0.47) and Satisfaction with Life Scale (r=-0.42).	Adequate (Successful clinical shortening with good validity indices, but lacks Confirmatory Factor Analysis and the Application subscale showed borderline internal consistency).

	Three items are reverse scored (Items 2, 4, 11). Higher scores indicate greater self-stigma.						
Self-Stigma of Stuttering Scale (4S-TR), Turkish adaptation Tiryaki et al. (2023) Turkey Translation, validation and reliability study.	28 items (reduced from 33 after EFA). 3 Subscales: 1. Stigma Application (12 items) 2. Stigma Awareness (12 items) 3. Stereotype Agreement (4 items). 5-point Likert scale (1 = Strongly disagree to 5 = Strongly agree) Includes 11 reverse-scored items. Higher scores indicate greater	Total Participants: 350 AWS (after excluding 35 incomplete and no non-stuttering control group used). Gender: 241 males (68.9%), 109 females (31.1%). Age Range: 18 to 65+ years (56% between 18-25). Mean Age: NE (Only categorized age percentages reported). Subsamples: EFA group (n=120)	Rigorous forward-backward translation by bilingual SLPs. Focus group of 10 PWS rated each statement on a 1-5 scale for clarity, relevance, and intelligibility, leading to minor instruction changes.	EFA (n=120) eliminated 5 cross-loaded items, extracting a 3-factor structure explaining 74.76% of total variance. CFA (n=230) confirmed acceptable model fit (chi square/df = 2.38, RMSEA = .078, CFI = .98).	Excellent internal consistency. Overall Cronbach's α = 0.898 Subscales: Application (α) = 0.974 Awareness (α) = 0.886 Agreement (α) = 0.889	Reliability (Test-Retest): Excellent stability at 2 weeks (n=32), r = 0.969 Split-half: Split-half Spearman-Brown = 0.965. Construct validity showed significant negative correlations with self-esteem (r = -0.41) and life satisfaction (r = -0.38).	Very Good (Robust multi-phase structural validation using EFA and CFA, excellent internal and temporal reliability, and strong correlation proofs; slightly limited by convenience sampling bias).

	self-stigma.	CFA group (n=230) Test-retest group (n=32).					
Self-Stigma of Stuttering Scale (4S)-Polish adaptation Wesierska et al. (2025) Poland Adaptation and validation study	33 items 3 Subscales: 1. Stigma Awareness (14 items) 2. Stereotype Agreement (8 items; Q15-22) 3. Stigma Application (11 items; Q23-33) 5-point Likert scale (1 = Strongly disagree to 5 =Strongly agree) Scores are averaged for subscales and total. Higher scores	Total Participants: 108 adults who stutter (AWS) and no non-stuttering control group used. Gender: 56 males (51.9%), 52 females (48.1%). Age Range: 18 to 67 years. Mean Age: 30.56 years (SD=10.86).	Translation checked by a fluency specialist, back-translated, and confirmed by original scale author. Pilot tested via interviews with 7 "double experts" (PWS who are also SLPs/group leaders) who confirmed clinical usefulness and accurate translation.	NE (Explicitly stated by authors that the sample size of 108 was too small to conduct a factor analysis).	High internal consistency. Overall Cronbach's α = 0.93 Subscale: Awareness (α) = 0.89 Agreement (α) = 0.77 Application (α) = 0.92	Construct Validity: 4S scores showed significant negative correlations with life satisfaction and positive correlations with perceived difficulties associated with stuttering. Test-Retest Reliability: NE.	Adequate (Strong translation methodology, internal consistency, and concurrent construct evaluations; limited by small sample size preventing structural validation and lacking test-retest data).

		indicate higher internalized stigma.						
BAB	Behavior Assessment Battery (BAB)-Polish Version Węsierska et al. (2018) Poland Normative investigation and psychometric	204 items across 4 independent subtests: 1. Speech Situation Checklist-Emotional Reaction (SSC-ER): 51 items. 2. Speech Situation Checklist-Speech Disruption (SSC-SD): 51 items. 3. Behavior Checklist (BCL): 68 items. 4. Communication Attitude Test for Adults Who Stutter (BigCAT): 34 items. Mixed systems	Total Participants: 274 adults. PWS Group (n=123): 93 males, 30 females; Age range 17-70 years; Mean age = 29.83 (SD=10.59). PWNS Control (n=151): 76 males, 75 females; Age range 18-67 years; Mean age = 29.98 (SD=10.97).	Rigorous translation process utilizing WHO guidelines. Included forward translation by two Polish bilinguals, committee synthesis, back-translation by a blinded native English speaker, and final verification by the original BAB author.	NE (The authors explicitly relied on the previously established structural/factor validity of the original English BAB; no new EFA/CFA was performed on the Polish data).	Excellent internal consistency using Cronbach's α for both groups. PWS: SSC-ER=0.97 SSC-SD=0.98, BCL=0.95, BigCAT=0.93. PWNS: SSC-ER=0.96, SSC-SD=0.97, BCL=0.83, BigCAT=0.83.	Test-Retest Reliability: NE Known-Groups & Construct Validity: PWS scored significantly higher than PWNS on all 4 subtests ($p<0.001$) with massive effect sizes ($d = 1.44$ to 2.62). Strong inter-test correlations confirmed construct alignment.	Good (Extremely robust normative sample, rigorous translation, and internal consistency/known-groups validity; limited by a lack of new structural factor analysis and temporal reliability data).

	validation.	depending on subtest. SSC-ER & SSC-SD: 5-point Likert scale (1 = not at all to 5 = very much). Scores: 51-255 BCL: Binary "Yes/No" indicating presence of coping response. (Total score is the number of 'Yes' responses: 0-68). A secondary 5-point frequency scale is used clinically but not scored in the total. BigCAT: Binary True/False forced-choice. Scores: 0-34.						
CAT	CAT Vanryckeghem & Brutten (1992) United	35 items Unidimensional scale Binary true or false response format (1-	44 Dutch speaking school aged children diagnosed as stuttering (39 male and 5 females with the age range	NE	NE	NE	Test-retest Reliability: by using Pearson product moment correlation coefficient was computed	Doubtful

States/ Belgium Reliability study	negative attitude and 0- positive attitude towards speech)	of 7 to 13 years)				between the first and second, second and third and first and third test scores. the scores were +0.83 (1 week interval), +0.81(11-week interval) and +0.76 (12 week interval). all were significant at the 0.0001 level. Good inter-item reliability.	
BigCAT Vanryckeghem & Brutten (2011) United States Normative and Comparative study	35 items Unidimensional scale Binary true or false response format (1 = negative belief and 0 = does not imply negativity)	96 adults who stutter (PWS: 76 male and 20 female in the age range of 18-61) and 216 adults who do not (PWNS: 111 male and 105 female in the age range of 18-72).	NE	NE	Total α: 0.89 for AWS and 0.86 for AWNS	Discriminant validity: Significant mean difference between AWS (M=28.01, SD=5.79) and AWNS (M=3.32, SD=3.56), t = 44.13, p = 0.000, with a very large effect size (d = 5.36).	Inadequate.

Dutch KiddyCAT Vanryckeghem et al., 2015 Belgium Test-retest reliability study	12 items Unidimensional Binary yes/or no response format	34 CWS and 42 CWNS	NE	NE	NE	Test-retest Reliability: Pearson r=0.90 for 34 CWS and 0.67 for 42 CWNS with an interval of 7-12 days.	Doubtful
English BigCAT Vanryckeghem & Muir (2016) USA Test-retest reliability investigation.	35 items unidimensional Binary true or false response format (1 = negative attitude and 0 = positive attitude towards speech)	33 PWS (18-54 yrs) and 50 PWNS (18-58yrs)	NE	NE	NE	Test-retest Reliability: Pearson r=0.80 for 33 PWS and 0.76 for 50 PWNS with an interval of 5-7 days.	Doubtful due to less sample size for assessing the test-retest reliability (33 PWS).

Communication Attitude Test for Adults Who Stutter (BigCAT), Persian version Valinejad et al. (2018) Iran Reliability, validation, and normative investigation.	34 items (Initially 35 items, but Item 6 was deleted due to low item-total correlation) No subscales Unidimensional Dichotomous True/False scale. 1 = A negative attitude/belief & 0 = A positive attitude/belief. Total scores range from 0 to 34.	Total Participants: 180 adults. PWS Group (n=90): 62 males, 28 females; Age range 18-51 years; Mean age = 26.18 (SD=6.11) PWNS Control Group (n=90): 45 males, 45 females; Age range 18-50 years; Mean age = 24.86 (SD=5.41). Test-Retest Subsample: n=17 PWS. Criterion Validity Subsample: n=86 PWS.	8 Persian-speaking SLPs evaluated items. Content Validity Ratio (CVR) was calculated for each item; all scored >0.75, confirming strong expert content validity.	NE (Exploratory or Confirmatory Factor Analysis was not performed to confirm unidimensionality).	Kuder-Richardson Formula 20 (KR-20) indicated good internal consistency: 0.88 for PWS and 0.83 for PWNS.	Test-Retest Reliability: ICC = 0.89 over a 7-10 day interval (n=17). Criterion Validity: Pearson r = 0.79 ($p < 0.001$) against the Erickson S-24 scale in 86 PWS. Known Groups Validity: Significantly differentiated PWS (Mean 24.33) from PWNS (Mean 7.05), $p < 0.001$.	Good (Strong normative group size, criterion validity, and internal consistency; limited by severely underpowered test-retest sample and lack of structural factor analysis).
CAT-K (Kannada)	30 items and no subscale.	100 CWS and 293 CWNS	Forward-Backward translation and	Not evaluated (no factor analysis was	Cronbach's α: 0.86 for CWS and 0.58 for	Test-retest Reliability: ICC= 0.92 for CWS and	Doubtful as there was small sample

Veerabhadrappa et al., 2020 India Cross cultural adaptation nad validation study	Binary True or false response format.		expert opinions were taken.	conducted and item no 6 was deleted purely on item total corrections.	CWNS.	0.86 for CWNS (10% of the sample with 7=10 days of interval). Construct validity: Discriminant analysis were done and it classified 98% of CWS accurately and 1005 CWNA (overall 99.5%).	size for the test-retest reliability (10).
BigCAT-K Veerabhadrappa et al., 2021 India Adaptation and validation study	34 items Unidimentionality scale Binary true or false response format (1- negative attitude and 0- positive attitude towards speech)	317 AWS (163 male and 154 female with the mean age of 23.69) and 100 AWNS (89 male and 11 female with the mean age of 23.64)	NE	NE	AWS Total α = 0.72 AWNS Total α = 0.86	Construct validity: Correctly identified 98.6% of the original grouped cases through discriminant analysis. Test-retest reliability: High for AWNS (r=0.75) and excellent for AWS (r=0.94). Inter rater	Doubtful

							reliability: Good (r=0.90)	
CBS-PCWS	Caregiver Burden Scale for Parents of Children Who Stutter (CBS-PCWS) Mehdizadeh Behtash et al. (2022) Iran Development and psychometric evaluation conducted in 2 phases.	22 items total. 5 subscales: 1. Child care challenges (5 items) 2. Psychological & emotional burden (4 items) 3. Feelings of guilt & anxiety (4 items) 4. Personal & social burden (6 items) 5. Financial burden (3 items). 5-point Likert scale (1 = Never to 5 = Always). Total score ranges from 22 to 110.	Total Participants: 230 parents of children who stutter (CWS) and 12 SLPs. & no non-stuttering control group used. Phase 1 (Item Generation): 15 parents (10 F, 5 M); Age range 28–48; Mean age = 34.2. Phase 1 (Face Validity): 10 parents of CWS. Phase 1 (Content Validity): 12 expert SLPs. Phase 2 (Factor Analysis): 205 parents of CWS (155 F, 50 M); Age range 22–63; Mean age = 35.88 (SD=6.3). Phase 2 (Reliability): Subsample of 40	Qualitative Face Validity: 10 parents confirmed appropriate difficulty, relevancy, and lack of ambiguity. Quantitative Face Validity: Impact Score evaluated; all retained items scored > 1.5. Qualitative Content Validity: 12 SLPs reviewed wording, scaling, and grammar, leading to minor revisions. Quantitative	EFA (KMO = 0.902) extracted 5 factors that cumulatively explained 63.89% of the total variance.	Cronbach's α = 0.93 for total scale; subscale alphas ranged from 0.73 to 0.84.	Test-Retest Reliability: ICC = 0.97 total; 2 weeks apart; n=40 Measurement Error (SEM): 2.11. Minimal Detectable Change (MDC): 5.84.	Adequate (Strong multi-phase qualitative and quantitative item generation; adequately powered EFA sample; however, lacks Confirmatory Factor Analysis and test-retest n=40 is slightly underpowered).

			parents.	Content Validity: 12 SLPs assessed items; items retained if Content Validity Ratio (CVR) > 0.56 and Content Validity Index (CVI) > 0.79.				
ICA	Inventory of Communication Attitudes (ICA) Watson et al. (1987) USA Development and psychometric	144 total ratings The inventory consists of 36 situational items across 13 subscales (e.g., Telephone, Authority figures, Strangers). Each situation is rated on 4 response scales: Frequency, Tension, Avoidance, and Enjoyment. Evaluated. 5-point	Total Participants: 189 adults. Phase I (AWS Group): n=107 (81 males, 26 females); Age range 17-69 years; Mean age = 28.5 (SD=10.9). Phase II (AWS Group): n=26 (20 males, 6 females); Age range 18-68 years; Mean age = 29.5 (SD=12.2). Phase II (AWNS	Evaluated. Items were generated based on a tripartite attitudinal model (affective, behavioral, cognitive) utilizing 100 speaking situations derived from literature and clinical interviews, refined by expert	Evaluated. Pearson correlations were calculated to examine interrelationships among the 4 response scales, showing they measure related but distinct components (r = 0.52 to 0.78).	Evaluated. Cronbach's α was excellent. Phase I alphas were 0.94, 0.97, 0.96, and 0.96 for Frequency, Tension, Avoidance, and Enjoyment respectively.	Test-Retest Reliability: (Phase I: r = 0.8$ to 0.94, 3-4 weeks apart, n=28. Phase II: r = 0.86 to 0.96). Discriminant/Known-Groups Validity: Evaluated. AWS scored significantly higher (more negative) than AWNS across all	Good (Robust multi-phase development, large samples, and comprehensive evaluation of multiple validity parameters).

	evaluation conducted in 2 phases.	Likert scales (1 to 5) Frequency: 1= never to 5 = always. Tension: 1 = none to 5 = very tense. Avoidance: 1 = never to 5=always Enjoyment: 1 = always to 5 = never. Higher scores indicate more negative attitudes/behaviors.	Control): n=56 (30 males, 26 females); Age range 18-56 years; Mean age = 30.2 (SD=9.1).	consensus.			4 response scales and all 13 situations ($p < 0.01$).	
ISACS	The Impact Scale for Assessment of Cluttering and Stuttering (ISACS) Kelkar et al. (2024) India	100 items total evaluated across 4 Subscales mapped to the WHO ICF: 1. Body functions (12 items) 2. Personal contextual factors (28 items) 3. Activities and Environmental factors (42 items) 4. Participation / QoL (18 items). Evaluated. 5-point	Total Participants: 149 adults. PWF Group (n=52): 47 PWS, 5 PWC; 41 males, 11 females; Mean age = 25.48 (SD=10.21). Typical Speaker Control (n=52): 38 males, 14 females;	Designed across 4 drafts evaluated subjectively by 2 SLPs and 1 Clinical Psychologist. Piloted on 30 individuals. Translated from English to Marathi via rigorous forward/back-translation with	NE (The authors explicitly stated that the "data was inadequate for a factor analysis" due to sample size constraints).	Internal consistency ranged from adequate to excellent depending on the subscale. Cronbach's α (PWF): I=0.64, II=0.96, III=0.95, IV=0.94.	Test-Retest Reliability: NE Other Measurement Properties (Known-Groups & Concurrent):. ANOVA confirmed significant total score differences between PWF,	Adequate (Strong content generation grounded in the ICF model, rigorous translation/cultural adaptation, and excellent known-groups discriminatio

	Development, translation, and preliminary psychometric evaluation.	Likert-type scales (1 to 5) with varying anchors. Generally, 1 = "low impact" and 5 = "high impact". Three items (1, 3, 7 of Subscale I) are reverse-scored. • Uses a "Percent Score" formula to account for items marked "not applicable".	Mean age = 26.55 (SD=10.01). Significant Other Person (SOP) Group (n=45): 16 males, 29 females; Mean age = 40.64 (SD=13.40).	native-speaker consensus. Wilcoxon signed ranks test proved statistical equivalence between language versions.		Split-half (PWF): I=0.60, II=0.91, III=0.81, IV=0.87.	Typical Speakers, and SOPs ($p<0.001$). ISACS(A) positively correlated with SSI-4 stuttering severity ($r=0.35$, $p=0.03$).	n; limited by small sample size preventing EFA/CFA and lack of test-retest data).
LCB	Locus of Control of Behaviour (LCB) Scale Craig et al. (1984) Australia Development, construct validation, and predictive reliability.	17 items total (reduced from an initial pool of 62 items). (No subscales; empirically proven as a unidimensional scale). Evaluated. 6-point Likert scale (0 to 5). • 0 = Strongly disagree • 1 = Disagree • 2 = Slightly disagree •	Total Participants: 303 adults. Stutterers (PWS): n=78; 62 males, 16 females; Mean age = 24.3 (SD=8.4). Normal Controls (PWNS): n=154 university students; 78 males, 76 females; Mean age = 22.8 (SD=8.1). Agoraphobics: n=40; 7 males, 33 females; Mean age = 36.3 (SD=9.8). Social Phobics: n=31; 19 males, 12	An initial pool of 62 items was submitted to 5 independent judges. Items were only retained if there was absolute consensus on their categorization (Internal vs. External), leaving 40 items. Item-whole correlations further refined	Principal factor analysis with varimax rotation was performed. Extracted one major factor accounting for 36.5% of the total variance, confirming the tool is unidimensional.	Strong internal consistency. Total Cronbach's α = 0.79. Split-half reliability = 0.73.	Reliability (Test-Retest): Very stable: $r=0.90$ over 1 week (n=67 normal); $r=0.73$ over 6 months without treatment (n=55 stutterers). Other Measurement Properties: Predictive: Correctly predicted post-therapy relapse in stutterers with	Good (Exceptionally thorough psychometric testing across multiple distinct clinical and normative populations; highly predictive clinical utility).

		3 = Slightly agree • 4 = Agree • 5 = Strongly agree 7 items are reverse-scored. Total scores range from 0 to 85. Higher scores indicate greater externality (belief that destiny/others control behavior). Lower scores indicate greater internality (perceiving personal control).	females; Mean age = 29.4 (SD=8.6).	the scale to the 17 most robust items.			84% accuracy. Known-Groups: Differentiated clinical groups from normal ($p<0.01$). Concurrent: Correlated r=0.54 with Rotter's I-E scale. Divergent: Zero correlation (r=-0.05) with social desirability bias.	
OASES	OASES Yaruss & Quesal, 2006 United States Development and validation study	100 items 4 dimensions: a)General Information b)Your Reactions to Stuttering c)Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores	173 adults who stutter	Items generation: based on WHO ICF framework, literature review, and focus groups. Panel expert: multidisciplinary review.	NE	Subscales α: 0.92-0.97 for 4 dimensions.	Test-retest reliability: subset of N=14.	Doubtful

	indicate that greater adverse impact of stuttering.					
OASES-A-D Koedoot et al., 2011 Netherlands Psychometric evaluation study	100 items 4 dimensions: a)General Information b)Your Reactions to Stuttering c)Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores indicate greater adverse impact of stuttering.	138 adults who stutter	Items generation: translated from original English version. Panel expert: translation and adaptation completed.	NE	Subscales α: 0.84-0.96 for 4 dimensions.	Construct validity: known-group validity limited by severely skewed distribution.	Doubtful
OASES-S-D Lankman et al., 2015 Netherlands Validation	60 items 4 dimensions: a)General Information b)Your Reactions to Stuttering	152 people. (101 school-age children who stutter & 51 control children) Age Range = 7-	Items generation: translated from original English version. Panel expert:	NE	Total α: 0.915 Subscales α: 0.54 for Section I (General Information).	Construct validity: significantly differentiated between children who stutter and controls.	Doubtful

and evaluation study	c)Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores indicate greater adverse impact of stuttering.	12yrs	backward and forward translation.				
OASES-A-J Sakai et al., 2017 Japan Translation and validation study	100 items 4 dimensions: a)General Information b)Your Reactions to Stuttering c)Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores indicate that greater adverse impact of stuttering.	200 adults who stutter (mean age 32.5)	Items generation: translated from original English version. Panel expert: cultural adaptation and translation	NE	Subscales α: 0.80-0.98 for 4 dimensions.	Test-retest reliability: (interval time not specified) subset of N=14. Construct validity: 1) positive correlation with modified Erickson scale.	Doubtful

Swedish OASES (OASES-A, T, S) Lindstrom et al., 2020 Sweden Validation study	OASES-S: 60 items OASES-T: 80 items OASES-A: 100 Items 4 dimensions: a) General Information b) Your Reactions to Stuttering c).Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores indicate that greater adverse impact of stuttering.	139 people who stutter (80 adults, 27 teenagers & 32 children with stuttering)	Items generation: translated from original English version. Panel expert: backward and forward translation.	NE	Subscales α: 0.29 (Teenager Section I) and 0.54 (Child Section I) for 4 dimensions.	Construct validity: International comparisons made with normative data.	Doubtful
OASES-PL(Polish)OASES-S, OASES-T AND	OASES-S: 60 items OASES-T: 80 items	154 people who stutter (55 children aged 5-12, 41 teenagers aged 13-17 and 58 adults	Items generation: translated from original English	EFA: conducted. Kaiser-Meyer-Olkin (KMO)	Total α: 0.93-0.97 Subscales α: 0.816-0.971	Construct validity: 1) comparisons between US and Polish normative	Doubtful

OASES-A Wesierska et al., 2023 Poland Translation and validation study	OASES-A: 100 items 4 Sub scales includes: General Information, Your Reactions to Stuttering, Communication in Daily Situations, and Quality of Life 5-point Likert scale Higher scores indicate that greater adverse impact of stuttering	aged 18+ with stuttering)	version. Panel expert: forward and backward translation Face validity: assessed with target population.	measure was suboptimal (0.462) for the teenage group.	for 4 dimensions.	data.	
OASES-A-K Mahesh et al., 2024 India Adaptation and validation study	100 items 4 dimensions: a)General Information b)Your Reactions to Stuttering c)Communication in Daily Situations d) Quality of Life 5-point Likert scale.	51 adults who stutter	Items generation: translated from original English version. Panel expert: backward and forward translation, reviewed for linguistic and	NE	Subscales α: 0.83-0.97 for 4 dimensions.	Test-retest reliability: subset of N=10.	Doubtful

	Higher scores indicate that greater adverse impact of stuttering.		cultural equivalence.				
OASES-A-R Tichenor & Yaruss, 2024 United States Development and validation study	25items 4 dimensions: a)General Information b)Your Reactions to Stuttering c)Communication in Daily Situations d) Quality of Life 5-point Likert scale. Higher scores indicate that greater adverse impact of stuttering.	315 development sample 156 validation sample(adults who stutter)	Items generation: selected from the originally validated OASES-A pool.	IRT (Item Response Theory): Graded response modelling used to identify discrimination values of each item.	Subscales α: 0.83-0.93 for 4 dimensions.	Construct validity: High concurrent validity correlations with original 100-item OASES-A.	Doubtful
(OASES-A-SC) Ma et al. (2025)	100 items total. 4 Subscales (Sections): I. General Information (20 items) I	Total Participants: 346 adults who stutter (AWS). No non-stuttering control group used. Gender: 201 males	Translation underwent expert panel review (10 experts), back-translation, and	NE (Factor analysis was not conducted; the study relied on the pre-established	Excellent internal consistency. Cronbach's α: Overall=.98, Sec I=.86, Sec	Test-Retest Reliability: ICCs ranged from .92 to .97 across sections over a 1 week to 3-month interval	Good (Strong reliability and translation methodology, but limited

	China Translation and psychometric evaluation.	I. Reactions to Stuttering (30 items) III. Communication in Daily Situations (25 items) I V. Quality of Life (25 items). 5-point Likert-type scale 1 = Least negative impact & 5 = Most significant negative impact. Total impact is categorized: Mild (1.0-1.49) Mild-to-moderate (1.5-2.24) Moderate (2.25-2.99) Moderate-to-severe (3.0-3.74) Severe (3.75-5.0).	(59.1%), 139 females (40.9%), 6 not reported. Age Range: 18 to 49 years. Mean Age: 27.1 years (SD=6.3). Test-Retest Subsample: n=28.	pilot pretesting with 7 AWS. Construct/Content validity was confirmed via a follow-up survey of 10 participants who unanimously agreed it captured the full scope of their stuttering experience.	4-section structure of the original English OASES-A).	II=.94, Sec III=.96, Sec IV=.96.	(n=28). Item-Level Analysis: 9 items showed ceiling effects (>30% scored 5); 6 items showed floor effects (>30% scored 1).	by lack of structural factor analysis and a slightly underpowered test-retest sample size).
9-point Self-Rating	9-point Self-Rating Scale (Stuttering	Single item (stuttering severity).	Total: 10 adults who stutter (AWS). Gender/Age/Mean:	NE	NE	NE	Reliability: High agreement between speaker and SLP for 9 of	Low (Small sample, descriptive, not a

Scale	Severity) O'Brian et al. (2004) Australia Reliability study (agreement between speaker and SLP).	Evaluated. 9-point scale. 1 = No stuttering to 9 = Extremely severe stuttering.	Not reported (NE).				10 speakers. Agreement between initial ratings and recordings 6 months later for 8 of 10 speakers.	comprehensive psychometric tool validation).
r-SASBS	Revised Scale of Avoidance and Struggle Behaviours in Stuttering (r-SASBS) Hancer & Tokgoz-Yilmaz (2025) Turkey Scale revision, validity, and reliability	14 items total (reduced from an initial pool of 42, to 30, to 16, and finally to 14). 2 Subscales (Factors): 1. Avoidance (8 items) 2. Struggle (6 items). 5-point Likert scale (1 = Never 5 = Always)	Total Participants (Main Study): 440 adults. PWS Group (n=365): 280 males (76.7%), 85 females (23.3%). Control Group / PWNS (n=75): 53 males (70.7%), 22 females (29.3%). Overall Gender: 333 males (75.7%), 107 females (24.3%). Age Range: 18 to 57 years (Ranges: 18-27 [36.8%], 28-	In the preliminary phase, items were generated from compositions by 15 PWS and literature review. Content Validity Ratio (CVR) and Content Validity Index (CVI) were calculated using 10	Exploratory Factor Analysis (EFA) on 165 PWS extracted 2 factors explaining 77.52% of the variance (KMO = 0.79, Bartlett's $p < 0.001$). Confirmatory Factor Analysis (CFA) on 200 PWS	Internal consistency was reported as excellent, with the total scale Cronbach's $\alpha = 0.98$.	Reliability (Test-Retest): Evaluated. The test-retest reliability coefficient over a 20-27 day interval was 0.97 (n=110). Discriminant Validity: Evaluated. The scale significantly differentiated between PWS (Mean=37.03) and	Very Good (Robust multi-phase development, excellent expert panel CVR/CVI, adequately powered EFA and CFA, strong discriminant and reliability testing).

	study conducted across multiple phases.		37 [42.1%], 38-47 [19.5%], 48-57 [1.6%]). Mean Age: NE (Only age ranges and percentages were provided for the main sample). Subsamples: EFA dataset included 165 PWS CFA dataset included 200 PWS Discriminant validity included 75 PWS and 75 PWNS Test-retest dataset included 110 participants.	experts (CVR=0.96, CVI=0.93). During revision, 9 experts yielded CVI=0.95 and CVR=0.98.	confirmed excellent model fit after 2 modifications (CMIN/DF = 2.112, CFI = 0.98, GFI = 0.902, RMSEA = 0.075).		PWNS (Mean=1.29) (t(438)=22.50, p < 0.001, Cohen's d = 3.67$).	
SA	Self-Assessment (SA) Scale Huinck & Rietveld (2007) Netherlands Validity of SA Scale.	Single-item (SA) scale. Evaluated. 10-point scale (1 = Fluency to 10 = Stuttering)	Total: 15 adults who stutter (AWS). Gender/Age/Mean: Not reported (NE).	NE	NE	NE	Criterion Validity: Correlations calculated between SA scale and objective measures (%SS and SPM) and self-evaluation/listener ratings.	Low (Very small sample size; study focused on relating scores to objective measures rather than full tool validation).

SNA / FNA- SA	Speech Naturalness (SNA) and Feel Naturalness (FNA) Self-Rating Scales Finn & Ingham (1994) USA Reliability and validity assessment of single-item scales during varied speaking conditions.	Single item for Speech Naturalness (SNA). Single item for Feel Naturalness (FNA). 9-point equal appearing interval scales (1 to 9). SNA: 1 = highly natural sounding to 9 = highly unnatural sounding FNA: 1 = highly natural feeling to 9 = highly unnatural feeling	Total Participants: 12 adults who stutter (AWS). No non-stuttering control group used. Gender: 10 males, 2 females. Age Range: 21 to 50 years. Mean Age: 31.8 years.	NE	NE	NE	Intra-judge (Test-Retest) Reliability: Exact/adjacent agreement across 3 rating occasions was 85-98% for SNA and 88-100% for FNA. Concurrent Validity: Self-ratings of SNA correlated significantly with independent listeners' ratings (r = 0.79 to 0.89).	Adequate (Strong experimental control, but limited by small n=12 sample and single-item scope).
PAiS	PAiS (Premonitory Awareness in Stuttering Scale) Cholin et al., 2016	12 items (originally 13) Unidimensional 4-point Likert scale (0 = not at all true, 3 = very much true).	21 adults who stutter (AWS) and 21 matched controls (ANS)	Content validity: Derived from PUTS and refined through pilot study (n = 15). Construct validity: Negative	NE	Total α = 0.76 (after excluding item 13).	Items 2, 5, and 9 correlated significantly with stuttering frequency.	Good (Pilot tested and formally validated against objective severity measures).

Germany Development and Preliminary Study			correlation with % stuttered syllables (r = -0.519). Discriminant validity: Significant difference between AWS (M = 17.71) and ANS (M = 3.67).				
PAiS-TR Uysar & Yasar (2018) Turkey Adaptation and measurement of psychometric properties study	Not mentioned in the article. but mentioned that an additional one question was included between the 10 and 11th question from the original item. so 13 items.	60 Children with stuttering(CWS: 48 male and 12 female) and 60 control group. age range of 6-16 yrs of age.	Low (+/L)	Low (+/L)	Total α = 0.98	Test-retest Reliability (2 week interval): r = 0.99 and p is less than 0.01. Construct validity: SSI-4 (r = 0.927) and SS% (r = 0.92)	Doubtful

	PAiS-R (English Premonitory Awareness in Stuttering Scale) Bies et al., 2025 United States Validation study in English	12 items initially and revised to 8 items 5-Point Likert scale (0 = Not at all true to 4 = Very much true)	78 Adults who stutter (AWS)	Sufficient (+): Items revised for relevance; "itchy" and "avoidance" items removed to ensure conceptual focus	Sufficient (+): PCA 1-factor solution explained 54.5% of variance; all loadings >_0.30	Total α = 0.88 (after removing 4 items)	Test-retest Reliability: ICC=0.86 (PAiS-R) Construct Validity: r=0.33(PAiS)	Good
Palin PRS	Palin PRS Millard & Davis (2016) UK Psychometric evaluation and tool refinement study	19 items refined from 25 items. 3 Factors: 1- Impact of stuttering on the child 2- Severity of stuttering and impact on parents 3- Parent knowledge and confidence 10-cm visual analogue scale (0-10) or online	A total of 259 questionnaires were identified and completed by Parents prior to the therapy (146 mothers and 113 fathers). 259 Children with a male to female ration of 3:1 (194 males and 65 female) in the age range of 2.6 yrs to	Derived from a prior Delphi study where parents were generated 129 statements and reduced based on importance and conscences.	Exploratory Factor Analysis / PCA with varimax and oblimin rotation. 3 factors explained 57% of variance. All retained items had factor loadings > 0.4	The internal consistency was very good. Cronbach's α = 0.865 for Factor 1, 0.863 for Factor 2, 0.838 for Factor 3.	The study established standard normative data. It calculated percentile ranks and stanine scores to help classify parental responses.	Adequate. The sample size of 259 was large enough to properly assess structural validity. However, the data came from looking back at old clinical files (retrospective file audit),

	selection (0=worst and 10=best)	14.6yrs.					which can cause some bias.
Palin PRS-TR (Turkish) Cangi et al. (2025) Türkiye Cross-cultural adaptation, validity, and reliability Study.	19 items 3 factors (identical to original). Parents used an 11-point rating scale ranging from 0 to 10.	Total Parents: 193 (111 mothers, 82 fathers); Age range = 23–62; Mean age = 39.63 (SD=5.72). Children (n=118): 83 males, 35 females; Age range = 3;0 to 14;1; Mean age = 7.21 (SD=2.84).	The tool was translated (forward and backward translation) into Turkish. In a pilot study, 5 parents gave the scale a perfect score for clarity and relevance.	Confirmatory Factor Analysis (CFA) fit indices: chi square/df = 2.67, CFI = 0.904, IFI = .905, RMSEA = .093 after minor item reordering (supported the original 3-factor structure after making minor changes to the order of a few items).	The reliability scores were excellent. All measures of internal consistency (Cronbach's α, McDonald's omega) were high across the three factors (Cronbach's α = 0.858, 0.833, 0.910; McDonald's ω = 0.879, 0.854, 0.920; Guttman's λ6 = 0.883, 0.866, 0.909 across the 3 factors).	Test-Retest Reliability: The scores were stable when tested again 3 weeks later (Pearson r = 0.636 for F1, 0.708 for F2, 0.724 for F3; ~3 weeks interval; n=118). Measurement Invariance: The scale worked equally well for both mothers and fathers (Supported configural, metric, and scalar invariance across mothers and fathers). Item Response Theory: The items showed a strong	Adequate. The study used strong statistical tests and a good sample size (n=193). However, the test-retest score for the first factor was slightly lower than the optimal threshold (CFA indices fall slightly short of optimal thresholds, and F1 test-retest r < 0.70)

ability to discriminate between different levels of parental concern (Graded Response Model showed discrimination parameters ranging from 0.707 to 6.44).

Convergent Validity: they calculated composite reliability to measure convergent validity. The composite reliability values ranged from 0.819 to 0.918, which provided strong evidence of convergent validity (Composite reliability = 0.842, 0.819, 0.918).

Palin Parent Rating Scales, Persian version (Palin PRS-P) Barzegar Bafrooei et al. (2025) Iran Scale translation, cross-cultural adaptation, and psychometric evaluation.	19 items total 3 Subscales (Factors): 1. Impact of stuttering on the child (7 items) 2. Severity of stuttering and its impact on parents (7 items) 3. Parents' knowledge and confidence in managing stuttering (5 items). Scored on an 11-point Visual Analogue Scale (VAS) from 0 to 10. For Factors 1 & 2: Higher scores indicate a higher negative effect/impact of stuttering on the child and parents. For Factor 3 (the final 5 items): These are positively worded,	Total Participants: 133 parents of children who stutter (CWS). No control group used. Parent Gender: 111 mothers (83.5%), 22 fathers (16.5%). Parent Age Range: 23 to 51 years (Mean=34.62, SD=5.25). Child Demographics: 95 boys, 38 girls; Age range 36 to 71 months (Mean=55.45 months, SD=10.59). Test-Retest Subsample: n=30 parents.	Standard forward-backward translation protocols were used. Evaluated by an expert panel of 11 SLPs with >5 years of experience. Content Validity Ratio (CVR) and Content Validity Index (CVI) were calculated, confirming excellent item necessity and relevance.	Exploratory Factor Analysis (EFA) extracted 3 factors explaining 61.16% of total variance (KMO=0.803). Confirmatory Factor Analysis (CFA) verified the model fit: RMSEA=0.086, CFI=0.963, TLI=0.957, SRMR=0.082.	Excellent internal consistency using Cronbach's α. • Total Scale: 0.82 • Factor 1: 0.89 • Factor 2: 0.88 • Factor 3: 0.77.	Reliability (Test-Retest): Excellent stability over a 2-week interval (n=30). Total ICC = 0.97. Subscale ICCs ranged from 0.94 to 0.96. Other Measurement Properties (Floor/Ceiling Effects): No significant floor or ceiling effects detected (all were $<15\%$).	Good (Robust cross-cultural adaptation using quantitative CVI/CVR metrics, excellent reliability; limited by using the same sample for both EFA and CFA, and lack of external criterion validity testing).

		where higher scores denote higher knowledge and confidence.						
S-Scale	Communication Attitudes Scale (S-Scale) Erickson (1969) USA Scale development and empirical derivation.	39 items total (empirically reduced from a 466-item inventory). (No subscales; unidimensional scale). Dichotomous True / False response format. Items are scored "1" if answered in the designated direction (the direction stutterers answered significantly more often than non-stutterers). • "0" for the opposite response.	Total Participants: 264 college students. PWS Criterion Group (n=120): 120 males, 0 females. Subdivided into Group A (n=50; Age range 18-39; Mean=22.14) and Group B (n=70; Age range 18-38; Mean=22.43). PWNS Normative Group (n=144): 144 males, 0 females. Subdivided into Group A (n=100; Age range 17-29; Mean=19.84) and Group B (n=44; Age range 18-37;	Initial 1060 items pooled from 14 published personality/attitude scales, clinical observations, and literature. Screened for 6th-grade reading level and social desirability balance, resulting in 466 items tested empirically to find the final 39.	NE (Relied on empirical criterion group differentiation rather than formal factor analysis).	Excellent internal consistency. Split-half reliability (Spearman-Brown) = 0.92; Kuder-Richardson (KR-20) coefficient = 0.91.	Concurrent Validity: Correlated with discomfort in social conversation (r=0.66), self-rated stuttering severity (r=0.46), and clinician severity ratings (r=0.35). Discriminant Validity: High S-Scale scorers identified with significantly different adjectives (e.g., "Insecure", "Tense") compared to low scorers ("Confident",	Adequate (Massive normative sample [N=264] and robust empirical derivation methodology; limited by the exclusion of females and lack of factor analysis/test-retest data).

		Scores range from 0 to 39. Higher scores indicate greater deviation from "normal" attitudes.	Mean=21.64).				"Relaxed"). Test-Retest Reliability: NE	
S-Scale / Erickson	Modified Erickson Scale (S24 Scale) Andrews & Cutler (1974) Australia Refinement, reliability, and responsiveness-to-change study.	24 items total (reduced from Erickson's 39). (No subscales; unidimensional scale). Dichotomous "True / False" system (Identical to Erickson 1969). Scores range from 0 to 24. Higher scores indicate greater deviation from "normal" attitudes.	Total Participants: 61 individuals. PWS Group (n=36): 36 young adult males, 0 females. Composed of Exp Group 1 (n=24) and Exp Group 2 (n=12). Mainly university students. PWNS Control (n=25): 25 young adult males, 0 females. Matched for age, sex, education, occupation. Age Range / Mean: Not explicitly defined, characterized as "young adult"	NE (Relied on Erickson's original item pool; items were removed purely based on statistical instability and lack of sensitivity to treatment).	NE (No factor analysis performed).	NE (Not reported in this study).	Test-Retest Reliability: Evaluated at item-level. 4 items from the original 39 were deleted because their test-retest correlations in the control group were < 0.70. Total scale ICC was not reported. Construct / Known Groups Validity: PWS pre-treatment mean (19.22) was significantly higher than PWNS (9.14). Responsiveness to Change: Scores	Adequate (Successfully refined the tool for therapeutic repeated measures and proved responsiveness to change; limited by small sample size and lack of factor analysis/total internal consistency metrics).

			university students.				dropped significantly across 3 therapy phases (19.22 to 14.27 to 9.11).	
SBIS	Short Behavioral Inhibition Scale (SBIS) Ntourou et al. (2020) United States Scale development, validation, and application conducted across 2 studies: Study 1: Development of the SBIS and testing of its validity and	The final scale contains 5 items (reduced from an initial 7 items) . The tool has no subscales; it measures a single, relatively homogeneous dimension/unitary construct of behavioral inhibition. The tool uses a 5-point Likert-type scale . • Each item presents both ends of the assessed Behavior continuum (e.g., 1 = "retreats immediately", 3 = "average", 5 =	Overall Study 1 Total: 468 participants (225 CWS, 243 CWNS); Age range 2;10 to 6;3 years; Mean age = 49.15 months (SD=9.84) . Study 1 EFA Subsample (n=234): 112 CWS (30 F, 82 M), 122 CWNS (55 F, 67 M) . Study 1 CFA Subsample (n=234): 113 CWS (30 F, 83 M), 121 CWNS (57 F, 64 M) . Study 1 Reliability Subsample: 76 children (46 CWNS,	In Study 1, the initial items were generated from a review of previous literature (41 descriptors grouped into 7 categories). To assess content/face validity, 10 parents of normally fluent preschool-age children rated the items for "readability and clarity" on a 5-point scale and suggested wording revisions, leading to	In Study 1, Exploratory Factor Analysis (EFA) on 234 participants yielded a single-factor solution, leading to the removal of 2 items with poor loadings (0.32 and 0.22). Confirmatory Factor Analysis (CFA) on a separate sample of 234 participants confirmed the	In Study 1, the internal consistency was strong with a Cronbach's alpha of α = .81 for the final 5-item scale based on the entire sample (N=468).	Test-Retest Reliability: Pearson's r = .79, p < .001; ICC = .88, p < .001; 8 months apart; n=76 Concurrent Validity: Significantly negatively correlated with direct behavioral observation (latency to 6th spontaneous comment; Spearman's rho = -.31, p=.022; n=53).	Very Good (Exceptionally large sample sizes, robust cross-validation using both EFA and CFA, and comprehensive evaluation of multiple construct validity parameters).

reliability . Study 2: Application to determine differences in BI between CWS and CWNS, and associations with stuttering severity, frequency, and attitudes.	"approaches easily") . Possible scores range from 5 to 25 . Note: Lower scores indicate higher Behavioral Inhibition (BI).	30 CWS) . Study 1 Concurrent Validity Subsample: 53 participants . Study 1 Convergent/Divergent Validity Subsample: 391 participants . Study 2 Total: 377 participants. Study 2 CWS Group (n=179): 49 girls, 130 boys; Age range 3;0 to 6;3; Mean age = 47.54 months (SD=9.40). Study 2 CWNS Control Group (n=198): 91 girls, 107 boys; Age range 3;0 to 6;3; Mean age = 51.01 months (SD=10.24).	minor changes.	5-item single-factor model had excellent fit ($X^2/2=3.32$, p=.51, SRMR =.02, TLI =1.00, CFI =1.00, RMSEA =0.00).	Convergent Validity: Significantly negatively correlated with BSQ approach/withdrawal scores (r = -.68, p < .001; n=391). Divergent Validity: No significant correlation with BSQ persistence scores (r = -.02, p = .76; n=391). Known Groups Validity (Study 2): CWS scored significantly lower (indicating higher BI) than CWNS (F=9.58, p=.002). Significantly more CWS fell into the extreme high BI group ($X^2=7.69$,

p=.006).

Predictive Validity (Study 2): Higher BI significantly predicted greater stuttering frequency (p=.049), stuttering severity (p=.003), TOCS Disfluency-Related Consequences (p=.003), and KiddyCAT negative communication attitudes for CWS >4 years old (p=.015).

SCESS	Satisfaction with Communication in Everyday Speaking Situations (SCESS) scale Karimi et al. (2018) Australia Scale development, reliability, and construct validity assessment.	The tool consists of a single item assessing overall satisfaction. No subscales. The tool uses a 9-point scale from 1 = "Extremely Satisfied" (Most positive) to 9 = "Extremely Dissatisfied" (Most negative) Note: As corrected in the 2021 Corrigendum, there are no text descriptors for scores 2 through 8.	Overall Total: 87 adults who stutter (AWS), plus 14 experts and 20 pilot AWS. No non-stuttering control group used. Phase 1 (Pilot/Content Validity): 14 expert SLPs/psychologists, 15 pilot AWS, and 5 evaluative AWS. Phase 2 (Main Validation Group): 87 AWS (40 not seeking treatment, 47 seeking treatment). Gender (Main Group): 67 males, 20 females. Age Range: 20 to 79 years. Mean Age: 32 years (SD=14).	Development involved an extensive literature review, testing on 15 pilot AWS, and review by a 14-member expert panel evaluating relevance, utility, and format on a Likert scale. Following a 1-hour interview with 8 experts, 5 AWS verified the wording as clear, easy to understand, and representative of both behavioral and nonbehavioral aspects of stuttering.	NE (Not applicable; cannot conduct factor analysis on a single-item scale).	NE (Not applicable; cannot calculate internal consistency for a single item).	Test-Retest Reliability: Relative reliability via Spearman's r = 0.77 ($p < 0.001$; 2-4 weeks apart; n=37). Absolute reliability via Standard Error of Measurement (SEM) was 1.75. Construct Validity (Convergent): SCESS showed significant moderate-to-high correlations with OASES (r=0.71), typical self-rated severity (r=0.65), UTBAS (r=0.63), SPAI (r=0.63), worst self-rated severity (r=0.59), avoidance (r=0.52), STAI (r=0.51), IPDEQ Anxious (r=0.50), and BFNE (r=0.47). It showed low	Adequate (Strong content validation process and construct validity sample; however, absolute reliability SEM=1.75 suggests high individual fluctuation, and a single item cannot capture multi-dimensional nuances).

	Phase 3 (Reliability Subsample): 37 AWS.	correlation with %SS (r=0.25). Construct Validity (Divergent): SCESS showed low non-significant or weak correlations with unrelated personality traits like IPDEQ Histrionic (r=0.03), Anankastic (r=0.11), Dissocial (r=0.12), Schizoid (r=0.14), Dependent (r=0.19), Impulsive (r=0.21), and URICA readiness to change (r=0.24).

SESAS	Self-Efficacy Scale for Adult Stutterers (SESAS) Ornstein & Manning (1985) USA Psychometric evaluation.	50 items. Two types of situations: approach (40 items) and maintenance (10 items). 10 to 100 numerical scale (10 = no confidence, 100 = complete confidence).	20 Adults who stutter (AWS) & 20 Adults who do not stutter (AWNS).	Initial items generated from input of 5 adult stutterers, pared down to 50	NE	NE	Test-Retest Reliability: Pearson r = 0.89 for total scale; n=11 AWS, 11 AWNS. Construct/Criterion Validity: Pearson r = -0.71 vs Erickson Scale; r = -0.52 vs PSI. Discriminant Validity: Significantly differentiated AWS from AWNS, p < .001	Doubtful (Underpowered sample size for test-retest reliability, n=11 AWS; lack of formal structural validity and internal consistency evaluation).
SGSM	Stuttering Generalization Self-Measure (SGSM) Alameer et al. (2017) Kuwait Preliminary	9 items (only 7 used in the study). No subscales. 5-point Likert (anxiety) and 9-point Likert (stuttering).	43 adults (22 PWS, 21 control)	Includes Face Validity (listener categories matched existing tools) and Content Validity (items compared to WTC, SPCC, IIS).	NE	Not evaluated because the scale contains only one item for each situation.	Test-Retest Reliability : Spearman's rho = 0.97 to 0.99) Intra-judge reliability: ICC = 0.95 Inter-judge reliability: ICC =	Doubtful (underpowered sample size with n=43 for Test-retest reliability; lacks formal cognitive interviews with patients for content

development of a self-measuring tool						0.95 Convergent validity: Correlated with SPCC r=0.79, WTC r=0.83 Discriminant validity: Significant difference between PWS and controls, p < 0.001	validity).
Persian Stuttering Generalization Self-Measure (P-SGSM) Hozeili et al. (2024) Iran Psychometric properties evaluation (cross-	9 items. No subscales. 5-point Likert (anxiety) and 9-point Likert (stuttering).	60 adults (30 AWS & 30 AWNS)	Includes Face Validity (10 AWS interviewed confirmed clarity) and Content Validity (CVR > 0.62 and CVI > 0.79 evaluated by 10 experts).	NE	Cronbach's α = 0.98 for all subtests).	Test-Retest Reliability: ICC = 0.93 to 0.99 Inter-judge reliability: Spearman-Brown = 0.98 Discriminant validity: Significantly differentiated AWS from AWNS, p < 0.001	Doubtful (Test-retest reliability underpowered with n=30 AWS; structural validity not proven prior to internal consistency).

	cultural adaptation)						Internal responsiveness: MSR = 1.09 for SS%	
SSC	SSC-R (SSC-ER AND SSC-SD) Vanryckeghem et al., 2017 USA Psychometric, normative and comparative study.	38 items (revised from 51 items). Two sections: Emotional reaction(SSC-ER) & speech disruption(SSC-SD). Factor analysis were done and it yielded 7 factors for SSC-ER and 6 factors for SSC-SD. 5 Point rating scale (1=not at all to 5=very much)	297 adults including 88 PWS (69 male and 19 female in the age range of 18-61 with a mean age of 28.5yrs) and 209 PWNS (108 males and 101 female with the age range of 18-62 with a mean age of 31.4yrs).	Items were derived from the clinical files and reports over decades and 13 redundant items were removed.	Evaluated by using principal component method with varimax rotation;7 factors explained 70.46% variance for SSC-ER and 6 factors explained 70.50% variance for SSC-SD.	Evaluated using Cronbach's alpha (α)= 0.96 for SSC-ER and 0.97 for SSC-SD in PWS.	Discriminant/construct validity: evaluated and found that there is a significant difference between PWS and PWNS, P<0.0001. Accurately identified 96.7% for SSC-ER and 96.6% for SSC-SD. Item-to-total correlation: Found that all items are correlated with the total score.	Doubtful as there was less sample number while doing structural validity and also internal consistency were calculated based on total score rather than calculating based on each dimensions.
	SSC-ER-K (Kannada) Veerabhadrappa et al.,	SSC consists of originally 55 items and revised to 40 speaking situations to eliminate the	100 CWS (12 female and 88 males with the mean age of 10.34) and 257 CWNS (128 female	Items generation: translated from original English	NE	Evaluated using Cronbach's alpha (α)= 0.96 for CWS	Test-retest reliability: Intra-class correlation coefficient (ICC): CWS= r= 0.98	Doubtful (Test-retest sample size of n=10 CWS is

2021 India Normative and comparative study	redundancy. 5-point rating scale ranging from "Not afraid" (score 1) to "very much afraid" (score=5). Minimum score= 40 and maximum score= 200	and 147 males with mean age of 10.80) in the age range of 7-14 yrs .	version. Panel expert: backward and forward translation.	(Excellent) and 0.76 for CWNS (Good).	and 0.95 for CWNS (determined by re-administering the test to random 10% of the participants in both group). Discriminant validity: Evaluated (Linear discriminant analysis correctly distinguished 81% of CWS and 100% of CWNS, overall 95% accuracy; significantly differentiated mild vs. moderate/severe CWS). Item Analysis: Evaluated (All items correlated significantly with total score for CWS).	severely underpowered; structural validity was not proven before calculating internal consistency).

SSS	Subjective Screening of Stuttering (SSS): Research Edition Riley et al. (2004) USA Screening tool development and preliminary reliability/validity study.	8 items total. 3 Subscales (Areas): 1. Stuttering Severity (2 items) 2. Locus of Control (3 items) 3. Avoidance (3 items). Evaluated. 9-point equal appearing interval scale (1 to 9). 1 = Normal or target level (e.g., "Relatively fluent", Never, "0%", "A great deal") and 9 = Most severe (e.g., "Severe stuttering", "Constantly", "100%", "Very little"). Ratings are made for three distinct audiences: close friend, authority figure, and telephone (close friend ratings are excluded from total	Total Participants: 16 adults who stutter (AWS) participated in the reliability/validity phase. Another sample of 32 was used for item-total correlations, and 23 for a medication trial. No non-stuttering control group was used. Gender: Not reported (NE). Age Range: Not reported (NE). Mean Age: Not reported (NE).	Evaluated. Items were selected based on 10 years of clinical experience by the authors and feedback from other clinicians using earlier versions. No formal quantitative CVI/CVR indices were calculated.	NE (Structural validity/factor analysis was not performed).	(Item-Total Correlations). While Cronbach's alpha was not reported, item-to-area correlations were calculated on 32 PWS, ranging from 0.81 to 0.97. Subtest-to-total correlations were high: SEV (0.92), LOC (0.92), AVD (0.95).	Reliability (Test-Retest): Evaluated. 2 weeks apart (n=16). Severity agreement = 88% (r=0.90); LOC agreement = 89% (r=0.93); Avoidance agreement = 84% (r=0.79). Criterion-Related / Concurrent Validity: Evaluated (n=16). SSS Severity correlated with %SS (r=0.75, p<0.01) and duration (r=0.69). SSS LOC correlated with PSI LOC items (r=0.70, p<0.05). SSS Avoidance correlated with PSI Avoidance items (r=0.83, p<0.01). The abstract notes	Low (Sufficient as a preliminary screening tool, but lacking the sample size [n=16-32], demographic reporting, and formal structural validation required for a high-quality psychometric instrument).

		score).					%SS correlated with LOC (r=0.43) but NOT with Avoidance.	
Safety Behaviors Questionnaire	Safety Behaviors Questionnaire (Persian translated version) Azarinfar et al. (2022) Iran Transcultural adaptation, construct validity, and reliability assessment.	The final scale contains 29 items (reduced from 34 initial items after face validity and factor analysis). It includes 4 subscales (factors): 1. General avoidance (12 items) 2. Practice and control (10 items) 3. Rehearsal (4 items) 4. Choosing safe and easy people (3 items). The tool uses a 5-point Likert scale to report the frequency of safety Behavior use. 1 = "never" 2 = "rarely", 3 =	Overall Total: 167 adults who stutter (AWS), 17 SLPs, and 5 additional AWS. No non-stuttering control group used. Phase 1 (Face Validity Delphi Panel): 17 expert SLPs and 5 AWS. Phase 2 (Factor Analysis & Reliability): 167 AWS. Gender (Main Group): 116 males (69.5%), 51 females (30.5%). Age Range: 18.0 to 59.0 years.	A Delphi method was used. 17 SLPs and 5 AWS judged the 34 initial items on a 1-6 Likert scale for clarity and cultural relevance (1 = "not understandable/ culturally relevant", 6 = "completely clear/relevant"). Items required >=80% of judges scoring them >= 4 to be approved; 11 items were initially modified/combined, resulting in 32 items	Exploratory Factor Analysis (EFA) was conducted (KMO = 0.834, Bartlett's $p < 0.001$). Four factors were extracted, explaining 50% of the total variance (Factor 1=25%, Factor 2=15%, Factor 3=5%, Factor 4=5%). Three items did not load and were removed.	Internal consistency was high with a total Cronbach's α = 0.89. Subscale α were: General avoidance (0.91), Practice and control (0.81), Rehearsal (0.85), and Choosing safe/easy people (0.76).	Inter-item / Item-total Correlation: Inter-item correlations were within acceptable ranges (Factor 1=0.46, Factor 2=0.30, Factor 3=0.58, Factor 4=0.51) and all item-total correlations were >0.30. Test-Retest Reliability: NE (Authors explicitly stated this needs assessment in future studies). Construct Validity (Convergent / Discriminant / Known Groups):	Adequate (Strong sample size for EFA [n=167] and high internal reliability, but lacks Confirmatory Factor Analysis (CFA), test-retest reliability, and external construct validity evaluations).

		"sometimes", 4 = "almost always" (or "often") and 5 = "always".	Mean Age: 28.1 years (SD=7.1).	entering factor analysis.				NE (Authors explicitly stated the relationship between safety behavior types and social anxiety severity is yet to be explored).
UTBAS	Unhelpful Thoughts and Beliefs about Stuttering (Original UTBAS) St Clare et al. (2009) Australia Measure development and preliminary validity study conducted across 3	66 items. Includes 39 general items (no reference to stuttering) and 27 stuttering-specific items. Original measure focused on a single scale of negative thought frequency. The tool uses a 5-point Likert scale ranging from 1 to 5. A score of 1 indicates "never have the thought"	Phase 2 Total: 57 participants. Phase 1 Control Group (n=31): 16 males, 15 females; Age range 19 to 56 years; Mean age = 29.5 (SD=10.9). Phase 2 AWS Group (n=26): 21 males, 5 females; Age range 16 to 65 years; Mean age = 27.0 (SD=14.5). Phase 3 Total: 18 AWS (15 males, 3	In Phase 1, items were generated from a comprehensive 10-year retrospective audit of clinical files belonging to AWS undergoing Cognitive Behavior Therapy.	NE	In Phase 3, Cronbach's α = 0.98 (Time 1) and α = 0.96 (Time 2) indicating very high internal consistency.	Reliability (Test-Retest): Evaluated in Phase 3. Correlation was r = 0.89 (1 month apart; n=18). A paired t-test confirmed no significant difference between Time 1 (mean=135.6) and Time 2 (mean=135.5) scores (p=0.99). Known Groups Validity: Evaluated in	Doubtful (While it has an excellent psychometric design across 3 phases, the Phase 3 reliability assessment utilized a severely underpowered sample of n=18 AWS, and structural validity was not tested).

phases: • Phase 1: Initial development of the measure. • Phase 2: Construct validity and sensitivity of the UTBAS. • Phase 3: Reliability of the UTBAS.	A score of 2 indicates "rarely have the thought" A score of 3 indicates "sometimes have the thought" A score of 4 indicates "often have the thought" A score of 5 indicates "always have the thought". Total scores range from 66 to 330.	females); Age range 21 to 61 years; Mean age = 39.7 (SD=14.4).	Phase 2. The AWS group scored significantly higher than the control group on the 39 general items ($t(38)=8.06$, $p<0.0001$, effect size $d=2.3$). Convergent Validity: Evaluated in Phase 2. The UTBAS strongly correlated with social anxiety measures including SPAI ($r=0.72$), SADS ($r=0.68$), and FNE ($r=0.53$). Discriminant Validity: Evaluated in Phase 2. Weak, non-significant correlations were found with general anxiety

						(BAI, r=0.24) and depression (BDI-II, r=0.27). Sensitivity to Change: Evaluated in Phase 2. Pre-CBT scores (mean=172.0) decreased significantly post-CBT (mean=99.9) (t(25)=10.13, p<0.0001, effect size d=2.5).	
Unhelpful Thoughts and Beliefs about Stuttering (UTBAS) scales (I, II, and III) Iverach et al. (2011) Australia Further validation of original	66 items. Includes three distinct scales: UTBAS-I (frequency), UTBAS-II (belief), and UTBAS-III (anxiety). 5-point Likert scale (1 to 5) for all three sections. UTBAS I (Frequency): 1="never have the thought" to 5="always have the	Total: 140 adults who stutter (AWS). Gender: 93 males (66%) and 47 females (34%). Age Range: 18 to 68 years. Mean Age: 36.6 years (SD = 12.8).	Refined from a 10-year clinical audit.	Principal Components Analysis (PCA) with Varimax rotation confirmed a three-factor structure corresponding to the frequency, belief, and anxiety scales.	Cronbach's alpha values were α = 0.98 (UTBAS I), α = 0.98 (UTBAS II), and α = 0.97 (UTBAS III).	Reliability (Test-Retest): r=0.91 to 0.94; 2 weeks apart; n=40 Construct Validity: Correlated with Social Phobia Anxiety Inventory and Fear of Negative Evaluation scale. Known Groups Validity: Significantly	Adequate (Comprehensive development and validation of all three subscales; strong psychometric properties).

UTBAS-I and development/ validation of UTBAS-II and III.	thought". UTBAS II (Belief): 1="don't believe this at all" to 5="believe this totally". UTBAS III (Anxiety): 1="not anxious at all" to 5="extremely anxious".					higher scores in AWS with social anxiety disorder vs. those without; $p < 0.001$	
Brief version of the Unhelpful Thoughts and Beliefs About Stuttering Scales (UTBAS-6) Iverach et al. (2016) Australia Item reduction and development of a brief	18 items total (reduced from 198). 3 subscales with exactly 6 items each: 1. UTBAS-1 (Frequency) 2. UTBAS-2 (Belief) 3. UTBAS-3 (Anxiety). The tool uses a 5-point Likert scale ranging from 1 to 5 for all three	Total Participants: 337 adults who stutter (AWS). (Note: No non-stuttering control group was included in this specific study). Gender: 268 males (79.5%), 69 females (20.5%). Age Range: 18 to 80 years. Mean Age: 35.0 years (SD=14.0).	Items were selected from the previously validated full scale based on highest item-total correlations and their ability to cover the broadest clinical content.	Evaluated (via Item Reduction). Item reduction analysis was conducted on the full scales. Findings included: Inter-item correlation ranges: -.005 to .756 (UTBAS-1), .043 to .737 (UTBAS-2), and -.032 to .755 (UTBAS-	The brief UTBAS-6 yielded an acceptable Cronbach's α = 0.82. (For context, prior to item reduction, Cronbach's α for the full scales was 0.98 across all three subscales).	Test-Retest Reliability: NE. Convergent/Discriminant/Known Groups Validity: NE (Authors explicitly state these parameters must be evaluated in future research). Diagnostic Accuracy: The brief scale demonstrated a	Adequate (Strong sample size of n=337 for item reduction modeling utilizing comprehensive inter-item and multicollinearity analyses, but lacks formal factor analysis and external construct validity for

screening tool.	sections. A score of 1 indicates "never or not at all" A score of 2 indicates "rarely or a little". A score of 3 indicates "sometimes or somewhat" A score of 4 indicates "often or a lot" A score of 5 indicates "always or totally". Brief scale scores range from 6 to 30; brief total scores range from 18 to 90.	3). Item-total correlation ranges: .46 to .77 (UTBAS-1), .39 to .79 (UTBAS-2), and .45 to .79 (UTBAS-3). tem-rest correlation ranges: .46 to .77 (UTBAS-1), .39 to .77 (UTBAS-2), and .41 to .79 (UTBAS-3). Item multicollinearity (multiple R) ranges: .77 to .90 (UTBAS-1), .72 to .91 (UTBAS-2), and .80 to .92 (UTBAS-3).	chance-corrected sensitivity of 93.4% and specificity of 89.8% in classifying patients at or above the 5th decile clinical cutoff of the full scale.	the new brief form).

Unhelpful Thoughts and Beliefs about Stuttering (UTBAS) scales, Japanese version (UTBAS-J) Chua et al. (2017) Japan Scale translation, normative data, and psychometric evaluation.	66 items total evaluated across 3 separate scales: 1. UTBAS-J I (Frequency of unhelpful thoughts) 2. UTBAS-J II (Belief in unhelpful thoughts) 3. UTBAS-J III (Anxiety associated with unhelpful thoughts). (No subscales within the 66 items; treated as unidimensional lists). 5-point Likert scale (1 to 5) Scale I (Frequency): 1 = "never", 5 = "all the time/frequently" Scale II (Belief): 1	Total Participants: 141 adults who stutter (AWS). No non-stuttering control group used. Gender: 111 males (78.7%), 30 females (21.3%). Age Range: 18 to 76 years. Mean Age: 36.5 years (SD=13.0). Test-Retest Subsample: n=42 AWS.	Evaluated. Rigorous translation process utilizing forward translation by two bilingual speakers, synthesis, back-translation by an independent bilingual speaker, and final verification by original authors and a Japanese speech-language pathologist to ensure conceptual equivalence.	NE (The authors explicitly stated they did not perform a factor analysis on the UTBAS-J, relying on the pre-established structure of the original English version).	Exceptionally high internal consistency. Cronbach's α = 0.98$ for all three scales (UTBAS-J I, II, and III).	Test-Retest Reliability: ICC = 0.89 for Scale I, 0.89 for Scale II, and 0.87 for Scale III over a 4-week interval (n=42). Other Measurement Properties (Convergent Validity): All three UTBAS-J scales correlated significantly and positively with the OASES-A-J (r=0.69-0.77), the Liebowitz Social Anxiety Scale (r=0.48-0.54), and the Kessler Psychological Distress Scale (r=0.44-0.51) at p<0.001.	Good (Strong normative sample size, exceptional internal consistency, temporal stability, and excellent convergent validity proofs; limited by a lack of new structural factor analysis and slight underpowering in the test-retest subsample).

	= "not at all", 5 = "totally/completely " Scale III (Anxiety): 1 = "not at all", 5 = "extremely" Total scores for each scale range from 66 to 330. Higher scores indicate greater frequency, belief, or anxiety.						
Unhelpful Thoughts and Beliefs about Stuttering-Turkish version (UTBAS-TR) Aydın Uysal & Ege (2020) Türkiye Psychometric evaluation (Reliability	66 items. 3 subscales: UTBAS I (frequency), UTBAS II (belief), UTBAS III (anxiety). 5-point Likert scale (1 to 5) for all sections. UTBAS I (Frequency): 1="never" to 5="always" UTBAS II (Belief): 1="don't believe"	Total: 100 adults who stutter (AWS). Gender: 74 males (74%) and 26 females (26%). Age Range: 17 to 41 years. Mean Age: 23.5 years (SD=5.5).	NE	NE	Cronbach's α = 0.94 (Total score).	Reliability (Test-Retest): ICC = 0.92 to 0.95 across scales; 1 month interval; n=30 Construct/Concurrent Validity: Correlated with S-24 scale and Erickson Scale; significantly differentiated AWS scores, p <	Doubtful (Structural validity not evaluated).

and Validity)	to 5="believe totally" UTBAS III (Anxiety): 1="no anxiety" to 5="extreme anxiety".				0.05		
Italian version of the Unhelpful Thoughts and Beliefs about Stuttering (UTBAS-ITA) Bernardini et al. (2024) Italy Translation, normative data, and psychometric evaluation.	The scale contains 66 items. Includes 39 general items (not directly stuttering-specific) and 27 stuttering-specific items. It consists of 3 subscales: UTBAS I (frequency of negative thoughts), UTBAS II (how realistic/correct thoughts are considered), and UTBAS III (anxiety associated with thoughts). The tool uses a 5-point Likert scale ranging from 1 to 5	Total: 196 adults. AWS Group: 98 Adults who stutter (57 males/58.1%, 41 females/41.84%); Age Range: 18 to 68 years; Mean Age: 37.02 (SD = 11.94). AWNS Group: 98 Adults who do not stutter (57 males/58.1%, 41 females/41.84%); Age Range: 19 to 68 years; Mean Age: 37.01 (SD = 12.00).	The tool underwent a forward-backward translation process, resolving discrepancies via email with the original authors to avoid offensive terms in Italian.	NE	The internal consistency was extremely high with a Cronbach's alpha of 0.99 for the total score in both AWS and AWNS groups, and 0.97 for the AWS subscales.	Reliability (Test-Retest): Spearman correlations were 0.93 to 0.95 across scales (tested 15 days apart; n=76). Construct Validity (Convergent): Showed moderate to strong positive correlations with the Fear of Negative Evaluation Scale (r=0.730 for Total) and STAI-Trait anxiety (r=0.466 for Total).	Doubtful (Convergent validity was underpowered with only n=20 AWS. No formal cognitive interviews with patients were conducted for content validity. Structural validity [factor analysis] was not evaluated prior to calculating internal

for all three sections. UTBAS I (Frequency): 1 indicates "never have the thought" and 5 indicates "always have the thought". UTBAS II (Belief): 1 indicates "I don't believe this at all" and 5 indicates "I believe this totally". UTBAS III (Anxiety): 1 indicates "does not make me anxious at all" and 5 indicates "makes me extremely anxious". The total combined score ranges from 198 to 990.	Note: A subsample of 76 AWS [mean age 38.29] was used for test-retest reliability, and a subsample of 20 AWS [16 males, age 19-48, mean age 29.25] was used for construct validity.	Construct Validity (Discriminant / Known Groups): AWS scored significantly higher than AWNS across all scales and total scores ($p < 0.001$).	consistency).

Persian version of the Unhelpful Thoughts and Beliefs about Stuttering (UTBAS-P) Farpour et al. (2025) Iran Translation, cross-cultural examination, and psychometric evaluation.	The final scale contains 71 items. 3 subscales: UTBAS I (frequency of thoughts), UTBAS II (belief in thoughts), and UTBAS III (anxiety associated with thoughts). The tool uses a 5-point Likert scale ranging from 1 to 5. A score of 1 indicates "never or not at all". A score of 2 indicates "rarely or a little". A score of 3 indicates "sometimes or somewhat". A score of 4 indicates "often or a lot". A score of 5 indicates "always or totally". The total combined score ranges from 213 to 1065.	Total: 62 adults who stutter (AWS). Pilot Group: 10 AWS. Main Group: 52 AWS. Gender (Main Group): 43 males (82.7%) and 9 females (17.3%). Age Range: 18 to 51 years. Mean Age: 26.4 years (SD = 6.4).	Face and content validity were tested by 6 experts and 10 AWS, leading to the addition of 6 culturally relevant items and the omission of 1 item.	NE	The internal consistency was extremely high with a Cronbach's α of 0.99 for the total scale.	Reliability (Test-Retest): The Pearson correlation was 0.87 (tested 15 days apart; n=10). Construct Validity (Convergent Validity): Showed significant positive correlations with anxiety measures (BAI r=0.58, BFNE r=0.48, SPAI Diff r=0.58, STAI-T r=0.57, NEO-N r=0.49). Post hoc power analysis indicated a high achieved power of 0.99 for detecting an average correlation coefficient of 0.53. Construct Validity (Divergent/Discriminant Validity):	Doubtful (The sample size for test-retest reliability is underpowered at n=10. Structural validity was not conducted prior to calculating internal consistency. Divergent validity achieved low statistical power).

Demonstrated weak or negative correlations with unrelated personality constructs like extraversion, openness, and agreeableness (NEO-E (r=-0.50), NEO-O (r=-0.16), NEO-A (r=-0.14), and NEO-C (r=-0.14). Post hoc power analysis revealed an achieved power of 0.59 for detecting an average correlation coefficient of 0.26, indicating potential underpowering for detecting weaker correlations.

VRYCS	Vanderbilt Responses to Your Child's Speech (VRYCS) Rating Scale Singer et al. (2022) USA Development, refinement, and exploration of psychometric properties (Item reduction, Content validity, Factor analysis).	18 items total (reduced from an initial pool of 40, to 30, to 21, and finally to 18). 5 subscales (Factors) & Findings: 1. Requesting Change (4 items; e.g., "tell child to relax"). Mean score = 2.36. 2. Speaking for the Child (4 items; e.g., "fill in words"). Mean score = 2.69. 3. Supporting Communication (3 items; e.g., "let child lead"). Mean score = 3.04 (Highest frequency). 4. Slowing & Simplifying (4 items; e.g., "ask simple questions"). Mean score = 2.13 (Lowest frequency).	Total Participants: 214 parents of young children who stutter (2.4 years to 5.11 years old) (CWS) . (Note: No non-stuttering control group was evaluated in this study) . Groups: 176 from a clinical sample (mean age 49.97 months) and 38 from a laboratory sample (mean age 47.29 months) . Gender of CWS: 158 males (74%), 56 females (26%) . Age Range: 2;4 to 5;11 (years; months) . Mean Age: 48.67 months (SD=10.27).	An online survey was distributed to 10 international stuttering experts (SLPs) to judge 30 items. 21 items achieved 90% or greater agreement for inclusion and were retained. 9 items with 80% or less agreement were excluded.	A Principal Component Analysis (PCA) with varimax rotation was used (KMO = 0.776, Bartlett's $p < 0.001$). Five factors with eigenvalues >1 were extracted, accounting for 59.3% of the total variance. Two items cross-loaded or failed criteria and were removed.	Overall scale Cronbach's $\alpha = 0.645$. Factor alphas: F1 = 0.880, F2 = 0.772, F3 = 0.516, F4 = 0.615, F5 = 0.528. All met the minimum >0.5 threshold for scales with few items.	Test-Retest Reliability: Authors stated the stability of responses over time needs to be examined in future research. Construct Validity (Concurrent): Authors stated the lack of valid English-language measures similar in purpose prevented concurrent validity assessment. Construct Validity (Known Groups): Authors explicitly stated comparing parents of CWS to parents of CWNS should be done in future studies.	Adequate (Strong sample size N=214 for PCA, excellent formal content validity using 10 international experts, but lacks Confirmatory Factor Analysis (CFA), test-retest reliability, and external construct validity evaluations).

5. Responding Emotionally (3 items; e.g., "worry about talking"). Mean score = 2.38.

The tool uses a 5-point rating scale (0 to 4) assessing frequency over the past 2 months . • 0 = "never" . • 1 = "rarely" . • 2 = "sometimes" . • 3 = "often" . • 4 = "always" . Ten items are reverse-scored so that a higher final score always indicates a more supportive/positive parent response. The mean Total Score was 2.50 (SD=0.39).

WASSP	Wright & Ayre Stuttering Self-Rating Profile (WASSP) Wright et al. (1998) UK Tool introduction and conceptual clinical description	21 items explicitly listed in this early text. 4 Subscales described: 1. Behaviour (8 items) 2. Avoidance (4 items) 3. Feelings (6 items) 4. Disadvantage (3 items). 7-point scale from 1 = "None" to 7 = "Very severe".	Total Participants: Fewer than 20 adults who stutter (AWS) . (No non-stuttering control group). Gender/Age/Mean Age: Not reported (NE).	Evaluated Qualitatively: Content generation involved PWS discussion groups. Breadth and relevance were informally confirmed by clinician feedback and subsequent client reactions.	NE	NE	Test-Retest Reliability: NE (Proposed for future research at a 24-hour interval, but not done here). Other Measurement Properties: Evaluated qualitatively. Authors noted the tool demonstrated change in clinically predicted areas, offering initial theoretical construct validity, but provided no statistical correlations.	Low (As this is an introductory conceptual paper presenting the tool's clinical framework; it lacks the empirical sample sizes and statistical data required of a formal psychometric validation).

WASSP-Turkish version (WASSP-TR) Uysal & Köse (2021) Turkey Scale adaptation, validity, and reliability evaluation.	24 items total. 5 Subscales: 1. Stuttering behaviours (8 items) 2. Thoughts about stuttering (3 items) 3. Feelings about stuttering (5 items) 4. Avoidance (4 items) 5. Disadvantage (4 items). 7-point Likert-type scale. (1=None to 7=Very Severe, consistent with original tool).	Total Participants: 120 AWS . No non-stuttering control group. Gender: 107 males (89.2%), 13 females (10.8%) . Age Range: 18 to 54 years . Mean Age: 26.42 years (SD=7.92) . Stuttering Severity Groups: Mild (n=36), Moderate (n=48), Severe (n=36) . Test-Retest Subsample: n=30 AWS.	Translated per WHO guidelines. 8 experts reviewed translations, achieving a Krippendorff's alpha of 0.74 for agreement. Pre-tested on 20 AWS for comprehensibility and spelling.	Confirmatory Factor Analysis (CFA) performed. Standard factor loadings ranged from 0.46 to 0.49. Showed perfect model-data fit (chi square/df=2.00, RMSEA=0.059, SRMR=0.064, CFI=0.99).	Excellent internal consistency. Total Cronbach's α = 0.949 Subscales: Behaviours=0.891 Thoughts=0.932 Feelings=0.919 Avoidance=0.858 Disadvantage=0.875	Test-Retest Reliability: Evaluated. Total r = 0.972 ($p<0.01$) over a 2-week interval (n=30). Criterion & Known-Groups validity: Negative significant correlations found against SF-36 quality of life metrics (e.g., feelings vs. mental health r=-0.747). ANOVA confirmed WASSP-TR scores increased significantly alongside objective stuttering severity groups ($p<0.01$).	Adequate (Robust CFA, exceptional reliability metrics, and strong criterion evaluation; slightly limited by reliance on the minimum recommended sample size for CFA [n=120] and no fluent control group).

Note. This table details the Studies included in review and psychometrics properties of the included tools. CFA = Confirmatory Factor Analysis; EFA = Exploratory Factor Analysis; CVI = Content Validity Index; CVR = Content Validity Ratio; PWS = People Who Stutter; CWS = Children Who Stutter; AWS = Adults Who Stutter, Communication Attitude Test (CAT), Overall Assessment of Speakers Experience in Stuttering (OASES), Pre-monitory Awareness in Stuttering (PAiS), Speech Situation Checklist (SSC), Self-Efficacy Scale for Adults who Stutter (SESAS), Stuttering Generalization Self Measure Tool (SGSM), Speech Situation Checklist (SSC), Palin Parent Rating Scale (Palin PRS), Unhelpful Thoughts and Beliefs About Stuttering (UTBAS), Caregiver Burden Scale for Parents of Children with Stuttering (CBS-PCWS), Short Behavioural Inhibition Scale (SBIS), Safety Behavior Questionnaire (SBQ), Satisfaction with Communication in Everyday Speaking Situations (SCESS), The Wright and Ayre Stuttering Self-rating Scale (WASSP), Vanderbilt Response to Your Child's Speech rating scale for Parents of Young Children who Stutter (VRYCS), Scale of Avoidance and Struggle Behaviours in Stuttering (SASBS), 9-Point self-rating of Severity Scale, Self-Assessment (SA) Scale, Subjective Screening of Stuttering Severity (SSS), Inventory of Communication Attitudes (ICA), Self Stigma of Stuttering Scale (4S), S-Scale, Behavior Assessment Battery (BAB), Locus Control Behavior (LCB) Scale and Impacts Scale for Assessment Cluttering and Stuttering (ISACS)

Table S2

Methodological Weakness in the Evaluation of Content Validity Among the Studies Included in the Review

SI NO	Main Tool	Tool with studies	Methodological Weakness	Specific Examples	Impact on Content validity
1	4S	4S (Original version) Boyle (2013)	Structural validity relied on EFA without CFA.	The original 3-factor structure was established using EFA. A subsequent CFA was not conducted in this 2013 foundational study to independently verify the model fit (though Boyle later conducted one in 2015).	Not directly impacting content validity, but limits the highest rating for structural validity for this specific publication.
			Low total variance explained by extracted factors.	The extracted 3 factors (Awareness, Agreement, Application) accounted for 40.5% of the total variance. Psychometric guidelines generally desire >50%.	Not directly impacting content validity, but suggests that self-stigma encompasses other dimensions not captured by these specific 33 items.
			Content Validity: Exceptional qualitative patient generation.	Instead of just pulling items from the literature, Boyle conducted semi-structured qualitative interviews with 15 PWS specifically exploring their lived experiences of stigma, ensuring the raw	Very Good: The qualitative patient-driven item generation represents the "gold standard" for COSMIN comprehensiveness and relevance.

		items matched real-world patient taxonomy.	
	Low total variance explained by extracted factors.	While the EFA successfully replicated the 3-factor structure, those factors combined accounted for only 39.5% of the overall variance. COSMIN and general psychometric guidelines usually desire factors to explain at least 50% of the variance.	Not directly impacting content validity, but suggests that perceptions of stigma may be heavily influenced by domain-specific traits rather than just the underlying core construct.
4S-J S Limura et al. (2022)	Structural validity lacked Confirmatory Factor Analysis (CFA).	The researchers derived the 16-item short version using Exploratory Factor Analysis (EFA). However, they did not perform a CFA on an independent sample to confirm that this new 16-item structure is a good fit, which is required by COSMIN when creating a short form.	Not directly impacting content validity, but downgrades the overall quality of evidence for structural validity from High to Moderate.
	Borderline internal consistency on a subscale.	The Stigma Application subscale yielded a Cronbach's alpha of 0.61. While >/= to 0.60 is sometimes accepted in preliminary/short scales, But COSMIN prefers >/= to 0.70 for strong internal	Not directly impacting content validity, but indicates the items within that specific subscale may not be perfectly uniform.

		consistency.	
4S-Turkish Tiryaki et al. (2023)	Convenience sampling bias affecting normative representation.	The demographic breakdown showed 56% of the 350 participants were aged 18-25, and a massive 89.1% were receiving or had received therapy. The authors note this limits generalizability as self-stigma tends to naturally reduce with age and non-therapy seekers might be systematically different.	Not directly impacting content validity, but limits the generalizability of the statistical findings to the broader Turkish stuttering population.
	Strength (Content Validity): Rigorous pilot-testing.	Instead of just translating, the researchers subjected the new Turkish items to a 10-person target-population focus group, forcing participants to rate statements strictly on clarity and intelligibility.	Very Good: Confirms the comprehensibility dimension of content validity using the target demographic.
4S-Polish Wesierska et al. (2025)	Insufficient sample size for structural validation.	The study included only 108 participants, which the authors explicitly noted was too small to conduct a formal factor analysis (EFA/CFA) to verify if the 3-factor structure held true in the Polish context.	Not directly impacting content validity, but results in a "Not Evaluated" (NE) rating for structural validity, lowering the scale's overall current psychometric proof.
	Sampling bias and absence of test-retest	Participants were mostly recruited via SLPs and self-	Indeterminate: Does not negate content validity,

			reliability.	help group leaders, potentially omitting PWS isolated from the clinical community. Test-retest stability over time was completely omitted.	but restricts the generalizability of the normative scores.
2	9-point Self Rating Scale	O'Brian et al. (2004) 9-point Self Rating Scale	Lack of comprehensive psychometric evaluation.	This study serves as a limited reliability investigation of a generic scale rather than a full tool development study.	Indeterminate: Content validity was not explicitly assessed, as the scale is a standard clinical measure of severity, not a validated tool.
3	BAB	BAB-Polish Węsierska et al. (2018)	Absence of Test-Retest Reliability testing.	While internal consistency was rigorously evaluated, the temporal stability (test-retest reliability via ICC) of the Polish adaptation was not assessed.	Not directly impacting content validity, but lowers the comprehensive psychometric profile of the translated tool.
			Strength (Content Validity): Exceptional translational rigor.	The study utilized strict World Health Organization (WHO) translation protocols, including bilingual committees and back-translation confirmed directly by the original author	Very Good: Guarantees that the translated items accurately mirror the theoretical scope and clinical intent of the original instrument.

				(Vanryckeghem), ensuring semantic and conceptual equivalence.	
4	BigCAT	Big CAT-K (Veerabhadrappa et al., 2021)	Lack of patient involvement in elicitation of concept	Scale was translated from English to Kannada, but there is no new qualitative interviews were conducted with the target population.	instrument may not show the proper cultural lived experience of Kannada speakers.
			No formal assessment of content overload	during backward and forward translation, there was no report of interviewing patients until no new concepts emerged.	Weakens evidence for construct coverage in the translated language.
		BigCAT(Vanryckeghem & Brutten. 2011)	Not Evaluated (Lack of formal cognitive interviews with patients during initial tool construction.)	Not Evaluated (No specific patient-involvement data reported for item elicitation (relied on adaptation from paediatric CAT).)	Not Evaluated (May limit the assurance that all items perfectly capture the adult lived experience without formal cognitive debriefing.)
		Persian BigCAT (Valinejad et al., 2018)	Lack of patient involvement in the assessment of content validity.	the content validity ratio(CVR) was calculated based only from the opinion of 8 SLPs and no formal cognitive interviews were conducted with Persian-speaking AWS to confirm if the items were relevant,	As there no assessment was done to the target population, there is a risk that the SLPs perspective does not match fully with the patient lived experience, potentially missing the culturally

				comprehensive and comprehensible to them.	relevant concepts.
5	CAT	CAT (Vanryckeghem & Brutten (1992))	Study focuses strictly on reliability; content validity not	No assessment of relevance or comprehensiveness in this specific paper.	Content validity cannot be established from this study alone.
		CAT-K (Veerabhadrappa et al., 2020)	Study relied only on the translation methodology without formal qualitative patient assessment.	The study utilized forward and backward translation evaluated by bilingual SLPs and the original author of the study. it lacked formal cognitive interviews with Kannada speaking children to systematically ask them about the relevance or comprehensiveness of the translated constructs.	Doubtful as it cannot be guaranteed that the translated items are fully comprehensible or fully relevant for the specific age groups without qualitative feedback from the children.
		Dutch KiddyCAT (Vanryckeghem et al., 2015)	Content validity is not a part of study design.	Study mainly focused on evaluating the test=retest reliability of the Dutch Kiddy CAT (7-12 days interval) and compared the score differences between the groups.	Not evaluated
		English BigCAT(Vanryckeghem & Muir, 2016)	Content validity is not a part of study design.	Study was strictly designed to establish the test-retest reliability of the English-CAT by evaluating the test twice (5-7 days interval). so	Not evaluated

				no content validity assessments were done.	
6	CBS-PCWS	Mehdizadeh Behtash et al. (2022) CBS-PCWS	Structural validity relied on Exploratory Factor Analysis (EFA) without Confirmatory Factor Analysis (CFA).	In Phase 2, researchers extracted 5 factors explaining 63.89% of the variance using an EFA on 205 parents. However, a new factor structure should also be tested with a CFA in an independent sample to strictly confirm structural validity.	Not directly impacting content validity, but it downgrades the overall quality of evidence for structural validity from High to Moderate.
			Reliability sample size falls slightly below optimal COSMIN standards.	In Phase 2, test-retest reliability over a 2-week interval was calculated using 40 parents. COSMIN criteria require a sample size >= 50.	Not directly impacting content validity, but results in downgrading the rating for test-retest reliability due to minor imprecision (n < 50).
			Strength (No Weakness in Content Validity): Rigorous, explicit qualitative and quantitative evaluations.	The study separated validity into specific steps: Qualitative Face Validity (10 parents checked ambiguity), Quantitative Face Validity (Impact scores > 1.5), Qualitative Content Validity (12 experts checked wording), and Quantitative Content Validity (CVR > 0.56 and CVI > 0.79	Adequate / Very Good: Because both target populations (parents) and experts (SLPs) were heavily involved using strict quantitative cutoffs and qualitative feedback, content validity is highly robust.

				established by 12 experts).	
7	ICA	Inventory of Communication Attitudes (ICA) Watson et al. (1987)	Structural validity lacked Factor Analysis.	The researchers used inter-scale Pearson correlations to verify that the 4 response scales (behavior, affect, cognition) were related, but did not perform Exploratory or Confirmatory Factor Analysis (EFA/CFA) to strictly prove one-dimensionality.	Not directly impacting content validity, but downgrades the overall quality of evidence for structural validity from High to Moderate.
8	ISACS	ISACS Kelkar et al. (2024)	Insufficient sample size for structural validation.	The study included 52 PWF, which the authors explicitly acknowledged was inadequate to conduct a formal Exploratory or Confirmatory Factor Analysis (EFA/CFA) to verify if the 100 items statistically load onto the 4 proposed ICF domains.	Not directly impacting content validity, but results in a "Not Evaluated" (NE) rating for structural validity, lowering the scale's current psychometric proof.
			Absence of Test-Retest Reliability testing.	While internal consistency was evaluated, the temporal stability (test-retest reliability via ICC) of the new ISACS over time was completely omitted from the study.	Not directly impacting content validity, but lowers the comprehensive psychometric profile of the tool.

9	LCB	LCB Craig et al. (1984)	Low total variance explained by the single extracted factor.	The structural factor analysis successfully confirmed one-dimensionality (one major factor), but this factor only accounted for 36.5% of the total variance. General psychometric guidelines often desire factors to explain at least 50% of the variance.	Not directly impacting content validity, but suggests that some variance in responses is influenced by unmeasured traits outside the core "locus of control" construct.
			Lack of Confirmatory Factor Analysis (CFA).	As an older (1984) study, the authors relied exclusively on Exploratory Factor Analysis (EFA). A modern CFA on an independent sample was not performed to statistically confirm the model fit.	Not directly impacting content validity, but limits the structural validity rating.
10	OASES	OASES Original English (Yaruss & Quesal, 2006)	No formal assessment of concept saturation	Lack of explicit reporting on whether concept saturation was reached during initial focus groups and item generation	Weakens content validity per strict COSMIN standards
		OASES-A-D (Koedoot et al., 2011)	Unspecified sample size for cognitive testing during translation	No details on methods used for pilot testing or cognitive debriefing of the translated items	Limited evidence for content validity

OASES-A-J (Sakai et al., 2017)	Lack of transparency in cognitive interview procedures	No formal structured interview guides reported for the target population during cultural adaptation	Low-quality evidence for content validity regarding cultural comprehensibility
OASES-A-K (Mahesh et al., 2024)	Small sample and inadequate procedures for cognitive testing	A very small subset was used for linguistic equivalence review with limited detail on the cognitive debriefing process	Inadequate evidence of content validity in the target demographic
OASES-A-R (Tichenor & Yaruss, 2024)	Lack of patient-generated input for the shortened version	Items were selected based purely on statistical modeling (IRT) rather than new qualitative concept elicitation involving patients	May compromise comprehensiveness of the tool's content validity
OASES-PL (Wesierska et al., 2023)	Limited documentation of cognitive testing procedures across all age groups	Lack of explicit involvement of children and teenagers in concept elicitation; reliance on original English constructs without re-evaluating cultural relevance	Weakens content validity evidence, especially for youth populations
	Inadequate description of cognitive testing and lack of concept elicitation	No concept elicitation involving school-age children directly to verify if the translated items reflect their lived experiences	Instrument may not fully capture the perspective of Dutch children who stutter
Swedish OASES (Lankman et al., 2015)	Inadequate cognitive testing leading to comprehensibility	No formal concept elicitation phase for the child and teenager translations; poor	Weakens evidence for content validity and compromises reliability

			issues	understanding of items led to low internal consistency	
		OASES-A-SC (simplifies Chinese version) Ma et al. (2025)	Structural validity lacked Factor Analysis.	The researchers utilized the OASES-A-SC based on its original 4-section English structure. They did not perform Exploratory or Confirmatory Factor Analysis (EFA/CFA) to statistically confirm that the 100 items load onto the same four factors within the Chinese cultural context.	Not directly impacting content validity, but results in a "Not Evaluated" (NE) rating for structural validity parameters.
			Underpowered sample size for test-retest reliability.	Test-retest reliability was calculated using only 28 participants. COSMIN criteria require a sample size greater than or equals to 50 for adequate precision in correlation metrics like the ICC.	Not directly impacting content validity, but downgrades the overall quality of evidence for test-retest stability from High to Moderate.
11	PAiS	PAiS (Cholin et al., 2016)	Limited sample size and demographic range.	Validated on a smaller group of 21 German speaking adults.	Limits generalizability across different languages and age groups.
			Lack age-based norms	Suggestions that awareness develops with age lack experimental evidence	Uncertainty regarding utility for paediatric populations.

	Lack of differentiation between types of anticipation	Does not differentiate between prospective and immediate type of anticipation	May require qualitative follow-up to draw a precise individual profile
PAiS-R (Bies et al., 2025)	Conflicts in constructs	the items 10 and 11 measuring the avoidance behaviours like word swapping or silence	Reduced structural validity
	Weak trait association	the Item 4 had an item difficulty (p) of 0.08, outside the acceptable range (0.2-0.8).	weakened reliability
	Lack of objective clinical data	study did not administer the SSI-4 or didn't collect any speech samples for objective evaluation	Uncertain convergent validity
	Undefined discriminant boundaries	this study did not formally assess if the scale distinguishes anticipation from anxiety.	Potential construct overlap
PAiS-TR (Uysar & Yasar, 2018)	Small sample size for validation	only 60 CWS were included in the analysis	Limited evidence for the stability of the scale across age groups.
	No formal concept elicitation	The items were adapted from the existing scale rather than generated from interviews with Turkish CWS.	It weakens the evidence that the tool reflects the full lived experience of the target population.

12	Palin PRS	Palin PRS (Millard & Davis, 2016)	Content validity based on a Delphi study with high attrition.	The items were derived from an earlier Delphi study (Millard et al., 2009) where 32 parents were asked to participate, but only 22 returned the initial statements, and only 14 returned the final ratings for importance, representing a significant attrition rate.	Adequate: While attrition was high in the foundational Delphi study, the original content was generated directly by the target population (parents of children who stutter) rather than relying solely on expert opinion.
		Palin PRS-TR (Turkish) Cangi et al. (2025)	Lack of formal cognitive interviewing to assess comprehensiveness in the translated version.	The study conducted a pilot with 5 parents who rated the translated version 5/5 for clarity, relevance, and understanding. However, it did not employ formal cognitive interviews to explicitly ask Turkish parents if any important culturally specific concepts regarding stuttering impact were missing.	Because of the parents were only asked to rate the existing items and not probed for missing items, the comprehensiveness of the tool in the Turkish context cannot be completely guaranteed.
			Confirmatory Factor Analysis (CFA) fit indices do not meet strict optimal COSMIN thresholds.	The CFA resulted in a Comparative Fit Index (CFI) of 0.904 and RMSEA of 0.093. The strict COSMIN criteria for structural validity require CFI > 0.95 or RMSEA < 0.06.	Not directly impacting content validity, but it results in an "Insufficient" rating for structural validity according to strict COSMIN guidelines,

			despite the authors deeming the model acceptable.
	Test-retest reliability coefficient for Factor 1 falls below the optimal threshold.	The Pearson correlation for test-retest reliability (3-week interval) was r = 0.636 for Factor 1 (Impact on Child). COSMIN dictates that reliability correlation coefficients should be greater than or equals to 0.70.	Not directly impacting content validity, but degrades the overall quality of evidence for test-retest reliability.
Palin PRS-P Barzegar Bafrooei et al. (2025)	Utilizing the same sample for EFA and CFA.	The researchers conducted both the Exploratory Factor Analysis (EFA) and the Confirmatory Factor Analysis (CFA) on the exact same sample of 133 participants. COSMIN dictates that EFA and CFA should be conducted on independent split-samples to properly verify structural validity.	Not directly impacting content validity, but downgrades the overall quality of evidence for structural validity from High to Moderate.
	Absence of Criterion or Convergent Validity testing.	The study did not correlate the newly translated Palin PRS-P scores against any other established quality-of-life or psychological measures for parents (e.g., anxiety/stress scales or	Not directly impacting content validity, but leaves the tool's concurrent construct alignment mathematically unproven in the Persian context.

				stuttering severity ratings by clinicians).	
			Slight underpowering of test-retest sample size.	Temporal stability was calculated using an Intraclass Correlation Coefficient (ICC) on 30 parents. COSMIN criteria prefer a sample size greater than or equals to 50 guarantee high precision in stability metrics.	Not directly impacting content validity, but slightly lowers the grade of the reliability evidence.
			Lack of formal discriminant validity testing against social desirability.	The authors noted that future research should correlate PATCS scores with measures of social desirability bias to ensure children aren't just answering what they think the researchers want to hear.	Indeterminate: Without this, there is a minor risk that some "positive" attitude variance is driven by social compliance rather than true content agreement.
13	r-SASBS	Revised Scale of Avoidance and Struggle Behaviours in Stuttering (r-SASBS) Hancer & Tokgoz-Yilmaz (2025)	Lack of cognitive interviewing with patients regarding final item phrasing.	While PWS wrote initial compositions to generate items, the final wording refinements, deletions, and structural validations were dictated by expert SLPs and statistical analyses (EFA/CFA) rather than formal cognitive interviews with PWS to confirm	Doubtful: Because the final item pool was not explicitly pilot-tested via qualitative interviews with the target population to ensure phrasing was interpreted exactly as intended, there is a minor risk of misinterpretation.

				comprehension.	
			Strength (No Weakness in Content Validation): Exceptional multi-phase expert agreement.	The authors utilized strict Lawshe's criteria. Initial CVR/CVI by 10 experts was 0.96 and 0.93. During the revision phase, 9 new experts provided a CVI of 0.95 and CVR of 0.98, mathematically proving necessity and relevance.	Very Good: The quantitative rigor applied to expert panels guarantees comprehensive content coverage.
14	S-Scale	S-Scale-39 Items Erickson (1969)	Total exclusion of females in the sample.	The sample of 264 participants consisted exclusively of males to prevent a "sex differential in responses to personality inventory items".	Because no females were involved in the empirical derivation of the items, the tool's content comprehensiveness regarding the female stuttering experience cannot be assumed.
			Item deletion based solely on therapy outcomes rather than construct relevance.	The authors deleted 15 items to create the S24. Five items were deleted simply because "they showed no change with therapy" or were "couched in the past tense".	Removing items because they are stable (unchanging) traits might improve the tool's use as a therapy outcome measure, but it artificially restricts the scale's content validity regarding lifelong stuttering attitudes.

			Insufficient sample size and demographic exclusion.	The refinement was based on a small sample of 36 PWS, all of whom were "young adult males," continuing the exclusion of females and older adults.	Not directly impacting content validity, but limits the generalizability of the tool.
15	Safety Behaviors Questionnaire	Azarinfar et al. (2022) Safety Behaviors Questionnaire	Content validity lacked formal quantitative evaluation via strict Content Validity Indices (CVI/CVR).	The face and content validity relied on a Delphi method where 17 SLPs and 5 AWS rated items on a 1-6 scale, requiring 80% consensus >= 4. While this effectively established clarity, it lacked rigorous formal calculations of CVI or CVR to strictly confirm comprehensiveness and necessity.	Doubtful: Because standard quantitative indices were omitted, the rigorous mathematical proof of comprehensive content validity is slightly weakened.
			Structural validity relied on Exploratory Factor Analysis (EFA) without Confirmatory Factor Analysis (CFA).	The researchers extracted 4 factors explaining 50% of the variance using an EFA on 167 AWS. COSMIN standards suggest that a new factor structure in a translated tool should be tested with a CFA in an independent sample, which was not done.	Not directly impacting content validity, but it downgrades the overall quality of evidence for structural validity.
			Absence of external construct validity	The authors explicitly stated that comparing safety	Not directly impacting content validity, but

			evaluation.	behavior frequency against external social anxiety measures (e.g., FNE or UTBAS) "is yet to be explored".	results in a "Not Evaluated" rating for construct validity parameters.
16	SBIS	Ntourou et al. (2020) SBIS	Content validity lacked formal quantitative evaluation via Content Validity Index (CVI) or expert panel reviews.	The initial items were derived from literature and given to 10 parents of fluent children to rate for "readability and clarity". While this establishes basic face validity and qualitative comprehensibility, the study did not employ formal quantitative content validity ratios (CVR) or expert clinician panels to strictly judge the relevance and comprehensiveness of the items.	Doubtful: The lack of structured quantitative validation by experts regarding item relevance slightly diminishes the robust proof of comprehensiveness.
			Extended test-retest reliability interval.	The test-retest reliability was measured over a prolonged 8-month interval (M=8.28 months). COSMIN guidelines generally recommend a 1-to-2-week interval to ensure the construct has not naturally changed over time.	Not directly impacting content validity, but could introduce true variance in the child's temperament over 8 months rather than just measuring tool stability.

			Strength (No Weakness in Structural Validity): Exceptional Factor Analysis methodology.	Unlike many other studies, this study rigorously split a large sample (N=468) in half to perform Exploratory Factor Analysis (EFA, n=234) to establish the structure, and then Confirmatory Factor Analysis (CFA, n=234) to independently verify the model fit (X^2=3.32, p=.51, SRMR =.02, TLI =1.00, CFI =1.00, RMSEA =0.00).	Very Good: The use of both EFA and CFA on adequately powered samples provides exceptionally high-quality evidence for the structural validity of the tool.
17	SCESS	Karimi et al. (2018) SCESS	Reliability sample size falls severely below optimal COSMIN standards.	The test-retest reliability over a 2-4 week interval was calculated using only 37 AWS. COSMIN criteria require a sample size >= 50 for adequate precision.	Not directly impacting content validity, but results in downgrading the rating for test-retest reliability due to severe imprecision.
			Low absolute reliability (high Standard Error of Measurement).	The Standard Error of Measurement (SEM) was 1.75 on a 9-point scale. The authors admit this indicates a potential test-retest fluctuation of nearly 2 full points, warning that the SCESS should be used cautiously for detecting small changes within individuals.	Not directly impacting content validity, but it highlights a substantial flaw in measurement stability.

			Single-item scales inherently lack comprehensive scope.	The SCESS distils treatment satisfaction into a single 9-point question. While efficient, a single item cannot comprehensively cover all distinct nuances of behavioral and psychosocial satisfaction.	Indeterminate: The tool intentionally trades comprehensiveness for brevity.
18	Self-Assessment Scale	Huinck & Rietveld (2007) Self-Assessment Scale	Lack of formal psychometric validation procedures.	The scale was used as an outcome measure without reporting formal internal consistency, structural validity, or comprehensive content validity.	Indeterminate: Content validity was not explicitly assessed.
19	SESAS	SESAS Ornstein & Manning (1985)	Insufficient formal patient involvement in evaluating the final scale for relevance, comprehensiveness, and comprehensibility.	While the researchers initially asked 5 adult stutterers to write down speaking situations to generate a pool of 100 items, they independently pared the list down to 50 items. There is no report of conducting cognitive interviews with the target population on the final 50 items to ensure all items were comprehensible, relevant, and comprehensive.	As the final selection of items was not pilot-tested or formally evaluated by the target patient population via cognitive interviewing, it cannot be guaranteed that the 50 items comprehensively represent all relevant speaking situations from the patient's perspective.
20	SGSM	Persian SGSM (P-	Patients were asked	The study involved 10 AWS	While comprehensibility

SGSM) Hozeili et al. (2024)	about clarity, but not explicitly about comprehensiveness.	to comment on the "clarity, fluency, and comprehensibility" of the SGSM items for face validity. However, they were not explicitly asked to identify missing concepts (comprehensiveness).	is adequate, the lack of formal patient interviews specifically addressing whether any important daily speaking situations were missing limits the evidence for comprehensiveness.
	Structural validity was not evaluated before calculating internal consistency.	Cronbach's alpha was calculated for the subtests (α = 0.98) without first conducting a factor analysis to prove one-dimensionality in the Persian-speaking population.	Not directly impacting content validity, but degrades the quality of evidence for internal consistency.
Alameer et al. (2017) SGSM	Insufficient formal patient involvement in evaluating the tool for comprehensiveness and comprehensibility.	Content validity was established by comparing the listener categories of the SGSM to existing tools (WTC, SPCC, IIS). There is no report of conducting formal cognitive interviews with people who stutter to ask if all relevant speaking situations were included or if the items were clearly understood.	Doubtful: Without qualitative feedback from the target population regarding missing items, it cannot be guaranteed that the tool comprehensively captures all relevant real-life speaking situations from the patient's perspective.
	Reliability sample size falls below optimal COSMIN standards.	The test-retest reliability was calculated using a combined sample of 43 participants (22	Not directly impacting content validity, but results in a "Doubtful"

				PWS and 21 controls).	rating for reliability due to imprecision (n < 50).
21	SNA & FNA	Speech Naturalness (SNA) and Feel Naturalness (FNA) self-rating scale Finn & Ingham (1994)	Use of single-item scales and severely small sample size.	The SNA and FNA are single 9-point questions. The entire reliability and concurrent validity analysis was conducted on only 12 participants. COSMIN dictates greater than or equals to 50 is required for robust reliability metrics.	Indeterminate / Doubtful: A single item cannot comprehensively cover the multidimensional construct of speech "naturalness", and the sample is too small for generalizable validation.
22	SSC	SSC-ER-K Veerabhadrappa et al., 2021	Relied on translation methodology without formal qualitative patient interviews for relevance and comprehensiveness.	The study utilized forward and backward translation evaluated by bilingual SLPs and the original test author. It lacked formal cognitive interviews with Kannada-speaking children to systematically ask them about the relevance or if any daily speaking situations were missing from the 40 items.	Without formal qualitative feedback from the children regarding missing concepts, it cannot be guaranteed that the translated items comprehensively cover all relevant speech-associated anxiety situations for this specific cultural group.
			Structural validity was not evaluated before calculating internal consistency.	A high Cronbach's alpha (α = 0.96) was calculated for CWS. However, COSMIN dictates that a factor analysis must be conducted first to	Not directly impacting content validity, but results in downgrading the quality of evidence for internal consistency.

		prove the scale's one-dimensionality in the translated population, which was not done.	
	Reliability sample size falls below optimal COSMIN standards.	The test-retest reliability (7-10 day interval) was calculated using only 10% of the participants (n = 10 CWS).	Not directly impacting content validity, but results in a "Doubtful" rating for reliability due to severe imprecision (n < 50).
SSC-R (Vanryckeghem et al., 2017)	Lacks formal cognitive interviews with the target population for the revised tool.	The 38 items were updated and selected from clinical files and client reports over decades. However, the study does not report conducting formal cognitive interviews with adult PWS to explicitly verify if the newly revised 38 items are fully comprehensive, relevant, and comprehensible from the patient's perspective.	While the items are clinically grounded, the absence of direct, formal qualitative feedback from patients regarding potential missing concepts restricts the evidence for comprehensiveness and relevance.
	While the items are clinically grounded, the absence of direct, formal qualitative feedback from patients regarding potential missing concepts	A Principal Component Analysis was conducted on 38 items using the sample of 88 PWS. According to the COSMIN Guidelines a sample size greater than or equals to 5 * the number of	Not directly impacting content validity, but results in a "Doubtful" rating for structural validity due to an underpowered sample size.

			restricts the evidence for comprehensiveness and relevance.	items (which would require n greater than or equals to 190) or at least n > 100 for adequate factor analysis.	
23	SSS	Subjective Screening of Stuttering (SSS) Riley et al. (2004)	Content validity lacked formal quantitative evaluation via strict Content Validity Indices (CVI/CVR).	The items were derived from the clinical observations of a single primary author across 10 years and informal feedback from other clinicians. The study lacked structured expert panels, patient generation techniques, or calculation of CVR/CVI to prove comprehensiveness and necessity.	Doubtful: The lack of structured quantitative validation regarding item relevance slightly diminishes the proof of comprehensive content validity.
			Severe sample size imprecision for reliability and validity testing.	Test-retest reliability and criterion validity (against %SS and PSI) were established using an extremely small sample of only 16 PWS. COSMIN criteria require a sample size greater than or equals to 50 to consider correlation parameters adequately powered.	Not directly impacting content validity, but results in a "Low" rating for the quality of the psychometric evidence generated.
24	UTBAS	St Clare et al. (2009) Original UTBAS	Content validity relied on clinician audits	In Phase 1, the measure was derived from a 10-year	Doubtful: Because the final item pool was not

without direct cognitive interviewing of patients.	retrospective audit of clinical treatment files. The study lacks direct, formal cognitive interviewing of AWS to confirm if the 66 items were completely comprehensive from the patient's perspective or if any key unhelpful thoughts were missing.	qualitatively pilot-tested with the target population to explicitly probe for missing concepts, there is a minor risk that some patient-perceived important thoughts are missing.
Content validity relied on clinician audits without direct cognitive interviewing of patients.	In Phase 1, the measure was derived from a 10-year retrospective audit of clinical treatment files. The study lacks direct, formal cognitive interviewing of AWS to confirm if the 66 items were completely comprehensive from the patient's perspective or if any key unhelpful thoughts were missing.	Doubtful: Because the final item pool was not qualitatively pilot-tested with the target population to explicitly probe for missing concepts, there is a minor risk that some patient-perceived important thoughts are missing.
Reliability sample size falls severely below optimal COSMIN standards.	In Phase 3, the test-retest reliability over a 1-month interval was calculated using only 18 AWS participants. COSMIN criteria require a sample size>= 50 for adequate reliability testing.	Not directly impacting content validity, but results in a "Doubtful" rating for reliability due to severe imprecision (n < 50).

	Structural validity was not evaluated before calculating internal consistency.	Extremely high Cronbach's alpha scores (0.96 and 0.98) were reported. However, COSMIN dictates that a factor analysis must be conducted first to prove unidimensionality, which was not done in this preliminary study.	Not directly impacting content validity, but downgrades the overall quality of evidence for internal consistency.
Iverach et al. (2011) UTBAS	Content validity based on clinical record audit rather than cognitive interviewing.	Items were refined from 10 years of clinical therapy records. While this ensures clinical grounding, the study did not report conducting formal cognitive interviews with AWS to assess if the finalized 66 items were fully comprehensive or interpreted as intended.	Doubtful: Because the final selection was not qualitatively pilot-tested with the target population, there is a minor risk that patient-perceived important thoughts are missing.
Iverach et al. (2016) UTBAS-6	Comprehensiveness is inherently limited by the nature of a brief screening tool.	The scale was reduced from 66 items down to just 6 items to prioritize rapid clinical screening. While it accurately predicts the total score, a 6-item scale cannot comprehensively assess all nuances of speech-related anxiety.	Indeterminate / Doubtful: As a brief screener, the tool purposefully sacrifices comprehensiveness for efficiency. It is valid for screening, but doubtful for a comprehensive content assessment.
	Structural validity	The 6 items were selected	Not directly impacting

relied on regression and item-total correlations rather than Factor Analysis.	using multiple regressions, item-total correlations (.39 to .79), item-rest correlations (.39 to .79), and multicollinearity ranges (.72 to .92) to maximize prediction of the full 66-item score. COSMIN guidelines prefer Confirmatory Factor Analysis (CFA) or Item Response Theory (IRT) to formally prove structural validity.	content validity, but degrades the methodological quality of the structural validity evidence.
Lack of external criterion evaluation for Construct Validity.	The authors explicitly acknowledged: "the validity of the UTBAS-6 was not evaluated against an external criterion... it remains to be seen whether the brief UTBAS-6 is capable of achieving known-groups validity... convergent validity and discriminant validity... need to be explored".	Not directly impacting content validity, but results in a "Not Evaluated" rating for construct validity for this specific paper.
Structural validity lacked Factor Analysis.	The researchers assumed the structure of the UTBAS scales based on the original	Not directly impacting content validity, but results in a "Not

	English version. They did not perform Exploratory or Confirmatory Factor Analysis (EFA/CFA) to statistically confirm that the 66 items behave as a single construct within the Japanese cultural context.	Evaluated" (NE) rating for structural validity parameters.
Slightly underpowered sample size for test-retest reliability.	Test-retest reliability (ICC) was calculated using 42 participants. COSMIN criteria require a sample size greater than or equals to 50, to guarantee high precision in temporal stability metrics.	Not directly impacting content validity, but downgrades the overall quality of evidence for test-retest stability from High to Moderate.
Absence of a non-stuttering control group.	The normative data was collected exclusively on 141 adults who stutter. The study did not test fluent individuals, meaning discriminant (known-groups) validity could not be mathematically evaluated.	Indeterminate: Does not directly harm content validity, but restricts the generalizability of the normative clinical cut-offs compared to tools that prove discrimination against healthy controls.
Content validity lacked formal qualitative assessment with the target population.	The translation process followed standard adaptation procedures, but did not formally conduct cognitive interviews with Turkish-speaking adults who stutter to	Doubtful: The absence of qualitative patient input means the comprehensiveness of the tool in the Turkish context is not

		ensure all 66 items were relevant and comprehensive.	guaranteed.
Bernardini et al. (2024) UTBAS-ITA	Structural validity not assessed prior to internal consistency.	Cronbach's alpha was reported without first performing factor analysis to confirm that the items cluster into the theorized three-factor structure in the Turkish sample.	Not directly impacting content validity, but degrades the quality of evidence for internal consistency.
	Content validity lacked formal cognitive interviews with the target population.	The translation process involved native speakers and professional translators, ensuring wording avoided offensive terms in Italian. However, the study does not report conducting any formal pilot testing or cognitive interviews with Italian adults who stutter to verify if the translated items were completely relevant, comprehensive, and comprehensible from the patient's own perspective.	Doubtful: Because the target patient population was not directly consulted about the translated wording or missing concepts, there is no qualitative guarantee that the tool captures all aspects of the construct strictly within the Italian cultural context.
	Structural validity was not evaluated before calculating internal consistency.	A very high Cronbach's alpha score (0.99) was reported for the total scale. However, COSMIN dictates that a factor analysis must be	Not directly impacting content validity, but downgrades the overall quality of evidence for internal consistency.

	conducted first to prove one-dimensionality in the translated population, which was not done.	
Construct validity (Convergent validity) sample size falls severely below optimal COSMIN standards.	The correlations between the UTBAS-ITA and the other anxiety measures (STAI and FNE) were calculated using a small subsample of only 20 AWS. COSMIN criteria require a sample size >= 50 for adequate hypothesis testing.	Not directly impacting content validity, but results in a "Doubtful" rating for convergent validity due to severe imprecision and lack of statistical power (n = 20).
Strength: Test-retest reliability was adequately powered.	Unlike many other stuttering validation studies, this study used an adequately powered sample size of 76 AWS to evaluate test-retest reliability over a 15-day interval, successfully exceeding the COSMIN optimal threshold of n >= 50.	Not directly impacting content validity, but provides high-quality evidence for the stability of the tool.
Structural validity was not evaluated before calculating internal consistency.	A very high Cronbach's alpha score (0.99) was reported for the scale. However, COSMIN dictates that a factor analysis must be conducted first to prove one-dimensionality. The authors explicitly stated	Not directly impacting content validity, but downgrades the overall quality of evidence for internal consistency.

		that factor analysis should be left for future studies.	
Farpour et al. (2025) UTBAS-P	Reliability sample size falls severely below optimal COSMIN standards.	The test-retest reliability over a 15-day interval was calculated using only 10 AWS from the pilot group.	Not directly impacting content validity, but results in a "Doubtful" rating for reliability due to severe imprecision (n < 50).
	Divergent validity assessment was potentially underpowered.	The post hoc power analysis for divergent validity revealed an achieved power of 0.59, which is considerably lower than the conventional threshold of 0.80.	Not directly impacting content validity, but increases the risk of Type II errors, meaning weaker relationships might have gone undetected.
	Strength (No Weakness in Content Validity): Rigorous qualitative patient involvement.	This study explicitly included 10 AWS to assess face and content validity and specifically asked if any unhelpful thoughts and beliefs were missing. This resulted in the removal of an irrelevant item and the addition of 6 new culturally specific items.	Adequate / Very Good: Because the target population was directly involved in confirming the relevance and comprehensiveness, the content validity is highly robust and tailored to the cultural context.
Aydın Uysal & Ege (2020) UTBAS-TR	Content validity lacked formal qualitative assessment with the target population.	The translation process followed standard adaptation procedures, but did not formally conduct cognitive	Doubtful: The absence of qualitative patient input means the comprehensiveness of

				interviews with Turkish-speaking adults who stutter to ensure all 66 items were relevant and comprehensive.	the tool in the Turkish context is not guaranteed.
			Structural validity not assessed prior to internal consistency.	Cronbach's alpha was reported without first performing factor analysis to confirm that the items cluster into the theorized three-factor structure in the Turkish sample.	Not directly impacting content validity, but degrades the quality of evidence for internal consistency.
25	VRYCS	Singer et al. (2022) VRYCS	Absence of external concurrent/construct validity evaluation.	The authors explicitly stated that "the lack of valid English-language measures similar in purpose to the VRYCS prevented an assessment of the concurrent validity".	Not directly impacting content validity, but results in a "Not Evaluated" (NE) rating for construct validity parameters.
			Structural validity relied on PCA without Confirmatory Factor Analysis (CFA).	The researchers established the 5-factor structure using a Principal Component Analysis (PCA). COSMIN standards suggest that this newly discovered factor structure should subsequently be tested with a CFA in an independent sample, which was not performed.	Not directly impacting content validity, but it downgrades the overall quality of evidence for structural validity.
			Strength (Content	The study utilized 10	Very Good: Using a

			Validity): Rigorous quantitative expert agreement.	international experts who voted on the inclusion of items. The researchers set a strict threshold, only retaining items that achieved 90% or greater agreement among the experts, resulting in 9 items being robustly discarded.	highly specialized international panel with strict agreement thresholds guarantees that the retained items are clinically relevant and comprehensive.
26	WASSP	WASSP Wright et al. (1998)	Total lack of quantitative psychometric validation.	The paper describes the rationale and creation of the tool but does not provide any statistical analyses, factor extraction, internal consistency metrics, or reliability coefficients, explicitly stating it has only been used with "fewer than 20 clients".	While patient discussion groups were utilized for qualitative item generation (a strength), the lack of any quantitative verification means the content validity cannot be statistically confirmed by this 1998 paper alone.
		WASSP-TR Uysal & Köse (2021)	Content validity lacked strict quantitative CVI/CVR indices.	While the translation was reviewed by an expert panel of 8 professionals, the study relied on Krippendorff's alpha (0.74) to measure translation agreement rather than calculating formal Content Validity Ratios (CVR) to strictly prove item necessity.	Doubtful: As standard quantitative necessity indices were omitted, the rigorous mathematical proof of comprehensive content validity in the Turkish context is slightly weakened.
			Minimum threshold	The researchers ran the	Not directly impacting

	sample size for structural validation.	Confirmatory Factor Analysis (CFA) on 120 participants. While they noted this meets the minimum "5 participants per item" rule (5 x 24 = 120), COSMIN guidelines generally recommend much larger sample sizes ($n \ge 200$) to guarantee robust structural factor modeling.	content validity, but slightly limits the statistical power of the structural validity findings.

Note. This table details the methodological weaknesses, specific examples, and their impact on content validity for each tool evaluated. CFA = Confirmatory Factor Analysis; EFA = Exploratory Factor Analysis; CVI = Content Validity Index; CVR = Content Validity Ratio; PWS = People Who Stutter; CWS = Children Who Stutter; AWS = Adults Who Stutter.

Table S3

Article Exclusion by Full-Text with Reasons

Sl. No.	Reference	Reason for Exclusion
1	Boyle, M. P. (2012). Self-stigma of stuttering: implications for self-esteem, self-efficacy, and life Satisfaction.	Dissertation study
2	Jackson, E. S., Gerlach, H., Rodgers, N. H., & Zebrowski, P. M. (2018, September). My client knows that he's about to stutter: How can we address stuttering anticipation during therapy with young people who stutter?. In Seminars in speech and language (Vol. 39, No. 04, pp. 356-370). Thieme Medical Publishers.	Single subject case study
3	Johannisson, T. B., Wennerfeldt, S., Havstam, C., Naeslund, M., Jacobson, K., & Lohmander, A. (2009). The Communication Attitude Test (CAT-S): normative values for 220 Swedish children. International journal of language & communication disorders, 44(6), 813-825.	Study focused on normative testing and done in normal children.
4	Klarin, E., Leko Krhen, A., & Jelčić Jakšić, S. (2018). A preliminary study on the Unhelpful Thoughts and Beliefs about Stuttering (UTBAS) questionnaire in Croatia.	Normative study and not done any psychometric evaluations.
5	Langevin, M. (2009). The Peer Attitudes Toward Children who Stutter scale: Reliability, known groups validity, and negativity of elementary school-age children's attitudes. *Journal of Fluency Disorders*, *34*(2), 72–86. 10.1016/j.jfludis.2009.05.001	Wrong outcome
6	Langevin, M., Kleitman, S., Packman, A., & Onslow, M. (2009). The Peer Attitudes Toward Children who Stutter (PATCS) scale: an evaluation of validity, reliability and the negativity of attitudes. *International Journal of Language & Communication Disorders*, *44*(3), 352–368. https://doi.org/10.1080/13682820802130533	Wrong outcome
7	Veerabhadrappa, R. C., Krishnakumar, J., Vanryckeghem, M., & Maruthy, S. (2021). Communication attitude of Kannada-speaking adults who do and do not stutter. Journal of fluency disorders, 70, 105866.	Duplicate article
8	Watson, J. B., Gregory, H. H., & Kistler, D. J. (1987). Development and evaluation of an inventory to assess adult stutterers' communication attitudes. Journal of fluency disorders, 12(6), 429-450.	Duplicate article

Table S4

Evaluation of Content Validity and Other Measurement Properties Among the Studies Included in the Review

SI NO	Main Tool	Citation Details	Relevance	Comprehensiveness	Comprehensibility	Overall Content Validity	Structural Validity	Reliability	Internal Consistency	Other Measurement Properties	Recommendation
1	4S	Boyle (2013) Original 4S	(+/H) (Derived directly from qualitative interviews with 15 PWS)	(+/H) (Comprehensive 33-item evaluation covers awareness, agreement, and application)	(+/H) (Expert panel reviewed initial items specifically for clarity and redundancy)	(+/H) (Exceptional, patient-driven content generation phase)	(+/M) (Adequately powered EFA extracted 3 factors; downgraded to Moderate as they only explained 40.5% of variance, and no CFA)	(+/H) (Strong r=0.80 over 2-3 weeks and adequately powered with n=58)	(+/H) (Good alpha=0.87 total & consistent across adequately powered sample N=291)	Construct Validity: (+/H) (Proved strong theoretical correlations, establishing the tool's foundational relationship with self-esteem, self-efficacy, and life satisfaction)	Recommended for use as the psychometric study establishing the 4S as a highly reliable and valid measure of internalized stuttering stigma.
		Boyle (2015) Original 4S	(+/H) (Theoretical model	(+/H) (Comprehensive 33-item scale	(+/H) (Simple 5-point Likert system)	(+/H) (Rigorous development phase	(+/M) (Adequately powered EFA	NE (Explicitly reference	(+/H) (Excellent alpha=0.89 total &	Construct Validity: (+/H) (Extensively	Recommended as it has robust validation

	generated via explicit PWS qualitative interviews)	evaluates multidimensional psychological constructs)		established in previous 2013 foundational study)	replicated 3 factors; downgraded to Moderate as they only explained 39.5% of variance)	d prior 2013 data; not reassessed here)	highly consistent across adequately powered sample N=354)	evaluated against 7 psychometric scales. Proved highly expected correlations with depression, anxiety, hope, empowerment, and quality of life)	study establishing the tool's psychological construct framework.
Limura et al. (2022) 4S-J-16 (Short Japanese)	(+/H) (Translated from the highly relevant original 4S)	(?/M) (As a 16-item short scale, it sacrifices some comprehensive nuance for clinical speed)	(+/H) (Bilingual experts ensured conceptual comprehensibility in Japanese)	(+/M) (Strong translation, but lacks formal qualitative pre-testing with Japanese patients)	(?/M) (EFA successfully extracted 16 items explaining 49.3%; downgraded due to lack of CFA)	(+/M) (Excellent Total r=0.82 over 1 month; downgraded slightly due to minor sample imprecision n=40)	(+/M) (Good total alpha = 0.85, but Application subscale was borderline at 0.61)	Construct Validity: (+/H) (Significantly correlated with RSES self-esteem and SWLS life satisfaction exactly as hypothesized)	Potential to be recommended, but further research is required. (A highly efficient clinical screener demonstrating good validity, but requires

									CFA validation of the short structure and investigation into the Application subscale's reliability)
Tiryaki et al. (2023) Turkish 4S-TR	(+/H) (Target-population pilot testing ensured cultural alignment)	(+/M) (28-item scale reduced from 33 effectively captures the 3 core dimensions)	(+/H) (Explicitly tested by 10 PWS rating clarity 1-5, triggering instruction edits)	(+/H) (Excellent, verified multi-phase content adaptation)	(+/H) (EFA [n=120] extracted 3 factors; independent CFA [n=230] confirmed model fit)	(+/M) (Exceptional r=0.969 over 2 weeks & downgraded slightly due to minor sample imprecision n=32)	(+/H) (Excellent alpha=0.898 total & highly consistent across subscales)	Construct & Split-Half Validity: (+/H) (Scale massively differentiated along theoretical predictions against self-esteem and life satisfaction; strong split-half indices)	Recommended for use as demonstrates good structural and temporal reliability, internal consistency, and robust theoretical construct alignment.
Wesierska et	(+/H) (Evaluat	(+/H) (Includes the	(+/H) (Clarity	(+/H) (Robust,	NE (Sample	NE (Test-	(+/H) (Excellent	Construct Validity:	Potential to be

		al. (2025) Polish 4S	ed and confirmed highly relevant by 7 "double experts" [SLPs who stutter] during pilot)	full 33-item multidimensional spectrum of the original tool)	confirmed during double-expert qualitative interviews)	multi-step translation with expert target-population verification)	size n=108 was statistically inadequate to perform factor analysis)	retest correlation was not conducted)	alpha=0.93 total & subscales range 0.77-0.92)	(+/M) (Total scores correctly correlated with theoretical measures like life satisfaction and speech assessment; downgraded to Moderate due to small n)	recommended, but further research is required. (A promising translation demonstrating good initial reliability and construct correlation, but structurally unverified due to small sample size; requires future EFA/CFA and test-retest)
2	9-point Self	O'Brian et al.	(?/L)	(?/L)	(+/H)	(?/L)	NE	(+/M)	NE	(+/M)	Potential to be

	Rating Scale	(2004) 9-point Severity Scale									recommended, but further research is required.
3	BAB	Węsierska et al. (2018) Polish BAB	(+/H) (Items rigorously back-translated and verified by original author for clinical relevance)	(+/H) (Exceptionally comprehensive 204-item battery covering emotion, speech disruption, attitude, and coping)	(+/H) (Language successfully adjusted to match the linguistic nuances of Polish adults)	(+/H) (Strong, rigorous cross-cultural translation process complying with WHO guidelines)	NE (Relied on pre-established English framework; no EFA/CFA performed on Polish data)	NE	(+/H) (Exceptional α values ranging from .83 to .98 across all four subtests for both PWS and PWNS; N=274)	Known Groups & Construct Validity: (+/H) (Massively differentiated AWS from AWNS on all 4 subtests [p<0.001] with huge effect sizes [d>1.44]; no age/gender bias)	Recommended as it demonstrates excellent normative power, exceptional internal consistency, and massive clinical discriminative validity.
4	BigCAT	BigCAT	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	(+/L) low evidence as the structural validity was not proven	Not evaluated	Potential to be recommended for use, but further research is required.

BigCAT-K	(+/L) AND _	(+/L)	(+/L)	(+/L) AND _	(?/L) no factor analysis was conducted.	(+/L) Low because of small sample size i.e., 10% original data for test-retest reliability.	(+/L) low evidence as the structural validity was not proven	Construct validity: (+/L)	Potential to be recommended for use, but further research is required.
English BigCAT	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	(+/L) Low du to less sample size (n=33 and r=0.80)	Not evaluated	Not evaluated	Potential to be recommended, but further research is required to establish the quality (low quality evidence for reliability due to less sample

											size).
		Valinejad et al. (2018) Persian BigCAT	(+/H) (Rigorous back-translation and expert CVR confirmed clinical necessity of items)	(+/H) (34-item single dimension thoroughly covers speech attitude)	(+/H) (Clarity explicitly endorsed by 8 expert SLPs)	(+/H) (Solid quantitative proof via CVR > 0.75 for all items)	NE (No factor analysis performed)	(+/L) (Strong ICC = 0.89 over 7-10 days, but severely downgraded due to underpowered n=17)	(+/H) (Good KR-20 metrics for both PWS [0.88] and PWNS [0.83])	Criterion / Known Groups Validity: (+/H) (Excellent criterion r=0.79 against S-24 scale; massively differentiated AWS [Mean 24.3] from AWNS [Mean 7.0] at $p < 0.001$)	Recommended for use. (Highly recommended. Demonstrates excellent criterion validity, normative power, and internal consistency)
5	CAT	CAT	Not evaluated	Not evaluated	Not evaluated	Not evaluated	Not evaluated	(+/L) Low as under COSMIN standards, Pearson's r is considered an	Not evaluated	Not evaluated	Potential to be recommended for use, but further research is required.

						inadequate statistical method for establishing test-retest reliability.			
CAT-K	(?/L) Evaluated based on the translation process as there is no quantitative data from the patients	(?/L) as there was no formal interviews conducted with the children to verify all the aspects covered	(+/M)	(?/L) as it lacks quantitative data from the patients.	Not evaluated	(+/VL) due to less sample size(10)	(+/L) low evidence as the structural validity was not proven	Construct validity: (+/M) (Discriminant analysis classified 98% of CWS)	Potential to be recommended, but further research is required to establish the quality by examining the structural validity and adequate reliability testing.
Dutch KIddy	Not evaluate	Not evaluated	Not evaluated	Not evaluated	Not evaluated	(+/L) Low due	Not evaluated	Not evaluated	Potential to be

		CAT	d					to less sample size (n=34 and r=0.90)			recommended, but further research is required to establish the quality (low quality evidence for reliability due to less sample size).
6	CBS-PCWS	Mehdizadeh Behtash et al. (2022) CBS-PCWS	(+/H) (Phase 1 Quantitative Content Validity established via CVI > 0.79 for relevancy; CVR > 0.56 for necessity	(+/H) (Phase 1 Qualitative Item Generation directly from parent interviews ensured all dimensions were captured)	(+/H) (Phase 1 Qualitative Face Validity evaluated by 10 parents for ambiguity; Quantitative Impact Score > 1.5 proved clarity)	(+/H) (Strong, rigorous qualitative and quantitative generation /validation)	(+/M) (Phase 2 EFA conducted on sample N=205; explained 63.89% of variance, but downgraded to Moderate as CFA was not performed)	(+/M) (ICC = 0.97 total; downgraded to Moderate evidence due to minor sample imprecision n=40)	(+/H) (Phase 2 showed excellent α = 0.93 for total scale)	Measurement Error & MDC: (+/H) (SEM = 2.11; MDC = 5.84).	Can be recommended for use (Strong evidence of excellent content validity, internal consistency, structural validity, and

			)								reliable measurement error thresholds)
7	ICA	Watson et al. (1987) Inventory of Communication Attitudes (ICA)	(+/H) (Based on robust tripartite model of stuttering)	(+/H) (13 distinct social/speaking situations rigorously evaluated across 4 dimensions)	(+/H) (Pilot tested extensively in Phase I)	(+/H) (Strong, rigorous qualitative content generation)	(?/M) (Intercorrelations established relationships, but formal EFA/CFA was NE)	(+/M) (Excellent r>0.85 over 3-4 weeks; downgraded to Moderate due to slight sample imprecision n=28)	(+/H) (Excellent α>0.90 across all 4 scales in Phase I; adequately powered n=107)	Construct Validity (Discriminant): (+/H) (Scale massively differentiated AWS from AWNS controls across all 4 scales and all 13 situations at p<0.01)	Recommended for use. Despite its age, it demonstrates excellent internal consistency, reliability, and discriminant validity across well-powered samples.
8	ISACS	elkar et al. (2024) ISACS	(+/H) (Items intentionally	(+/H) (Exceptionally comprehensi	(+/H) (Language equivalence between	(+/H) (Strong culturally-adapted	NE (Sample size too small to	(+/M) (Subscales II, III, and IV	NE (Test-retest correlation	Known Groups & Concurrent: + (M)	Potential to be recommended, but

			adapted for Indian collectivistic culture, incorporating extended family/societal attitudes)	ve 100-item evaluation spanning all 4 major WHO ICF domains)	English and Marathi formally verified via Wilcoxon signed ranks test)	generation phase with dual-perspective [Self vs. Other] utility)	perform factor analysis)	showed excellent $\alpha > 0.90$; downgraded to Moderate because Subscale I was borderline $\alpha= 0.64$)	was not conducted)	(Massively differentiated pathology from normal fluency [p<0.001] and showed expected modest correlation with SSI-4; downgraded due to severe lack of cluttering sample data).	further research is required (A highly promising, culturally vital tool demonstrating good initial reliability and discrimination, but requires adequately powered EFA/CFA, test-retest, and a larger cluttering sample before broad recommendation)
9	LCB	Craig et al. (1984)	(=/H) (Relevant for	(+/M) (17 items comprehensi	(+/H) (Successfully	(+/H) (Flawless quantitativ	(+/M) (EFA successfully	(+/H) (Strong alpha=0.	(+/H) (Exceptional r=0.90	Predictive / Discriminant: (+/H)	Recommended for use. A

		LCB Scale	evaluating treatment adherence and relapse potential)	vely address personal responsibility, though general to all behavior rather than stuttering-specific)	administered across diverse clinical populations including phobics and PWS	e 5-judge consensus during item generation)	confirmed unidimensionality, but explained variance was low [36.5%] and no CFA was performed)	79 and split-half r=0.73 over a very robust mixed sample N=303)	over 1 week [n=67] and $r=0.73$ over 6 months [n=55]; adequately powered)	(Massively predicted stuttering therapy relapse at 10-months [84% accuracy]; significantly differentiated pathology from normal controls).	foundational, rigorously validated psychometric tool with profound predictive clinical utility for stuttering relapse)
10	OASES	OASES (Original) (Yaruss 2006)	(+/M)	(+/M)	(+/M)	Content validity (focus groups & panel) (+/M)	NE	Test-retest reliability (-/VL)	Cronbach's Alpha (+/H)	1. Construct validity (inter-domain correlations) (+/M) 2. Concurrent validity (correlated with S-24 scale) (+/M)	Potential to be recommended for use, but further research is required.
		OASES-A-D (Koedoot 2011)	(+/L)	(+/L)	(?/L)	Translation process evaluation (?/L)	NE	NE	Cronbach's Alpha (+/M)	1. Construct validity / Known-groups (severity subgroups) (-/L)	Potential to be recommended for use, but further research is

								2. Concurrent validity (correlated with S-24 scale) (+/M)	required.
OASES-S-D (Lankman 2015)	(+/L)	(+/L)	(?/VL)	Translation process evaluation (?/VL)	NE	NE	Cronbach's Alpha (-/M)	1. Construct validity / Known-groups (stuttering vs. control) (+/M) 2. Concurrent validity (+/M)	Potential to be recommended for use, but further research is required.
OASES-A-J (Sakai 2017)	(+/L)	(+/L)	(?/L)	Cultural adaptation & translation evaluation (?/L)	NE	Test-retest reliability: (-/VL)	Cronbach's Alpha (+/M)	1. Construct validity (inter-domain correlations) (+/M) 2. Concurrent validity (correlated with Erickson scale) (+/M)	Potential to be recommended for use, but further research is required.
Swedish OASES	(+/L)	(+/L)	(-/L)	Translation process evaluation	NE	NE	Cronbach's Alpha (-/L)	1. Construct validity (cross-	Not recommended for use

(Lindström 2020)				(-/L)				cultural comparisons) (+/L) 2. Concurrent validity / Criterion (correlated with CAT-S) (+/L)	(for youth versions) / No conclusion (adults)
OASES-PL (Węsierska 2023)	(+/L)	(+/L)	(?/VL)	Face validity (?/VL)	EFA (Exploratory Factor Analysis) (-/L)	NE	Cronbach's Alpha (+/L)	1. Construct validity (normative comparisons) (+/L) 2. Concurrent validity (correlated with other measures) (+/L)	Potential to be recommended for use, but further research is required.
OASES-A-K (Mahesh 2024)	(+/L)	(+/L)	(?/VL)	Translation & panel expert review (?/VL)	NE	Test-retest reliability (-/VL)	Cronbach's Alpha (+/L)	1. Construct validity (inter-domain correlations) (+/L) (Concurrent validity not performed)	Potential to be recommended for use, but further research is required.

		OASES-A-R (Tichenor 2024)	(+/M)	(?/L)	(+/M)	Content validity (derived from original) (?/L)	IRT (Item Response Theory) (+/M)	NE	Cronbach's Alpha (+/H)	1. Concurrent validity (correlated with original OASES-A) (+/M)	Potential to be recommended for use, but further research is required.
		OASES-A-SC (2025)	(+/H) (Carefully reviewed by 10 bilingual stuttering experts for cultural appropriateness)	(+/H) (Participants confirmed the 100 items reflect the full scope of their experience)	(+/H) (Pretested with PWS; no items were reported as difficult to understand or skipped)	(+/H) (Strong, rigorous cross-cultural translation process with patient involvement)	NE (Relied on pre-established English subscales; no EFA/CFA performed)	(+/M) (Excellent ICCs ≥ .92, but downgraded to Moderate due to slight sample imprecision n=28)	(+/H) (Excellent α ranging from .86 to .98 across sections; adequately powered N=346)	Item-Level / Construct Validity: (+/H) (Clear identification of cultural differences via floor/ceiling effect analysis)	Potential to be recommended for use, but further research is required.
11	PAiS	PAiS	(+/H) (Highly	(+/M) (Covers	(+/H) (Refined via	(+/M) (Strong	NE (No factor	(+/M) (Good	NE (Not	Known Groups &	Potential to be

	relevant qualitative generation mapping premonitory urges to stuttering	immediate anticipation well, though authors note prospective anticipation is missing)	feedback from 15 AWS during pilot phase)	conceptual foundation, but limited by small initial sample size	analysis performed)	alpha=0.76, but sample size n=21 is underpowered)	evaluated)	Concurrent: + (H) (Successfully differentiated AWS from controls; highlighted a negative correlation with %SS).	recommended, but further research is required. (A strong foundational tool, but requires structural validation and test-retest reliability)
PAiS-R	(+/H) (Removed irrelevant items like "itching" to vastly improve clinical relevance)	(+/M) (Refined to 8 items targeting immediate anticipation exclusively)	(+/H) (Endpoint-only 5-point Likert scale reduces cognitive load)	(+/H) (Exceptional empirical refinement of the scale's content)	(+/M) (PCA confirmed a single 8-item component [54.5% variance], but lacks CFA)	(+/H) (Strong $\alpha=0.88$ for the revised 8-item scale; N=78)	(+/H) (Strong ICC = 0.86 over 2 weeks; adequately powered N=49)	Concurrent: + (M) (Positive correlation with subjective severity; downgraded due to lack of objective fluency metrics or control group).	Recommended for use. The PAiS-R resolves the structural flaws of the original scale, demonstrating excellent internal and temporal

											reliability.
		PAiS-TR	(+/L)	(+/L)	(+/L)	(+/L)	(?/L) no factor analysis was conducted.	(+/L) Low because of small sample size of 20 for test-retest reliability.	(+/L) low evidence as the structural validity was not proven	Construct validity: (+/L)	Potential to be recommended for use, but further research is required.
12	Palin PRS	Palin PRS	(+/H) As the items were generated by parents in a Delphi study.	(+/M) As parents generated 129 statements, ensuring thorough coverage of constructs.	(+/H) Pilot study found to have minimal respondent burden and completed unaided.	(+/H) Strong qualitative generation directly from the target population.	(+/H) PCA/EFA conducted on adequate sample N=259; explained >50% of variance [57%] and all item factor loadings >0.30 [all >0.40].	NE	(+/H) Cronbach's $\alpha > 0.83$ across all three factors; supported by prior proven structural validity.	Normative Data: (+/H) Calculated standard percentile and stanine ranks based on the sample of 259	Recommended for Use as there is a strong evidence of sufficient content validity, structural validity and internal consistency.

Palin PRS-TR (Turkish)	(+/M) 5 Parents confirmed the relevance.	(?/L) as no formal cognitive interviews with parents to assess the items which are missed.	(+/M) As 5 parents confirmed about the clarity and understanding.	(?/L)Downgraded due to lack of comprehensive qualitative testing in the target cultural context.	(-/H) CFA conducted on adequate sample N=193; however, mathematically insufficient per COSMIN as CFI = 0.904 [Not > 0.95] and RMSEA = 0.093 [Not < 0.06].	(+|-/H) Insufficient for F1 [r=0.636], but sufficient for F2 [r=0.708] and F3 [r=0.724]; adequately powered N=118.	(+/M) α > 0.83 and ω > 0.85 across all factors; downgraded because CFA structural validity indices were sub-optimal.	Measurement Invariance: (+/H) Supported configural, metric, and scalar across genders. IRT: (+/H) Most items showed very high discrimination. Convergent Validity: (+/H) CR > 0.81 for all factors.	Potential to be recommended, but further research is required to establish the quality as it Shows strong psychometric potential, but requires further structural validity refinement to meet strict fit index thresholds and deeper qualitative content validity testing.
Barzeg	(+/H)	(+/H)	(+/H)	(+/H)	(+/M)	(+/H)	(+/M)	Floor/Ceiling	Potential

ar Bafrooei et al. (2025) Palin PRS-P	(Quantitative CVR metrics by 11 expert SLPs proved strict relevance/necessity)	(19 items thoroughly cover child impact, parent impact, and parent confidence)	(Pilot tested on 10 parents specifically to refine item clarity and simplicity)	(Highly rigorous translation process utilizing strict quantitative expert consensus)	(Both EFA and CFA showed good fit indices, but downgraded to Moderate because they were run on the same sample N=133)	(Excellent alpha metrics ranging from 0.77 to 0.89 across all factors; adequately powered sample)	(Outstanding ICC = 0.97 over 2 weeks, but downgraded to Moderate due to minor sample imprecision n=30)	Effects: (+/H) (Zero significant floor or ceiling constraints found). Convergent/ Criterion: NE (No external scale correlations performed).	to be recommended, but further research is required to establish the quality (A highly reliable and content-valid clinical tool for Iranian parents; requires independent CFA and convergent validity testing against external benchmarks for an A rating)

13	r-SASBS	Hancer & Tokgoz-Yilmaz (2025) r-SASBS	(+/H) (Items derived directly from PWS compositions and validated by expert CVR=0.98)	(+/H) (Expert CVI=0.95 ensured thorough coverage of both avoidance and struggle)	(?/M) (Experts checked intelligibility, but lacked formal patient cognitive interviews)	(+/H) (Strong, rigorous quantitative content generation)	(+/H) (EFA [n=165] followed by independent CFA [n=200] confirming excellent two-factor fit)	(+/H) (Excellent ICC = 0.97 over 20-27 days; adequately powered with n=110)	(+/H) (Cronbach's α = 0.98; highly consistent)	Construct Validity (Discriminant): (+/H) (Scale massively differentiated PWS [Mean 37.03] from PWNS [Mean 1.29]; Cohen's d = 3.67)	Can be recommended for use as it demonstrates excellent structural validity, internal consistency, reliability, and discriminant validity across highly powered samples)
14	S-Scale	Erickson (1969) S-Scale (39 Items)	(+/H) (Items rigorously vetted for readability and social desirabili	(?/M) (Broad coverage of attitudes, but completely excluded females from derivation)	(+/H) (Filtered specifically to guarantee a 6th-grade reading level)	(+/H) (A massive, 1060-item empirical generation process yielding strong	NE	NE	(+/H) (Excellent Split-half = 0.92 and KR-20 = 0.91; adequately powered N=264)	Concurrent / Discriminant: (+/H) (Solid correlations against self-rated severity [0.46] and discomfort	Recommended for use as it provides excellent empirical derivation and internal

			ty)			quantitative inclusion criteria)				[0.66]; clearly distinguished personality adjectives)	consistency, despite needing later factor analysis.
15	S24 Scale	Andrews & Cutler (1974) S24 Scale	(+/H) (Items retained specifically for their sensitivity to stuttering therapy)	(?/L) (Deleted 15 items, sacrificing broad trait comprehensiveness to create a state-based outcome measure)	(+/H) (Inherited Erickson's 6th-grade reading level parameters)	(?/M) (Content generation relied entirely on Erickson's prior work)	NE	NE	(+/M) (Removed items with r < 0.70; ensuring the final 24 items are highly stable in nonstutterers)	Responsiveness / Known Groups: (+/H) (Proved massively sensitive to clinical change [19.2 to 9.1] and significantly differentiated AWS from AWNS)	Recommended for use as the clinical outcome measure have good dynamic sensitivity to therapeutic change.
16	SA Scale	Huinck & Rietveld (2007) SA Scale	(?/L)	(?/L)	(+/H)	(?/L)	NE	NE	NE	Construct Validity: (?/L)	Should not be recommended.

17	Safety Behaviors Questionnaire	Azarinfar et al. (2022) Persian Safety Behaviors Questionnaire	(+/M) (Delphi panel of 17 SLPs & 5 AWS evaluated relevance; >80% scored >= 4)	(?/L) (No formal CVR/CVI calculated to strictly verify comprehensiveness)	(+/M) (11 items modified for clarity during Delphi phase)	(+/M) (Adequate qualitative face validation but lacks rigorous quantitative indices)	(+/M) (EFA conducted on N=167 [explained 50% variance]; downgraded to Moderate as CFA was not performed)	NE (Explicitly omitted; reserved for future study)	(+/H) (Excellent α = 0.89 total; inter-item correlations were optimal)	NE (Not evaluated against external criteria like social anxiety measures)	Potential to be recommended, but further research is required. (Promising cultural translation with excellent internal consistency, but requires CFA, test-retest reliability, and construct validity testing against external criteria)
18	SBIS	Ntourou et al. (2020)	(+/M) (Items derived	(?/L) (No formal expert panel	(+/M) (Tested by 10 parents	(+/M) (Sufficient qualitative	(+/H) (Excellent methodolog	(+/H) (ICC = .88 and r	(+/H) (Cronbach 's α = .81,	Construct Validity: (+/H)	Can be recommended for use

		Short Behavioral Inhibition Scale (SBIS)	directly from established psychological literature regarding BI traits)	or CVR testing to guarantee no missing concepts)	specifically for readability, clarity, and wording improvements)	generation, but downgraded due to lack of formal quantitative expert evaluation)	y: EFA [n=234] followed by independent CFA [n=234] confirming an excellent single-factor fit)	= .79; adequately powered with n=76)	adequately supported by high-quality structural validity)	(Extensively evaluated on large samples. Excellent Convergent validity with BSQ [r=-.68, n=391]; Excellent Divergent validity [r=-.02, n=391]; Significant Concurrent validity with behavioral latency [rho=-.31, n=53]; and robust Known-Groups differentiation [p=.002]).	(Highly recommended. Demonstrates excellent structural validity, internal consistency, reliability, and broad construct validity testing across highly powered samples)
19	SCESS	Karimi et al. (2018) SCESS Scale	(+/H) (Confirmed highly relevant by 14	(?/L) (Single item intentionally sacrifices comprehensi	(+/H) (Qualitative interviews with AWS confirmed	(+/M) (Excellent clinical relevance but	NE (Not applicable for single item)	(-/VL) (Spearman r=0.77, but	NE (Not applicable for single item)	Construct Validity: (+/H) (Adequately powered	Potential to be recommended, but further

			experts and multiple AWS groups)	ve scope for brevity)	the phrasing was universally understood)	inherently limited comprehensiveness)		absolute SEM was very high [1.75], and sample size was severely underpowered at n=37)		[N=87] evaluation proved hypothesized convergent [OASES, UTBAS, SPAI] and divergent relationships).	research is required. (Demonstrates excellent construct correlations and clinical utility as a fast screener, but severely lacking in absolute test-retest stability for individual clinical tracking)
20	SESAS	SESAS	(?/(L) (Items derived from 5 patients, but no formal quantitative or qualitativ	(?/L) As no formal cognitive interviews to check for missing concepts in the final scale	NE	(?/L) Downgraded due to lack of rigorous qualitative patient involvement.	NE	Test-retest reliability: (+/VL)	NE	Construct validity: (+/M) (r=-0.71 vs Erikson scale and r=-0.52 vs PSI and effectively discriminated between	Potential to be recommended, but further research is required to establish the quality by

			e rating of the final 50 items)							AWS and AWNS.	reassessing structural validity, content validity and adequately powered reliability.
21	SGSM	Persian SGSM	(+/M) As CVR > 0.62 by 10 experts and face validity was checked by 10AWS.	(?/L) as no formal cognitive interviews with children to check the life situations which are missed.	(+/M) As 10 AWS confirmed that the items were clear and comprehensible.	(?/L) Downgraded due to lack of rigorous qualitative patient involvement.	NE	Test-retest reliability: (+/VL) (ICC > 0.93 across measures and severe imprecision due to less sample of n=30 CWS)	(+/L) (Total α=0.98; downgraded as structural validity was not)	Convergent validity and responsiveness: (+/M) As the MSR= 1.09 for SS%.	Potential to be recommended, but further research is required to establish the quality by assessing the structural validity.
		SGSM	(?/L) as no patient rating	(?/L) as no formal cognitive interviews with children	NE	(?/L) Downgraded due to lack of rigorous	NE	Test-retest reliability: (+/L) Due to	NE as there were only one items with respect to	Convergent and discriminant validity: (+/M) As it is	Potential to be recommended, but further

				to check the real life situations which are missed.		qualitative patient involvement.		less sample size	each situations.	correlated with SPCC nad WTC.	research is required to establish the quality by assessing the content validity and adequately powered reliability.
22	SNA & FNA	Finn & Ingham (1994) SNA / FNA Scales	(+/M) (Clinical utility for naturalness tracking)	(-/VL) (Single items cannot capture multidimensional scope)	(+/M) (Simple 9-point anchored scale)	(?/L) (Adequate for a screener, but severely restricted comprehensiveness)	NE (Not applicable for single items)	(+/L) (Exact/adjacent agreement was 85-100%, but severely downgraded due to underpowered sample n=12)	NE (Not applicable for single items)	Concurrent Validity: (+/L) (Self-ratings correlated highly with listener judgments [r=0.79−0.89], but severely underpowered with n=12)	Potential to be recommended, but further research is required. (A useful clinical metric that demonstrates reliability against listener judgments, but

											requires adequately powered reliability testing)
23	SSC	SSC-R (SSC-ER &SSC-SD)	(?/L) as it is addressed via expert translation process, but lacks quantitative patient ratings	(?/L) as no formal cognitive interviews with children to verify if all anxiety-inducing situations are covered	(?/L) As no formal testing done with the patients were reported.	(?/L) Downgraded due to lack of qualitative patient involvement for the revised tool.	(+/VL) Due to less sample number.	NE	(+/L) Quantitatively excellent but it doesn't measured factor basis rather than it measures as a whole.	Discriminant/ construct validity: (+/H) As it differentiating PWS from PWNS.	Potential to be recommended, but further research is required to establish the quality by doing formal content validity and an adequate structural validity.
		Veerabhadrappa et al. (2021) (SSC-ER-K)	(?/L) as it is addressed via expert translatio	(?/L) as no formal cognitive interviews with children to verify if all	(+/M) as Pilot tested with 2 practice questions; 5 children	(?/L) Downgraded due to lack of rigorous qualitative	NE	(+/VL) (ICC = 0.98 for CWS; severe imprecisi	(α = 0.96 for CWS; mathematically excellent but	Discriminant validity: (+/M) As it accurately classified the 95% of the	Potential to be recommended, but further research is

			n process, but lacks quantitative patient ratings	anxiety-inducing situations are covered	required neutral explanations, indicating instructions were monitored for clarity	patient involvement for comprehensiveness		on due to underpowered sample of n=10 CWS)	downgraded to Low as structural validity was not	cases.	required
24	SSS: Research Ed.	Riley et al. (2004) SSS: Research Ed.	(+/M) (Clinically derived relevance)	(?/L) (As an 8-item screener, it inherently sacrifices comprehensive scope for brevity)	(+/M) (Designed with simple, direct questions)	(+/M) (Adequate for a screener, but lacks quantitative CVR/CVI proof)	NE (No factor analysis performed)	(+/VL) (Excellent Pearson r=0.79-0.93, but severely downgraded due to underpowered sample n=16)	(+/L) (Cronbach's alpha not reported, but item-to-area correlations were all ≥ 0.81; downgraded due to small n=32)	Criterion/Concurrent Validity: (+/L) (Subscales correlated in expected directions against %SS and PSI; but severely underpowered with n=16)	Potential to be recommended, but further research is required. (Although it is a useful clinical screener that demonstrates the independence of severity and avoidance, but requires

											adequately powered reliability testing and formal structural validation)
25	UTBAS	St Clare et al. (2009) UTBAS (Original)	(+/M) (Phase 1 clinical audit ensured high relevance)	(+/M) (The 10-year audit ensured a broad and thorough capture of thoughts)	(+/M) (Items are clinically derived from actual patient statements)	(+/M) (Strong clinical basis, but lacks formal qualitative cognitive interviews)	NE (Not evaluated in this preliminary study)	(+/VL) (r=0.89 and paired t-test confirmed stability; severe imprecision due to underpowered sample of n=18)	(+/L) (α=0.98; mathematically excellent but downgraded to Low as structural validity was NE)	Construct Validity (Convergent/ Discriminant): (+/M) (Phase 2 demonstrated strong convergent validity with SPAI/FNE, excellent discriminant validity, and large known-groups differentiation; adequately powered with n=57)	Potential to be recommended, but further research is required. (A highly promising foundational study that demonstrates excellent construct validity and responsiveness, but requires structural

									validity and an adequately powered reliability trial)
Iverach et al. (2011) UTBAS scales	(+/M) (Based on long-term clinical audit)	(+/M) (Extensive 10-year audit ensured depth of items)	(+/M) (Clinically derived wording)	(+/M) (Strong clinical basis)	(+/H) (PCA confirmed 3-factor structure)	(+/H) (r=0.91–0.94; n=40)	(+/H) (α=0.97–0.98)	Construct Validity (Convergent/ Discriminant) : (+/H) (Significant correlation with FNE and SPAI; distinguished social anxiety groups)	Can be recommended for use (Highly recommended for research and clinical use)
Iverach et al. (2016) Brief UTBAS-6	(+/M) (Items selected to cover the most relevant clinical content)	(?/L) (Sacrifices comprehensiveness to serve as an efficient 6-item screener)	(+/M) (Derived from the highly comprehensible original scale)	(?/L) (Sufficient as a screener, but downgraded for comprehensive content validity)	(?/L) (Item reduction successfully achieved via item-total correlations and multicollinearity modeling	NE (Not evaluated in this paper)	(+/M) (α = 0.82; acceptable for a brief clinical screening tool)	Construct Validity (Convergent/ Discriminant/ Known Groups): NE (Authors explicitly stated these parameters were not evaluated and	Potential to be recommended, but further research is required. (Demonstrates excellent diagnostic accuracy

					on a large sample N=337, but lacked formal EFA/CFA validation)			are required in future research)	and internal consistency as a brief screener, but requires formal structural validity, test-retest reliability, and construct validity testing against external criteria)
UTBAS-J Chua et al., 2017	(+/H) (Items rigorously back-translated and verified by original authors and local SLPs for	(+/H)	(+/H)	(+/H)	NE	(+/M)	(+/H)	Convergent / Construct: (+/H) (Proved strong, highly significant concurrent validity against Japanese versions of the OASES,	Potential to be recommended, but further research is required. A highly efficient and reliable clinical

	clinical relevance)							LSAS, and K6 metrics).	translation demonstrating excellent convergent validity, but requires CFA validation to achieve level A status.
Aydın Uysal & Ege (2020) UTBAS-TR	(?/L)	(?/L)	(+/M)	(?/L)	NE	(+/M) (ICC = 0.92–0.95)	(+/L) (α = 0.94)	Construct Validity (Convergent/ Discriminant) : (+/M) (Significant construct/concurrent validity correlations)	Potential to be recommended, but further research is required.
Bernardini et al. (2024) Italian UTBAS	(?/L) (Translated carefully, but no formal quantitati	(?/L) (No formal cognitive interviews with the target population to	(?/L) (Translated by experts to ensure clear wording, but	(?/L) (Downgraded due to a lack of direct, formal qualitative	NE	(+/L) (α = 0.99; mathematically excellent but	(+/M) (Spearman r = 0.93 to 0.95; adequately powered with an	Construct Validity (Discriminant / Known-Groups): (+/H) (Successfully	Potential to be recommended, but further research is required.

(UTBAS-ITA)	ve or qualitative patient rating of relevance reported)	check for missing concepts)	comprehensibility was not pilot-tested directly on patients)	involvement of patients who stutter during the cross-cultural adaptation)		downgraded to Low as structural validity was NE)	optimal sample size of n=76; downgraded slightly because structural validity was NE)	differentiated AWS from AWNS with a highly powered overall sample of N=196). Construct Validity (Convergent): (+/VL) (Significantly correlated with FNE [r=0.730] and STAI-Y2 [r=0.466], but severely downgraded for imprecision due to underpowered sample of n=20).	(Demonstrates excellent test-retest reliability and discriminant validity, but requires structural validity assessment, adequately powered convergent validity testing, and qualitative content validity testing with the target population)
Farpour et al. (2025)	(+/H) As Rated >= 3 by	(+/H) As the target patients were	(+/H) (97% of items were	(+/H) (Strong qualitative	NE	(+/VL) (Pearson r = 0.87;	(+/L) (α = 0.99; mathemati	Construct Validity (Convergent)	Potential to be recommen

Persian UTBAS (UTBAS-P)	experts and patients; 1 irrelevant item was successfully identified and removed	explicitly interviewed for missing concepts, resulting in 6 new cultural items added	rated as completely understandable and clear by both patients and experts)	generation and validation directly involving the target population)		severe imprecision due to underpowered sample of n=10 AWS)	cally excellent but downgraded to Low as structural validity was NE)	: (+/H) (Positively correlated with established anxiety measures: BAI, STAI-T, SPAI Diff, BFNE, NEO-N; post hoc power analysis confirmed a high power of 0.99). Construct Validity (Divergent): (+/M) (Negatively or weakly correlated with unrelated personality traits; downgraded to Moderate quality evidence	ded, but further research is required. (Demonstrates excellent content validity and strong construct correlations, but requires structural validity assessment and an adequately powered test-retest reliability trial before achieving a Level A recommendation)

										because post hoc power analysis revealed a low achieved power of 0.59).	
26	VRYCS	Singer et al. (2022) VRYCS Rating Scale	(+/H) (10 international stuttering experts voted on 30 items; only those with >= 90% agreement were kept)	(+/M) (Items derived from comprehensive clinical practices, theories, and research on parent behaviors)	(?/L) (Lack of formal qualitative cognitive interviews with the parents to verify wording interpretation)	(+/M) (Strong expert-driven content validity, but downgraded slightly due to lacking parent cognitive interviews)	(+/M) (Adequately powered [N=214] PCA identified 5 factors explaining 59.3% variance; downgraded to Moderate as CFA was not performed)	NE (Explicitly identified as a necessary step for future research)	(+/M) (Overall α = 0.645; subscales ranged from 0.516 to 0.880, acceptable for factors with 3-4 items)	NE (Explicitly omitted due to a lack of similar English-language tools; known-groups testing against CWNS reserved for future research)	Potential to be recommended, but further research is required. (Demonstrates solid content validity and structural development, but requires CFA, test-retest reliability, and construct validity

											testing against external criteria to achieve Level A)
27	WASSP	Wright et al. (1998) Original WASSP	(+/M) (Informal feedback from clinicians and PWS discussion groups confirmed clinical relevance)	(+/M) (The initial framework covered behaviors, avoidance, feelings, and disadvantage)	(?/L) (Assumed comprehensible by users, but no formal quantitative clarity metrics reported)	(?/L) (Qualitative foundation is present, but lacks any quantitative validation data in this specific paper)	NE	NE	NE	Construct: NE (Authors noted clinical changes aligned with predictions, but no statistical construct testing was performed)	Not be recommended for use. (Based strictly on this 1998 introductory tutorial alone, the tool cannot be recommended without the subsequent formal psychometric validations that followed

									in later years)
Uysal & Köse (2021) WASSP-TR (Turkish)	(+/H) (Translation approved by 8 experts achieving a 0.74 Krippendorff alpha for appropriateness)	(+/H) (Contains the full, finalized 24-item/5-subscale structure mapping to WHO ICF domains)	(+/H) (Pre-tested on 20 PWS specifically to resolve comprehensibility and spelling issues)	(+/H) (Strong, rigorous cross-cultural translation process with target population pilot-testing)	(+/H) (CFA confirmed perfect theoretical model-data fit indices [CFI=0.99, RMSEA=0.059])	(+/H) (Exceptional total r=0.972 over 2 weeks; adequately powered with n=30 for clinical tests)	(+/H) (Excellent α=0.949 total; highly consistent across subscales >0.85)	Criterion & Known Groups: (+/H) (Proved strong concurrent validity against SF-36 QOL metrics; successfully differentiated stuttering severity groups via ANOVA)	Recommended for use as it demonstrates good structural validity, excellent internal and temporal reliability, and robust theoretical construct alignment within the Turkish demographic.

Note. This table details the evaluation of content validity, structural validity, reliability, and other measurement properties for each tool. (+/H) = Adequate/High; (+/M) = Adequate/Moderate; (+/L) = Adequate/Low; (?/L) = Doubtful/Low; NE = Not Evaluated. CFA = Confirmatory Factor Analysis; EFA = Exploratory Factor Analysis; PWS = People Who Stutter; AWS = Adults Who Stutter.