Differential Diagnosis of Childhood Apraxia of Speech Compared to Other Speech Sound Disorders: A Systematic Review

Article *in* Journal of Speech Language and Hearing Research · November 2020

DOI: 10.1044/2020_AJSLP-20-00063

Elizabeth Murray
The University of Sydney
22 PUBLICATIONS 443 CITATIONS
SEE PROFILE

Jenya Iuzzini-Seigel
Marquette University
23 PUBLICATIONS 331 CITATIONS
SEE PROFILE

Edwin Maas
Temple University
55 PUBLICATIONS 1,523 CITATIONS
SEE PROFILE

Hayo Terband
Utrecht University
43 PUBLICATIONS 549 CITATIONS
SEE PROFILE

Differential diagnosis of Childhood Apraxia of Speech compared to other Speech Sound Disorders: A Systematic Review.

Authors: Elizabeth Murray[1,2], Jenya Iuzzini-Seigel[3], Edwin Maas[4], Hayo Terband[5], Kirrie J. Ballard[1]

Affiliation of all authors:

[1] The University of Sydney, Sydney, Australia

[2] Remarkable Speech + Movement, Sydney, Australia

[3] Marquette University, Milwaukee WI, USA

[4] Temple University, Philadelphia PA Pennsylvania, USA

[5] Utrecht Institute of Linguistics - OTS, Utrecht University, Utrecht, the Netherlands

Corresponding author:

Elizabeth Murray
Faculty of Health Sciences
The University of Sydney
PO Box 170
Lidcombe 1820
Phone: +61 2 9351 9780
Email address: elizabeth.murray@sydney.edu.au

ORCID IDs:

Elizabeth Murray - https://orcid.org/0000-0002-0883-155X

Jenya Iuzzini-Seigel - https://orcid.org/0000-0002-7679-2556

Edwin Maas - https://orcid.org/0000-0003-4452-2196

Hayo Terband - https://orcid.org/0000-0001-7265-3711

Kirrie J. Ballard - https://orcid.org/0000-0002-9917-5390

Keywords: (not in title)

Dyspraxia, Assessment, Sensitivity/specificity, Phonology, Dysarthria

Conflict of Interest Statement:

There are no conflicts of interest. Authors did not review their own manuscripts in this systematic review.

Funding Statement:

There was no funding for this work.

Abstract

Purpose: To determine the discriminative features that might contribute to differentiation of childhood apraxia of speech (CAS) from other speech sound disorders (SSD).

Method: A comprehensive literature search was conducted for articles or doctoral dissertations that included ≥1 child with CAS and ≥1 child with SSD. Of 2071 publications screened, 53 met the criteria. Articles were assessed for (a) study design and risk of bias; (b) participant characteristics and confidence in diagnosis and (c) discriminative perceptual, acoustic or kinematic measures. A criterion was used to identify promising studies: AAN study design (Class III+), replicable participant descriptions and adequate confidence in diagnosis (≥3), and ≥1 discriminative and reliable measure.

Results: Over 75% of studies were retrospective, case-control designs and/or assessed English-speaking children. Many studies did not fully describe study design and quality. No studies met the Class I (highest) quality rating according to AAN guidelines. CAS was most compared to speech delay/phonological disorder. Only 6 studies had diagnostic confidence ratings of 1 (best). Twenty-six studies reported discriminative perceptual measures, 14 reported discriminative acoustic markers, and 4 reported discriminative kinematic markers. Measures were diverse and only two studies directly replicated previous findings. Overall 7 studies met the quality criteria and another 8 nearly met the study criteria to warrant further investigation.

Conclusions: There are no studies of the highest diagnostic quality. There are 15 studies that can contribute to further diagnostic efforts discriminating CAS from other SSD. Future research should utilize careful diagnostic design, support replication and adhere to standard reporting guidelines.

Plain language summary

Childhood apraxia of speech (CAS) is a motor speech disorder where children have difficulties planning movement to speak clearly. CAS is hard to differentiate from other speech problems. This review looked for evidence-based measures to help contribute to determining CAS from other speech problems.

The systematic review reviewed 53 articles found after a thorough search and screening process. Authors rated studies based on their study quality, descriptions of the children who participated in the studies and the measures that discriminated CAS from other speech problems.

The results showed no studies met the highest quality rating. Most studies used data collected in the past and deliberately compared select groups rather than looking across a sample representative of the population. Many studies did not fully describe study design and quality. CAS was most compared to speech delay/phonological disorder, a meaning-based speech disorder. Overall 15 studies reported measures that discriminated groups with enough quality to warrant further testing and research.

Children with speech sound disorders (SSDs) have "*gaps or simplifications in their speech sound systems that can make what they say difficult to understand*" (Bowen, 2015 p. 3). SSDs are one of the most prevalent types of communication difficulties of early childhood (Eadie et al., 2015). They constitute between 30-75% of a pediatric speech-language pathologists' (SLPs) caseload (ASHA, 2016; Broomfield & Dodd, 2004; McLeod & Baker, 2014). SSD is an umbrella term for a wide range of difficulties due to known etiology (e.g., craniofacial anomalies, hearing impairment, down syndrome and other neurodevelopmental or genetic anomalies), or unknown etiology (IEPMCS, 2012). The focus here is on one specific SSD - childhood apraxia of speech (CAS; also known as developmental apraxia of speech and developmental verbal dyspraxia). CAS is a developmental, neurological SSD that affects motor planning and/or programming (ASHA, 2007). The impairment in children with CAS can be described as a deficit in transforming phonological codes into motor speech commands (ASHA, 2007; Terband, Maassen, Guenther, & Brumberg, 2009) essentially the early stages of planning the movements that will generate the intended speech sounds and sound movement sequences. The outcome of this deficit is unintelligible, inconsistent and robotic speech due to difficulties timing the articulators and transitioning from sound to sound and syllable to syllable (ASHA, 2007). Like other SSDs, the impact of CAS is not only on speech production. It adversely affects social communication and increases the risk of bullying, mental health concerns, and difficulties with phonological awareness and literacy skill development (Lewis, Freebairn, Hansen, lyengar, & Taylor, 2004; Murray & Iuzzini-Seigel, 2017; Rusiewicz, Maize, & Ptakowski, 2018). It also can negatively impact academic and occupational success (Carrigg, Parry, Baker, Shriberg, & Ballard, 2016; Lewis, Freebairn, Hansen, Taylor, et al., 2004).

An initial assessment battery (Macrae, 2016) is needed to assess a child suspected of any SSD to determine (a) if they have a SSD, (b) what type of SSD they have and (c) what evidence-based treatments may be of greatest benefit to them. SLPs have effective means to determine if a child has an SSD compared to typical development, using speech samples assessed with published developmental norms and considering functional needs (Storkel, 2019). Where the client matches the population sample, standardized tests are used (e.g., Goldman & Fristoe, 2015). However, *differential* diagnosis of CAS from other SSDs remains a clinical challenge.

There are currently no evidence-based diagnostic guidelines, criteria, or markers for reliably differentiating CAS from a range of other SSDs (Dodd, 2014; Iuzzini-Seigel & Murray, 2017). The key challenges are the current reliance on perceptually-based feature lists for differential diagnosis that lack operational definitions of each feature or critical thresholds for degree, frequency, and context of these features; as well as a lack of a psychometrically robust assessment tool (e.g., ASHA, 2007). This is in the face of overlapping symptoms across SSDs (McCabe, Rosenthal, & McLeod, 1998; Rupela, Velleman, & Andrianopoulos, 2016), high rates of comorbidity (Iuzzini-Seigel, 2019; Liégeois & Morgan, 2012) and variability across and within children over time (Maassen, Nijland, & Terband, 2012). Despite these challenges, we need effective differential diagnosis to provide appropriate information to families and provide treatments that target the specific underlying impairment(s).

There are many descriptive and diagnostic studies that have been published since CAS was first mentioned in the literature by Morley, Court, and Miller (1954). As a starting point, the aim of the current study is to systematically review this literature to evaluate the rigor of this body of work and determine any potential markers, procedures or tools that can support future research and clinical efforts to accurately differentiate children with CAS from other SSDs.

Existing literature reviews on the diagnosis of CAS

There have been three previous literature reviews specifically focused on diagnostic methods involving children with CAS, all of which focused on published assessment tools. In the first, McCauley and Strand (2008) assessed six standardized tests of oral non-verbal and speech motor performance for their reliability and methods of interpreting results. Their threshold of acceptability included acceptable normative data, clear-cut behavioral standards for decision making, adequate information describing participants and statistics used, a reliability coefficient (e.g., intra-class correlation coefficient) of 0.90 or higher for inter-rater reliability and clear methods of content, construct and validity. The Verbal Motor Production Assessment for Children (VMPAC, Hayden & Square-Storer, 1999) met more criteria than the other tests, yet still failed to meet six of the eight criteria. For example, the VMPAC only contained norms for typically developing children, lacked statistical examination of construct validity and provided limited guidance on interpreting the test for therapy planning or measuring change over time. Overall, none of the tools met the criteria to be recommended for clinical use. Secondly, Gubiani, Pagliarin, and Keske-Soares (2015) reviewed published assessment tools for CAS used in any journal article for either initial diagnosis or differential diagnosis purposes to determine tools for Brazilian Portuguese speakers. They identified 5 assessment tools: VMPAC, Dynamic Evaluation of Motor Speech Skills (DEMSS), Kaufman Speech Praxis Test, The Orofacial Praxis Test, and the Madison Speech Assessment Protocol (MSAP) and described each of these without critical appraisal. They concluded that none were adapted and standardized for use with clients who speak Brazilian Portuguese, but that the DEMSS is suitable for adaptation as it has a defined protocol and known reliability and validity. Finally, Sayahi and Jalaie (2016) used a brief systematic search strategy of four databases and identified 13 studies fitting their search criteria. The authors agreed with McCauley and Strand (2008) that the VMPAC was the most reliable test to use. They also assessed the frequency of clinical markers in the diagnosis of CAS and found inconsistency was the most frequently used marker in the studies they reviewed. However, there were multiple journal articles and tools that were not identified (e.g., the DEMSS was not found in their search). They also did not critically review the methodology or outcomes of the studies, which is essential for determining which protocols and assessment tools are sufficiently reliable, valid and sensitive/specific to be implemented clinically. Thus, there is a need for an extensive review of the peer-reviewed literature for procedures or markers under evaluation. A review of experimental studies is warranted to determine which protocols and behavioral markers show promise for sensitive and specific differential diagnosis of CAS, and which can be implemented both in clinical and research practice with greater confidence.

Methodology and quality rating of differential diagnostic studies

Critical evaluation of the CAS versus SSD diagnostic literature requires examination of methodology with respect to study design, internal validity, and risk of bias. Study design can be categorized by levels of evidence (Merlin, Weston, & Tooher, 2009; OCEBM, 2016). Besides systematic reviews (level 1), the best individual diagnostic designs (level 2) are cross-sectional (aka cohort studies that do not require a longitudinal component) studies. The intention is to assess a large group of people with a defined set of clinical features (e.g., any speech errors) and further classify participants using data-driven methods. To qualify as level 2, a study needs to apply a clinical gold standard (reference test) to compare to the measures being evaluated (index test) to determine measures and diagnosis, with blinding and consecutive entry into the study; otherwise, it is classified as level 3. Case-control designs where one group of people with an a priori established diagnosis (e.g., children with CAS) are compared to another group without that diagnosis (e.g., children with phonological disorder) or with a different diagnosis are level 4. Finally, mechanism-based studies and early diagnostic yield studies without an established reference test provide the lowest level of evidence (level 5). These classification systems are aimed at medical and instrumental diagnostic accuracy studies and the levels of evidence are driven from medical research. Specifically taking behavioral studies of neurological disorders into consideration, the American Academy of Neurology (AAN, 2011) developed their own class system for diagnostic studies (reported in Table 1). We have adopted this system for classifying overall study quality as the broader categories are better suitable for evaluating behavioral studies rather than instrumental diagnostic studies at this stage. Nonetheless, the medical levels of evidence provide us direction on how to improve diagnostic methods in the future.

Regarding internal validity and risk of bias, there is only one critiquing tool to assess a published study (i.e., Quality Assessment of Diagnostic Accuracy Studies – 2 or QUADAS-2; Whiting et al., 2011) and one reporting guideline for writing up a diagnostic study for publication (i.e., Standards for Reporting of Diagnostic Accuracy Studies (i.e., Standards for Reporting of Diagnostic Accuracy Studies or STARD, Cohen et al., 2016). These again focus on medical instrumental assessments and have only been available this decade. Both report the need for (a) an index test (i.e., a diagnostic test that is being evaluated) to be compared to an established reference test with an appropriate interval between tests; (b) blinding of results between the index and reference tests; (c) inclusion of methods for reducing bias in participant enrollment, including details of enrollment method (e.g., consecutive, random, or convenience sampling);

and (d) statement of whether the data collection was planned prospectively or retrospectively, relative to when the reference test was completed. The focus regarding index versus reference tests is, importantly, on establishing whether testing is biased by circularity; that is, whether the index and reference tests are non-independent because they are based on the same measures. These tools also recommend authors provide clear descriptions of the tests/measures used and reporting the outcome of the assessment, indicating who was diagnosed with what disorder following the test and the statistical approach applied. Ideally, studies provide cross-tabulation of results across tests and diagnostic accuracy (e.g., 95% confidence intervals, calculation of sensitivity and specificity, odds ratios and/or likelihood ratios).

The AAN (2011) also provides valuable information on critiquing behavioral studies. Here, we have used the AAN data extraction process to examine each study identified in our systematic search supplemented with the above risk of bias features from the QUADAS-2 and STARD. To further evaluate the replicability of study procedures by others, we have also reported on the level of detail provided for dependent measures and the inter-rater reliability of measurements, given that many measures of speech production rely on perceptual judgment which is prone to several biases (Kent, 1996). Finally, we also considered the level of detail on participants' characteristics, which is crucial for all diagnostic studies (Whiting et al., 2011).

Insert Table 1 about here

Participant descriptions are especially relevant for this review considering that (a) the current gold-standard in diagnosing CAS is expert judgment on presence of a set of perceptually judged speech features (ASHA, 2007; Murray, McCabe, & Ballard, 2015) and (b) CAS can be comorbid with other speech and language disorders, which can complicate interpretation of performance (e.g., Murray, Thomas, & McKechnie, 2019). For these reasons, we evaluated participant descriptions of each participant group in terms of sex, severity, age, prior therapy, comorbidity, and other assessment results including genetic testing. For each study, we also specified the inclusion criteria applied and the initial diagnostic criteria used (i.e., reference test), and we made a judgment as to whether these descriptions were enough to allow replication. Finally, we also considered the level of confidence in the gold standard diagnosis (i.e., outcome of the contemporary ASHA (2007) three core-features of CAS reference test). This is a difficult reference test because of the lack of operational definitions provided in the ASHA (2007) technical report (see Terband, Maassen, & Maas, 2019). The level of confidence in the diagnosis was therefore determined through the method of Murray et al. (2015), adapted from (Wambaugh, Duffy, McNeil, Robin, & Rogers, 2006) using discriminative features of CAS based on ASHA (2007). These detailed analyses will assist identification of discriminative measures or markers that warrant further exploration in a future population-based prospective study.

Aims of the present study

The systematic review was completed by the Apraxia of Speech Writing Group, which is part of the Evidence-Based Clinical Research Committee of the Academy of Neurological Communication Disorders and Sciences. This systematic review evaluated studies published through December 2018 with the overall aim of identifying measures that differentiate CAS from at least one other SSD. To meet this overall aim, our specific aims were to determine the following:

1. Study design quality: the design type, replicability and inter-rater reliability of measures, circularity (reference vs. index test), and statistical methods used (AAN, 2011).
2. Adequacy of participant descriptions: the level of detail in descriptions of the participants and level of confidence in the initial diagnoses made by authors (i.e. outcome of reference test).
3. Discriminative diagnostic variables (features/markers): the perceptual, acoustic and/or kinematic measures that statistically differentiated CAS from the comparison SSD(s).

Method

This systematic review is registered with PROSPERO, registration number: CRD42017056616. This review was completed using the PRISMA (Moher, Liberati, Tetzlaff, & Altman, 2009) and PRISMA for Diagnostic Accuracy studies (McInnes et al., 2018) which recommend minimal reporting guidelines for systematic reviews.

Systematic search strategy

We followed the multi-step process of PRISMA (Moher et al., 2009) to find studies that met the inclusion criteria (see eligibility below). Figure 1 presents the flow chart of the study selection procedure.

Insert Figure 1 about here

Identification

A comprehensive search of nine databases related to speech-language pathology was completed. These were Allied and Complementary Medicine (AMED), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Education Resources Information Center (ERIC), Linguistic Language Behavior Abstracts (LLBA), Medline, PsycINFO, Scopus, Web of Science and Google Scholar.

Search terms were developed based on the inclusion criteria and in accordance with MeSH headings so they could be used across databases. The terms were: [“apraxi*” or “dyspraxi*” or “childhood apraxia of speech” or “motor speech” AND “child*” or “develop*” AND “diagnos*” or “assess*” or “different*” or “classif*” or “evalua*” or “suspect*” or “marker” or “discrimin* AND “speech” or “communicat*” or “language” or “articulat*” or “inconsisten*” or “intelligib*” or “prosod*” or “kinemat*” or “speak*”. Specific MeSH keywords used were: “apraxia, developmental verbal”, “apraxia, verbal”, “dyspraxia, verbal”, “Developmental Verbal Dyspraxia” – Qualifier – Diagnosis (DI), classification (CL)]. A total of 4121 references were found.

Screening

References were exported to Endnote X8.2 (2014) for screening. Duplicate references were removed (n= 2050), leaving 2071 for screening. References were checked and excluded in the following order: (1) CAS not present, (2) animal subjects only, (3) theoretical / review studies with no primary data, (4) adult participants only, (5) treatment (not a diagnostic study) and (6) qualitative and/or survey study. Searches were completed within Endnote of title, abstract and keywords for the terms within the list above. References were screened by reading the title and then the abstract and for some that were not clear, obtaining and checking the full text as needed. Inter-rater reliability was completed by the first author and a research assistant for 241 articles (11% of the articles) with 100% agreement. Screening removed 1878 articles (see Figure 1), leaving 193 studies to be assessed for eligibility.

Eligibility

The full texts were obtained and assessed against the final inclusion criteria before being reviewed. This review included studies that met all the inclusion criteria, as follows. Participants included (a) people from birth to 18 years of age; (b) at least one participant diagnosed with CAS (diagnosis could be made before or during the study) and (c) at least one other participant with a different speech sound disorder (SSD) for comparison. Synonyms were included for CAS (e.g., Developmental Apraxia of Speech, Developmental Verbal Dyspraxia, and Verbal Dyspraxia) and a full range of terms relating to SSDs (e.g., speech/ articulation/ phonological delay or disorder or impairment, inconsistent phonological disorder, dysarthria (including paediatric or subtypes), and oral apraxia[1]. Articles included (d) were written in any language with participants and researchers speaking any language and (e) were either published articles or doctoral dissertations not published as a journal article. Studies included were designed to (a) assess CAS compared to other speech sound disorders for differential diagnostic purposes, or (b) determine markers or measures for speech and/or genotype-phenotype relationships.

From the 193 articles, 53 articles were identified for review. Intra-rater reliability (first author) for inclusion of 20% of these studies was assessed again at four months after the first decision had been made. Percent agreement was 98% (n= 49/50). Inter-rater reliability with an independent rater (research assistant) on 100% of the articles was 97% (n = 187/193; see Supplemental Appendix 2 for the list of excluded articles and the reason for exclusion). The eight articles with intra- or inter-rater disagreements were assessed by the entire author team for consensus, with 7/8 excluded. Excluded articles were not further analyzed.

Data processing & analysis

[1] Please note in this study we use ‘phonological disorder’ to refer to phonological impairment/ disorder and developmental language disorder to refer to language delay for readability.

The authors each rated a subset of the 53 included studies, using a procedural manual developed by the authors, entering ratings into an Excel spreadsheet (Supplemental Appendix 1). The manual was based on Wambaugh et al (2006) and the variables are presented according to each aim below. Raters were not blinded to authors of the study as raters were already familiar with the research area. Raters did not assess any paper they authored or were involved in. Where possible, an author rated studies by the same research team to assist with determining the evolution of research programs over time and to assist with methodological details that may have been reported in previous studies.

Five of the 53 studies were rated for reliability by all five authors; two initially after training, another two mid-way in rating to ensure consistency and then a final one towards the end of the rating process. . Inter-rater reliability across the five studies for all coding was 84% (SD= 5.1%, range: 79 to 89%). Discrepancies were resolved by consensus and the manual and spreadsheet notes were reviewed to ensure any issues determined did not affect ratings. If an aspect of a study was difficult to rate, we discussed this at team meetings and resolved by consensus.

Study Quality (aim 1).

The included studies were assessed for the quality of the diagnostic research design used using definitions of research studies (cohort studies, case-control studies or other) from (AAN, 2011 pg. 10). To distinguish further among studies, raters assessed whether there was a clear index (assessment being tested) compared to a reference (currently used comparison measure) and whether group assignment was made a priori (diagnosed beforehand) or not (suspected group or had risk factors associated with CAS or SSD). Raters assessed if each study provided inclusion criteria and diagnostic criteria for CAS and the other SSDs. Methodological quality was also assessed, in regards to whether the outcome or diagnostic measures tested were replicable (i.e. could be used again clinically or in future research), reliable (could be completed by other people with similar results), blinded and were assessed with appropriate statistical tests/methods.

Statistical analysis, where completed, was assessed by determining whether measures were parametric or non-parametric, whether testing of assumptions was reported, and whether the analysis was appropriate for the data and question (e.g., use of t-test or ANOVAs to compare between groups). If raters were unsure in a methodology used, they investigated the method and/or discussed this as a group. From this analysis, it was possible to compute an overall rating of the diagnostic studies based on American Academy of Neurology guidelines (AAN, 2011; see Table 1).

Participant descriptions (aim 2).

Raters assessed each study's description of CAS and SSD participant characteristics. Variables recorded were number of participants, age, sex, severity and etiology of speech disorder, comorbid diagnoses, country of origin and language spoken, the inclusion criteria of the studies and diagnostic indicators. The following was also assessed as to whether or not the information was reported for participants in the studies: any medications taken, neuroimaging data, genetic data, socio-economic status, cognitive functioning, language functioning, hearing and vision, and if so, described the participant’s characteristics.

Confidence in CAS diagnosis.

The confidence of diagnosis was assessed using procedures from (Murray, Mc Cabe, Heard, & Ballard, 2015) based on (Wambaugh et al., 2006) (See Supplemental Appendix 4). For the CAS participants, confidence was established using the criteria in Murray et al (2015) based on description of primary features using the three consensus-based CAS features listed in the ASHA technical report (2007). This was compared to non-discriminative features, which were those shared with other SSDs, such as poor intelligibility, slow progress, or delayed language (ASHA, 2007; McCabe et al., 1998). For the SSD participants, the criteria covered the same five levels as CAS with generic descriptors that were flexible across SSDs. Clear cases of comorbid disorders were also noted. A rating of 1 indicated high confidence in CAS or SSD diagnosis and a rating of 5 indicated no confidence. Intra-rater reliability (first author) on CAS diagnosis was 100% and SSD diagnosis was 100% from the four studies used for calculating reliability. Inter-rater reliability between all raters for the CAS diagnosis was 90% and SSD diagnosis was 90% for the same four studies. Discrepancies were resolved by consensus and the manual was reviewed to ensure any issues determined did not affect subsequent ratings.

Discriminative diagnostic variables (features/markers) (aim 3).

Raters classified the experimental measures from each study as either perceptual, acoustic and/or kinematic measures, and documented the specific measures collected. Further assessment determined whether any measures were discriminative in differentiating CAS from any type of SSD/s, and whether both discriminative and non-discriminative measures had acceptable reliability. Our benchmarks for reliability were: percent agreement > 85% (Kratochwill et al., 2010), Kappa statistics >60% (McHugh, 2012) and intra-class correlation coefficients > 0.60 (Hallgren, 2012). Discriminative measures were also examined for specificity (true positives) and sensitivity (true negatives) values of ≥ 0.90 as per the standard in clinical medicine (Shriberg & et al., 1997a)

Data analysis

Descriptive statistics were documented for the assessments of interest reported previously. A statistical meta-analysis of data was not possible with this dataset due to the differing measures, participants and methodologies across studies. Instead, studies with adequate quality and that demonstrated variables of diagnostic promise were collated using the criteria in Figure 2. As many promising papers were close but did not meet the full set of criteria, a second table of papers that (1) failed to meet one quality criterion (e.g., replication of participants) and/or (2) papers that reported up to two quality criteria in another source (e.g., reliability of perceptual and acoustic measures were reported in a previous study by the same research group) were also collated as potential measures for further research and clinical use. These papers were then described in greater detail within aim 3 to determine promising discriminative features that differentiated CAS from other SSDs. Quality ratings were computed by the first and second authors with (n = 53/55) 95% agreement and the two disagreements were resolved by consensus.

Insert Figure 2 about here

Results

For readability, each criterion assessed is presented in order of most to least rigorous and references are given in the text for no more than 5 studies. The 53 articles reviewed are listed in Supplemental Appendix 3. Key findings are presented in summary tables, while the full analyses per article assessed can be found in Supplemental Appendix 1 (Excel spreadsheet).

Study Quality (aim 1)

Of the 53 studies, 11 (20.8%) used a cohort study design with three (5.7%) using a statistical multivariate data-driven, prospective method to derive subgroups from a larger, unselected sample of children with SSD (Peter, 2006; Strand, McCauley, Weigand, Stoeckel, & Baas, 2013; Vick et al., 2014). Of these three, two compared the resulting subgroups against clinical diagnosis (Peter, 2006; Strand et al., 2013), whereas the other compared the resulting subgroups against classification criteria reported in the literature (Vick et al., 2014). Additionally, 41 (77.4%) employed a case-control design and one (1.9%) involved a case description (Hayden, 1994).

Of the 41 case-control design studies, all but three divided children into subgroups a priori (before analysis) and then compared these groups on the index tests (i.e. the measure/s of interest being assessed - e.g., the DEMSS [Strand et al., 2013]). The basis for classification into subgroups varied considerably with different subgroups and criteria being used. Two studies (Williams, Ingham, & Rosenthal, 1981; Yoss & Darley, 1974) formed their groups after analysis differentiating ‘functional’ articulation disorder (with no organic disability) versus developmental apraxia of speech (i.e. CAS), and for one study it was not stated whether groups were formed a priori (Barry, 1995).

In 23 studies (43.4%), there were indications of diagnostic circularity (in which aspects of the index test were also used in classification by the reference test). In 26 studies (49.1%) there was no diagnostic circularity, and in the remaining 4 studies (7.5%), it could not be determined whether there was diagnostic circularity.

Eleven of the 53 studies (20.8%) were prospective; prospective studies were defined as studies in which data analysis was planned before data collection on index and reference tests (Cohen et al., 2016).

The remaining 42 studies were either retrospective (n = 9; 17.0%) or did not clearly state whether the study was retrospective or prospective (n = 33; 62.3%).

Participants were selected via consecutive sampling in two studies (3.8%; Strand et al., 2013; Williams et al., 1981) and via convenience sampling in 12 others (22.6%). Most studies (n = 39; 73.6%) did not explicitly state their sampling method, although descriptions of recruitment methods suggest convenience sampling in virtually all these studies.

Ten studies (18.9%) clearly indicated that the index test and reference test were administered by separate, blinded examiners. In two studies (3.8%), the assessors were explicitly not blinded (Aziz, Shohdi, Osman, & Habib, 2010; Strand et al., 2013). In the remaining cases (n = 41; 77.4%), we were unable to determine whether the assessors who administered or scored the index test were blinded to the group classification status based on the reference test; this was in part due to the tendency for articles to use passive voice construction in reporting procedures, without specifying who conducted the testing (e.g., “Children were evaluated in a quiet room”).

Only three studies (5.7%) indicated the time period between the index and reference tests (Lewis, Freebairn, Hansen, lyengar, et al. (2004): approximately 4 years within a longitudinal study; Mei et al. [2018]: hours; Strand et al. [2013]: weeks). The remaining 50 studies (94.3%) did not state the time period between index and reference tests.

Overall, of the 53 studies, none met study quality criteria for a Class I rating (AAN, 2011), three (5.7%) met criteria for a Class II rating (Shriberg et al., 2017a, 2017b; Thoonen, Maassen, Gabreëls, & Schreuder, 1999) 16 (30.2%) met criteria for a Class III rating, and the remaining 34 (64.2%) were rated as Class IV studies (see Table 1 for the criteria).

Participant Descriptions (aim 2)

A total of 4,213 participants were included across studies, with an average sample size of 79 participants (median = 41), and a range from 3 (two studies: Betz & Stoel-Gammon, 2005; Smith, Marquardt, Cannito, & Davis, 1994) to 665 (Lewis et al., 2018). Of these 4,213 participants, 845 (20.1%) were children with CAS. It should be noted that these numbers do not necessarily reflect the number of *unique* individuals, as different studies sometimes report data from the same children (e.g., Bradford, Murdoch, Thompson, & Stokes, 1997; Lewis et al., 2018; Shriberg et al., 2003). Given the lack of details in some reports, the number of unique children reported in the 53 studies cannot be determined. The average sample size of children with CAS was 16 (median = 11), with a range from 1 (four studies: Betz & Stoel-Gammon, 2005; Hayden, 1994; Smith et al., 1994; Terband, Zaalen, & Maassen, 2012) to 63 (Lewis et al., 2018).

The comparison groups varied across studies, and several studies included multiple comparison groups. All but five studies (48/53, 90.6%) included a comparison group consisting of children with speech delay/phonological disorder (total of 1,802 children across studies; sample size mean = 38, median = 13, range = 1 – 251). Other comparison groups included were children with dysarthria (seven studies; total 99 children; mean sample size = 14, median = 9, range = 1 – 36), children with developmental language disorder without SSD (four studies; total 54 children, mean sample size = 14, median = 12, range = 7 – 23), children with developmental language disorder and SSD (three studies; total 226 children, mean = 75, median = 14, range = 14 – 170), children who stutter (one study; sample size = 31), children with autism spectrum disorder (one study; sample size = 46), and adults with acquired apraxia of speech (two studies; total 53 participants, range = 22 – 31). Thirty-one studies (58.5%) also included a comparison group of typically developing children (total 1,041 children; mean sample size = 34, median = 12, range = 1 – 255), and three also included adults without speech/language disorders (total 46 participants, mean = 15, median = 18, range 10 – 18).

Only 34 studies (64.2%) reported the sex of participants; together, there were 2,822 participants in these studies. Of the 519 children with CAS reported in these 34 studies, 148 were girls (28.5%) and 371 were boys (71.5%), for a F:M ratio of 1:2.51. In three studies, the sex distribution of the non-CAS participants was not specified for all children (Lewis, Freebairn, Hansen, lyengar, et al., 2004; Shriberg & et al., 1997b; Shriberg, Potter, & Strand, 2011); sex was not specified for a total of 63 children across these studies. Thus, the sex distribution of the non-typical non-CAS comparison group below is based on a total of 1,567 participants. Among these participants, there were 504 girls (32.2%) and 1063 boys

(67.8%), for a F:M ratio of 1:2.11. The language background of the children in the studies was predominantly English (41/53, 77.4%), with five studies involving Dutch-speaking children (9.4%), and one study each (1.9%) for Arabic, French, German, and Portuguese; three studies (5.7%) did not report the language of the children.

Only 24 studies (45.3%) reported disorder severity, and severity in these studies ranged from mild to severe (for both CAS and non-CAS comparison groups). The ages of the children ranged from 3 to 18 years (in both CAS and non-CAS comparison groups); mean age could not be calculated because multiple studies reported only age ranges. Etiology for both CAS and non-CAS groups was not reported or was unknown in most studies, with only ten studies reporting genetic origins for CAS from genetic test results.

Twenty-seven studies (50.9%) reported that the children with speech disorders had received prior speech therapy. Thirty-two studies (60.4%) reported on cognitive function, 39 studies (73.6%) reported on language status, 44 studies (83.0%) reported on hearing status, and 10 studies (18.9%) reported on vision status of participants. Only four studies (7.5%) reported on socioeconomic status or education, and no studies (0.0%) reported medication history or neuroimaging findings. Twenty-eight studies (52.8%) reported comorbid conditions for children with CAS; in most cases, the comorbid condition was language disorder, but several studies also reported dysarthria, cognitive impairments, atypical oral motor function (e.g., oral apraxia, fine motor), and hearing loss.

Diagnostic Criteria and Confidence

Seventeen studies (32.1%) failed to explicitly report criteria for diagnosis of CAS; seven of these studies (13.2% of all studies) referred to other articles for the diagnostic criteria used, in some cases to other articles in a series (Shriberg et al., 2017a, 2017b). For comparison groups, 16 studies (30.2%) failed to provide criteria; five of these (9.4% of all studies) referred to other sources for diagnostic criteria used. Some studies that did report diagnostic criteria were not sufficiently specific or clear to allow replication. In total, only 26 studies (49.1%) provided sufficient information to enable replication with respect to sample characteristics; 27 studies (50.9%) were deemed unreplicable based on the diagnostic information provided (one of which would be considered replicable based on reference to other sources (Shriberg et al., 2011).

Confidence in accuracy of CAS diagnosis spanned the range from 1 to 5 (see Appendix for definitions), with only seven studies (13.2%) receiving a rating of 1 (Iuzzini-Seigel, Hogan Tiffany, & Green Jordan, 2017; Iuzzini-Seigel, Hogan, Guarino, & Green, 2015; Maas & Mailend, 2017; Mei et al., 2018; Murray, Mc Cabe, et al., 2015; Peter, 2006; Zuk, Iuzzini-Seigel, Cabbage, Green, & Hogan, 2018). No studies received a rating of 2; 18 studies (34.0%) received a rating of 3; 12 (22.6%) received a rating of 4; and the remaining 16 (30.2%) received a rating of 5. The average rating across studies was 3.57. For the comparison groups, confidence in diagnosis also spanned the range. For the non-typical comparison groups, nine studies (17.0%) included groups with a confidence rating of 1; zero (0.0%) with ratings of 2; seven (13.2%) with ratings of 3; 21 (39.6%) with a rating of 4; and 16 (30.2%) with a rating of 5. The average rating across studies was 3.66.

Statistical Analysis

Of the 53 studies, 20 (37.7%) did not report reliability data for their indexmeasures. For three additional studies, reliability of index measures was not reported because measurement did not depend on human judgment such as transcription or signal segmentation; one involved the Iowa Oral Performance Instrument (Potter, Nievergelt, & Shriberg, 2013) and the others involved tasks of speech perception where children with CAS responded to speech tasks (e.g., rhyming discrimination or differentiating d/g sounds presented on a computer (Nijland, 2009; Zuk et al., 2018). Five further studies (9.4%) referred to reliability data reported elsewhere. In all, 25 studies (47.2%) reported reliability of index measures or the basis for their index measures; most studies reported reliability of transcription or acoustic segmentation rather than the reliability of the specific measures derived from these raw data (e.g., errors, formant values).

Inferential statistics were used in 38 of the 53 studies (71.7%) and not in the remaining 15 (28.3%). The nature of these statistical tests varied considerably across studies, although ANOVAs, t-tests, and chi-square tests were common. In many cases, it could not be determined from the information provided in the paper whether statistical assumptions were met, but statistical approaches were considered generally appropriate for the research design employed in 30/38 studies (78.9%).

Discriminative diagnostic variables (features/markers) (aim 3)

For the 53 studies examined, study variables were categorized as perceptual, acoustic, or kinematic. Most studies examined perceptual variables (n = 43), including a large number (n = 14) of experiments that also investigated acoustic and/or kinematic variables; 31 studies reported perceptual features that were discriminative. Twenty-four studies investigated acoustic variables including 18 that also examined perceptual and/or kinematic variables; 14 of these studies reported acoustic variables that were discriminative. Finally, seven studies examined kinematic variables, four of which also investigated perceptual and/or acoustic variables; four studies reported discriminative kinematic variables. Variables were further classified based on the type of variable they explored (e.g., auditory perception, prosody) to demonstrate the breadth and focus of the literature.

Of the 53 studies, 7 met the quality criteria for promising studies based on (1) adequate study quality (AAN, 2011 Class III or better), (2) adequate diagnostic confidence for participants with CAS (ratings ≤3), (2) replicable participants, (3), and (4) report of at least 1 reliable discriminative feature/marker/variable. Table 2 summarizes the seven papers that met these criteria. Table 3 includes an additional eight papers that were close to meeting all our criteria but (1) failed to meet one quality criterion (e.g., replicability of participants) and/or (2) papers that referenced up to two quality criteria from another source (e.g., reliability of perceptual and acoustic measures), rather than explicitly reporting them within the paper. Both tables report sensitivity and specificity for the discriminating markers/variables where available. Results below summarize the findings from these 15 papers, organized by the type of feature or measure found to be discriminative for CAS (e.g. auditory perception; articulation, speech movements and rate; literacy; maximum performance; nonspeech and prosody) for comparison. Sensitivity and specificity data where calculated are reported in Table 2 and 3.

Insert Tables 2 and 3 about here

Articulation, Speech Movements, and Rate (including pausing).

Twelve studies (80%) reported discriminative variables that reflected articulation, speech movements, rate or pausing. Most of these variables investigated articulatory accuracy and speech inconsistency/consistency. Findings revealed that children with CAS tended to perform poorer than children with typical development on tasks that assessed percent phonemes or vowels correct (Murray, McCabe, et al., 2015; Ziethe, Springer, Willmes, & Kröger, 2013) and standardized tests of articulation and phonology (Peter, 2006). Murray et al. (2015) found that children with CAS could be differentiated from those with phonological disorder and dysarthria when a combination of measures were used including an articulation measure – percent phonemes correct. Children with CAS had a lower percentage of lexical stress matches, a high occurrence of syllable segregation, a relatively higher percentage of phonemes correct [than children with dysarthria] and low accuracy on the diadochokinetic task ‘peteke’. Ziethe et al (2013) found that percentage vowels correct in complex, multisyllabic pseudo-words differentiated children with CAS who performed more poorly than children with phonological disorder. Similarly, Bradford and Dodd (1996) demonstrated CAS and inconsistent PD performed worse on learning novel words than PD, SD and TD; however CAS performed more poorly than inconsistent PD, PD, SD and TD on imitation tasks. Shriberg, Lohmeier, Strand, and Jakielski (2012) showed that children with CAS demonstrated a higher percentage of within-class manner substitutions and responses containing one or more additions compared to controls with typically developing speech and speech delay (with or without comorbid developmental language disorder). Iuzzini-Seigel et al. (2017) found children with CAS evidenced higher inconsistency on repeated production of monosyllabic words and the phrase “buy Bobby a Puppy” compared to children with typically developing speech and those with speech delay. Acoustic analyses revealed that relative to children with speech delay, children with CAS produced fewer optimal voice onset times for /p/ targets, and reduced vowel space area when auditory feedback was attenuated with noise masking (Iuzzini-Seigel et al., 2015). Children with CAS and Galactosemia also demonstrated slower speaking and articulation rates and less stable F2 and vowel duration relative to children with Galactosemia and speech delay (Shriberg et al., 2011). Finally, Shriberg et al. (2017a) found that children with CAS demonstrated inappropriate pausing (i.e., pause marker) compared with individuals with speech delay.

Maximum Performance.

Three studies (Murray, McCabe, et al., 2015; Thoonen et al., 1999; Thoonen, Maassen, Wit, Gabreëls, & Schreuder, 1996) reported maximum performance tasks that discriminated between groups. Compared to typically developing controls and children with spastic dysarthria, children with CAS evidenced significantly shorter maximum fricative durations and slower trisyllabic repetition rate (Thoonen, 1996). In addition, children with CAS evidenced more failed attempts at producing the sequence ‘pataka’ and required more attempts to produce a correct trisyllable sequence compared to children with typical development or spastic dysarthria. A follow-up study was conducted to cross-validate Thoonen and colleagues’ findings and revealed that CAS can be sensitively and specifically differentiated from TD, speech disorder of unknown origin, and dysarthria on the basis of maximum repetition rate of ‘pataka’ in combination with maximum fricative duration of /f:/ (Thoonen et al., 1999). Likewise, Murray and colleagues (2015) found that articulatory accuracy on repetitions of /pətəkə/ were sensitive and specific against expert diagnosis when differentiating CAS from phonological disorder and dysarthria. This was in combination with syllable segregation, lexical stress matches, and percentage phonemes correct on a polysyllabic picture-naming task.

Nonspeech.

Three studies (20%) (Bradford & Dodd, 1996; Bradford et al., 1997; Peter, 2006) reported discriminative nonspeech variables. Bradford and colleagues (1996) used the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978) and demonstrated that (1) children with CAS and inconsistent PD performed more poorly than children with TD, PD or speech delay on the upper limb strength and dexterity subtest and (2) children with CAS performed poorer than inconsistent PD, PD, SD and TD on the Visual-Motor Control subtest. Similarly, Peter (2006) used a clustering procedure to reveal that children with CAS fell into two different clusters characterized by (1) low scores on the majority of nonspeech timing tasks (clapped rhythm and repetitive tapping) and language tasks (as well as across speech and phonological testing) and (2) a second cluster that had average language scores and intermediate scores on timing tasks, but low scores on speech and phonological testing. Peter also found that undetermined handedness was associated with CAS. Finally, Bradford and colleagues (1997) found that tongue strength and endurance were lower in children with CAS compared to those with inconsistent PD and deviant PD.

Prosody.

Five studies (33%) reported prosodic variables that discriminated between groups. Relative to control groups of children with phonological disorder and dysarthria, children with CAS demonstrated poorer performance on tasks that assessed percent of lexical stress matches and syllable segregation on the Single-Word Test of Polysyllables (Murray et al., 2015). Within this limited sample, Murray et al., (2015) found prosodic accuracy accounted for 80% diagnostic accuracy alone but performed better in a combination of measures to achieve 100% diagnostic accuracy. Relative to children with speech delay or those with non-CAS speech disorder, those with CAS evidenced lower stress accuracy (Shriberg et al., 2017a) or use of inappropriate stress on assessments including the DEMSS, Syllable Repetition Task, and Prosody-Voice Screening Profile (Shriberg et al., 2012; Shriberg et al., 2011; Shriberg et al., 2017a, 2017b; Strand et al., 2013).

Auditory perception.

One study (7%) found that auditory speech perception ability as measured by a syllable discrimination task differentiated groups (Zuk et al., 2018). Specifically, children with CAS and normal language tended to have appropriate speech perception whereas those with CAS and comorbid developmental language disorder evidenced poor speech perception per this task. Speech perception therefore helped identify children with comorbid developmental language disorder but poor speech perception was not considered a core feature of CAS.

Literacy.

There were no studies that met or nearly met the criteria which assessed literacy measures.

Discussion

The differential diagnosis of paediatric speech sound disorders (SSDs) is one of the ongoing challenges in the field of speech-language pathology. Although there is consensus about CAS as a neurological SSD of motor planning and/or programming, CAS has been associated with a wide variety of symptoms that overlap with other types of SSD (ASHA, 2007). The primary aim of the present study was to evaluate the existing literature to determine the discriminative features that might guide and contribute to the development of a diagnostic protocol to differentiate CAS from other SSDs for future clinical and research use.

To this end, this systematic review provides a comprehensive summary of the existing research regarding differential features of CAS. Fifty-three studies were identified that examined speech-language characteristics that sought to differentiate CAS from at least one other type of SSD. The studies cover a wide range of speech characteristics that have been investigated to date as well as how they have been operationalized in concrete index measures. Additionally, studies show large variation in how methodology and results are reported. As such, studies were difficult to combine and compare, and a meta-analysis of data was not possible.

Study Quality (aim 1)

None of the studies met the quality criteria for a Class I AAN rating and only three met the criteria for a Class II AAN rating; consequently, the impact of findings is limited. An important factor herein is consistency and transparency in reporting. Important methodological aspects were often absent or not clearly reported. This applies to the full range of the studies' methodology and includes information regarding design (e.g., whether the study, including analysis, was planned retrospectively or prospectively), sampling method (e.g., what method was used: random, consecutive, or convenience sampling), sample size (e.g., whether the sample size was backed up by a power calculation), inclusion criteria, test administration (e.g., who administered the assessments; were the assessors for the reference and the index test blinded or not), testing procedure (e.g., what was the time between the administration of the reference and the index test), data processing and analysis (e.g., who scored the tests; were the assessors for the reference and the index test blinded or not), and statistics (e.g., whether statistical assumptions were met or not).

Regarding inclusion criteria, diagnostic criteria, assessment procedures, and data analysis, the reader is regularly referred to previous work for methodological details without even a brief description. Although this is not uncommon, motivated by a desire for brevity or by editorial concerns regarding copyright or plagiarism, extensive reporting is important. Although these issues might not affect the impact in the research community, which usually has access to the referred sources, this is generally not the case for clinicians. Furthermore, policy makers and insurance companies are guided by numeric assessments of scientific evidence. For purposes of health care management and insurance coverage, clarity and integrity/completeness of reporting is therefore crucial. We therefore strongly recommend that researchers explicitly adopt existing guidelines such as PRISMA (Moher et al, 2009; McInnes et al, 2018), QUADAS-2 (Whiting et al, 2011), and STARD (Cohen et al., 2015) to ensure study design and reporting meet these quality and reporting criteria and thus maximize impact.

Participant descriptions and confidence in diagnosis (aim 2)

The size of the CAS participant samples varied widely from 1-63 with a mean of 16, and half of the studies featured a sample size below 11. Such small samples are problematic, especially combined with the large age ranges involved. Across studies, the ages of children in the CAS groups ranged from 3 to 18 years and within studies most age ranges that spanned > 5 years. CAS is notorious for its heterogeneity in symptomatology, and symptoms are subject to change during development and resultant from treatment (Maassen, Nijland, & Terband, 2010; Maassen, 2015; Strand & McCauley, 2008). The potential consequences of sampling errors are serious. CAS is slowly but certainly starting to become a well-studied disorder (at least for the English language) and there is a need for larger sample sizes. This issue is complicated further by the fact that multiple studies involve the same participants. The total number of 845 participants with CAS and 1,802 participants with other SSDs included across studies does not represent the same number of unique children. It is paramount that it is explicitly acknowledged if a study utilizes an existing dataset or if it includes the same children as previous studies. Reporting on these aspects is improving, but it also appears the practice of studying the same children is becoming more frequent. Retrospectively exploiting datasets to their maximum and performing multiple studies on the same participant pool constitute large risks that need to be recognized and not underestimated. The risks include that we gain most of our understanding about a disorder from a small set of participants that may not be representative of the larger population and that family-wise statistical errors can overestimate significant differences and associations.

Like what has been noted above regarding the operationalization of the investigated differential markers of CAS (index test), we note that there is no consistent diagnostic method (reference test). Rather, each research group seems to rely on their own protocol with their own inclusion criteria, diagnostic criteria, and assessment procedures. These differences make it difficult to verify diagnoses and compare results between studies. It should also be noted that 26 of the 53 papers predated the publication of the ASHA Technical Report and thus reflect a time of ongoing controversy about CAS and its specific characteristics. A number of older, influential papers that shaped the three core-features of CAS presented in the ASHA Technical Report (2007) are not included as promising papers in the current review, in part due to having poor specification of CAS diagnosis (e.g., Barry, 1995; Lewis, Freebairn, Hansen, lyengar, et al., 2004; Shriberg & et al., 1997b). Future systematic reviews which focus on studies performed well after 2007 may identify more studies with diagnosis explicitly utilizing the ASHA characteristics.

Another source of variation in diagnostic protocols is the fact that the ASHA CAS Technical Report (2007) did not report operational definitions or standardized procedures for measuring the three diagnostic features. This requires researchers to use their own operational definitions and protocols (e.g., Iuzzini-Seigel & Murray, 2017; Terband et al., 2019). An important factor herein is crosslinguistic differences. Certain features appear to have greater differential diagnostic capacity in some languages and less so in others. For example, English features a very strong expression of lexical stress. Languages like Dutch in which lexical stress is expressed less strongly, or French which has a less intricate system of stress assignment compared to English, do not lend themselves to measuring the production accuracy of word stress in the same way as English. There is a need for more uniform, explicit diagnostic criteria that are more replicable, cross linguistically validated using specific tests or scores.

However, also with respect to highly language-neutral tasks such as maximum repetition rate (MRR) or diadochokinesis (DDK) with non-meaningful stimuli (e.g., “pataka”), a variety of methods are used across languages and research groups (but see Diepeveen, van Haaften, Terband, de Swart, & Maassen, 2019; Rvachew, Hodge, & Ohberg, 2005 for standardized protocols). What the studies do have in common though, is that the investigated measures address aspects of an underlying motor impairment, such as temporal control of the articulators, whether that be lexical stress contrasts in English (Murray et al, 2015) or for syllabic organization and phonetic context in Dutch (Nijland, Maassen, & Der Meulen, 2003). This highlights that while it may manifest differently across languages, the underlying motor deficit across people and languages is considered the same.

Discriminative diagnostic variables (features/markers) (aim 3)

Several features were found to be discriminative. A complicating factor in this respect is that relevant clinical comparison groups varied strongly across studies and comprised six different diagnostic classifications (that in turn varied in definition between studies), with some comparing to typically developing children additionally. Furthermore, over half of the studies only reported discriminative measures on the group level and did not report sensitivity and specificity of these measures in the identification of individuals with CAS; not due to methodological neglect but simply because these studies had different aims than validating a marker for differential diagnosis. This included three out of seven studies that met the quality criteria for promising studies and five out of eight studies that were close to meeting all our criteria. For example, percentage vowels correct in pseudo-words was found to differentiate between groups with CAS, PD, and TD for German children, but the index measures were not analyzed as a differential diagnostic marker (Ziethe et al., 2013). In the following section, we will discuss the diagnostic measures that have been found to reliably differentiate individuals with CAS per comparison group, i.e. of which sensitivity and specificity figures are reported. It must be stated, however, that all sensitivity and specificity measures reported across all studies did not include confidence intervals, meaning the sensitivity and specificity estimates reported could in reality be lower if random error is considered (Deeks & Altman, 1999; Dollaghan, 2007). Future diagnostic studies should ensure reporting of confidence intervals with diagnostic accuracy data.

The strongest evidenced differentiation is between CAS and developmental dysarthria, in which maximum performance tasks play a central role. Thoonen et al. (1999) showed in a replication of Thoonen et al (1996), CAS in Dutch can be reliably differentiated from spastic dysarthria based on maximum repetition rate of ‘pataka’ in combination with maximum fricative duration (100%/91% sensitivity/specificity). For Australian English, Murray et al. (2015) showed reliable differentiation between CAS and dysarthria based on accuracy on repetition of /pətəkə/ in combination with percent phonemes correct, lexical stress matches and the presence of syllable segregation on the polysyllable test (100%/100% sensitivity/specificity). This study found that the same set of measures also served to reliably differentiate between CAS and speech delay/phonological disorder/non-CAS speech disorder. For American English, the DEMSS motor speech assessment combining inconsistency, segmental accuracy and prosody measures showed high specificity (97%) but lower sensitivity (65%) (Strand et al., 2013). Other studies in American English have mainly focused on single diagnostic markers, of which many measures address prosody. Word level measures show low to reasonable accuracy, with sensitivity and specificity of stress accuracy and use of inappropriate stress both varying from around 50 to 75 % (Shriberg et al., 2012). The phrase/sentence level prosody measure--inappropriate pausing--showed sensitivity and specificity in differentiating CAS from SD, 87% and 99% respectively (Shriberg et al., 2017a). Other measures that differentiated between CAS and speech delay/phonological disorder/non-CAS speech disorder in American English-speaking children all showed lower diagnostic accuracy. Inconsistency on five repeated productions of the phrase “buy Bobby a Puppy” had 70% sensitivity and 80% specificity (Iuzzini-Seigel et al., 2017). Single segmental error measures such as percentage of within-class manner substitutions and percentage of additions showed sensitivity and specificity levels varying 53-74% and 52-75% respectively (Shriberg et al., 2012). Among children with Galactosemia, speaking and articulation rate and stability of F2 and vowel duration showed reasonable to high sensitivity and specificity in differentiating children with CAS from children with Galactosemia without CAS and children with speech delay (71-88% / 84-100% sensitivity/specificity; Shriberg et al., 2011).

Measures that differentiated CAS and developmental language disorder with reported sensitivity and specificity were perceptual. Percentage of within-class manner substitutions and percentage of additions showed low to reasonable accuracy (53-74% and 52-75% sensitivity and specificity respectively) in differentiating between CAS and language disorder, both with and without concomitant SSD (Shriberg et al., 2012). With respect to comparison groups of children who stutter, had literacy difficulties or autism spectrum disorder, none of the studies met or were close to meeting the quality criteria for promising studies.

In summary, the results of this review indicate that to date there are no individual markers that are sufficiently sensitive and specific (> 90%; Shriberg et al., 1997; Thoonen et al., 1999) in differentiating between CAS and other SSDs. Combinations of measures are more efficacious than single diagnostic markers, with sensitivity and specificity estimates reaching 100%. It can be concluded that reliable differential diagnosis requires a combination of measures (e.g., Murray et al., 2015; Thoonen et al., 1999) to assess across a range of SSDs. The set of diagnostic measures that would be best for use in clinical practice is further discussed below.

Implications for clinical practice

Children with suspected CAS continue to present to clinics for differential diagnosis and for evidence-based treatment planning and management. The primary question addressed in this review is ‘what tools *do* clinicians have that establish whether a child has CAS?’ Clinicians are recommended to complete an initial, comprehensive test battery, including a hearing test, oral-musculature assessment, single-word production (including polysyllable words) and connected speech sampling (see Terband et al., 2019). This assessment can serve to generate a hypothesis as to which speech sound disorder(s) a child may have.

To subsequently test such hypotheses, clinicians need methods and measures that can reliably and validly determine the presence or absence of CAS. This review evaluated several methods that have been used in the literature, some of which hold promise in this regard in the interim. Further development and validation are needed to arrive at definitive methods and measures. Until then, a clinician's best course of action would be to identify papers in which the children share characteristics with the client for whom a differential diagnosis is sought, and to use the measures faithfully to help determine if the child has CAS. The promising measures identified in this review can provide rigorous methods to this end (see Tables 2 and 3). For example, for a four year old child suspected of having CAS or a phonological disorder, a clinician could use the DEMSS (i.e., to provide inconsistency, sequencing and prosody information, Strand & McCauley, 2019; Strand et al., 2013); Iuzzini-Seigel et al.'s (2017) inconsistency measure supplemented with either the Murray et al (2015) protocol or a combination of measures from the Robbins and Klee oral musculature assessment (Robbins & Klee, 1987), and polysyllable test (Gozzard, Baker, & McCabe, 2004) using the formula in the paper or Thoonen et al.'s (1996, 1999) maximum performance tasks using a tutorial to assist (Diepeveen et al., 2019; Rvachew et al., 2005). If the differential diagnostic possibilities include CAS or a speech delay, one could use Shriberg et al (2017)'s pause marker using the protocol from Tilkens et al. (2017) and/or Iuzzini-Seigel et al (2017)'s inconsistency measure. To differentiate between CAS and dysarthria, there are robust methods from few studies available. A thorough oral musculature assessment (OMA) investigating muscle tone, strength and range of movement would be required (Hodge & Hancock, 1994; Murray, McCabe, et al., 2015). If the clinician suspects spastic dysarthria, Thoonen et al (1999)'s maximum performance tasks will help provide a dysarthria score (Diepeveen et al., 2019; Rvachew et al., 2005). Murray et al (2015)'s combination of measures will also help identify dysarthria versus CAS. Although further research is clearly needed to develop, refine, and validate diagnostic measures, there are tools available to the informed clinician that can help with differential diagnosis in the interim, especially when used in combination, to build greater confidence in diagnosis.

Limitations and implications for future research

Despite attempting to find papers with participants speaking any language and being written in any language, the results of this review are very English-dominant. There were five articles that report children who spoke Dutch and four other articles in which children spoke Arabic, French, German, and Portuguese; but most of the studies (89%) assessed only English speakers. The results thus predominantly relate to English speakers.

An adjacent limitation is our use of the English-dominant ASHA (2007) features in determining confidence in the original CAS diagnosis made. As mentioned above, internationally there appears to be a consensus that CAS is a pediatric disorder of speech motor planning and/or programming. However, the way in which the underlying neurological deficit manifests itself symptomatically is likely to differ across languages, driven by phonetic and phonological characteristics (such as sound inventory, phonotactics, rules governing phonological alternation, prosody). As a result, the ASHA (2007) features may not directly apply to studies completed in other languages, which could have affected some studies' ability to meet our current quality criteria and be included as promising papers. CAS is starting to become a well-studied disorder *for English-speaking participants and researchers*. On the one hand, cross-linguistic differences limit the generalizability of findings and raise the need to replicate studies and validate measures in other languages. On the other hand, there is great potential in exploiting cross-linguistic differences to tease out the underlying mechanisms and specify the underlying deficit, its symptoms and its preconditions.

Furthermore, none of the findings based on the existing literature have been replicated with a new, different sample of participants apart from the prospective validation study by Thoonen et al. (1999). Replication studies featuring larger sample sizes are warranted to determine the reproducibility of earlier studies' results. We note that replication studies are not always appreciated or valued by employers and funding institutions, nor by journals and researcher colleagues. It is often difficult to get a replication study funded and published, and it is often frowned upon from a researcher's career perspective. However, replication is an important part of the scientific endeavor, as fundamental from a methodological point of view, it is the ultimate and only real test of the validity of a study's results.

Furthermore, it is important to establish the methods and circumstances required for reproducibility beyond exact replications in order to determine the scope of validity and establish the requirements and conditions for implementation in clinical practice. We thus call upon employers and funding institutions to encourage replication studies, and upon journals to be more open to publishing them, including well-designed and -executed studies with null results.

In this respect, researchers are recommended to register trials and (pre-)publish study protocols. This custom would minimize possible publication bias and, through unique trial participant registration numbers, allow identification of participants and data sets across studies. It would also allow for greater replication, learning and potentially, for collaboration across research groups as publishing of protocols allows for early feedback and discussion, rather than waiting for the end of trials and providing feedback through the peer-review process. A data-driven approach featuring theoretically substantiated cluster analyses allows us to circumvent the issue of reliable initial group assignment and minimize diagnostic circularity.

Summarizing these points, recommendations for future assessment and diagnostic studies include careful and prospectively planned studies using AAN and STARD reporting guidelines (Cohen et al., 2016; Neurology), 2011) and QUADAS-2 quality criteria (Whiting et al., 2011) It is helpful to register trials on PROSPERO to share protocols and collaborate prior to data collection.

Conclusion

This systematic review of 53 papers sought to discriminate CAS from other SSDs. Results indicate that no study in the existing literature meet the highest level of study quality (AAN level 1). Most studies were retrospective, used case-control designs, and lacked reporting of several design aspects. In terms of diagnostic methods, there was no consistent reference test used across studies. Seven studies demonstrated promising index tests (examined measures or tools) for future clinical and research use. Another eight almost met the quality criteria, also demonstrating some promise. Future studies should replicate existing methods and use larger sample sizes. In addition, research needs to address the presentation and diagnosis of CAS in languages other than English (e.g., Wong, Lee, & Tong, 2020).

Research designs ideally should include cohort studies, prospectively designed according to STARD and QUADAS-2 guidelines and have data-driven analyses for the greatest rigor. Clinicians are recommended to complete routine speech assessment and then use promising tools in this study to refine if a child has CAS compared to or comorbid with another SSD diagnosis.

Acknowledgments

We acknowledge support from Lauren Rusedski, the research assistant who supported finding the articles in the review; the Academy of Neurologic Communication Disorders and Sciences (ANCDS) and the Apraxia of Speech Writing Committee (Acquired AOS members) in providing the manual and excel spreadsheet templates we adapted for this CAS review.

References

AAN (American Academy of Neurology). (2011). *Clinical Practice Guideline Process Manual*. St. Paul, MN: The American Academy of Neurology.

ASHA (American Speech-Language-Hearing Association). (2007). Childhood Apraxia of Speech: Technical Report. Retrieved from http://www.asha.org/policy/TR2007-00278/

ASHA (American Speech-Language-Hearing Association). (2016). 2016 Schools Survey report: SLP caseload characteristics. Retrieved from https://www.asha.org/uploadedFiles/2016-Schools-Survey-SLP-Caseload-Characteristics.pdf

Aziz, A. A., Shohdi, S., Osman, D. M., & Habib, E. I. (2010). Childhood apraxia of speech and multiple phonological disorders in Cairo-Egyptian Arabic speaking children: language, speech, and oro-motor differences. *International Journal of Pediatric Otorhinolaryngology, 74*(6), 578-585. https://doi.org//10.1016/j.ijporl.2010.02.003

Barry, R. M. (1995). The Relationship between Dysarthria and Verbal Dyspraxia in Children: A Comparative Study Using Profiling and Instrumental Analyses. *Clinical Linguistics & Phonetics*, 9(4), 277-305. https://doi.org/10.3109/02699209508985338

Betz, S. K., & Stoel-Gammon, C. (2005). Measuring articulatory error consistency in children with developmental apraxia of speech. *Clinical Linguistics & Phonetics, 19*(1), 53-66. https://doi.org/10.1080/02699200512331325791

Bowen, C. (2015). *Children's Speech Sound Disorders: Second Edition*. Sydney, Australia: John Wiley & Sons, Ltd.Bradford, A., & Dodd, B. (1996). Do All Speech-Disordered Children Have Motor

Deficits? *Clinical Linguistics & Phonetics*, *10*(2), 77-101. https://doi.org/10.3109/02699209608985164

Bradford, A., Murdoch, B., Thompson, E., & Stokes, P. (1997). Lip and Tongue Function in Children with Developmental Speech Disorders: A Preliminary Investigation. *Clinical Linguistics & Phonetics*, 11(5), 363-387. https://doi.org/10.1080/02699209708985201

Broomfield, J., & Dodd, B. (2004). The Nature of Referred Subtypes of Primary Speech Disability. *Child Language Teaching and Therapy, 20*(2), 135-151. https://doi.org/10.1191/0265659004ct267oa

Carrigg, B., Parry, L., Baker, E., Shriberg, L., & Ballard, K. (2016). Cognitive, Linguistic, and Motor Abilities in a Multigenerational Family with Childhood Apraxia of Speech. *Archives of Clinical Neuropsychology, 31*. https://doi.org/doi:10.1093/arclin/acw077

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., . . . Bossuyt, P. M. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open, 6*(11), e012799. https://doi.org/10.1136/bmjopen-2016-012799

Deeks, J. J., & Altman, D. G. (1999). Sensitivity and specificity and their confidence intervals cannot exceed 100%. *British Medical Journal*, *318*(7177), 193. https://doi.org/10.1136/bmj.318.7177.193b

Diepeveen, S., van Haaften, L., Terband, H., de Swart, B., & Maassen, B. (2019). A Standardized Protocol for Maximum Repetition Rate Assessment in Children. *Folia Phoniatrica et Logopaedica, 71*(5-6), 238-250. https://doi.org/10.1159/000500305

Dodd, B. (2014). Differential Diagnosis of Pediatric Speech Sound Disorder. *Current Developmental Disorders Reports, 1*(3), 189-196. https://doi.org/10.1007/s40474-014-0017-3

Dollaghan, C. A. (2007). *The Handbook for Evidence-based Practice in Communication Disorders.* Baltimore: Paul H. Brookes Publishing Co.

Eadie, P., Morgan, A., Ukoumunne, O. C., Ttofari Eecen, K., Wake, M., & Reilly, S. (2015). Speech sound disorder at 4 years: prevalence, comorbidities, and predictors in a community cohort of children. *Developmental Medicine and Child Neurology, 57*(6), 578-584. https://doi.org/10.1111/dmcn.12635

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation* (3rd ed.): Pearson Inc.

Gozzard, H., Baker, E., & McCabe, P. (2004). *Single Word Test of Polysyllables*: Unpublished work.

Gubiani, M. B., Pagliarin, K. C., & Keske-Soares, M. (2015). Tools for the assessment of childhood apraxia of speech. *CODAS, 27*(6), 610-615. https://doi.org//10.1590/2317-1782/20152014152

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hayden, D. (1994). Differential diagnosis of motor speech dysfunction in children. *Clinics in communication disorders, 4*(2), 119-141.

Hayden, D., & Square-Storer, P. (1999). *VMPAC: verbal motor production assessment for children*. San Antonio, TX: Psychological Corporation.

Hodge, M. M., & Hancock, H. R. (1994). Assessment of children with developmental apraxia of speech: a procedure. *Clinics in communication disorders, 4*(2), 102-118.

Iuzzini-Seigel, J. (2019). Motor Performance in Children With Childhood Apraxia of Speech and Speech Sound Disorders. *Journal of Speech, Language, and Hearing Research, 62*(9), 3220-3233. https://doi.org/10.1044/2019_JSLHR-S-18-0380

Iuzzini-Seigel, J., Hogan Tiffany, P., & Green Jordan, R. (2017). Speech Inconsistency in Children With Childhood Apraxia of Speech, Language Impairment, and Speech Delay: Depends on the Stimuli. *Journal of Speech, Language, and Hearing Research, 60*(5), 1194-1210. https://doi.org/10.1044/2016_JSLHR-S-15-0184

Iuzzini-Seigel, J., Hogan, T. P., Guarino, A. J., & Green, J. R. (2015). Reliance on auditory feedback in children with childhood apraxia of speech. *Journal of Communication Disorders, 54*, 32-42. https://doi.org /10.1016/j.jcomdis.2015.01.002

Iuzzini-Seigel, J., & Murray, E. (2017). Speech assessment in children with childhood apraxia of speech. *Perspectives of the ASHA Special Interest Groups, 2*(2), 47-60. https://doi.org/10.1044/persp2.SIG2.47

Kent, R. D. (1996). Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. *American Journal of Speech-Language Pathology, 5*(3), 7-23. https://doi.org/

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Storkel H. L. (2019). Using Developmental Norms for Speech Sounds as a Means of Determining Treatment Eligibility in Schools. *Perspectives of the ASHA Special Interest Groups, 4*(1), 67-75. https://doi.org/10.1044/2018_pers-SIG1-2018-0014

Lewis, B. A., Freebairn, L., Tag, J., Benchek, P., Morris, N. J., Iyengar, S. K., . . . Stein, C. M. (2018). Heritability and longitudinal outcomes of spelling skills in individuals with histories of early speech and language disorders. *Learning and Individual Differences, 65*, 1-11. https://doi.org/10.1016/j.lindif.2018.05.001

Lewis, B. A., Freebairn, L. A., Hansen, A., Taylor, H., Iyengar, S. K., & Shriberg, L. D. (2004). Family pedigrees of children with suspected childhood apraxia of speech. *Journal of Communication Disorders, 37*(2), 157-175. https://doi.org//10.1016/j.jcomdis.2003.08.003

Lewis, B. A., Freebairn, L. A., Hansen, A. J., lyengar, S. K., & Taylor, H. (2004). School-age follow-up of children with childhood apraxia of speech. *Language, Speech, and Hearing Services in Schools, 35*(2), 122-140. https://doi.org//10.1044/0161-1461%282004/014%29

Liégeois, F. J., & Morgan, A. T. (2012). Neural bases of childhood speech disorders: Lateralization and plasticity for speech functions during development. *Neuroscience & Biobehavioral Reviews, 36*(1), 439-458. doi: https://doi.org//10.1016/j.neubiorev.2011.07.011

Maas, E., & Mailend, M. L. (2017). Fricative contrast and coarticulation in children with and without speech sound disorders. *American Journal of Speech-Language Pathology, 26*(2), 649-663. https://doi.org/10.1044/2017_AJSLP-16-0110

Maassen, B., Nijland, L., & Terband, H. (2010). Developmental models of Childhood Apraxia of Speech In B. M. P. V. Lieshout (Ed.), *Speech motor control: New developments in basic and applied research* (pp. 243-258). Oxford, UK: Oxford University Press.

Maassen, B., Nijland, L., & Terband, H. (2012). Developmental models of childhood apraxia of speech. In *Speech Motor Control: New Developments in Basic and Applied Research*.

Maassen, B. A. M. (2015). Developmental models of childhood apraxia of speech. In *Routledge handbook of communication disorders* (pp. 124-133). New York, NY: Routledge/Taylor & Francis Group; US.

Macrae, T. (2016). Comprehensive Assessment of Speech Sound Production in Preschool Children. *Perspectives of the ASHA Special Interest Groups, 1*(1), 39-56.https://doi.org/10.1044/persp1.SIG1.39

McCabe, P., Rosenthal, J. B., & McLeod, S. (1998). Features of developmental dyspraxia in the general speech-impaired population? Clinical Linguistics & Phonetics, 12(2), 105-126. https://doi.org/ 10.3109/02699209808985216

McCauley, R. J., & Strand, E. A. (2008). A review of standardized tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology, 17*(1), 81-91. https://doi.org/10.1044/1058-0360(2008/007)

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemica Medica (Zagreb), 22*(3), 276-282.

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., Clifford, T., . . . Willis, B. H. (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *Jama, 319*(4), 388-396. https://doi.org/10.1001/jama.2017.19163

McLeod, S., & Baker, E. (2014). Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for children with speech sound disorders. *Clin Linguist Phon, 28*(7-8), 508-531. https://doi.org/10.3109/02699206.2014.926994

Mei, C., Fedorenko, E., Amor, D. J., Boys, A., Hoeflin, C., Carew, P., . . . Morgan, A. T. (2018). Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *European Journal of Human Genetics, 26*(5), 676-686. https://doi.org/10.1038/s41431-018-0102-x

Merlin, T., Weston, A., & Tooher, R. (2009). Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology, 9*(1), 34. https://doi.org/10.1186/1471-2288-9-34

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ, 339*. https://doi.org/10.1136/bmj.b2535

Morley, M., Court, D., & Miller, H. (1954). Developmental dysarthria. *British Medical Journal, 1*(4852), 8-10. https://doi.org/10.1136/bmj.1.4852.8

Murray, E., & Iuzzini-Seigel, J. (2017). Efficacious Treatment of Children With Childhood Apraxia of Speech According to the International Classification of Functioning, Disability and Health. *Perspectives of the ASHA Special Interest Groups 2: Neurogenic Communication Disorders, 2*(2). https://doi.org/10.1044/persp2.SIG2.61

Murray, E., Mc Cabe, P., Heard, R., & Ballard, K. J. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 58*(1), 43-60. https://doi.org/10.1044/2014_JSLHR-S-12-0358

Murray, E., McCabe, P., & Ballard, K. J. (2015). A randomized controlled trial for children with childhood apraxia of speech comparing rapid syllable transition treatment and the nuffield dyspraxia programme–third edition. *Journal of Speech, Language, and Hearing Research, 58*(3), 669-686. https://doi.org/10.1044/2015_JSLHR-S-13-0179

Murray, E., Thomas, D., & McKechnie, J. (2019). Comorbid morphological disorder apparent in some children aged 4-5 years with childhood apraxia of speech: findings from standardised testing. *Clinical Linguistics & Phonetics, 33*(1-2), 42-59. https://doi.org/10.1080/02699206.2018.1513565

Nijland, L. (2009). Speech perception in children with speech output disorders. *Clinical Linguistics & Phonetics*, *23*(3), 222-239. https://doi.org/10.1080/02699200802399947

Nijland, L., Maassen, B., & Der Meulen, S. v. (2003). Evidence of motor programming deficits in children diagnosed with DAS. *Journal of Speech, Language, and Hearing Research*, *46*(2), 437-450. https://doi.org/10.1044/1092-4388(2003/036)

OCEBM Levels of Evidence Working Group*. *The Oxford Levels of Evidence 2*. Oxford Centre for Evidence-Based Medicine. https://www.cebm.net/index.aspx?o=5653

PDF™ Tron Systems (2014). Endnote Software. X8.2.

Peter, B. (2006). *Multivariate characteristics and data-based disorder classification in children with speech disorders of unknown origin.* (PhD), University of Washington, Unpublished dissertation. Retrieved from Linguistics and Language Behavior Abstracts (LLBA) database.

Potter, N. L., Nievergelt, Y., & Shriberg, L. D. (2013). Motor and speech disorders in classic galactosemia. *JIMD Reports*, *11*, 31-41. https://doi.org//10.1007/8904_2013_219

PROSPERO. (2019). Retrieved from https://www.crd.york.ac.uk/prospero/

Robbins, J., & Klee, T. (1987). Clinical Assessment of Oropharyngeal Motor Development in Young Children. *Journal of Speech and Hearing Disorders, 52*, 271-277. https://doi.org/10.1044/jshd.5203.271

Rupela, V., Velleman, S. L., & Andrianopoulos, M. V. (2016). Motor speech skills in children with Down syndrome: A descriptive study. *International Journal of Speech Language Pathology, 18*(5), 483. https://doi.org/10.3109/17549507.2015.1112836

Rusiewicz, H. L., Maize, K., & Ptakowski, T. (2018). Parental experiences and perceptions related to childhood apraxia of speech: Focus on functional implications. *Int J Speech Lang Pathol, 20*(5), 569-580. https://doi.org/10.1080/17549507.2017.1359333

Rvachew, S., Hodge, M., & Ohberg, A. (2005). Obtaining and interpreting maximum performance tasks from children: a tutorial. *Journal of Speech-Language Pathology & Audiology, 29*(4), 146-157.

Sayahi, F., & Jalaie, S. (2016). Diagnosis of Childhood Apraxia of Speech: A Systematic Review. *Journal of Diagnostics, 3*(1), 21-26. https://doi.org/10.18488/journal.98/2016.3.1/98.1.21.26

Shriberg, L. D., Campbell, T. F., Karlsson, H. B., Brown, R. L., McSweeny, J. L., & Nadler, C. J. (2003). A diagnostic marker for childhood apraxia of speech: the lexical stress ratio. Clinical Linguistics & Phonetics, 17(7), 549-574. https://doi.org/10.1080/0269920031000138123

Shriberg, L. D., Aram, D.M., Kwiatkowski, J. (1997a). Developmental Apraxia of Speech: I. Descriptive and Theoretical Perspectives. *Journal of Speech, Language, and Hearing Research, 40*(2), 273-285. https://doi.org/10.1044/jslhr.4002.273

Shriberg, L. D., Aram, D.M., Kwiatkowski, J. (1997b). Developmental Apraxia of Speech: II. Toward a Diagnostic Marker. *Journal of Speech, Language, and Hearing Research, 40*(2), 286-312. https://doi.org/10.1080/0269920031000138123

Shriberg, L. D., Lohmeier, H. L., Strand, E. A., & Jakielski, K. J. (2012). Encoding, memory, and transcoding deficits in Childhood Apraxia of Speech. *Clinical Linguistics & Phonetics, 26*(5), 445-482. https://doi.org/10.1044/jslhr.4002.286

Shriberg, L. D., Potter, N. L., & Strand, E. A. (2011). Prevalence and Phenotype of Childhood Apraxia of Speech in Youth With Galactosemia. *Journal of Speech, Language & Hearing Research, 54*(2), 487-519. https://doi.org/1092-4388(2010/10-0068)

Shriberg, L. D., Strand, E. A., Fourakis, M., Jakielski, K. J., Hall, S. D., Karlsson, H. B., . . . Wilson, D. L. (2017a). A Diagnostic Marker to Discriminate Childhood Apraxia of Speech From Speech Delay: II. Validity Studies of the Pause Marker. *Journal of Speech, Language & Hearing Research, 60*(4), S1118-s1134. https://doi.org/10.1044/2016_jslhr-s-15-0297

Shriberg, L. D., Strand, E. A., Fourakis, M., Jakielski, K. J., Hall, S. D., Karlsson, H. B., . . . Wilson, D. L. (2017b). A Diagnostic Marker to Discriminate Childhood Apraxia of Speech From Speech Delay: III. Theoretical Coherence of the Pause Marker with Speech Processing Deficits in Childhood Apraxia of Speech. *Journal of Speech, Language & Hearing Research, 60*(4), S1135-s1152. https://doi.org/10.1044/2016_jslhr-s-15-0298

Smith, B., Marquardt, T. P., Cannito, M. P., & Davis, B. L. (1994). Vowel Variability in Developmental Apraxia of Speech. In J. A. Till, K. M. Yorkston, & D. R. Beukelman (Eds.), *Motor Speech Disorders: Advances in Assessment and Treatment.* Baltimore, MD: Brookes Publishing.

IEPMCS (International Expert Panel on Multilingual Children's Speech) (2012). *Multilingual children with speech sound disorders: Position paper.* Retrieved from http://www.csu.edu.au/research/multilingual-speech/position-paper

Strand, E. A., & McCauley, R. J. (2008). Differential diagnosis of severe speech impairment in young children. *ASHA Leader, 13*(10), 10-13. https://doi.org/10.1044/leader.FTR1.13102008.10

Strand, E. A., & McCauley, R. J. (2019). *Dynamic Evaluation of Motor Speech Skill (DEMSS) Manual*. Baltimore, MD: Brookes Publishing.

Strand, E. A., McCauley, R. J., Weigand, S. D., Stoeckel, R. E., & Baas, B. S. (2013). A Motor Speech Assessment for Children With Severe Speech Disorders: Reliability and Validity Evidence. *Journal of Speech, Language & Hearing Research, 56*(2), 505-520. https://doi.org/1092-4388(2012/12-0094)

Terband, H., Maassen, B., Guenther, F. H., & Brumberg, J. (2009). Computational neural modeling of speech motor control in childhood apraxia of speech (CAS). *Journal of Speech, Language & Hearing Research, 52*(6), 1595-1609. https://doi.org/10.1044/1092-4388(2009/07-0283)

Terband, H., Maassen, B., & Maas, E. (2019). A psycholinguistic framework for diagnosis and treatment planning of developmental speech disorders. *Folia Phoniatrica et Logopaedica, 71*(5-6), 216-227. https://doi.org/10.1159/000499426

Terband, H., Maassen, B., van Lieshout, P., & Nijland, L. (2011). Stability and composition of functional synergies for speech movements in children with developmental speech disorders. *Journal of Communication Disorders, 44*(1), 59-74. https://doi.org/10.1016/j.jcomdis.2010.07.003

Terband, H., Namasivayam, A., van Brenk, F., Diepeveen, S., Mailend, M.-L., Maas, E., . . . Maassen, B. (2019). Assessment of Childhood Apraxia of Speech: a review/tutorial of objective measurement techniques. *Journal of Speech, Language and Hearing Research, 62*(8S), 2999–3032. https://doi.org/10.1044/2019_JSLHR-S-CSMC7-19-0214

Terband, H., Zaalen, Y. V., & Maassen, B. (2012). Lateral jaw stability in adults, children, and children with developmental speech disorders. *Journal of Medical Speech-Language Pathology, 20*(4), 112-118.

Thoonen, G., Maassen, B., Gabreëls, F., & Schreuder, R. (1999). Validity of maximum performance tasks to diagnose motor speech disorders in children. *Clinical Linguistics and Phonetics, 13*(1), 1-23. https://doi.org/10.1080/026992099299211

Thoonen, G., Maassen, B., Wit, J., Gabreëls, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnostic evaluations among children with motor speech disorders. Clinical Linguistics and Phonetics, 10(4), 311-336. https://doi.org/10.3109/02699209608985178

Tilkens, C. M., Karlsson, H. B., Fourakis, M., Hall, S. D., Mabie, H. L., McSweeny, J. L., . . . Shriberg, L. D. (2017) *A Diagnostic Marker to Discriminate Childhood Apraxia of Speech (CAS) from Speech Delay (SD): The Pause Marker. Technical Report No. 22*.

Vick, J. C., Campbell, T. F., Shriberg, L. D., Green, J. R., Truemper, K., Rusiewicz, H. L., & Moore, C. A. (2014). Data-Driven Subclassification of Speech Sound Disorders in Preschool Children. *Journal of Speech, Language & Hearing Research, 57*(6), 2033-2050. https://doi.org 10.1044/2014_JSLHR-S-12-0193

Wambaugh, J. L., Duffy, J. R., McNeil, M. R., Robin, D. A., & Rogers, M. A. (2006). Treatment guidelines for acquired apraxia of speech: A synthesis and evaluation of the evidence. *Journal of Medical Speech-Language Pathology, 14*(2), xv-xxxiii.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., . . . Bossuyt, P. M. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annuals of Internal Medicine, 155*(8), 529-536. https://doi.org 10.7326/0003-4819-155-8-201110180-00009

Williams, R., Ingham, R. J., & Rosenthal, J. (1981). A further analysis for developmental apraxia of speech in children with defective articulation. *Journal of Speech & Hearing Research, 24*(4), 496-505. https://doi.org http://dx.doi.org/10.1044/jshr.2404.496

Wong, E. C. H., Lee, K. Y. S., & Tong, M. C. F. (2020). The Applicability of the Clinical Features of English Childhood Apraxia of Speech to Cantonese: A Modified Delphi Survey. *American Journal of Speech-Language Pathology*. https://doi.org 10.1044/2019_AJSLP-19-00118

Yoss, K. A., & Darley, F. L. (1974). Developmental apraxia of speech in children with defective articulation. *Journal of Speech & Hearing Research, 17*(3), 399-416. https://doi.org /10.1044/jshr.1703.399

Ziethe, A., Springer, L., Willmes, K., & Kröger, B. J. (2013). Study of core features of children with childhood apraxia of speech aged between 4 and 7 years. *Sprache Stimme Gehor, 37*(4), 210-214. https://doi.org 10.1055/s-0032-1323786

Zuk, J., Iuzzini-Seigel, J., Cabbage, K., Green, J. R., & Hogan, T. P. (2018). Poor speech perception is not a core deficit of childhood apraxia of speech: Preliminary findings. *Journal of Speech, Language, and Hearing Research, 61*(3), 583-592. https://doi.org 10.1044/2017_JSLHR-S-16-0106

Figures and Figure Legends

Figure 1.

Flow diagram of study selection (adapted from PRISMA, Moher et al., 2009).

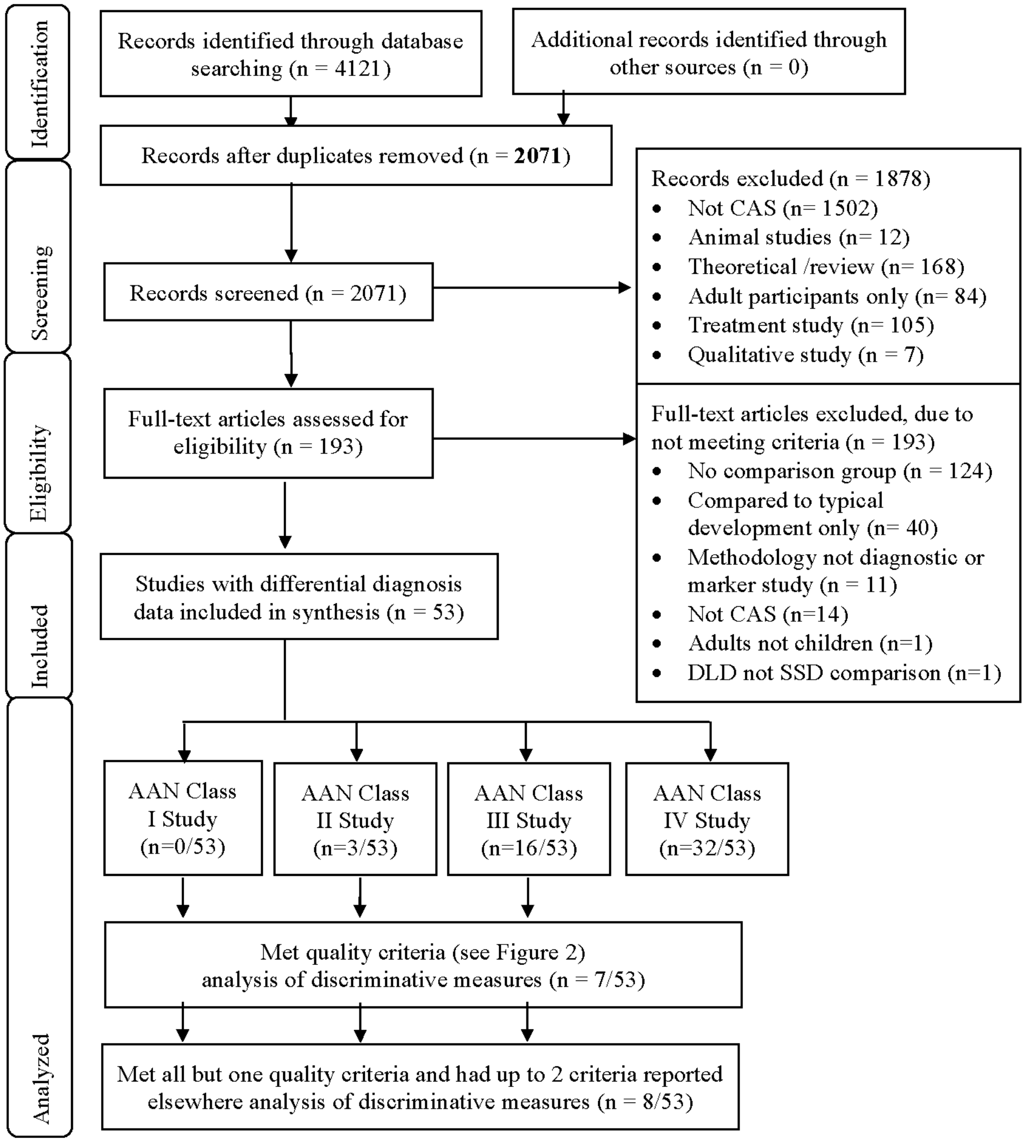
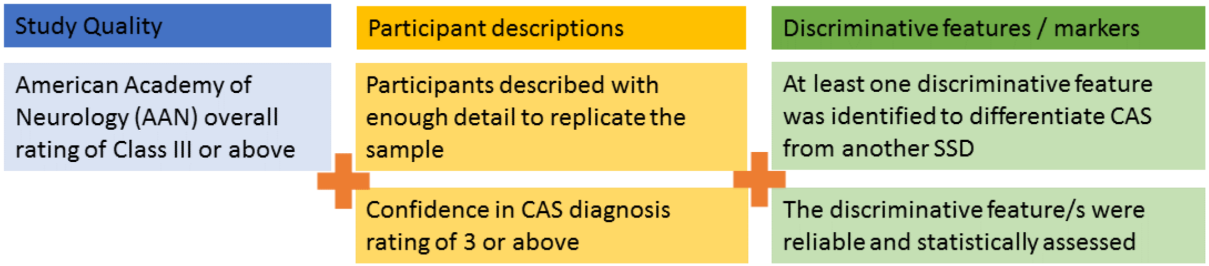

Identification
Screening
Eligibility
Included
Analyzed
Records identified through database searching (n = 4121)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = **2071**)
Records screened (n = 2071)
Records excluded (n = 1878)
• Not CAS (n= 1502)
• Animal studies (n= 12)
• Theoretical /review (n= 168)
• Adult participants only (n= 84)
• Treatment study (n= 105)
• Qualitative study (n = 7)
Full-text articles assessed for eligibility (n = 193)
Full-text articles excluded, due to not meeting criteria (n = 193)
• No comparison group (n = 124)
• Compared to typical development only (n= 40)
• Methodology not diagnostic or marker study (n = 11)
• Not CAS (n=14)
• Adults not children (n=1)
• DLD not SSD comparison (n=1)
Studies with differential diagnosis data included in synthesis (n = 53)
AAN Class I Study (n=0/53)
AAN Class II Study (n=3/53)
AAN Class III Study (n=16/53)
AAN Class IV Study (n=32/53)
Met quality criteria (see Figure 2) analysis of discriminative measures (n = 7/53)
Met all but one quality criteria and had up to 2 criteria reported elsewhere analysis of discriminative measures (n = 8/53)

Figure 2

Quality criteria for analysis



Study Quality

American Academy of Neurology (AAN) overall rating of Class III or above

Participant descriptions

Participants described with enough detail to replicate the sample

Confidence in CAS diagnosis rating of 3 or above

Discriminative features / markers

At least one discriminative feature was identified to differentiate CAS from another SSD

The discriminative feature/s were reliable and statistically assessed

Tables

Table 1.

American Academy of Neurology Classes for Rating Diagnostic Articles (AAN [American Academy of Neurology], 2011).

Class	Description
Class I	A cohort [cross-sectional] study with prospective data collection of a broad spectrum of persons with the suspected condition, using an acceptable reference standard for case definition. The diagnostic test is objective or performed and interpreted without knowledge of the patient's clinical status. Study results allow calculation of measures of diagnostic accuracy.
Class II	A case control study of a broad spectrum of persons with the condition established by an acceptable reference standard compared to a broad spectrum of controls or a cohort study where a broad spectrum of persons with the suspected condition where the data was collected retrospectively. The diagnostic test is objective or performed and interpreted without knowledge of disease status. Study results allow calculation of measures of diagnostic accuracy.
Class III	A case control study or cohort study where either persons with the condition or controls are of a narrow spectrum. The condition is established by an acceptable reference standard. The reference standard and diagnostic test are objective or performed and interpreted by different observers. Study results allow calculation of measures of diagnostic accuracy.
Class IV	Studies not meeting Class I, II or III criteria including consensus, expert opinion or a case report.