Intelligibility and comprehensibility: A Delphi consensus study

Timothy Pommée, Mathieu Balaguer, Julie Mauclair, Julien Pinquier, Virginie Woisard

Intelligibility and comprehensibility: A Delphi consensus study

Timothy Pommée[1]*, Mathieu Balaguer[1,2], Julie Mauclair[1], Julien Pinquier[1], Virginie Woisard[2,3,4]

[1] *IRIT, CNRS, Paul Sabatier University Toulouse III, Toulouse, France*

[2] *ENT department, University Hospital of Toulouse Larrey, Toulouse, France*

[3] *Oncorehabilitation unit, University Cancer Institute of Toulouse Oncopole, Toulouse, France*

[4] *Laboratoire Octogone Lordat, Jean Jaurès University Toulouse II, Toulouse, France*

Correspondence to:

*Timothy Pommée

timothy.pommee@irit.fr

+33 6 03 03 51 48

ORCID : 0000-0001-7846-7282

IRIT Institut de Recherche en Informatique de Toulouse

118 Route de Narbonne

31062 TOULOUSE CEDEX 9

Running head: Intelligibility and comprehensibility

Keywords: intelligibility, comprehensibility, speech, terminology, assessment, Delphi

Acknowledgments: The authors thank all the volunteers who gave their time to participate in this study, as well as Prof. Renée Speyer (University of Oslo) for her guidance regarding the Delphi methodology. This project was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 766287.

Declaration of interest: The authors declare that there is no conflict of interest.

Intelligibility and comprehensibility: A Delphi consensus study

Abstract

Background: Intelligibility and comprehensibility in speech disorders can be assessed both perceptually and instrumentally, but a lack of consensus exists regarding the terminology and the related speech measures, both in the clinical and in the scientific fields.

Aims: To draw up a more consensual definition of intelligibility and comprehensibility and to define which assessment methods relate to both concepts, as part of their definition.

Methods & Procedures: A three-round modified Delphi consensus study was carried out among clinicians, researchers and lecturers engaged in activities in speech disorders.

Outcomes & Results: Forty international experts from different fields (mainly clinicians, linguists and computer scientists) participated in the elaboration of a comprehensive definition of intelligibility and comprehensibility and their assessment. While both concepts are linked and both contribute to functional human communication, they relate to two different reconstruction levels of the transmitted speech material. Intelligibility refers to the acoustic-phonetic decoding of the utterance, while comprehensibility relates to the reconstruction of the meaning of the message. Consequently, the perceptual assessment of intelligibility requires the use of unpredictable speech material (pseudo-words, minimal word pairs, unpredictable sentences), whereas comprehensibility assessment is meaning- and context-related and entails more functional speech stimuli and tasks.

Conclusion & Implications: This consensus study provides the scientific and clinical communities with a better understanding of intelligibility and comprehensibility. A comprehensive definition was drafted, including specifications regarding the tasks that best fit their assessment. The outcome has implications both for clinical practice and for scientific research, as the disambiguation improves communication between professionals and thereby

increases the efficiency of patient assessment and care and benefits the progress of research as well as research translation.

What this paper adds

What is already known on the subject

- Intelligibility and comprehensibility in speech disorders can be assessed both perceptually and instrumentally, but a lack of consensus exists regarding the terminology and the related speech measures, both in the clinical and in the scientific fields.

What this paper adds to existing knowledge

- This consensus study allowed for a more consensual and comprehensive definition of intelligibility and comprehensibility and their assessment, for clinicians and researchers. The terminological disambiguation helps to improve communication between experts in the field of speech disorders and thereby benefits the progress of research as well as research translation.

What are the potential or actual clinical implications of this work?

- Unambiguous communication between professionals, e.g. in a multidisciplinary team, allows to improve the efficiency of patient care. Furthermore, this study allowed to specify the assessment tasks that best fit the definition of both intelligibility and comprehensibility, thereby providing valuable information to improve speech disorder assessment and its standardization.

Introduction

The assessment of speech disorders aims to evaluate several dimensions to allow for a comprehensive and individualized overview of each patient's speech. These dimensions include an examination of the orofacial sensitivity and motor functions, a functional assessment of respiration, phonation and resonance, articulation (motor planning, programming and

execution), intelligibility (acoustic-phonetic decoding), comprehensibility (understandability), as well as the psychosocial impact of the speech disorder (Dykstra *et al.* 2007, Rumbach *et al.* 2019).

Both perceptual and instrumental measures can be used, the first of which still seem to be the most common option in clinical practice (Altaher *et al.* 2019, Pommée *et al.* 2021, Rumbach *et al.* 2019). However, there appears to be a lack of consensus regarding the terminology of the perceptual concepts related to speech, as well as how to assess them. A recent clinician survey in French-speaking countries indeed revealed a lack of standardization of the speech assessment, regarding its overall structure, but also the assessment tasks and stimuli used for each dimension (Pommée *et al.* 2021). Furthermore, the terms used by the speech-and-language pathologists in this study indicated a lack of clarity regarding their definitions, more specifically regarding intelligibility and comprehensibility. This ambiguity in the use of professional terminology is also observed in existing assessment batteries such as the French Batterie d'Évaluation Clinique de la Dysarthrie (Auzou and Rolland-Monnoury 2006), as well as in the scientific literature (Denman *et al.* 2019, Pommée *et al.* 2020, Walsh *et al.* 2006, Walsh 2005). Wood already expressed the "terminology problem" in 1971: "Many terms and their meanings are not well crystallized because the subject matter is always changing; concepts themselves are often tentative and fluid …. This growth of speech pathology … has generated hundreds of terms, some of which are interchangeable, some of which have different means to different people" (Wood 1971). A more recent report by the Australian Institute of Health and Welfare confirmed that this issue has not yet been resolved: "Classification and terminology used to describe speech impairments are particularly fraught with inconsistency, in particular the use of different interpretations for the same terminology or different terminologies for the same meaning" (Australian Institute of Health and Welfare 2003). This issue can lead to communication issues between professionals and impede the efficiency of patient care (e.g. in

a multidisciplinary team) as well as the progress of research (e.g. by hampering scientific debates and inducing difficulty to compare and combine research results) (Denman *et al.* 2019, Walsh *et al.* 2006, Walsh 2005). In addition to its impact in the clinical and in the scientific fields, the lack of consensual terminology also impedes the link between these two fields by affecting research translation (Denman *et al.* 2019, Roulstone 2015).

In light of the important clinical and scientific impact of terminological ambiguity, the main aim of this study is to draw up a more consensual definition of intelligibility and comprehensibility. It also addresses a secondary aim that is closely linked to the first, to define which assessment methods relate to both concepts (as part of their definition).

The three most commonly used methods in health-related research that target consensus are the consensus development conference, the expert panel (or "nominal group") and the Delphi method (Jones and Hunter 1995, McMillan *et al.* 2016). The consensus development conference was originally developed by the National Institutes of Health (NIH) to validate the safety and effectiveness of health-related technologies and facilitate their transmission into clinical practice (Perry and Kalberer 1980). A panel of experts typically gathers for a few days (including all-night sessions) to draft and modify a consensus statement of recommendations. This method is rather tedious to organize and does not control for issues related to the group setting (e.g. social pressure to comply); also, all-night sessions lead to the likelihood of an agreement deriving from the panel's tiredness. Furthermore, no formal decision-making criteria or voting processes are used, and only qualitative estimations can be made (Letrilliart and Vanmeerbeek 2011). Conversely, the nominal group and Delphi methods are structured and systematic qualitative approaches that allow for quantitative results. Like the consensus development conference, the nominal group technique is also carried out in a face-to-face meeting of experts (McMillan *et al.* 2014), but uses highly structured consecutive rounds to rate

and discuss a series of questions (see [Jones and Hunter 1995] for a description of the different stages). Finally, the Delphi method is also a multi-stage process, but uses self-completed questionnaires instead of a face-to-face setting and allows for the use of larger panels, as compared to a recommended panel size of seven experts for the nominal group survey (McMillan *et al.* 2016). Originally developed in a military context, the Delphi method has since the 1980s been applied to various fields of research (von der Gracht 2012) to make forecasts or make decisions about present issues (Chalmers and Armour 2019), generate ideas or determine priorities (McMillan *et al.* 2016). It is nowadays commonly used in health-related topics, such as guideline development (McMillan *et al.* 2016), assessment of treatment appropriateness (Beers *et al.* 1991), disease prevalence forecasts (Chin *et al.* 1990) and improvement of education and training in health professions (Fasser *et al.* 1992). One of the many advantages of this method, in addition to the cheap cost and to the absence of geographical limitations (the rounds can be carried out by mail or online), is its quasi-anonymous nature (von der Gracht 2012, Sinha *et al.* 2011). The participants' identity remains unknown to each other, which allows for freedom of expression without any social or professional pressure from peers. The quasi-anonymous nature, the use of multiple rounds and the provision of structured feedback between the rounds allow to reduce bias in the consensus-aiming process (Chalmers and Armour 2019).

In light of its many advantages, the Delphi method was used in this study to draw up a more consensual definition of intelligibility and comprehensibility and their assessment.

Materials and Methods

Ethics Approval

This research was registered with the data protection officer of the Centre National de la Recherche Scientifique (CNRS) and was also approved by the computer science ethics

advisory board (Comité consuLtatif d'Éthique concernant la Recherche en Informatique de Toulouse — CLERIT). An information sheet describing the purpose of the study, the participant's rights and the data privacy policy was provided to all participants prior to the first round of the survey. Participants gave consent for their answers to be used in an anonymized and aggregated manner to derive consensus statements. Only e-mail addresses were known to the moderator, in order to enable round-to-round survey monitoring.

Study Design

The Delphi methodology is conducted over several consecutive rounds (usually three) (Birko *et al.* 2015, Linstone and Turoff 2002). In a "modified Delphi study", statements to be rated are directly provided in the first round, based on a preliminary literature search (Cunningham *et al.* 2019, Denman *et al.* 2019). After each round, the panel's responses are synthesized, and areas of agreement/disagreement are identified. Aggregated controlled feedback (von der Gracht 2012) is then provided to the panel in the following round to explain modifications that have been made to facilitate consensus, and participants give their opinion on the new assertions (Denman *et al.* 2019, Diamond *et al.* 2014). The iterations are usually carried out until consensus is reached or until stability in the answers is observed (see [Chalmers and Armour 2019, von der Gracht 2012] for examples of stability measures).

The stages of the present modified Delphi study are summarized in figure 1. The main steps which will be described hereafter are the problem definition, the panel selection, the literature search and the construction of the three consecutive Delphi rounds.

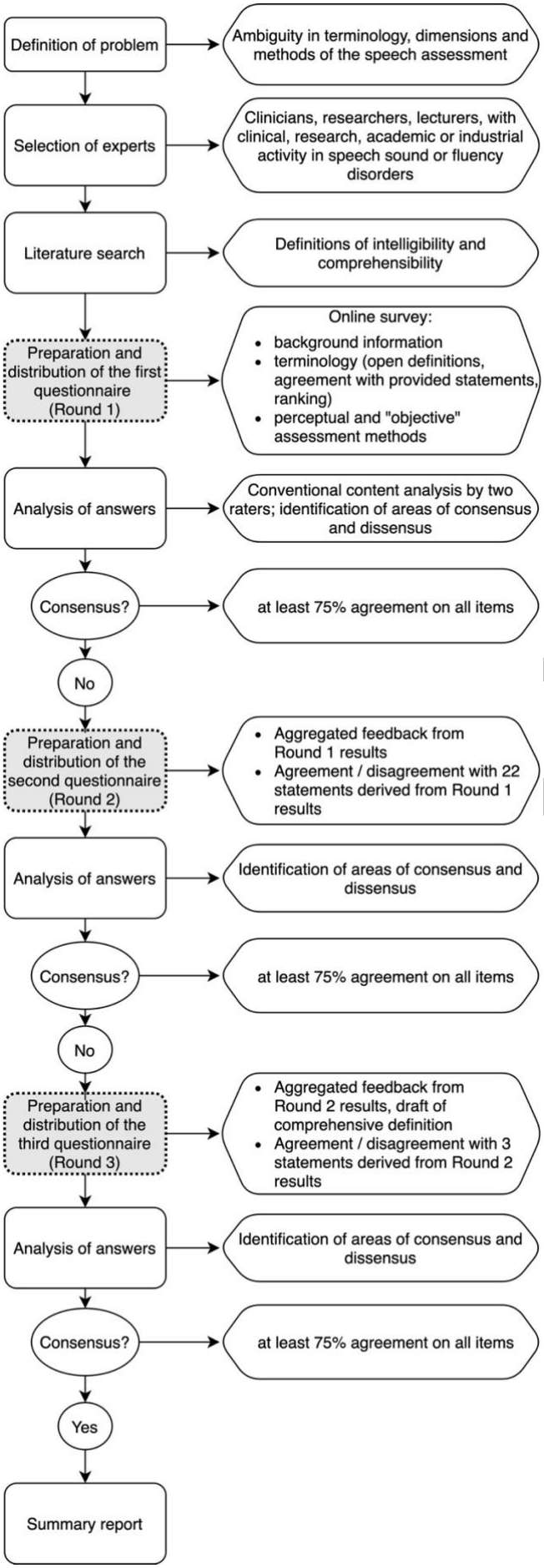
Definition of problem
Ambiguity in terminology, dimensions and methods of the speech assessment
Selection of experts
Clinicians, researchers, lecturers, with clinical, research, academic or industrial activity in speech sound or fluency disorders
Literature search
Definitions of intelligibility and comprehensibility
Preparation and distribution of the first questionnaire (Round 1)
Online survey:
background information
terminology (open definitions, agreement with provided statements, ranking)
perceptual and "objective" assessment methods
Analysis of answers
Conventional content analysis by two raters; identification of areas of consensus and dissensus
Consensus?
at least 75% agreement on all items
No
Preparation and distribution of the second questionnaire (Round 2)
Aggregated feedback from Round 1 results
Agreement / disagreement with 22 statements derived from Round 1 results
Analysis of answers
Identification of areas of consensus and dissensus
Consensus?
at least 75% agreement on all items
No
Preparation and distribution of the third questionnaire (Round 3)
Aggregated feedback from Round 2 results, draft of comprehensive definition
Agreement / disagreement with 3 statements derived from Round 2 results
Analysis of answers
Identification of areas of consensus and dissensus
Consensus?
at least 75% agreement on all items
Yes
Summary report

Figure 1 Flowchart of the Delphi process used in the present study

1. *Definition of problem*

As explained in the introduction, ambiguity exists in the speech-related terminology, particularly with regard to intelligibility and comprehensibility. This ambiguity is noted both in the clinical field and in the literature. It can lead to a communication problem between professionals and impede the efficiency of patient care as well as the progress of research. It also leads to a lack of consensus on the speech tasks and measures to be included in a standard speech assessment. Therefore, the aim of this Delphi study is to draw up a more consensual definition of intelligibility and comprehensibility and their assessment.

2. *Selection of Experts*

We targeted professionals (clinicians, researchers, lecturers) engaged in activities in speech sound disorders[1] and/or fluency disorders (stuttering/stammering). "Activities" were defined as clinical activity, research, academic or industrial activity, or a combination of these, if at least approximately 20% related to speech (self-estimated by the participants). These professionals were required to be able to read English at an intermediate level.

Recruitment was carried out via:

- national professional associations (speech-language-hearing, phoniatrics, voice, acoustics, computer science/signal processing, linguistics/phonetics); over 200 organizations were contacted worldwide by e-mail
- social networks (Twitter, Facebook), where targeted professions and associations were also solicited in private groups

[1]See the ASHA's definition of speech sound disorders: https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

- e-mail to over 50 hand-selected speech experts identified in literature searches on PubMed, who had at least three publications in the field of speech, were authors of a reference book, or participated in research projects linked with pathological speech

Non-respondents for each round were excluded from subsequent rounds. An *a posteriori* analysis using descriptive statistics was carried out to verify that the expert panel characteristics did not change. It was also verified that the increasing consensus throughout the Delphi rounds was not biased by the dropouts of predominantly disagreeing experts. To that end, quantitative data in all three rounds with and without the dropouts were compared, also using descriptive statistics.

3. Literature Search

Prior to launching the survey, a targeted non-systematic literature search was carried out by the first author to identify various definitions of intelligibility and comprehensibility. This search was carried out using PubMed as well as reference lists of known articles on the topic. The definitions had to be explicitly mentioned in the research paper, with the terms "intelligibility" and "comprehensibility". Five definitions, which were considered to best reflect the different interpretations of both terms, were retained to be presented in the first round of this Delphi study:

- Ghio et al. (Ghio *et al.* 2018) (translated from French[2]):

> "The perception of speech is a complex process that integrates both an ascending flow of information from the speech signal and a descending flow based on high-level information held by the listener. The bottom-up flow is mainly an acoustic-phonetic decoding operation that consists in

[2]Original definition: "La perception de la parole est un processus complexe qui intègre à la fois un flux ascendant d'informations provenant du signal vocal mais aussi un flux descendant fondé sur les informations de haut niveau détenues par l'auditeur. Le flux ascendant (« bottom-up ») est principalement une opération de décodage acoustico-phonétique qui consiste à identifier les phonèmes à partir du signal de parole. Les phonèmes, pouvant être considérés comme les plus petites unités permettant d'opposer du sens, sont les éléments de base de l'intelligibilité du discours. […] Le décodage acoustico-phonétique est donc le processus fondamental pour mesurer perceptivement l'intelligibilité d'un locuteur."

identifying phonemes from the speech signal. Phonemes, which can be considered as the smallest units for opposing meaning, are the basic elements of speech **intelligibility**. […] Acoustic-phonetic decoding is therefore the fundamental process for perceptually measuring a speaker's **intelligibility.**"

- Hodge et al. (Hodge and Whitehill 2010):

"**Intelligibility**, or how understandable one's speech is to another, is a functional indicator of oral communication competence. It reflects a talker's ability to convert language to a physical signal (speech) and a listener's ability to perceive and decode this signal to recover the meaning of the talker's message."

- Hustad (Hustad 2008):

"**Intelligibility** refers to how well a speaker's acoustic signal can be accurately recovered by a listener."

- Yorkston et al. (Yorkston *et al.* 1996):

"The term **intelligibility** refers to the degree to which the acoustic signal (the utterance produced by the dysarthric speaker) is understood by a listener. […] The concepts of comprehensibility and intelligibility may be distinguished by the fact that **comprehensibility** incorporates signal-independent information such as syntax, semantics, and physical context."

- Barefoot et al. (Barefoot *et al.* 1993):

"[…] comprehensibility is defined as the extent to which a listener understands utterances produced by a speaker in a communication context. In our view, comprehensibility pertains to the domains of both speech and language, whereas intelligibility pertains principally to the domain of speech. The primary distinction between comprehensibility and intelligibility is that comprehensibility is intended to account for communication features of utterances that extend beyond the auditory-acoustic domain. **Comprehensibility**, in our use of the term, explicitly incorporates contextual features such as syntax, semantics, and pragmatics, and involves face-to-face communication activity in which meaningful utterances are produced by talkers and processed by listeners."

The following statement was also added to stimulate the participants' reflection:

"**Intelligibility** and **comprehensibility** can be used as synonyms."

4. *Delphi Rounds and Data Analysis*

This Delphi consensus survey was conducted in three consecutive rounds, between July and December 2020. Round 1 was available for two months and a half; rounds 2 and 3 were available for 1 month. The online questionnaires are still available on the LimeSurvey platform[3]. The first questionnaire was open access, with built-in duplicate checking and security procedures. The subsequent questionnaires were restricted to previous participants and required a token provided by the moderator. All questionnaires were piloted by five researchers to get an estimate of the response time and to detect and correct potential execution problems (glitches and logical structure issues).

In each round, participants who did not agree with a statement were encouraged to explain the reason of their disagreement, but comments were not mandatory so as not to bias towards positive answers.

The qualitative and quantitative data were analyzed using Stata/MP software (version 14, StataCorp, College Station, TX).

Round 1

The first round was constructed in three parts (31 questions):

1. Background information (13 questions): demographic data about the participant, educational and professional status, experience in the field of speech, study/practice population…

[3] Round 1: https://enquetes.univ-tlse3.fr/index.php/623792?lang=en – Round 2: https://enquetes.univ-tlse3.fr/index.php/372685?lang=en – Round 3: https://enquetes.univ-tlse3.fr/index.php/526489?lang=en

Descriptive statistics were carried out on quantitative and qualitative demographic data to provide a global picture of the expert panel.

2. Terminology (13 questions):
    - open-ended definitions of intelligibility and comprehensibility[4]
    - listing of any other terms used when assessing/describing speech disorders
    - degree of agreement with six definitions/statements from the literature regarding intelligibility and comprehensibility, on a 6-point scale (1: Disagree Strongly - 6: Agree Strongly) and optional comment
    - ranking of the same definitions/statements in decreasing order of preference

Two raters (first and second authors), blinded to the identity of the participants, carried out conventional content analysis (Denman *et al.* 2019, Hsieh and Shannon 2005) of the open-ended definitions as well as of the definitions from the literature, to identify the main recurring themes and concepts. Frequency analysis was used to identify trends by quantifying the number of experts mentioning each of these concepts in their open definitions of both intelligibility and comprehensibility. The degrees of agreement and preference rankings regarding the definitions from the literature were also analyzed using frequency analysis, taking into account the concepts included in each of the definitions.

Together, all of these results were then used to draft 22 statements for round 2, targeting each of the identified main concepts relating to intelligibility and comprehensibility.

Open answers on other terms used to describe and assess speech disorders were semantically grouped into generic and specific terms by the two raters, and frequency analysis was applied.

[4] Participants were repeatedly given the possibility to change their open definitions throughout the questionnaire, e.g. after being asked to rate existing definitions.

3. Assessment methods (five questions):
    - perceptual assessment of intelligibility and comprehensibility (multiple choice with "Other" and "None" options) and listing of any other perceptual speech measures
    - "objective" assessment of intelligibility and comprehensibility (multiple choice with "Other" and "None" options)

    Frequency analysis was used to identify the main trends regarding perceptual and "objective" assessment methods of intelligibility and comprehensibility.

Round 2

Responses and comments from round 1 were synthesized and fed back to the participants for contextualization in round 2, together with the 22 new statements based on the previous responses. This second round was constructed in two parts; all statements were rated using binary answers (Agree/Disagree), with optional comments:

1) Terminology (14 statements): the new statements were grouped into the six concepts identified in the content analysis from round 1 (cf. "Terminology – Intelligibility and Comprehensibility" in the Results section)
2) Speech assessment methods (eight statements): the new statements regarding the perceptual and "objective" speech assessment of intelligibility and comprehensibility were also based on results from round 1

Round 3

A third round was necessary to clarify three statements, which were found to be somewhat ambiguous in round 2. A draft definition paragraph of intelligibility and comprehensibility, integrating all the consensual elements from the previous rounds, was also provided.

5. *Consensus and Stop Criteria*

The threshold for consensus was defined before data analysis as the agreement of at least 75% of the expert panel (Denman *et al.* 2019, Diamond *et al.* 2014).

The planned maximum number of rounds was four, and the stop criterion was the obtention of consensus on all items from each of the two main investigated parts (terminology and speech measures) or stability in the responses.

Results

Delphi Panel

Forty experts completed round 1; thirty-four experts completed round 2 (85%); thirty-three experts completed round 3 (97%, 83% of round 1). A total dropout of 17% from round 1 to round 3 was observed.

Detailed data for the participants in each round are available in Appendix A. The trends described hereafter are constant throughout the three rounds despite the dropouts. Percentages between brackets are ranges across the three rounds.

A majority of the expert panel are speech and language pathologists (SLPs) (70–73%) working in the fields of speech, fluency and voice disorders. Other major groups of participants are linguists (23–24%), ENT/phoniatricians (20–21%) and computer scientists (18–21%). More than half of the experts (58–61%) have at least 10 years of experience working in the speech and voice domains. Their main activity is research for 35–42%, clinical practice for 27–33% and academic activity for 27–28% of the experts (40–46% are associate professors); only two initial participants are engaged in industrial activity. Eighty-five to ninety-four percent of the experts are engaged in at least two main activities; clinical activity and research are combined

in 53–55%. More than half of the participants have a third-cycle diploma (PhD, 58–64%) obtained on average in 2009–2010 (±8 years).

France, the United Kingdom and Germany are the most represented countries (18–21%, 18% and 15% respectively), while the most frequent main language spoken at work is English (63–67%), followed by French (30–33%) and German (12–13%).

The patient/study populations are rather balanced regarding the age groups, with a slight prevalence for the elderly (32–37%) population. Also, the most encountered are acquired and degenerative neurological pathologies (38%-44%).

Preliminary note on dropouts

The *a posteriori* verification revealed that removing the seven dropouts from the analysis in all three rounds did not significantly change the conclusions and consensus values regarding the terminology and the measures of intelligibility and comprehensibility.

The six dropouts after round 1 agreed with the majority of the other experts. The percentages of agreement with the proposed definitions in round 1, for example, decreased by 0–3% when the dropouts were excluded. After round 2, only one additional dropout was counted. The participant was one of three participants who agreed to all of the proposed statements in round 2. The impact on the consensus rate was therefore considered to be minimal, if not positive regarding the reliability of the final results.

Terminology – Intelligibility and Comprehensibility

The conventional content analysis on the open definitions of intelligibility and comprehensibility revealed six main concepts, which also featured in the subsequently presented definitions from the literature:

- Synonymy: mentions of intelligibility and comprehensibility being synonyms

- Message reconstruction: definitions of intelligibility or comprehensibility as the accuracy of the reconstruction of the message by the listener, either at the level of the acoustic signal or at the semantic level
- Phonetic-acoustic production: with regard to intelligibility, mentions of the contribution of the low-level production abilities of the speaker to the message reconstruction
- Acoustic-phonetic decoding: with regard to intelligibility, mentions of the contribution of the low-level decoding abilities on the listener's side to the message reconstruction
- Functional communication: emphasis of the contribution of comprehensibility to functional communication
- Contextual elements: mentions of linguistic, extra-linguistic and para-linguistic elements contributing to comprehensibility

As a reminder, in round 1, participants first had to provide open definitions of intelligibility and comprehensibility. They were then asked to rate their agreement with six provided statements (cf. "Literature search" in the Materials and methods section) and to rank them in decreasing order of preference. The results are presented for each of the identified main concepts:

1. *Synonyms*

When providing spontaneous definitions, no participant mentioned that intelligibility and comprehensibility are synonyms.

Seventy-eight percent (31/40) of the experts disagreed with the statement "Intelligibility and comprehensibility are synonyms" (mean degree of agreement: 2.18/6, mode: 1/6).

Seventy-three percent (29/40) ranked this statement last in preference relative to the other five. Only five percent (2/40) ranked it at the first place. Fifteen percent (6/40) of participants highlighted in their comments that intelligibility refers to the speaker rather than to the listener.

2. *Message Reconstruction*

The majority of participants, in their spontaneous definitions, noted that intelligibility and comprehensibility allow the reconstruction of a message by a listener (intelligibility: 63%, 25/40; comprehensibility: 85%, 34/40). Regarding intelligibility, 90% (36/40) specified that the message is conveyed by the sound signal.

Ninety-eight percent (39/40) agreed with Yorkston et al.'s definition (mean degree of agreement: 5.48/6, mode: 6/6), which states in relation to intelligibility that the information is carried by the acoustic signal.

3. *Functional communication*

In the context of comprehensibility, participants emphasized the functional aspect of communication in their open definitions.

Ninety-three percent (37/40) of them agreed with Barefoot et al.'s definition, which identifies comprehensibility as an indicator of functional communication, as opposed to intelligibility, which is not a direct indicator of functional communication according to the respondents' comments. Thirteen percent (5/40) of the participants who agreed however found it too restrictive to talk about face-to-face communication only.

Seventy-eight percent (31/40) ranked this definition in the top three and no participant ranked it last.

4. *Phonetic-acoustic production*

In their open-ended definitions, with regard to intelligibility, some participants indicated that the reconstruction of the message is allowed by the speaker's phonetic-acoustic production ability (10%, 4/40), in order to obtain a message that is clear (clarity: 15%, 6/40) and easily understood by the speaker (ease of understanding: 8%, 3/40).

Ninety percent (36/40) agreed with Hodge et al.'s definition, which is the only one to take into account the concept of phonetic-acoustic production relating to intelligibility (alongside the concepts of acoustic-phonetic decoding, communication and the recovery of the meaning of the message).

5. *Acoustic-phonetic decoding*

Spontaneously, participants indicated that in the context of intelligibility the reconstruction of the message is based on the acoustic-phonetic decoding abilities of the listener (35%, 14/40), and that it is linked to the sensory capacities of the listener (5%, 2/40).

Ninety-five percent (38/40) agreed with Ghio et al.'s definition, and 100% with Hustad's, which both exclusively link the concept of acoustic-phonetic decoding to intelligibility. However, some participants (8%, 3/40) raised doubts about the limitation to the phonemic level, as well as the exclusion of higher-level elements.

6. *Contextual elements*

In their spontaneous definition of comprehensibility, respondents indicated that the listener's reconstruction of the message combines acoustic-phonetic decoding (5%, 2/40) with contextual elements (48%, 19/40), relying for example on linguistic abilities (10%, 4/40) or on non-verbal language (15%, 6/40).

Ninety-eight percent (39/40) agreed with Yorkston et al.'s definition, which equates intelligibility with acoustic-phonetic decoding and comprehensibility with the reconstruction of the meaning of the message using syntactic, semantic and contextual information. Fifteen percent (6/40), while agreeing with this definition, felt that intelligibility also incorporates signal-independent information. Both participants who agreed and who disagreed with Hodge et al.'s definition highlighted that it describes comprehensibility more than intelligibility.

Together, all of these results were used to draft 14 terminology-related binary assertions (Agree/Disagree) for round 2, targeting each of the six previously described main concepts. These statements, together with the resulting percentages of agreement in round 2, are shown in Appendix B.1.

The respondents did not reach agreement on one statement in round 2: only 41% (14/34) of them agreed with the statement "The assessment of intelligibility and comprehensibility should not take into account the perceptual abilities of the listener." In their comments, participants who disagreed pointed out that perception is part of communication and should be taken into account (95%, 19/20), although mainly in comprehensibility (30%, 6/20). Furthermore, the respondents' comments indicated that perception could be interpreted at different levels of the communication loop: at the peripheral auditory level (hearing screening), but also at the level of receptive language skills of the listener, as well as with regard to the auditory context (e.g. background noise).

Therefore, in round 3, the assertion was specified as follows: "In the context of perceptual assessments, while listener's speech perception factors have to be controlled beforehand (i.e. listener's hearing, but also receptive language skills and auditory context), intelligibility and comprehensibility are used to assess the talker's speech production". Reformulated as such, 85% (28/33) of the participants agreed with this statement. Three of those who disagreed (60%) again indicated that both concepts, but more specifically comprehensibility, also include the listener's ability to reconstruct the utterance/message.

Terminology – Other Terms

Besides intelligibility and comprehensibility, 70% (28/40) of the experts also used other terms, which can be grouped into three main categories (with a total of 72 mentions):

1. Generic terms (43%, 31/72): e.g. articulation/articulatory precision (8), naturalness (6), severity (5)
2. Specific terms (36%, 26/72): e.g. voice-related (phonation, voice quality, intensity …; 7), prosody (6)
3. Others (21%, 15/72): e.g. relating to the pathological context/taxonomy (error consistency, typicality …; 10) or to the functional impact/quality of life (2)

Speech measures

1. Perceptual measures

According to the participants' answers in round 1 regarding the perceptual measures which best describe speech intelligibility and comprehensibility (see figure 2):

- Intelligibility is best measured using orthographic transcription scores (e.g. %-correct items):
    - using real words (50%, 20/40)
    - at phoneme-level (48%, 19/40)
    - using pseudowords/non-words (38%, 15/40)
    - using unpredictable sentences (38%, 15/40)
- Comprehensibility is best assessed using semantic-related measures:
    - Semantic content questions (60%, 24/40)
    - Semantic judgment on sentences (true/false) (50%, 20/40)
    - Sentence-based picture selection (43%, 17/40)
    - Overall subjective rating on Likert or visual analog scales (38%, 15/40)

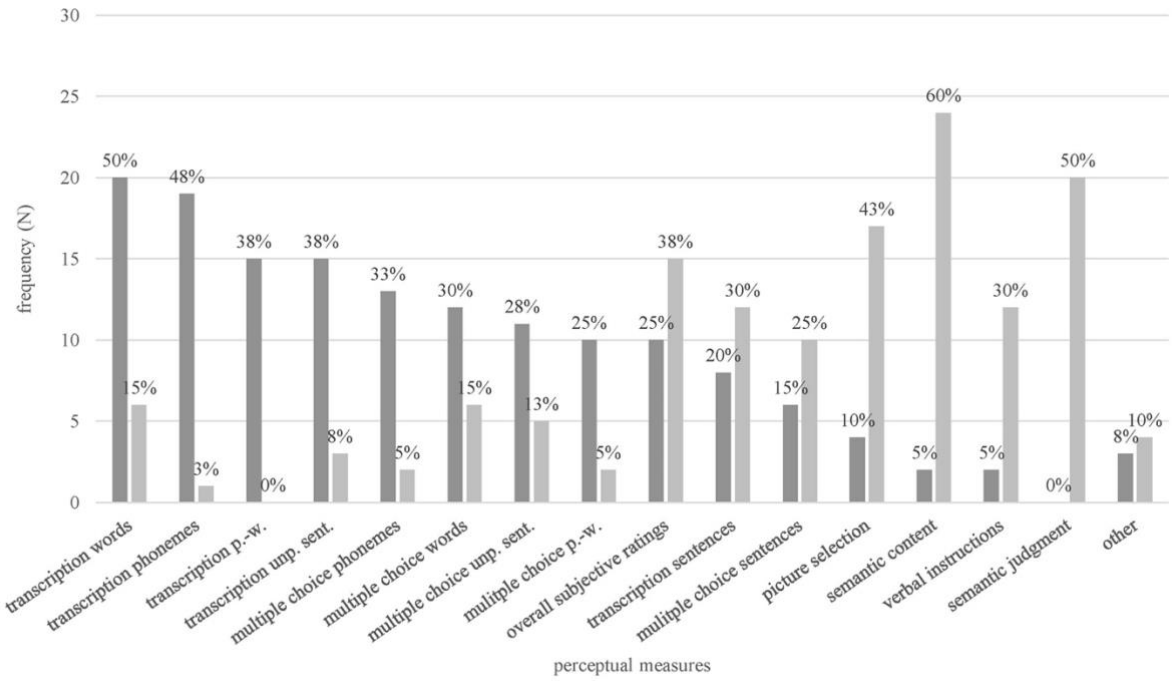

Figure 2 Perceptual measures which best describe speech intelligibility (dark gray) and comprehensibility (light gray); for easier visualization, results were ordered by decreasing order for intelligibility measures; p.-w.: pseudowords, unp. sent.: unpredictable sentences

Ratings on low-level linguistic units (phonemes, pseudo-words, words) are most commonly used for intelligibility (38% for pseudowords, 48% for phonemes, 50% for words).

Higher-lever ratings are preferred for the assessment of comprehensibility (e.g. 60% for semantic content questions). Word-level measures are associated with intelligibility more than with comprehensibility, consistent with the concern raised regarding the reduction of intelligibility to the phoneme-level. Three participants (7%) highlighted that word-level scores are more functional and allow to take into account coarticulation. Two others (5%) emphasized that the use of word-level ratings remains a challenge because of the memorization by the listener and compensation processes based on their linguistic knowledge. Furthermore, two other experts (5%), while agreeing that low-level units are of major interest to assess speech intelligibility, highlighted that phrase-level symptoms such as respiration in dysarthria are neglected.

2. *<u>"Objective" measures</u>*

According to the participants' answers in round 1 regarding the "objective" measures which best describe speech intelligibility and comprehensibility (see figure 3):

- Intelligibility is best assessed using acoustic measures:
    - on consonants (63%, 25/40) and vowels (53%, 21/40)
    - at the suprasegmental level (40%, 16/40)
    - of voice quality (35%, 14/40)
- Comprehensibility is not assessed objectively (68%, 27/40), or is assessed using suprasegmental measures (33%, 13/40)

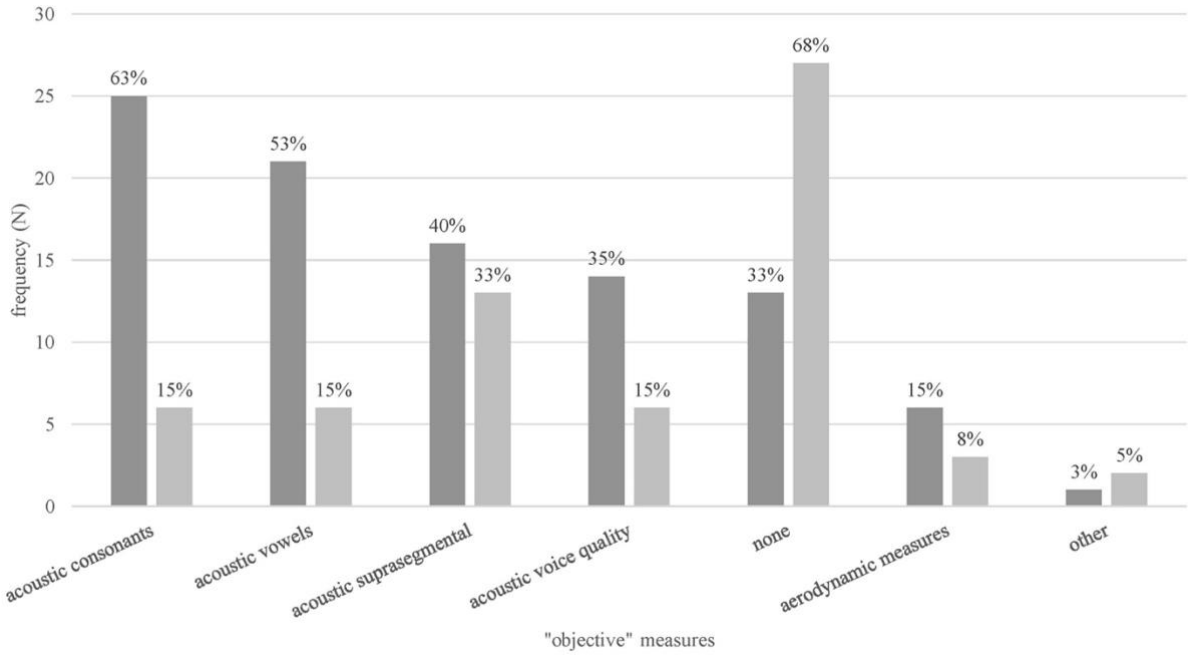

Figure 3 "Objective" measures which best describe speech intelligibility (dark gray) and comprehensibility (light gray); for easier visualization, results were ordered by decreasing order for intelligibility measures

Based on these results, six binary statements (Agree/Disagree) regarding the assessment of intelligibility and comprehensibility were constructed for round 2. These statements, as well as the resulting percentages of agreement in round 2, are shown in Appendix B.2.

The respondents did not reach consensus on one of these statements from round 2, relating to the granularity or level of analysis for the acoustic assessment of intelligibility: “Intelligibility is best assessed using phoneme-level acoustic measures.” Seventy-four percent (25/34) of the experts agreed with it. Those who disagreed specified that these measures are not exclusive, and that a combination of phoneme, word and sentence-level measures is recommended, taking into account various phonetic contexts. However, from the comments of the respondents, it appeared that some had interpreted the statement as referring to measures on isolated phonemes only, thus not taking into account the phonemic context. Therefore, this statement was reformulated in round 3: “Intelligibility is best assessed using consonant, vowel and glide acoustics (incl. inter-phoneme formant transitions), be they on isolated phonemes, or embedded in syllables, in (pseudo-)words or in sentences.” The consensus threshold was reached for this more specific assertion, with 79% (26/33) of the experts now agreeing (note that the participant who dropped out after round 2 agreed to the original assertion; his withdrawal did thus not impact the observed increase of consensus). Three of those who disagreed (43%) specified that there is no “best” way, but rather a necessity to consider several concepts and dimensions. The term “best” was consequently avoided in the integrated definition (see “Final outcome” hereafter).

Still pertaining to the assessment of intelligibility, a second statement of round 2 caught the authors’ attention: “Intelligibility also includes signal-independent elements.” While reaching the consensus cut-off (76% [26/34] agreed), this result was highly inconsistent with other responses (e.g. 97% agreement to the assertion “The intelligibility of a message is specifically carried by the acoustic signal.”). There seemed to be uncertainty about the term “signal-independent elements”, which was explicitly uttered in some respondents’ comments. The

intended meaning of "signal-independent elements" was: all information that is not carried by the acoustic signal, including the knowledge of the conversation topic, the general knowledge, the use of the linguistic context and of non-verbal communication… Hence, "signal-independent elements" referred to the top-down cognitive processes that are independent of the acoustic-phonetic decoding (bottom-up) processes. Accordingly, the statement was rephrased in round 3: "Intelligibility, as opposed to comprehensibility, does not include signal-independent elements (i.e. information from top-down cognitive processes: knowledge of the conversation topic, general knowledge, use of the linguistic context and of non-verbal communication….)". This new phrasing indeed yielded 91% (30/33) agreement. Hence, signal-independent elements were related to comprehensibility rather than to intelligibility in the above definition.

Final Outcome

The aim of this study was to draft a more consensual definition of intelligibility and comprehensibility. Throughout the Delphi process, it quickly appeared that SLPs/phoniatricians, computer scientists, linguists, audiologists, etc. have slightly different but complementary views of the concepts at hand. The following comprehensive passage includes all the consensual elements gathered throughout the Delphi process and tries to reconcile the points of view from the different fields of expertise:

> Intelligibility and comprehensibility are two terms relating to speech, but they are not synonyms. They both refer to the assessment of the speaker's production abilities and both contribute to communication. Hence, while speech production is targeted, the listener's speech perception factors cannot be dismissed (i.e. listener's hearing loss should be excluded at least).

Intelligibility refers to the reconstruction of an ***utterance*** at the acoustic-phonetic level, intelligibility-related information is thus carried by the acoustic signal (i.e. intelligibility focuses on signal-dependent information). This reconstruction is made possible both by the speaker's phonetic-acoustic production ability and by the listeners acoustic-phonetic decoding skills. Perceptually, intelligibility is best analyzed on low-predictability stimuli: phonemes, syllables, pseudo-words, but also words (in minimal pairs) and unpredictable sentences for a more functional assessment taking coarticulation and phrase-level symptoms into account (e.g. respiration and prosody), as long as top-down cognitive compensation processes of the listener are avoided (i.e. no help from semantic or linguistic context).

Objectively, intelligibility can be assessed using consonant, vowel and glide acoustics (incl. inter-phoneme formant transitions), be they on isolated phonemes, or embedded in syllables, in (pseudo-)words or in sentences. Furthermore, in some cases, voice quality also contributes to intelligibility, as it plays a role in certain phonemic contrasts. Supra-segmental parameters (e.g. objectively assessed by speech rate or stress) also contribute to intelligibility.

Comprehensibility refers to the reconstruction of a ***message*** at the semantic-discursive level, subsequent to the acoustic-phonetic reconstruction. Therefore, intelligibility is a component of comprehensibility. In addition to the acoustic-phonetic decoding, it also includes signal-independent, contextual elements such as the linguistic or the non-verbal context. However, one can be comprehensible without all low-level units necessarily being accurately decoded; therefore, while intelligibility affects comprehensibility, the latter is, however, not fully dependent on it.

Comprehensibility refers to the more functional dimension of communication and is perceptually best assessed using meaning-related ratings (i.e. taking into account top-down cognitive processes which might compensate for degraded acoustic-phonetic information).

Nowadays, no objective instrumental measure is yet suitable to assess comprehensibility per se (i.e. the transmission of the overall meaning of the message). However, some suprasegmental parameters contribute to comprehensibility and can be objectively assessed (e.g. timing and intonation measures).

Discussion

Expert panel

While the sample size in Delphi studies was shown to have a significant impact on consensus indexes (Birko *et al.* 2015), there are no clear guidelines on the recommended size. Expert panels can range from numbers as low as six to sample sizes higher than 1000 experts, depending on the topic at hand and on available resources (Powell 2003); a panel size of 12 to 15 experts has been reported in numerous studies (McPherson *et al.* 2018). In the present study, 40 experts participated in the first round. Minimizing attrition of participants throughout the Delphi process is important, as dropouts can lead to an overestimation of the final consensus and impede the reliability of the results (Sinha *et al*. 2011). Hence, in the invitation letter, participants in this study were invited to take part if participation in all of the three expected rounds was envisaged. This might have limited the number of initial participants but therefore resulted in a low dropout rate of 17% (7/40) from the first to the third round, while rates of 20 to 30% are usually expected (Chalmers and Armour 2019). While the sample size is thus satisfying, even more importantly, the characteristics of the expert panel are interesting, as it includes participants from various speech-related fields, backgrounds, and cultural and linguistic contexts. It is also to be noted that the expert panel profiles remained constant throughout the Delphi process (see Appendix A), without major changes in the represented professions, countries, level of education, languages, seniority and patient/study populations of the expert panel. Hence, the Delphi panel was considered as satisfying both in size and composition to address the main objective of this study, which is discussed hereafter.

Terminology

Many authors have debated the terminology of intelligibility and comprehensibility by underlining the lack of a clear-cut definition and by either suggesting their own (e.g. Munro

and Derwing 1995, Field 2005) or endorsing one from the literature (e.g. Berns 2008). Smith and Nelson (1985), for example, in addition to providing short definitions of intelligibility, comprehensibility and interpretability, also presented examples to illustrate the distinctions. Nelson (2008) provides an historical overview of intelligibility and comprehensibility in the study of world Englishes. Thomson (2018) also provides a literature review on the use of intelligibility and comprehensibility and stresses the need for greater consistency in their definition. However, to our knowledge, no study has used a methodological procedure to reach a consensus on these terms. Our Delphi study resulted in a comprehensive definition of intelligibility and comprehensibility, integrating all the consensual elements identified throughout the process. Figure 4 summarizes this definition and illustrates the relationship between the two concepts.

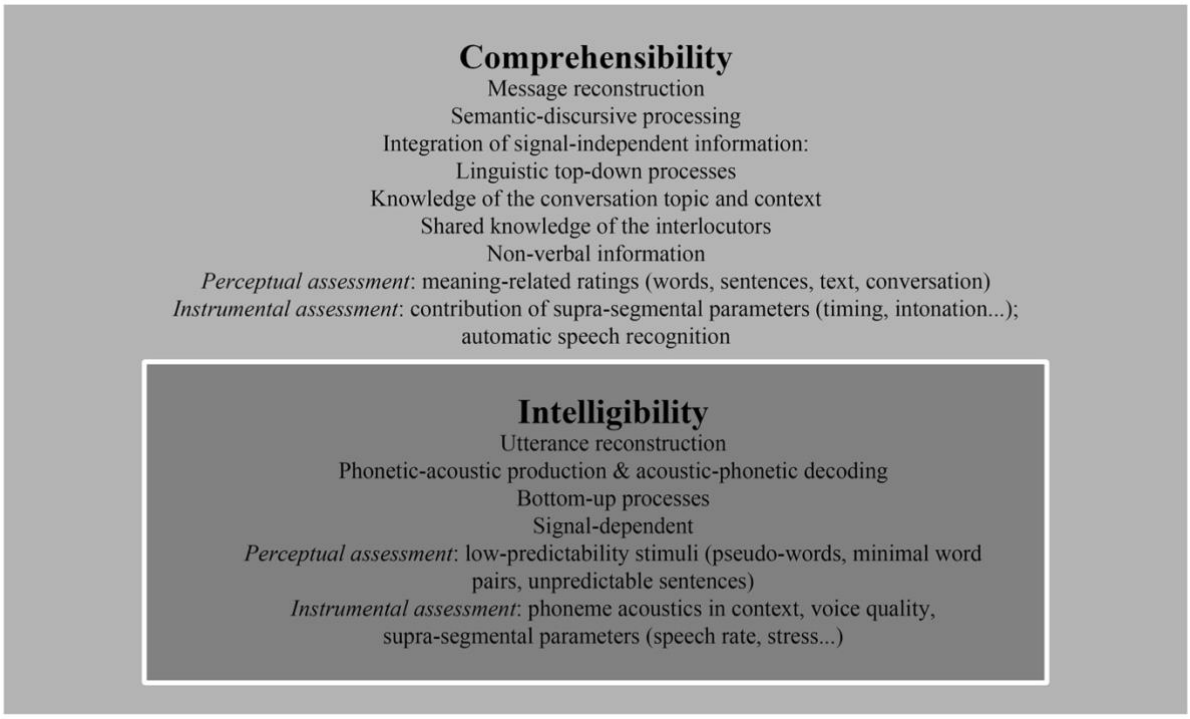

Figure 4 Intelligibility and comprehensibility in speech production

One way to differentiate intelligibility and comprehensibility in the proposed definition is by the term used to refer to the respective (re-)construction process: "*utterance*" is used to relate to the acoustic-phonetic speech material (and, thereby, to intelligibility), while "*message*" is

used as a broader term referring to the (re-)construction at the semantic level (i.e. comprehensibility, thus also including elements of intelligibility). Indeed, as two participants suggested in their comments, while the term "message" can be defined as referring to the "underlying theme or idea"[6] (thus, with a sense of "meaning"), the term "utterance" detaches itself from the communicated semantic content and rather relates to the transmitted acoustic signal.

Also, while intelligibility and comprehensibility are mostly meant to describe an individual's speech production, the listener side, from which both concepts are usually assessed and defined, also plays an important role. Indeed, it proved to be unequivocal in our Delphi study that intelligibility and comprehensibility both contribute to communication, and that functional human communication ("a process by which information is exchanged between individuals"[7]) by definition requires both a speaker and a listener (Schramm 1954). Intelligibility thus integrates not only the accuracy of the phonetic-acoustic production by the speaker, but also the acoustic-phonetic decoding by the listener. Comprehensibility additionally involves numerous higher-level factors, which are also both speaker-related (e.g. non-verbal cues and intonation to compensate for low intelligibility) and listener-related (e.g. listener's knowledge of the speaker, of their intentions and emotions).

Figure 5 further illustrates intelligibility and comprehensibility as part of the communication loop. At each of the latter's levels, variations can occur depending on natural or pathological characteristics of the speaker and of the listener (e.g. gender, age, regional accent, motor speech disorder…). The stages of this loop that form the concept of intelligibility are highlighted in gray. The remaining levels, which pertain to the semantic and pragmatic (re-)construction of the intended message, further account for the concept of comprehensibility. The latter, in

[6] https://www.merriam-webster.com/dictionary/message
[7] https://www.merriam-webster.com/dictionary/communication

addition to the linguistic content, also includes paralinguistic (e.g. sigh, grunt…) and extralinguistic cues (e.g. body language and facial expressions), as well as contextual elements (e.g. prior knowledge of the conversation topic, of the communication partner and field of common experience), which can facilitate the reconstruction of the transmitted message.

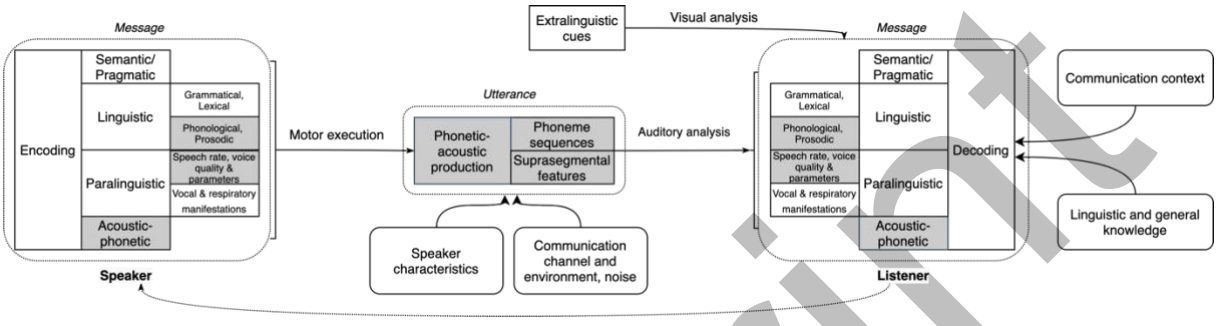

Figure 5 Spoken communication loop. In grey: stages referring to intelligibility

In addition to speaker and listener features, both intelligibility and comprehensibility can be affected by characteristics of the communication channel (e.g. due to an incomplete restitution of the frequency content of the signal in a telephone conversation) and of the communication context (e.g. due to background noise or environmental acoustics such as reverberation).

Speech measures

As discussed above, intelligibility and comprehensibility are most often defined from the listener's perspective, with regard to the reconstruction of the message rather than to its initial construction by the speaker. More specifically, definitions of both concepts usually refer to how they are assessed. Therefore, elements relating to perceptual and acoustic speech measures have also been included in the resulting definition of this Delphi study.

Perceptually, intelligibility can be assessed using low-predictability stimuli (phonemes, syllables, pseudo-words, minimal word pairs or unpredictable sentences)[8]. As mentioned by the

[8] Regarding the units of analysis of intelligibility, it has to be stressed that in some languages, such as in tonal languages, suprasegmental measures (eg. F0 contour) probably contribute to a higher degree to intelligibility than in Western languages, as compared to phoneme-level measures (Chen & Loizou, 2011).

participants, the use of real words and standard sentences, while more functional, can be subject to memorization by the listener if the assessment is carried out by the speech pathologist. Ideally, the speech should thus be assessed by an unfamiliar listener, to avoid any prior knowledge of the expected speech stimuli. Additionally, words and sentences are subject to compensation processes based on the listener's linguistic knowledge and hence border comprehensibility assessment, even if carried out by a lay person. Nonetheless, as illustrated in figure 5, in oral communication the acoustically encoded and decoded phoneme sequences are integrated in an utterance with its prosodic (suprasegmental) features. Both segmental and suprasegmental features thus contribute to intelligibility. For these reasons, minimal word pairs and unpredictable sentences were included in our definition, in addition to phonemes, syllables and pseudo-words. Minimal word pairs are more functional than isolated phonemes, syllables or pseudo-words, as they reflect the discrimination of meaningful units while minimizing the influence of linguistic information. Unpredictable sentences additionally allow coarticulatory and phrase-level phenomena to be taken into account. Preferably, the assessment should still be carried out by a lay person, unless the stimuli are randomly drawn from a large database with distractors to prevent prediction.

In order to assess intelligibility with less subjectivity, one might want to cast off the listener dimension. One way to do so is to turn to computer-aided measures. With regard to the acoustic assessment of intelligibility, while phoneme-level acoustic measures are a major instrumental indicator, the phonemic and suprasegmental contexts in which the target phonemes occur have to be taken into account (just like in perceptual measures, as previously discussed). Ideally, intelligibility measures should thus be carried out on running speech to allow for a more functional assessment, reflecting natural speech conditions. The automatic assessment of intelligibility, for example, is usually carried out on continuous speech but excludes any contextual clues from the algorithms to focus on the assessment of the phonetic-

acoustic production (Fredouille *et al.* 2019). The automatic assessment of comprehensibility, however, still remains problematic. The typical modus operandi of automatic speech recognition (ASR) systems for example can only partly be linked to the concept of comprehensibility. Some contextual cues are indeed used by the algorithms to reconstruct spoken messages. Comprehensibility is then assessed by examining the accuracy of the outputs of these ASR systems. Additionally, the lexical, syntactical and semantic aspects of the speaker's production can then be further analyzed. However, human communication is far more complex and involves numerous para- and extralinguistic dimensions that today's ASR systems do not yet take into account. Computer-aided measures therefore still only partially account for speech comprehensibility in human communication. Nonetheless, it is to be noted that the inclusion of extra- and paralinguistic and contextual information are being more and more investigated, particularly in the field of human-computer interactions, and that promising results have been observed (Kennington *et al.* 2015, Porzel 2011, Schuller *et al.* 2013, 2019).

To conclude, both acoustic and perceptual measures at different levels of granularity (i.e. segmental and suprasegmental) and on various speech materials (isolated phonemes and syllables, words, pseudowords and sentences) should be taken into account. Only their combination allows for a comprehensive assessment of both intelligibility and comprehensibility and thereby provides information on the patient's speech both at the segmental and functional levels.

Limits and Perspectives

The term "objective" was initially used in the online survey. However, the notion of "objectivity" is subject to discussion, some experts arguing that even acoustic measures remain somewhat subjective, as they are carried out by humans, with subjective biases remaining in the recording procedure, analysis settings, choice of the stimuli, of the window analysis… Therefore, this term was used in quotation marks throughout this manuscript. The initially

intended meaning was “reproducible, instrumental measures” as compared to perceptual, more subjective methods. In further studies targeting more consensual definitions of speech-related terms, more consideration needs to be given to the choice of vocabulary used throughout the process.

Several ambiguities persist regarding the terminology related to speech assessment, for example regarding the terms “objective”, “subjective”, “instrumental”, “perceptual” and “measure”, for which various interpretations can still be found in the literature. As one expert from the panel underlined, “… it seems cross-lingual semantics may be part of the tricky issue rather than the scientific principles which we probably agree upon but the semantics/terms are the most challenging part of this problem.” Therefore, further studies are needed to clarify terminology-related issues, generate more unified definitions and allow for an easier progress of research in speech and voice through better communication among experts. The Delphi methodology seems to be an appropriate medium to that end, in light of its features developed in the introduction. As participants pointed out, the Delphi process is “thought-provoking and intellectually challenging”, and sometimes calls into question ones sometimes long-standing daily used terminology. Through the use of multiple iterations, the Delphi process thus stimulates a more insightful problem-solving mindset (Hsu and Sandford 2007). Pending a more consensual speech-related terminology, it is highly recommended that authors clearly define targeted concepts when introducing their research, so as to avoid any ambiguities and allow people from various backgrounds to unequivocally understand their intended meanings.

Conclusion

This Delphi consensus survey has enabled the drafting of a comprehensive definition of intelligibility and comprehensibility, including all the consensual elements gathered throughout the process and thus taking into account the points of view from different fields of expertise. The result of this process allows clinicians and researchers to get a better understanding of these

two commonly used speech-related terms. It enabled us to specify their assessment by describing the tasks that best fit their comprehensive definition. While intelligibility and comprehensibility are linked and both contribute to functional human communication, they relate to the reconstruction of the transmitted speech material at two different levels. Intelligibility refers to the acoustic-phonetic decoding of the utterance, while comprehensibility relates to the reconstruction of the meaning of the message. Consequently, the perceptual assessment of intelligibility requires the use of unpredictable speech material (pseudo-words, minimal word pairs, unpredictable sentences), whereas comprehensibility assessment is meaning- and context-related and entails more functional speech stimuli and tasks.

The terminological disambiguation helps to improve communication between experts in the field of speech disorders and thereby benefits the progress of research as well as research translation. In a clinical perspective, less ambiguous communication between professionals (e.g. in a multidisciplinary team) allows to improve the efficiency of patient care. Furthermore, this study allowed us to specify, for clinicians and researchers, the assessment tasks that best fit the definition of both intelligibility and comprehensibility, thereby providing valuable information to improve speech disorder assessment and its standardization.

Acknowledgments

The authors thank all the participants who volunteered a significant amount of their time to participate in this survey, as well as Prof. Renée Speyer (University of Oslo) for her guidance regarding the Delphi methodology.

References

ALTAHER, A.M., CHU, S.Y., KAM, R. BINTI M., and RAZAK, R.A., 2019, A report of assessment tools for individuals with dysarthria. *The Open Public Health Journal*, **12**, 384–386. https://doi.org/10.2174/1874944501912010384.

AUSTRALIAN INSTITUTE OF HEALTH AND WELFARE, 2003, *Disability prevalence and trends*. Disability Series. AIHW Cat. No. DIS 34 (Canberra: AIHW).

AUZOU, P. and ROLLAND-MONNOURY, V., 2006, *BECD 2006 - Batterie d'Évaluation Clinique de la Dysarthrie* (Isbergues: Orthoédition).

BAREFOOT, S.M., BOCHNER, J. H., JOHNSON, B. A., and VOM EIGEN, B. A., 1993, Rating deaf speakers' comprehensibility: An exploratory investigation. *American Journal of Speech-Language Pathology*, **2**, 31–35. https://doi.org/10.1044/1058-0360.0203.31.

BEERS, M.H., OUSLANDER, J.G., ROLLINGHER, I., REUBEN, D.B., BROOKS, J., and BECK, J.C., 1991, Explicit criteria for determining inappropriate medication use in nursing home residents. *Archives of Internal Medicine*, **151**, 1825–1832. https://doi.org/10.1001/archinte.1991.00400090107019.

BERNS, M., 2008, World Englishes, English as a lingua franca, and intelligibility. *World Englishes*, **27**, 327–334. https://doi.org/10.1111/j.1467-971X.2008.00571.x.

BIRKO, S., DOVE, E.S., ÖZDEMIR, V., and DALAL, K., 2015, Evaluation of nine consensus indices in delphi foresight research and their dependency on delphi survey characteristics: A simulation study and debate on delphi design and interpretation. *PLoS ONE*, **10**, 1–14. https://doi.org/10.1371/journal.pone.0135162.

CHALMERS, J. and ARMOUR, M., 2019, The Delphi technique. In P. Liamputtong (eds), *Handbook of Research Methods in Health Social Sciences* (Singapore: Springer), pp. 715–735. https://doi.org/10.1007/978-981-10-5251-4_99.

CHIN, J., SATO, P. A., and MANN, J. M., 1990, Projections of HIV infections and AIDS cases to the year 2000. *Bulletin of the World Health Organization*, **68**, 1–11.

CUNNINGHAM, B. J., KWOK, E., TURKSTRA, L., and ORAM CARDY, J., 2019, Establishing consensus among community clinicians on how to categorize and define preschoolers' speech and language impairments at assessment. *Journal of Communication Disorders*,

82, 105925. https://doi.org/10.1016/j.jcomdis.2019.105925.

DENMAN, D., KIM, J. H., MUNRO, N., SPEYER, R., and CORDIER, R., 2019, Describing language assessments for school-aged children: A Delphi study. *International Journal of Speech-Language Pathology*, **21**, 602–612. https://doi.org/10.1080/17549507.2018.1552716.

DIAMOND, I.R., GRANT, R.C., FELDMAN, B.M., PENCHARZ, P.B., LING, S.C., MOORE, A.M., and WALES, P.W., 2014, Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, **67**, 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002.

DYKSTRA, A. D., HAKEL, M. E., and ADAMS, S. G., 2007, Application of the ICF in reduced speech intelligibility in dysarthria. *Seminars in Speech and Language*, **28**, 301–311. https://doi.org/10.1055/s-2007-986527.

FASSER, C. E., SMITH, Q. W., and LUCHI, R. J., 1992, Geriatrics fellows' perceptions of the quality of their research training. *Academic Medicine*, **67**, 696–8. https://doi.org/10.1097/00001888-199210000-00017.

FIELD, J., 2005, Intelligibility and the listener. The role of lexical stress. *TESOL Quarterly*, **39**, 399–423.

FREDOUILLE, C., GHIO, A., LAARIDH, I., LALAIN, M., and WOISARD, V., 2019, Acoustic-phonetic decoding for speech intelligibility evaluation in the context of Head and Neck Cancers. In *International Congress of Phonetic Sciences* (Melbourne), pp. 3051–3055.

GHIO, A., LALAIN, M., GIUSTI, L., POUCHOULIN, G., ROBERT, D., REBOURG, M., FREDOUILLE, C., LAARIDH, I., and WOISARD, V., 2018, Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In *XXXIIe Journées d'Études sur la* Parole (Aix-en-Provence), pp. 285–293.

VON DER GRACHT, H.A., 2012, Consensus measurement in Delphi studies. Review and implications for future quality assurance. *Technological Forecasting and Social Change*,

79, 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013.

HODGE, M. and WHITEHILL, T., 2010, Intelligibility impairments. In J. S. Damico, N. Müller, and M. J. Ball (eds), *The Handbook of Language and Speech Disorders* (Chichester: Blackwell Publishing), 99–114. https://doi.org/10.1002/9781444318975.ch4.

HSIEH, H.-F. and SHANNON, S. E., 2005, Three approaches to qualitative content analysis. *Qualitative Health Research*, **15**, 1277–1288. https://doi.org/10.1177/1049732305276687.

HSU, C.-C. and SANDFORD, B. A., 2007, The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, **12**. https://doi.org/10.7275/pdz9-th90.

HUSTAD, K. C., 2008, The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research*, **51**, 562–573. https://doi.org/10.1044/1092-4388(2008/040).

JONES, J. and HUNTER, D., 1995, Qualitative research: Consensus methods for medical and health services research. *Bmj*, **311**, 376. https://doi.org/10.1136/bmj.311.7001.376.

KENNINGTON, C., IIDA, R., TOKUNAGA, T., and SCHLANGEN, D., 2015, Incrementally tracking reference in human/human dialogue using linguistic and extra-linguistic information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver), pp. 272–282. https://doi.org/10.3115/v1/n15-1031.

LETRILLIART, L. and VANMEERBEEK, M., 2011, À la recherche du consensus : Quelle méthode utiliser ? *Exercer*, **99**, 170–177.

LINSTONE, H. A. and TUROFF, M., 2002, *The Delphi method: Techniques and applications* (Newark, NJ: New Jersey Institute of Technology).

MCMILLAN, S.S., KELLY, F., SAV, A., KENDALL, E., KING, M.A., WHITTY, J.A., and WHEELER, A.J., 2014, Using the nominal group technique: How to analyse across multiple groups.

Health Services and Outcomes Research Methodology, **14**, 92–108. https://doi.org/10.1007/s10742-014-0121-1.

MCMILLAN, S.S., KING, M., and TULLY, M.P., 2016, How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy*, **38**, 655–662. https://doi.org/10.1007/s11096-016-0257-x.

MCPHERSON, S., REESE, C., and WENDLER, M. C., 2018, Methodology update: Delphi studies. *Nursing Research*, **67**, 404–410. https://doi.org/10.1097/NNR.0000000000000297.

MUNRO, M.J. AND DERWING, T.M., 1995, Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, **45**, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x.

NELSON, C.L., 2008, Intelligibility since 1969. *World Englishes*, **27**, 297–308.

PERRY, S. and KALBERER, J. T., 1980, The NIH Consensus-Development Program and the Assessment of Health-Care Technologies. *New England Journal of Medicine*, **303**, 169–172. https://doi.org/10.1056/NEJM198007173030334.

POMMEE, T., BALAGUER, M., MAUCLAIR, J., PINQUIER, J., and WOISARD, V., 2021, Assessment of adult speech disorders: Current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, 1–15. https://doi.org/10.1080/14015439.2020.1870245.

POMMEE, T., BALAGUER, M., PINQUIER, J., MAUCLAIR, J., WOISARD, V., and SPEYER, R., (2021), Relationship between phoneme-level acoustic variables and speech intelligibility in healthy speech: A systematic review. *Speech, Language and Hearing*, **24**, 105-132. https://doi.org/10.1080/2050571X.2021.1913300.

PORZEL, R., 2011, *Contextual Computing* (Berlin, Heidelberg: Springer).

POWELL, C., 2003, The Delphi technique: Myths and realities. *Journal of Advanced Nursing*, **41**, 376–382. https://doi.org/10.1046/j.1365-2648.2003.02537.x.

ROULSTONE, S., 2015, Exploring the relationship between client perspectives, clinical expertise and research evidence. *International Journal of Speech-Language Pathology*, **17**, 211–221. https://doi.org/10.3109/17549507.2015.1016112.

RUMBACH, A.F., FINCH, E., and STEVENSON, G., 2019, What are the usual assessment practices in adult non-progressive dysarthria rehabilitation? A survey of Australian dysarthria practice patterns. *Journal of Communication Disorders*, **79**, 46–57. https://doi.org/10.1016/j.jcomdis.2019.03.002.

SCHRAMM, W. L., 1954, How communication works. In W. L. Schramm (eds), *The process and effects of mass communication* (Urbana: University of Illinois Press), pp. 3–26.

SCHULLER, B., STEIDL, S., BATLINER, A., BURKHARDT, F., DEVILLERS, L., MÜLLER, C., and NARAYANAN, S., 2013, Paralinguistics in speech and language - State-of-the-art and the challenge. *Computer Speech and Language*, **27**, 4–39. https://doi.org/10.1016/j.csl.2012.02.005.

SCHULLER, B., WENINGER, F., ZHANG, Y., RINGEVAL, F., BATLINER, A., STEIDL, S., EYBEN, F., MARCHI, E., VINCIARELLI, A., SCHERER, K., CHETOUANI, M., and MORTILLARO, M., 2019, Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge. *Computer Speech and Language*, **53**, 156–180. https://doi.org/10.1016/j.csl.2018.02.004.

SINHA, I. P., SMYTH, R. L., and WILLIAMSON, P.R., 2011, Using the Delphi technique to determine which outcomes to measure in clinical trials: Recommendations for the future based on a systematic review of existing studies. *PLoS Medicine*, **8**. https://doi.org/10.1371/journal.pmed.1000393.

SMITH, L.E. AND NELSON, C.L., 1985, International intelligibility of English: directions and resources. *World Englishes*, **4**, 333–342. https://doi.org/10.1111/j.1467-971X.1985.tb00423.x.

THOMSON, R., 2018, Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang and A. Ginther (eds), *Assessment in Second Language Pronunciation* (London : Routledge), pp. 11–29. https://doi.org/10.4324/9781315170756-2.

WALSH, I., MCNEILLY, L., KENT, R., FOTHERINGHAM, S., PATTERSON, A., BEHLAU, M., WALSH, I., ORMOND, T., ROBB, M., ENDERBY, P., BRADD, T., and WALSH, R., 2006, A history of the terminology of communication sciences and disorders. *Australian Federal Government Department of Education, Science and Training*, 1–29.

WALSH, R., 2005, Meaning and purpose: A conceptual model for speech pathology terminology. *International Journal of Speech-Language Pathology*, **7**, 65–76. https://doi.org/10.1080/14417040500125285.

WOOD, K. S., 1971, Terminology and nomenclature. In L. E. Travis (eds), *Handbook of speech pathology and audiology* (Englewood Cliffs, NJ: Prentice-Hall), p. 3.

YORKSTON, K.M., STRAND, E.A., and KENNEDY, M.R.T., 1996, Comprehensibility of dysarthric speech. *American Journal of Speech-Language Pathology*, **5**, 55–66. https://doi.org/10.1044/1058-0360.0501.55.

Appendix A. Description of the expert panel

Round	1	2	3
Participants	N= 40	N= 34	N= 33
Fields of activity	N (%[1])	N (%)	N (%)
Speech disorders (SD)	38 (95%)	32 (94.1%)	31 (93.9%)
SD only	13 (32.5%)	11 (32.4%)	11 (33.3%)
SD & voice/fluency	25 (62.5%)	21 (61.8%)	20 (60.6%)
Fluency disorders only	2 (5%)	2 (5.9%)	2 (6.1%)
Profession[2]			
Speech and language pathologist	29 (72.5%)	24 (70.6%)	23 (69.7%)
Linguist	9 (22.5%)	8 (23.5%)	8 (24.2%)
ENT/phoniatrician	8 (20%)	7 (20.6%)	7 (21.2%)
Computer scientist	7 (17.5%)	7 (20.6%)	7 (21.2%)
Psychologist/neuropsychologist	2 (5%)	2 (5.9%)	2 (6.1%)
Audiologist	2 (5%)	2 (5.9%)	2 (6.1%)
Neuroscientist	1 (2.5%)	1 (2.9%)	1 (3%)
Other (TEFL, performance/acting/singing)	2 (5%)	2 (5.9%)	2 (6.1%)
Years of experience in speech/voice			
>20 years	13 (32.5%)	12 (35.3%)	12 (36.4%)
15-20 years	6 (15%)	5 (14.7%)	5 (15.2%)
10-15 years	4 (10%)	3 (8.8%)	3 (9.1%)
5-10 years	10 (25%)	8 (23.5%)	8 (24.2%)
< 5 years	7 (17.5%)	6 (17.6%)	5 (15.2%)
Professional activity			
First rank (main activity)			
Research	14 (35%)	14 (41.2%)	14 (42.4%)
Clinical	13 (32.5%)	10 (29.4%)	9 (27.3%)
Academic	11 (27.5%)	9 (26.5%)	9 (27.3%)
Industrial	2 (5%)	1 (2.9%)	1 (3%)
Combinations	34 (85%)	32 (94.1%)	31 (93.9%)
Clinical activity only	4 (10%)	3 (8.8%)	2 (6.1%)
Clinical activity & Research	21 (53%)	18 (52.9%)	18 (54.5%)
Research & Academics	28 (70%)	26 (76.5%)	26 (78.8%)
Level of education			
Third cycle (PhD)	23 (57.5%)	21 (61.8%)	21 (63.6%)
Second cycle (Masters)	16 (40%)	13 (38.2%)	12 (36.4%)
of which PhD students	3 (7.5%)	3 (8.8%)	3 (9.1%)
First cycle (BA, BSc…)	1 (2.5%)	0 (0%)	0 (0%)
Year of diplomation			
Mean (std)	2009.6 (7.88)	2010.3 (7.72)	2010.1 (7.72)
Median (IQR)	2012 (11.5)	2014 (12)	2014 (12)
Academic title (US system equivalents)			
Associate Professor	16 (40%)	15 (44.1%)	15 (45.5%)
No academic activity	8 (20%)	6 (17.6%)	5 (15.2%)
Lecturer	5 (12.5%)	4 (11.8%)	4 (12.1%)

No academic title (e.g. PhD)	3 (7.5%)	3 (8.8%)	3 (9.1%)
Professor	3 (7.5%)	3 (8.8%)	3 (9.1%)
Clinical Instructor	2 (5%)	0 (0%)	0 (0%)
Assistant Professor	1 (2.5%)	1 (2.9%)	1 (3%)
Professor Emeritus	1 (2.5%)	1 (2.9%)	1 (3%)
Senior Lecturer	1 (2.5%)	1 (2.9%)	1 (3%)
Country			
France	7 (17.5%)	7 (20.6%)	7 (21.2%)
United Kingdom	7 (17.5%)	6 (17.6%)	6 (18.2%)
Germany	6 (15%)	5 (14.7%)	5 (15.2%)
Canada	3 (7.5%)	2 (5.9%)	2 (6.1%)
Australia	2 (5%)	2 (5.9%)	2 (6.1%)
Belgium	2 (5%)	2 (5.9%)	2 (6.1%)
Finland	2 (5%)	1 (2.9%)	0 (0%)
Malta	2 (5%)	2 (5.9%)	2 (6.1%)
Colombia	1 (2.5%)	1 (2.9%)	1 (3%)
Hong Kong	1 (2.5%)	0 (0%)	0 (0%)
India	1 (2.5%)	1 (2.9%)	1 (3%)
Netherlands	1 (2.5%)	1 (2.9%)	1 (3%)
Pakistan	1 (2.5%)	1 (2.9%)	1 (3%)
Spain	1 (2.5%)	1 (2.9%)	1 (3%)
Sweden	1 (2.5%)	1 (2.9%)	1 (3%)
Switzerland	1 (2.5%)	0 (0%)	0 (0%)
USA	1 (2.5%)	1 (2.9%)	1 (3%)
Main language[2]			
English	25 (62.5%)	22 (64.7%)	22 (66.7%)
French	12 (30%)	11 (32.4%)	11 (33.3%)
German	5 (12.5%)	4 (11.8%)	4 (12.1%)
Finnish	2 (5%)	1 (2.9%)	0 (0%)
Spanish	2 (5%)	2 (5.9%)	2 (6.1%)
Chinese	2 (5%)	1 (2.9%)	1 (3%)
Catala	1 (2.5%)	1 (2.9%)	1 (3%)
Dutch	1 (2.5%)	1 (2.9%)	1 (3%)
Hindi	1 (2.5%)	1 (2.9%)	1 (3%)
Kannada	1 (2.5%)	1 (2.9%)	1 (3%)
Malayalam	1 (2.5%)	1 (2.9%)	1 (3%)
Maltese	1 (2.5%)	1 (2.9%)	1 (3%)
Portuguese	1 (2.5%)	1 (2.9%)	1 (3%)
Swedish	1 (2.5%)	1 (2.9%)	1 (3%)
Urdu	1 (2.5%)	1 (2.9%)	1 (3%)
Main patient/study population[3]			
Age groups	Average % across participants		
Children (<10 years old)	24.8%	22.2%	20.6%
Adolescents (10-18 years old)	10.1%	7.8%	7.9%
Young adults (19-35 years old)	13.1%	11.5%	11.7%
Middle-aged adults (36-65 years old)	20.2%	22.3%	22.7%
Elderly (65+)	31.8%	36.2%	37.1%
Pathologies			
Neurodegenerative disorders	22.1%	25.7%	26.4%
Acquired neurological disorders	15.4%	17.1%	17.2%
Others (speech-/voice-related)	13.6%	8.8%	7.7%

Oncology	8.9%	10.4%	10.8%
Functional voice disorders	5.6%	5.4%	5.2%
Other structural speech deficits	5.2%	5.7%	5.9%
Fluency disorders	5.1%	4.7%	4.7%
Others (non speech-/voice-related)	4.9%	2.6%	2.1%
Structural voice disorders	4.6%	4.1%	4.1%
Hearing-impairment	4.5%	4.6%	4.7%
Neurogenic voice disorders	4.1%	4.6%	4.7%
Polyhandicap	3.8%	3.8%	3.9%
Neurological tumors	1.8%	2.1%	2.1%
Iatrogenesis	0.4%	0.4%	0.5%

[1] Due to rounding of numbers to one decimal, minimal deviations from 100% may occur

[2] Combinations were possible (e.g. in countries with more than one official language)

[3] Each participant distributed 100% over all given categories; results are average percentages across participants for each category

Appendix B. Statements presented in round 2 and percentages of agreement

In round 2, respondents have reached a consensus on 20 out of 22 statements (i.e., > 75% agreement between raters). A high agreement (> 90%) was even reached for 14 of these statements. Results for this second round are reported as frequencies of agreement for each of the binary assertions (Agree/Disagree), grouped by target concepts and in decreasing order of agreement. In bold, the statements that did not reach the consensus threshold or were inconsistent with other results. These statements were either rephrased in the subsequent round or discarded after round 3.

1. Terminology

Concept	Assertion	Agreement
Synonyms	- "Intelligibility and comprehensibility are two terms relating to speech but are not synonyms."	97% (33/34)
	- "Intelligibility and comprehensibility refer to the speech production abilities."	76% (26/34)
	- "The assessment of intelligibility and comprehensibility should not take into account the perceptual abilities of the listener."	**41% (14/34)**
Message reconstruction	- "Intelligibility and comprehensibility both allow the reconstruction of a message by the listener."	97% (33/34)
	- "More precisely, the intelligibility of a message is specifically carried by the acoustic signal."	97% (33/34)
Functional communication	- "Intelligibility and comprehensibility contribute to communication."	100% (34/34)
	- "Comprehensibility refers more to the functional dimension of communication than intelligibility."	94% (32/34)
	- "Comprehensibility refers to functional communication and is therefore not limited to the face-to-face context (e.g. audio telephone conversations)."	94% (32/34)
Phonetic-acoustic production	- "For intelligibility, the reconstruction of the message is made possible by the speaker's phonetic-acoustic production ability."	97% (33/34)
Acoustic-phonetic decoding	- "For intelligibility, the reconstruction of the message is made possible by the listener's acoustic-phonetic decoding capabilities."	97% (33/34)
	- "The most relevant unit of analysis for intelligibility is the phoneme."	85% (29/34)
Contextual elements	- "Comprehensibility combines, in addition to acoustic-phonetic decoding, contextual elements, such as the linguistic or the non-verbal context."	94% (32/34)
	- "Intelligibility, which relates to acoustic-phonetic decoding, is therefore a component of comprehensibility."	94% (32/34)
	- "Intelligibility also includes signal-independent elements."	**76% (26/34)**

2. Speech measures

Concept	Assertion	Agreement
Perceptual measures	- "Word-level ratings also relate to intelligibility."	100% (34/34)
	- "Perceptually, comprehensibility is best assessed using meaning-related ratings."	94% (32/34)
	- "Unpredictable sentences allow for a perceptual assessment of intelligibility."	91% (31/34)
	- "Perceptually, intelligibility is best assessed using %-correct scores on low-level items (phonemes, syllables, pseudowords)."	76% (26/34)
"Objective" measures	- "Supra-segmental objective measures (timing, intonation, stress…) relate to both intelligibility and comprehensibility."	91% (31/34)
	- "Voice quality also relates to intelligibility, as it contributes to the acoustic-phonetic decoding."	85% (29/34)
	- "No objective measure is suitable to directly assess comprehensibility."	76% (26/34)
	- "Intelligibility is best assessed using phoneme-level acoustic measures."	**74% (25/34)**