OPEN ACCESS

This is the German version.
The English version starts at p. 12.

Sprachtests im Online-Experiment am Beispiel des Oldenburger Satztests

Zusammenfassung

Während der COVID-19-Pandemie waren audiologische Untersuchungen im Labor nur unter sehr strengen Hygienemaßnahmen bzw. gar nicht möglich. Um dieses Hindernis zu umgehen, wurde ein Online-Experiment mit dem synthetischen Sprachmaterial des Oldenburger Satztests (OLSA) implementiert. Dadurch sollte geprüft werden, ob online mit synthetischer Sprache vergleichbare Ergebnisse erzielt werden wie in einem vorhergehenden Experiment im Labor. Mithilfe der Plattform Gorilla (https://gorilla.sc/) wurde ein Online-Experiment erstellt und durch einen Link über verschiedene Kommunikationskanäle (z. B. in Social Media) weitergegeben. Aufgerufen wurde dieser Link 182 Mal, was zu 81 vollständigen Datensätzen führte. Die ermittelten Sprachverständlichkeitsschwellen von 69 Probanden sind vergleichbar mit den Ergebnissen aus der Laborsituation. Allerdings wichen die Steigungen der Diskriminationsfunktionen signifikant voneinander ab. Grundsätzlich konnte gezeigt werden, dass der synthetische OLSA auch als Online-Experiment durchführbar ist. Hürden im Vergleich zur Labormessung zeigten sich jedoch bei Rekrutierung und Motivation der Probanden.

Schlüsselwörter: OLSA, Online-Experiment, SRT, synthetische Sprache

Anne Schlüter[1]
Elisabeth-Sophie Baumann[1]
Fatma Ben Ghorbal[1]
Nele Hauenschild[1]
Alina Kleinow[1]
Vanessa Mazur[1]
Julia Thomas[1]
Inga Holube[1]

1 Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg, Deutschland

1 Einleitung

In wissenschaftlichen Studien werden die meisten audiologischen Messungen, wie z. B. Sprachverständlichkeitsmessungen, in Laboren oder dafür vorgesehenen akustisch optimierten Räumen durchgeführt, die mit den dafür notwendigen Messapparaturen und -aufbauten ausgestattet sind. Während der COVID-19-Pandemie und den daraus folgenden Hygiene- und Schutzmaßnahmen waren Messungen im Labor nur schwierig bis gar nicht umsetzbar [1]. Um dennoch Untersuchungen realisieren zu können, bieten Messungen, die zuhause online durchgeführt werden, eine mögliche Alternative. Daraus ergeben sich die Fragen, ob bestimmte audiologische Untersuchungen wie ein Sprachtest auch online umgesetzt und so auch zuhause durchgeführt werden können und ob die Ergebnisse vergleichbar mit Labormessungen sind.

Nach Sauter et al. [2] sind für die Entwicklung einer Online-Studie folgende Schritte notwendig: a) das Programmieren des Experimentes z. B. zur Nutzung mit einem Browser, b) das Aufbauen eines geeigneten Testumfeldes z. B. mithilfe eines Online-Hosts auf einem Server und c) das Rekrutieren der Probanden. Für die grundsätzliche Umsetzung von Online-Verhaltens-Experimenten im Browser stellt Gorilla (https://gorilla.sc/) eine Online-Plattform zur Verfügung [3]. Sie ermöglicht die Umsetzung eigener Experimente über einfache graphische Benutzeroberflächen oder die Erstellung von JavaScript-Code. Dazu stellt Gorilla einen cloudbasierten Server und das Servermanagement bereit, die auch den Anforderungen der Datenschutz-Grundverordnung entsprechen [4]. So ist z. B. die grundsätzliche Funktionalität einer audiologischen Messung, die Darbietung und Veränderung von Audiosignalen sowie die Aufnahme der Reaktion des Probanden auf diese Signal über einen Browser möglich. Die Experimente können potentiellen Probanden über z. B. einen Link zugänglich gemacht werden oder die Probanden werden über Rekrutierungs-Portale wie z. B. Prolific oder MTurk (für eine Übersicht siehe [2], [5]) kontaktiert. Nach Abschluss der Experimente kann über eine Datenbank auf die erhobenen Daten zugegriffen, diese heruntergeladen und weiter ausgewertet werden.

Verschiedene Aspekte schränken die Auswahl des Sprachtests und seine Programmierung für ein Online-Experiment ein. Prinzipiell fehlen Einflussmöglichkeiten auf das Messequipment, wie z. B. der verwendete Kopfhörer und die Soundkarte. Über einen Antiphasentest [6] kann zwar die grundsätzliche Nutzung von Kopfhörern sichergestellt werden, aber eine exakte Kalibrierung des Messaufbaus ist nicht möglich. Deshalb können im Online-Experiment nur die Hörbarkeit sichergestellt und relative Änderungen der Pegel als Ergebnisse ermittelt werden. Der Einfluss möglicher Hintergrundgeräusche kann nur minimiert werden, indem die Signale deutlich überschwellig dargeboten werden. Für einen Online-Sprachtest eignet sich unter den genannten Aspekten deshalb ein Test im Störgeräusch, mit dem eine Sprachverständlichkeitsschwelle (engl. Speech Recognition Threshold, SRT) ermittelt wird. Diese beschreibt den Signal-Rausch-Abstand (engl. Signal-to-Noise Ratio, SNR),

bei dem 50% der Sprache verstanden wird. Außerdem ist es in einem Online-Test einfacher, geschlossene Antwortmöglichkeiten, also eine Auswahl aller Antwortmöglichkeiten, anzubieten, als ein freies Antworten zu ermöglichen. Die Antworteingabe im geschlossenen Test kann direkt in Verständlichkeits-Werte umgewandelt und im Verlauf der Messung weiter z. B. für eine adaptive Steuerung genutzt werden. Alternativen für eine offene Durchführung wie Audioaufnahmen oder Texteingaben müssten im laufenden Online-Experiment erst durch Sprach- oder Texterkenner als richtig oder falsch interpretiert werden, um sie dann in Verständlichkeitswerte zu überführen. Bei der Interpretation können Fehler entstehen, da Nachfragen durch den Untersucher wie im Labor nicht möglich sind. Um eine geschlossene Umsetzung für ein Online-Experiment zu ermöglichen, ist ein deutlich begrenztes Sprachmaterial notwendig. Gleichzeitig muss beachtet werden, dass das begrenzte Sprachmaterial selbst nicht erlernbar bzw. der Lerneffekt kontrollierbar ist.

Ein Sprachtest, der diese Kriterien erfüllt, ist der Oldenburger Satztest (OLSA, [7], [8], [9]). Seine Sätze besitzen eine einheitliche syntaktische Struktur: Name Verb Zahl Adjektiv Objekt. Je zehn Wörter stehen in jeder Wortgruppe zur Verfügung, die unterschiedlich miteinander kombiniert werden können. Dabei entstehen Sätze, die grammatikalisch richtig sind, aber nicht immer einen Sinn ergeben (z. B. „Peter kauft zehn nasse Sessel."). Das dazugehörige Rauschen wurde durch zufällige Überlagerung des Sprachmaterials erzeugt und besitzt dadurch das gleiche Langezeitspektrum wie die Sprache. Den Probanden werden die OLSA-Sätze im Rauschen vorgespielt und ihre Aufgabe ist es, die verstandenen Wörter zu wiederholen oder sie in der Eingabemaske auszuwählen. In Abhängigkeit von den verstandenen Wörtern wird der SNR adaptiv verändert, um den SRT zu ermitteln. Aufgrund des begrenzten Sprachmaterials ist der OLSA als geschlossener Test durchführbar und das Antwortverhalten der Probanden lässt sich wortgenau dokumentieren. Aufgrund der geringen Anzahl der Wörter pro Wortgruppe kann ein Lerneffekt auftreten [7], [8]. Deshalb muss vor der Messung das Sprachmaterial mit zwei Listen geübt werden [10].

Für die originale Version des OLSA wurden die Sätze von einem männlichen Sprecher eingesprochen [7], [8], [9]. Als weitere Variante wurde das Sprachmaterial mit einer weiblichen Sprecherin aufgenommen [11], [12]. Von Nuesse et al. [13] wurde das Sprachmaterial mit einem Text-to-Speech-System mit einer weiblichen Stimme synthetisiert und mit den Aufnahmen der natürlichen weiblichen Sprecherin im Labor verglichen. An diesen Messungen nahmen 48 Probanden (mittleres Alter: 21,8 Jahre) teil. Sie waren normalhörend, sprachen Deutsch als Muttersprache und ihnen war der OLSA unbekannt. Nuesse et al. [13] zeigten, dass bei Messungen mit der synthetisch erzeugten Sprecherin ein ähnliches Sprachverstehen wie mit der natürlichen Sprecherin erzielt wird. Das synthetisierte Sprachmaterial steht für Forschungszwecke frei zur Verfügung [14]. Deshalb wurde es für das Online-Experiment verwendet.

Bei der Rekrutierung und Durchführung von Online-Studien zeigen sich Vor- und Nachteile. Reips [15] beschreibt die Möglichkeit, eine größere Anzahl an Probanden in kürzerer Zeit zu rekrutieren. Die Stichprobe zeigt i. d. R. eine größere Diversität und ist damit der Allgemeinbevölkerung ähnlicher als Probanden, die für Labormessungen rekrutiert wurden. Diese Diversität zeigt sich auch im verwendeten Messequipment. Da die Probanden die Messungen selbständig durchführen, sind die Messungen kostengünstig und frei von Fehlern, die vom Untersucher erzeugt werden könnten. Durch den fehlenden persönlichen Kontakt bleibt aber auch die Beobachtung der Probanden hinsichtlich ihrer Motivation und ihres Antwortverhaltens aus. Mehrfachteilnahmen oder die Teilnahme von Bots sind möglich. Hinzu kommt, dass die Teilnahmebereitschaft und Konzentration von Probanden bei Online-Studien mit steigender Studiendauer sinkt (z. B. [16], [2]). Eine Studienlänge von über 30 min ist zwar durchführbar, kann aber durchaus demotivierend wirken [2]. Grundsätzlich muss mit größeren Abbrecherquoten als bei Labor-Studien gerechnet werden [15].

Um also die grundsätzliche Frage, ob Sprachtests online zuhause eingesetzt werden können, zu beantworten, wurde ein Online-Experiment mit dem OLSA über die Plattform Gorilla erstellt und durchgeführt. Aspekte, die die Randbedingungen des Experimentes beschreiben, wie z. B. Teilnehmerzahlen, Abbrecherquoten und Durchführungsdauern, werden dargestellt. Basierend auf der Hypothese, dass keine Unterschiede vorliegen, wurden die ermittelten SRT-Werte und Steigungen im Online-Experiment mit denjenigen verglichen, die von Nuesse et al. [13] im Labor erhoben wurden.

2 Methode

2.1 Messumgebung und -equipment

Die Probanden konnten über einen von Gorilla bereitgestellten Link das Online-Experiment jederzeit an jedem Ort mit einem Tablet oder Computer (Laptop und Desktop), der einen Internetzugriff erlaubte, aufrufen. Gorilla bietet verschiedene Möglichkeiten, die Rahmenbedingungen eines Experimentes einzuschränken. So kann z. B. das Medium ausgewählt werden, über das ein Online-Experiment vom Probanden durchgeführt werden kann. Tablets oder Computer wurden für dieses Experiment zugelassen. Mobiltelefone wurden ausgeschlossen, da angenommen wurde, dass sie häufiger als Computer oder Tablets in schnell veränderlichen Situationen eingesetzt werden. So sollte eine stabile und reliable Messumgebung gefördert werden. Browsertyp, Standort und Verbindungsgeschwindigkeit wurden nicht eingeschränkt. Außerdem wurden die Probanden gebeten, das Experiment nicht über Lautsprecher durchzuführen, sondern Kopfhörer zu tragen. Dies sollte ebenfalls die Stabilität und Reliabilität der Messsituation unterstützen.

2.2 Erstellung des Online-Experimentes

2.2.1 Aufbau des Experimentes

Online-Experimente auf der Plattform Gorilla [5] sind modular aufgebaut. Verschiedene Elemente wie sogenannte Questionnaires und Tasks werden separat für die Ausführung durch die Probanden erzeugt. Die Bestandteile der Questionnaires eignen sich zum Beispiel, um Formulare zu konstruieren oder auch Skalen und andere Arten von Fragebögen einzubinden. Tasks werden im Task Builder für die Probanden erstellt. So können verschiedene graphische Oberflächen kreiert werden, in denen die Aufgaben für Probanden integriert sind. Ebenso werden darin die Darstellungsreihenfolge der graphischen Oberflächen über das Spreadsheet und die Ergebnisstruktur definiert, sowie Stimuli (z. B. Abbildungen, Audiodateien oder Videos) bereitgestellt.
Der generelle Ablauf des hier verwendeten Experimentes ist in Abbildung 1 dargestellt. Gestartet wurde das Experiment mit einem Questionnaire, der auf seiner ersten Seite die Probanden begrüßte sowie über den Ablauf und die grundsätzlichen Anforderungen informierte. Zusätzlich wurde auf die ungefähre Messdauer von 30 min hingewiesen und die Verwendung eines Kopfhörers empfohlen sowie auf eine detaillierte Probandeninformation verlinkt. Auf weiteren Seiten wurden Kontaktinformationen eingeblendet, das Einverständnis zur Freiwilligkeit und Anonymität eingeholt, über den Datenschutz informiert und die Einwilligung zur Teilnahme erbeten. Um die Ergebnisse zu pseudonymisieren, folgte ein Task, mit dem ein persönliches Codewort erstellt wurde.
Im weiteren Verlauf des Experimentes wurden die Probanden aufgefordert, einen Anamnesefragebogen auszufüllen. Hier wurden allgemeine Daten wie das Alter, das Geschlecht, die Muttersprache und Angaben bezüglich des allgemeinen Gesundheitszustandes, einem möglichen Hörproblem sowie die Bekanntheit des OLSA mit weiblicher Sprecherin erfasst. Nach dem Gesundheitszustand wurde gefragt, um eine nachträgliche Auswahl der Daten vornehmen zu können und evtl. erkrankte Probanden, bei denen eine Beeinflussung des Messergebnisses aufgrund einer Erkrankung unklar ist, auszuschließen. Nach dem weiblichen OLSA wurde explizit gefragt, um ebenfalls eine nachträgliche Auswahl der Daten vornehmen zu können und gleichzeitig nicht zu viele Datensätze aufgrund der generellen Kenntnis der Probanden über den OLSA auszuschließen. Gaben die Probanden an, ein Hörproblem festgestellt zu haben, so wurden sie zu einem weiteren Fragebogen geleitet. Dieser erfragte eine Seitenungleichheit des Gehörs (Frage: Hören Sie mit beiden Ohren gleich gut?, Antwortmöglichkeiten: Ja, Nein, Weiß nicht) und den Grad des Hörverlustes (Frage: Bitte beschreiben Sie den Grad Ihres Hörverlustes (auf dem besseren Ohr); Antwortmöglichkeiten: sehr schwach, schwach, mittel, stark, sehr stark). An der Studie durften nur Personen teilnehmen, die zwischen 18 und 30 Jahre waren, als Muttersprache Deutsch und keinen oder einen sehr schwachen Hörverlust hatten. Entsprachen die Angaben nicht den festgelegten Einschlusskriterien der Studie, bekamen die Probanden eine Information zu ihrem Ausschluss angezeigt und das Experiment wurde automatisch beendet. Andernfalls konnten sie das Experiment fortsetzen.
Dann folgte ein Task zur Einstellung einer angenehmen Lautstärke der Stimuli (Sprache und Rauschen). Die Einstellung der Lautstärke erfolgte über den Hauptlautstärkesteller für die Audioausgabe des eigenen Computers bzw. Tablets. Zuerst sollte die Hauptlautstärke auf 0 gestellt und anschließend der Pegel des präsentierten Rauschens mit dem Hauptlautstärkesteller auf eine angenehme Lautstärke justiert werden. Im Weiteren sollte diese Einstellung mit einem präsentierten Sprachsignal vom Probanden nachjustiert werden, wenn die Sprache nicht verstanden wurde, und abschließend bestätigt werden. Diese Einstellung des Hauptlautstärkestellers des jeweiligen Endgerätes (Tablet oder Computer) war Grundlage für das weitere Experiment und konnte durch Gorilla nicht ausgelesen bzw. nicht in Gorilla verändert werden. Eine Übersteuerung der Signale im Online-Experiment war nicht zu erwarten, da Probanden mit einem normalen Gehör untersucht wurden und in Gorilla der Pegel angenehmer Lautheit lediglich durch die Veränderungen des SNR im OLSA angepasst wurde.
Nach der Einstellung der Lautstärke folgte ein Task, der die Nutzung von Kopfhörern mit Hilfe eines in Gorilla vorliegenden Antiphasentests [6] überprüfte. Hierzu wurden drei reine Töne in Stereo dargeboten, wovon ein Ton eine 180°-Phasenverschiebung auf einem Kanal aufweist. Die Probanden hatten die Aufgabe, den leisesten Ton auszuwählen. Diese Aufgabe sollte aufgrund der Phasenauslöschung bei Verwendung von Lautsprechern schwierig sein. Woods et al. [6] zeigten, dass Probanden zuverlässig nach der Nutzung von Kopfhörern und Lautsprechern unterschieden werden konnten. Nach der Durchführung des Antiphasentests wurde das Experiment für Probanden beendet, die keinen Kopfhörer nutzten. Beim letzten Schritt des Experimentes handelte es sich um den Task, der den OLSA beinhaltete (siehe Abschnitt 2.2.2).
In Gorilla wurde die Bearbeitungszeit für dieses Online-Experiment auf 2 h limitiert und nach dieser Zeit wurde es automatisch beendet. Diese Einschränkung war notwendig, um die Datenerfassung abschließen zu können und begonnene Experimente, die aber nicht fortgeführt wurden, abzubrechen. Nach Abschluss jedes Tasks und jedes Questionnaires wurden die dazugehörigen Daten automatisch von Gorilla in Tabellen gespeichert. Allerdings konnten nur für durch die Probanden komplett abgeschlossene Experimente die Messdaten abgerufen werden. Für jedes vollständig durchgeführte Experiment musste ein Betrag an Gorilla bezahlt werden [5].

2.2.2 OLSA in Gorilla

Für die Messung des OLSA in Gorilla wurde das Sprachmaterial genutzt, das mit einer synthetisch erzeugten weiblichen Stimme generiert wurde [13]. Zusätzlich wurde

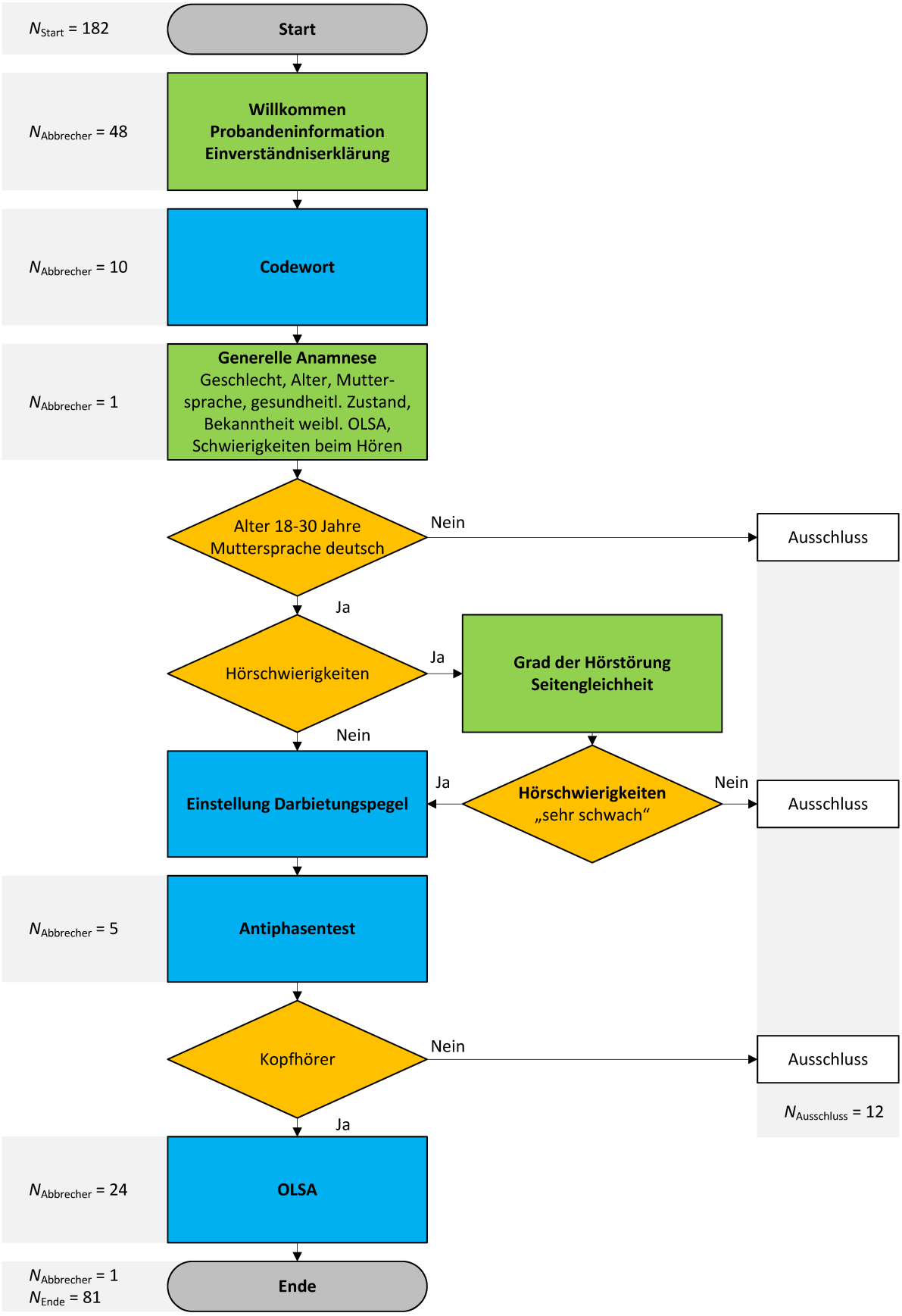

Abbildung 1: Ablaufdiagramm des Online-Experimentes zur Messung des OLSA erstellt in Gorilla. Grüne Prozesse wurden in Gorilla als Questionnaires umgesetzt. Blaue Prozesse wurden als Tasks gestaltet. Orangene Abschnitte waren Entscheidungen. Außerdem ist die Anzahl (*N*) der gestarteten Experimente, der innerhalb eines Abschnittes abgebrochenen Experimente, der Ausschlüsse und der am Ende vollständig abgeschlossenen Experimente angegeben.

das dazugehörige Rauschen verwendet. Jeweils 30 Sätze bildeten eine Liste. Insgesamt wurden 45 Listen als mp3-Dateien (Bitrate bei der Umwandlung von wav- in mp3-Format mit maximal möglicher Qualität von 320 kbits/s, Konvertierung durchgeführt mit der Webseite: https://online-audio-converter.com/de/) in Gorilla implementiert. Diese entsprechen den Listen aus Nuesse et al. [13]. Die Probanden hatten die Aufgabe, die verstandenen Wörter aus einer Antwortmatrix (10x5), die auf dem Bildschirm dargestellt war, auszuwählen. Während der Messung wurde ein fester Sprachpegel dargeboten und der Rauschpegel wurde abhängig von der Antwort des Probanden adaptiv angepasst. Der Sprachpegel wurde konstant gehalten, da dieser zuvor durch den Probanden in der Einstellung des Darbietungspegels so gewählt wurde, dass die Sprache verständlich war. Die adaptive Anpassung wird auch im originalen OLSA [7], [8], [11] angewendet und wurde von Brand und Kollmeier [17] beschrieben. Der SNR des ersten Satzes einer Liste betrug immer 0 dB. Die Probanden führten den OLSA dreimal nacheinander durch. Bei den ersten beiden Listen handelte es sich um Trainingslisten [7], [8], [9], [10].

In der Task Structure für den OLSA wurden vier Oberflächen erstellt. In der ersten Oberfläche wurde die Funktion Fullscreen eingebaut. Damit wurde automatisch der Bildschirm des Nutzers zu Beginn im Vollbildmodus angezeigt. Dadurch sollte die Konzentration der Probanden auf den Sprachtest erleichtert werden. Ein Textfenster mit einer ausführlichen Einweisung zum OLSA folgte. Diese beinhaltete das Ziel der Messung, den genauen Ablauf, eine Beschreibung der Stimuli, die Dauer und Anzahl der Sätze bzw. Satzlisten sowie die genaue Aufgabe des Probanden. Außerdem wurde darauf hingewiesen, dass der Sprachtest bewusst unter schwierigen Bedingungen stattfindet und dass Pausen jederzeit möglich sind. Nach dem Starten der Messung folgte eine Oberfläche, in der der Start einer Satzliste bzw. die Ankündigung des nächsten Messdurchgangs eingebettet waren. Beim Betätigen des Buttons „Nächste Messung“ wurde die Antwort-Oberfläche aktiviert. Sie zeigte zunächst für die Zeit der Satzpräsentation einen weißen Bildschirm, damit die Aufmerksamkeit des Probanden auf das Verstehen des Satzes gerichtet wurde und die Darstellung der Antwortmatrix und eines Fortschrittsbalkens von diesem nicht ablenkt. Erst nach der Satzpräsentation wurde die Antwortmatrix für die Eingabe der verstandenen Satzteile eingeblendet und ein Fortschrittsbalken dargestellt. Diese Oberfläche ist mit einem Spreadsheet verknüpft, das die Darstellung der Antwortoptionen ermöglicht und auch die Reihenfolge und Häufigkeit der zuvor erstellen Oberflächen aufgreift. Das Spreadsheet wurde auch dafür genutzt, die Ergebnisstruktur für jeden Probanden festzulegen. Es wurden u. a. die SNR-Werte sowie die Anzahl der richtig verstandenen Wörter pro Satz in die Tabelle mit aufgenommen. Der nächste Satz wurde dann mit dem Button „Weiter“ gestartet. Den Probanden wurde keine Rückmeldung zum Verständnis der Wörter gegeben. Die letzte Oberfläche des OLSA beinhaltete ein Kommentarfeld für die Personen, die am Experiment teilgenommen haben, und die Kontaktinformationen der Studienleiterinnen.

Der OLSA-Task wurde mithilfe des Skript-Editors in JavaScript bzw. HTML so erweitert, dass die OLSA-Stimuli abgespielt werden konnten. Dabei wurde berücksichtigt, dass die Listen und die Sätze innerhalb einer Liste randomisiert wiedergegeben wurden. Zusätzlich wurde der Abgleich der Probandeneingabe mit dem jeweils dargebotenen OLSA-Satz und die genutzte Pegelsteuerung des adaptiv angepassten Rauschpegels nach Brand und Kollmeier [17] implementiert.

Nach Abschluss der Messphase wurden alle auf der Gorilla-Plattform gespeicherten Ergebnisse von jedem Probanden, der die Messungen abgeschlossen hatte, von den Studienleiterinnen heruntergeladen. Für den OLSA lagen damit die Anzahl der richtig verstandenen Wörter sowie die dazugehörigen SNR-Werte für jeden dargebotenen Satz vor. Zur Schätzung der SRT- und Steigungswerte für jede dargebotene Testliste wurde eine Matlab-Implementierung der Maximum-Likelihood-Schätzung nach Brand und Kollmeier [17] genutzt, die auch im originalen OLSA verwendet wird. Für die Auswertung wurden, wie im originalen OLSA üblich, die Ergebnisse der dritten und damit letzten Messung verwendet, da für diese der Trainingseffekt als vernachlässigbar angesehen wird [10].

2.3 Studie

2.3.1 Teilnahmeoptionen

Die Teilnahme an der Studie erfolgte über den von Gorilla bereitgestellten Link. Da die Probanden zwischen 18 und 30 Jahren alt sein sollten, um den Probanden von Nuesse et al. [13] zu ähneln, wurde der Link überwiegend an Universitäten und Hochschulen verteilt.

2.3.2 Teilnahme

In einem Zeitraum vom 25.05.2021 bis 13.06.2021 wurde der Link insgesamt 182 Mal aufgerufen. 81 Mal wurde das Experiment bis zum Ende durchgeführt und 101 Mal abgebrochen. Davon erfolgten 89 Abbrüche aufgrund der Überschreitung des Zeitlimits von 2 Stunden. Das Zeitlimit wurde überschritten beim Lesen der allgemeinen Teilnahmeinformationen und der Einverständniserklärungen (N=48), bei der Erstellung des Codewortes (N=10), beim allgemeinen Anamnesefragebogen (N=1), beim Antiphasentest [6] zur Prüfung der Kopfhörernutzung (N=5), bei der Durchführung des OLSA (N=24) und nach Beendigung der OLSA-Messungen bei der Abschlussinformation (N=1). Zwölf weitere Ausschlüsse wurden erzeugt, da die Probanden die vorgegebenen Voraussetzungen (Alter, Muttersprache und Hörvermögen) nicht erfüllten.

2.3.3 Probanden

Die Altersspanne der 81 Probanden (47 weiblich, 33 männlich, 1 divers) lag zwischen 18 und 30 Jahren (Mit-

telwert: 23,7 Jahre). 91,0% der Probanden gaben an, keine Hörprobleme zu besitzen und 9,0% hatten sehr geringe Hörprobleme. 28,4% der Probanden beschrieben ihren allgemeinen Gesundheitszustand als „ausgezeichnet“, 54,3% als „sehr gut“ und 17,3% als „gut“.
Für den Vergleich der Ergebnisse wurden die Auswahlkriterien an die Kriterien von Nuesse et al. [13] angepasst. In diesen Vergleich wurden Ergebnisse von 69 Probanden (41 weiblich, 27 männlich, 1 divers) einbezogen, die angaben, den weiblichen OLSA nicht zu kennen. Diese Auswahl an Probanden war im Mittel 23,6 Jahre alt. 29,0% der Probanden beschrieben ihren allgemeinen Gesundheitszustand als „ausgezeichnet“, 53,6% als „sehr gut“ und 17,4% als „gut“. 89,9% der Probanden gaben an, keine Hörprobleme zu besitzen und 10,1% hatten sehr geringe Hörprobleme.

2.3.4 Statistische Analyse

Die statistische Analyse wurde mit Matlab von Mathworks (Version R2021a) durchgeführt und ein Signifikanzniveau von α=0,05 verwendet. Mit einem Shapiro-Wilk-Test wurden die SRT- und Steigungs-Werte auf Normalverteilung geprüft. Da die Daten nicht normalverteilt waren, wurden signifikante Unterschiede mit dem Mann-Whitney-U-Test untersucht. Um Unterschiede in den Verteilungen der Daten zu überprüfen, wurde ein F-Test angewendet.

3 Ergebnisse

3.1 Messdauer

Zur Auswertung der Messdauer wurden das Datum und die Uhrzeit des jeweils ersten und letzten Eintrags in den Ergebnistabellen der Probanden (N=81) genutzt (siehe Abbildung 2). Dabei zeigte sich, dass die Probanden im Median 32 min für die Messungen benötigten. Die kürzeste Messung dauerte 20 min, die längste 63 min.

3.2 SRT und Steigung

Abbildung 3a zeigt die SRT-Werte für alle Probanden, für die Probanden, die den weiblichen OLSA nicht kannten, und die Probandenergebnisse von Nuesse et al. [13]. Der mediane SRT liegt jeweils bei –9,1, –8,9 und –8,7 dB SNR. Die SRT-Werte der Probanden, die keinen weiblichen OLSA kannten, und diejenigen von Nuesse et al. [13] zeigen im Median nur geringe Unterschiede (siehe Abbildung 3a). Der Mann-Whitney-U-Test ergab, dass die SRT-Werte nicht signifikant unterschiedlich sind (U=3732, p=0,060). Die Streuung der online erhobenen SRT-Werte ist jedoch größer als diejenige von Nuesse et al. [13]. Ein F-Test bestätigt die signifikant andere Verteilung der SRT-Werte (p=0,002; F(68,47)=2,450).
Abbildung 3b zeigt die Steigung der Diskriminationsfunktionen gemessen mit allen Probanden, mit denjenigen, denen der weibliche OLSA unbekannt war, und die Ergebnisse von Nuesse et al. [13]. Die Steigungen liegen bei 16,7, 17,6 und 13,2%-Punkte/dB. Die Steigungen der Probanden, die den weiblichen OLSA nicht kannten und derjenigen von Nuesse et al. [13] unterscheiden sich signifikant (U=4995, p<0,001). Auch die Streuung der online ermittelten Steigungen ist größer als bei Nuesse et al. [13]. Ein F-Test bestätigt ebenfalls die signifikant andere Verteilung der Steigungs-Werte (p<0,001; F(68,47)=41,329).

4 Diskussion

4.1 Rekrutierung

Zur Rekrutierung der Probanden wurde der Link über private Kontakte, WhatsApp-Gruppen, Email-Verteiler, ein digitales schwarzes Brett und Social Media veröffentlicht. Trotz dieser großen Reichweite der Bekanntmachung des Links scheint die Resonanz und Teilnahmebereitschaft von 182 Aufrufen des Links in einem Zeitraum von 20 Tagen eher gering. Andere Alternativen der Rekrutierung könnten eine höhere Teilnehmerzahl in kürzerer Zeit im Vergleich zur Verteilung des Links an Universitäten und Hochschulen ermöglichen. Angebote von Probandendatenbanken wie z. B. Prolific oder MTurk (für eine Übersicht siehe [2]) könnten dabei eine Hilfe sein, jedoch muss der relativ geringe Anteil deutscher Muttersprachler in diesen Datenbanken berücksichtigt werden. Sie können in Gorilla eingebunden werden, sind aber i. d. R. kostenpflichtig. Die Datenbanken bieten einen gezielteren Kontakt zu Probanden an, die nach Kriterien gefiltert werden können. Die Probanden werden über diese Plattformen für ihren Aufwand entschädigt. Eine Aufwandsentschädigung könnte zusätzlich die Teilnahmebereitschaft erhöhen. Jedoch ist diese unter den gegebenen Datenschutzbestimmungen und evtl. Vorgaben der öffentlichen Träger oft schwierig umsetzbar.
Wird diese Teilnahmebereitschaft mit einer Laborstudie verglichen, so zeigt sich jedoch, dass in relativ kurzer Zeit (20 Tage) viele nutzbare Ergebnisse (N=81) erfasst werden konnten. Im Vergleich zur Laborsituation wurden die Ergebnisse ermittelt, ohne dass ein Laborraum mit Mess-Equipment genutzt wurde und die Messungen von einem Untersucher begleitet wurden. Eine Online-Studie ist somit kostengünstiger und zeitsparender als eine Labormessung [2].
Der Link zur Teilnahme am Online-Experiment wurde überwiegend an Universitäten und Hochschulen verteilt, damit die Probanden denen von Nuesse et al. [13] ähnelten. Es ist also davon auszugehen, dass überwiegend Studierende an der Untersuchung teilnahmen. Im Vergleich zu einer Laborstudie, an der i. d. R. nur Probanden teilnehmen, für die der Untersuchungsort erreichbar ist, konnten aber Studierende einer Hochschule an verschiedenen Studienorten angesprochen werden. Dies zeigt, dass verschiedene Auswahl-Faktoren die Diversität der Probanden einschränken aber auch wieder vergrößern können.

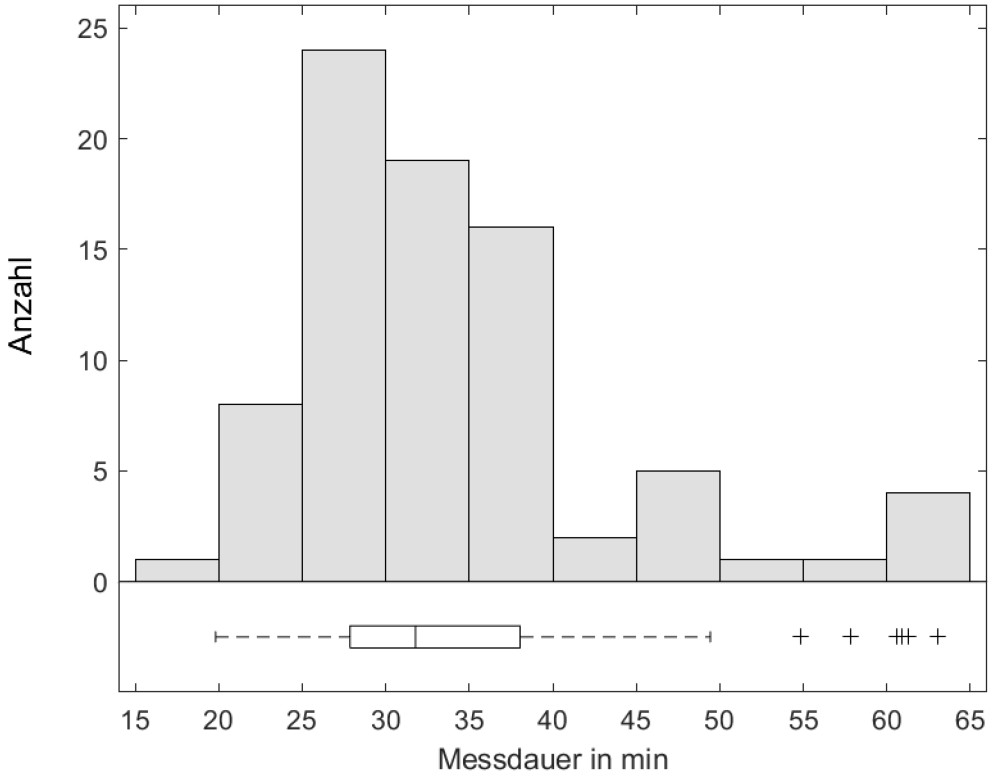

Abbildung 2: Dauer der Messungen, ermittelt für alle 81 Probanden; dargestellt ist sowohl die Häufigkeit der Dauer, gruppiert in 5-min-Abschnitten, als auch ein Boxplot zur Beschreibung der Daten.

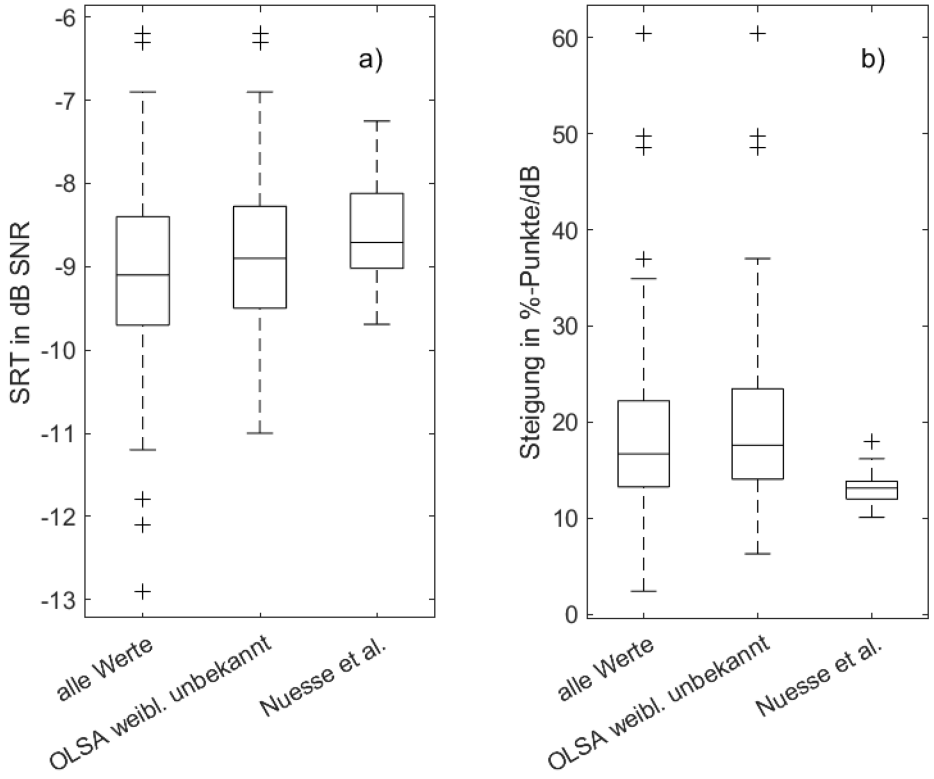

Abbildung 3: a) SRT- und b) Steigungs-Werte für alle Probanden, für die Probanden, denen der weibliche OLSA unbekannt war und die Ergebnisse von Nuesse et al. [13]

4.2 Abbrecherquoten und Messdauer

Insgesamt nutzten 182 Personen den Teilnahmelink, was darauf hinweist, dass die Zielgruppe für die Studie grundsätzlich erreicht wurde. Bei Betrachtung der Anzahl der Linkaufrufe scheint die Abbrecherzahl (*N*=101) verhältnismäßig hoch. Jedoch ist nach Reips [15] auch mit einer hohen Abbrecherquote zu rechnen. Die meisten Abbrüche kamen nicht aufgrund der Ausschlusskriterien zustande (*N*=12), sondern durch das Überschreiten der maximalen Messzeit von zwei Stunden (*N*=89). D. h., die Probanden hatten die Studie begonnen, aber nicht beendet. Bei vielen der Probanden wird die Zeitüberschreitung bereits bei der Einführung in das Experiment (Probandeninformation, Einverständniserklärung, Codeworterstellung und Kurzfragebogen) erreicht (*N*=59). Während des Lesens der ersten Informationen fühlten sich Probanden vermutlich nicht weiter angesprochen und setzten die Teilnahme nicht fort. Ebenfalls möglich ist, dass die Probanden ihre aktuelle Lebenssituation unpassend für das Experiment fanden und deshalb die Messung zu einem späteren Zeitpunkt fortsetzen wollten. Grundsätzlich waren mehrfache Durchführungen bzw. Abbrüche der Studie durch die Probanden möglich. Dies konnte aber bei der Aufnahme der Teilnehmer- bzw. Abbruchanzahlen und der Aufnahme der Daten nicht berücksichtigt werden. Das bedeutet, nur die begonnenen und die abgeschlossenen Experimente wurden gezählt. Ob ein Proband ein Experiment abbrach und zu einem späteren Zeitpunkt vollständig durchführte, kann aus den erhobenen Daten nicht nachvollzogen werden.
Weitere Probanden zeigten im Laufe der wesentlichen Untersuchungen (Antiphasentest und OLSA) eine Überschreitung des Zeitlimits (*N*=29). Ein möglicher Grund dafür kann die lange Durchführungsdauer des Experimentes gewesen sein. Nur etwa die Hälfte der Probanden schloss das Experiment innerhalb der empfohlenen Dauer von 30 min ab [2]. Andere Probanden benötigten länger. Um Trainingseffekte zu vermeiden, sind beim OLSA immer zwei Trainingslisten pro Termin am Anfang notwendig. Erst die dritte Liste kann als Ergebnis gewertet werden [8], [10]. Diese sich wiederholende Durchführung und die damit verbundene Monotonie des Ablaufes kann ebenfalls Abbrüche verursacht haben. Generell zeigt sich, dass die Anwendung des OLSA innerhalb der empfohlenen Messdauer eher schwierig ist. Wenn verschiedene Messsituationen miteinander verglichen werden sollen, ist durch das Training des Testes schnell die empfohlene Dauer erreicht. Sprachtests wie z. B. der Göttinger Satztest [18] oder der Ziffern-Tripel-Test [19], bei denen kein Training notwendig ist, könnten dann sinnvolle Alternativen sein.

4.3 Umsetzung des OLSA in Gorilla

Gorilla gibt über den Task Builder eine einfache Oberfläche zur Gestaltung eines Online-Experimentes vor. Jedoch reichten diese Bausteine nicht immer aus, um die komplexe Funktionalität eines OLSA online zu erreichen. Sie mussten durch Programmierungen in JavaScript ergänzt werden. Dies war z. B. der Fall für das Einlesen der Probandenantworten sowie ihr Abgleich mit den jeweils korrekten Antworten. Ebenso musste die adaptive Pegelsteuerung in Gorilla programmiert werden. Die hier gewählte Umsetzung des OLSA speicherte als Ergebnis in Gorilla nur die Verständlichkeit und den dargebotenen SNR für jeden Satz. Eine Berechnung des SRT mithilfe der Maximum-Likelihood-Methode [17] erfolgte nach dem Download der Ergebnisse in Matlab. Eine direkte Bestimmung des SRT im Online-Experiment wäre zwar grundsätzlich möglich gewesen, jedoch wurde aufgrund des Aufwands darauf verzichtet. Somit konnte der SRT nicht im weiteren Verlauf eines Online-Experimentes als Referenz genutzt werden, um z. B. einen bestimmten SNR darzubieten oder eine Auswahl von Probanden vorzunehmen.
Darüber hinaus konnten die Audiodateien des Satzmaterials nicht als verlustfreie wav-Dateien in Gorilla hinterlegt werden, sondern mussten in ein mp3-Format konvertiert werden. Nach Brandenburg [20] nimmt das mp3-Format eine Datenreduktion auf Grundlage der Psychoakustik vor. Das bedeutet, dass vor allem Signalanteile reduziert werden, die nicht wahrnehmbar sind. Dabei ist die verwendete Bitrate entscheidend für die Audioqualität. Je höher diese ist, desto besser ist die Audioqualität. Brandenburg beschreibt für mp3, dass bei einer Bitrate von z. B. 128 kbit/s für ein Stereosignal bei 48 kHz eine effiziente Kompression erreicht ist, bei der die Audioqualität hoch und die Datengröße gering sind. Größere Bitraten führen im Vergleich dazu nur zu geringfügigen Verbesserungen der Audioqualität bei größeren Dateien. Da in dieser Studie eine Bitrate von 320 kbit/s verwendet wurde, sind keine Verständlichkeitsunterschiede zwischen den Datenformaten für die hier getesteten jungen Probanden mit normalem Gehör zu erwarten. Außerdem wurden die Signale mit einem zusätzlichen Rauschen dargeboten, dessen energetische Maskierung vermutlich ausschlaggebend für die Verständlichkeit der Sprache ist.

4.4 Durchführung des Online-Experimentes

Grundsätzliche Hürden eines Online-Experimentes zeigen sich auch bei dieser Studie. Das von den Probanden genutzte Equipment und dessen Einstellung bleibt unbekannt. Ebenso kann nicht die Richtigkeit der Antworten, das Verständnis für die Aufgabe, die Motivation und die Aufmerksamkeit der Probanden überprüft, beobachtet oder direkt beeinflusst werden und z. B. noch einmal nachgefragt, erklärt oder Pausen vorgeschlagen werden. Die Hörfähigkeiten können nur über einen Fragebogen eingeschätzt werden und nicht durch ein Tonaudiogramm in einer Hörkabine mit kalibriertem Equipment geprüft werden. An dieser Studie nahmen Probanden teil, die subjektiv normalhörend waren oder einen sehr schwachen Hörverlust angaben. Von v. Gablenz et al. [21] wurde bei einem Vergleich von subjektiven Bewertungen mit dem nach den Kriterien der World Health Organisation beurteilten Hörverlust festgestellt, dass junge Probanden

ihre Hörprobleme eher überschätzten, während ältere Probanden ihre Probleme beim Hören eher unterschätzen. In dieser Studie und bei Nuesse et al. [13] nahmen junge Probanden teil. Bei Nuesse et al. [13] betrugen die Hörschwellen der Probanden bei maximal zwei Frequenzen 15 dB HL, was laut der World Health Organisation als normalhörend angesehen werden kann [22]. Die hier ermittelten SRT-Werte sind mit Nuesse et al. [13] vergleichbar. Nach Wagener und Brand [23] ändern sich die SRT-Werte im OLSA für Probanden mit normalem und beeinträchtigtem Gehör nicht, wenn sich der Präsentationspegel verändert. Personen mit einer Hörbeeinträchtigung zeigen bei unterschiedlichen Präsentationspegeln immer einen höheren (schlechteren) SRT-Wert als Normalhörende. Somit veranschaulichen die mit Nuesse et al. [13] vergleichbaren SRT-Werte dieser Studie, dass der Hörstatus für das Sprachverstehen der Probanden sehr ähnlich war. Die Selbstbewertung des Hörvermögens in dieser Studie scheinen die Probanden damit überwiegend zuverlässig vorgenommen zu haben.

4.5 SRT und Steigung

Der Vergleich der ermittelten Ergebnisse mit der Studie von Nuesse et al. [13] zeigt, dass der OLSA auch in einer Online-Studie zur Ermittlung des SRT genutzt werden kann. Die SRT-Werte beider Studien unterscheiden sich nur geringfügig und nicht signifikant voneinander. Die geringen medianen Unterschiede der SRT-Werte von 0,2 dB sind möglicherweise durch die Aufgabenstellung verursacht worden. Bei Nuesse et al. [13] wurde der OLSA als offene Variante durchgeführt. Im Gegensatz dazu sprachen die Probanden dieser Studie die verstandenen Wörter nicht laut nach, sondern nutzen eine geschlossene Antworteingabe und mussten die Wörter in einer Antwortmatrix auswählen. Holube et al. [24] stellten eine ähnliche Verbesserung der SRT-Werte um 0,2 dB beim Vergleich der geschlossenen und offenen Variante fest. Die geschlossene Variante ermöglicht den Probanden Rückschlüsse auf mögliche Antworten, die bei einer offenen Durchführung fehlen.
Außerdem unterschieden sich die Studien in der Art der Darbietung des Sprachmaterials. In Nuesse et al. [13] wurden drei feste SNR-Werte (−11,0, −8,5 und −6,0 dB SNR) dargeboten und anschließend eine Diskriminationsfunktion angepasst. Im Gegensatz dazu liegt den über Gorilla ermittelten Daten eine adaptive Steuerung des SNRs während jeder OLSA-Messung zu Grunde, sodass eine Änderung des SNR-Wertes bei jedem einzelnen Satz stattfand. Dies erfolgte nach der adaptiven Steuerung von Brand und Kollmeier [17]. Das adaptive Verfahren verwendete eine große Schrittweite zu Beginn einer Liste, um möglichst schnell SNR-Werte nahe des SRT darzubieten. Danach wurde die Schrittweite schnell verringert, um in geringer Entfernung um den SRT zu schwanken. Die so ermittelten Verständlichkeiten und SNR-Werte wurden dann in der Maximum-Likelihood-Methode [17] verwendet, um eine Diskriminationsfunktion sowie deren SRT und Steigung zu schätzen. Zeigen die Verständlichkeitswerte in Abhängigkeit vom SNR innerhalb einer Liste eine stetig steigende oder fallende Tendenz und schwanken sie nicht, wie es mit dem adaptiven Verfahren gewünscht ist, um den SRT, so kann es zu einer Über- oder Unterschätzung des SRT und der Steigung kommen. Dies kann auch in dieser Studie als Ursache für die hohe Streuung und die deutlichen Ausreißer gesehen werden. Vermutlich wurde die Aufmerksamkeit der Probanden zu einem frühen Zeitpunkt innerhalb einer Liste abgelenkt. Die resultierenden falschen Antworten führten zu einer schnellen Verringerung der Schrittweite, obwohl SNR-Werte nahe des SRT noch nicht erreicht wurden und die oben beschriebenen Verläufe entstanden. Wie bereits vermerkt, sind im Online-Experiment die Möglichkeiten der Ablenkung groß und schwer kontrollierbar und auch die Motivation des Probanden kann nicht überprüft werden.
Trotz der erhöhten Streuung der Daten und der vermuteten geringeren Aufmerksamkeit bzw. größeren Diversität der Probanden aber besonders der Teilnahmebedingungen (z. B. Kopfhörer, Soundkarten, Rechner, Browser, Messumgebung) ist es dennoch gelungen, vergleichbare SRT-Werte zu den im Labor gemessenen Werten zu erreichen. Die Diversität erhöht aber auch die Möglichkeit, die Ergebnisse auf realistischere Messsituationen im häuslichen Umfeld zu verallgemeinern.

5 Schlussfolgerungen

Grundsätzlich ist die Durchführung eines Sprachtests und die Ermittlung eines SRT in einer Online-Studie möglich. Berücksichtigt werden muss allerdings bei der Umsetzung des Experimentes, dass der Messaufbau und die Messumgebung unbekannt sind. Dies hat Folgen für die Auswahl des Tests, da nur relative Pegeländerungen als Ergebnisse aufgenommen werden können und die Messungen mit einem zusätzlichen Rauschen durchgeführt werden sollten, damit Hintergrundgeräusche überdeckt werden können. Grundsätzlich ist bei einer Online-Studie mit einer aufwändigeren Rekrutierung zu rechnen, da es zu höheren Abbrecherquoten kommt. Nichtsdestotrotz sind Online-Studien im Vergleich zu Labormessungen kostengünstiger und zeitsparender. Die Diversität der Probanden, des Equipments und der Messumgebung zeigt sich auch in der Varianz der Daten. Außerdem muss die Motivation der Probanden möglichst gleichbleibend sichergestellt werden, da sie nicht überprüft werden kann. Dies könnte durch kürzere Messdauern und abwechslungsreiche Aufgaben erreicht werden.

Anmerkungen

Interessenskonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Funding

Diese Studie wurde im Rahmen der Hochschullehre an der Jade Hochschule in Oldenburg durchgeführt. Sie war Bestandteil des Moduls Projektpraktikum 2 im 6. Semester des Studiengangs Hörtechnik und Audiologie. Die Bezahlung der erhaltenen Datensätze an Gorilla wurde aus einem Fonds der Jade Hochschule für diese Praktika vorgenommen.

Danksagung

Wir danken den Probanden für die Teilnahme am Online-Experiment, sowie Theresa Nüsse und Bianka Wiercinski für die Bereitstellung der Daten aus der Veröffentlichung [13].

Hinweis

Aus Gründen der besseren Lesbarkeit wird im Text verallgemeinernd das generische Maskulinum verwendet. Diese Formulierungen umfassen gleichermaßen weibliche, männliche und diverse Personen; alle sind damit selbstverständlich gleichberechtigt angesprochen.

ORCID der Autorin

Anne Schlüter: 0009-0006-1062-2702

Literatur

1. Bundesministerium für Gesundheit. Chronik zum Coronavirus SARS-CoV-2. Berlin: BMG; 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html
2. Sauter M, Draschkow D, Mack W. Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. Brain Sci. 2020 Apr;10(4):. DOI: 10.3390/brainsci10040251
3. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: An online behavioral experiment builder. Behav Res Methods. 2020 Feb;52(1):388-407. DOI: 10.3758/s13428-019-01237-x
4. Gorilla. Gorilla Experiment Builder: Gorilla Support, Due Diligence. 2023 [updated 2023 Feb 20; cited 2023 Feb 20]. Available from: https://support.gorilla.sc/support/due-diligence#overview
5. Gorilla. Gorilla Experiment Builder: Gorilla's Support Documentation. 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://support.gorilla.sc/support/
6. Woods KJP, Siegel MH, Traer J, McDermott JH. Headphone screening to facilitate web-based auditory experiments. Atten Percept Psychophys. 2017 Oct;79(7):2064-72. DOI: 10.3758/s13414-017-1361-2
7. Wagener K, Brand T, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(2):44-56.
8. Wagener K, Brand T, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(3):86-95.
9. Wagener K, Kühnel V, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(1):4-15.
10. Schlueter A, Lemke U, Kollmeier B, Holube I. Normal and Time-Compressed Speech: How Does Learning Affect Speech Recognition Thresholds in Noise? Trends in Hearing. 2016;20(1). DOI: 10.1177/2331216516669889
11. Wagener KC, Hochmuth S, Ahrlich M, Zokoll MA, Kollmeier B. Der weibliche Oldenburger Satztest. In: Deutsche Gesellschaft für Audiologie, ed. Abstracts der 17 Jahrestagung der Deutschen Gesellschaft für Audiologie; 2014. Abrufbar unter/Available from: https://www.dga-ev.com/fileadmin/daten/downloads/bisherige_Jahrestagung/dga2014_programm_final.pdf
12. Ahrlich M. Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit [Bachelorarbeit]. Oldenburg: Carl von Ossietzky Universität Oldenburg; 2013.
13. Nuesse T, Wiercinski B, Brand T, Holube I. Measuring Speech Recognition With a Matrix Test Using Synthetic Speech. Trends in Hearing. 2019;23. DOI: 10.1177/2331216519862982
14. Nuesse T, Wiercinski B, Holube I. Synthetic German matrix speech test material created with a text-to-speech system. Zenodo; 2021 [cited 2022 Jul 1]. DOI: 10.5281/zenodo.4501212
15. Reips UD. The Web Experiment Method. In: Birnbaum MH, editor. Psychological experiments on the internet. San Diego: Academic Pr; 2000. p. 89-117. DOI: 10.1016/B978-012099980-4/50005-8
16. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspect Psychol Sci. 2011 Jan;6(1):3-5. DOI: 10.1177/1745691610393980
17. Brand T, Kollmeier B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J Acoust Soc Am. 2002 Jun;111(6):2801-10. DOI: 10.1121/1.1479152
18. Kollmeier B, Wesselkamp M. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. J Acoust Soc Am. 1997 Oct;102(4):2412-21. DOI: 10.1121/1.419624
19. Van den Borre E, Denys S, van Wieringen A, Wouters J. The digit triplet test: a scoping review. Int J Audiol. 2021 Dec;60(12):946-63. DOI: 10.1080/14992027.2021.1902579
20. Brandenburg K. MP3 and AAC Explained. In: Audio Engineering Society, editor. Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding; 1999.
21. von Gablenz P, Otto-Sobotka F, Holube I. Adjusting Expectations: Hearing Abilities in a Population-Based Sample Using an SSQ Short Form. Trends Hear. 2018;22. DOI: 10.1177/2331216518784837

22. World Health Organization. Deafness and hearing loss. 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

23. Wagener KC, Brand T. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. Int J Audiol. 2005 Mar;44(3):144-56. DOI: 10.1080/14992020500057517

24. Holube I, Blab S, Fürsen K, Gürtler S, Taesler S. Einfluss des Maskierers und der Testmethode auf die Sprachverständlichkeit von jüngeren und älteren Normalhörenden. Zeitschrift für Audiologie (Audiological Acoustics). 2009;48(3):120-7.

Korrespondenzadresse:
Anne Schlüter
Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg, Deutschland
anne.schlueter@jade-hs.de

Bitte zitieren als
Schlüter A, Baumann ES, Ben Ghorbal F, Hauenschild N, Kleinow A, Mazur V, Thomas J, Holube I. Sprachtests im Online-Experiment am Beispiel des Oldenburger Satztests. GMS Z Audiol (Audiol Acoust). 2024;6:Doc05.
DOI: 10.3205/zaud000040, URN: urn:nbn:de:0183-zaud0000401

Artikel online frei zugänglich unter
https://doi.org/10.3205/zaud000040

Veröffentlicht: *16.04.2024*

OPEN ACCESS

This is the English version.
The German version starts at p. 1.

Research Article

Speech tests in online experiments using the example of the Oldenburg sentence test

Abstract

During the COVID-19 pandemic, audiological tests in the laboratory were only possible under very strict hygiene measures or not at all. To circumvent this obstacle, an online experiment was implemented using the synthetic speech material of the Oldenburg Sentence Test (OLSA). The aim was to test whether comparable results could be achieved online with synthetic speech as used in a previous experiment in the laboratory. With the help of the platform Gorilla (https://gorilla.sc/), an online experiment was created and shared through a link via various communication channels (e.g., in social media). This link was accessed 182 times, resulting in 81 complete data sets. The speech-recognition thresholds established for 69 participants were comparable to the results from the laboratory situation. However, the slopes of the discrimination functions differed significantly. In principle, it was shown that the synthetic OLSA is also feasible as an online experiment. Compared to the laboratory measurement, greater hurdles were encountered in the recruitment and motivation of the test participants.

Keywords: OLSA, online experiment, SRT, synthetic speech

Anne Schlüter[1]
Elisabeth-Sophie Baumann[1]
Fatma Ben Ghorbal[1]
Nele Hauenschild[1]
Alina Kleinow[1]
Vanessa Mazur[1]
Julia Thomas[1]
Inga Holube[1]

1 Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

1 Introduction

In scientific studies, most audiological measurements, such as speech-recognition tests, are carried out in laboratories or specially designed, acoustically optimised rooms that are equipped with the necessary measuring equipment and setups. During the COVID-19 pandemic and the resulting hygiene and protective measures, measurements in the laboratory were difficult or even impossible to realise [1]. In order to still be able to carry out tests, measurements conducted online at home offer a possible alternative. This raises questions as to whether certain audiological tests, such as a speech test, can also be implemented online and thus carried out at home, and whether the results are comparable with laboratory measurements.

According to Sauter et al. [2] the following steps are necessary for the development of an online study: a) programming the experiment, e.g., for use with a browser, b) setting up a suitable test environment, e.g., using an online host on a server, and c) recruiting the participants. Gorilla (https://gorilla.sc/) provides an online platform for the implementation of online behavioural experiments in a browser [3]. It allows users to implement their own experiments via simple graphical user interfaces or by creating JavaScript code. Gorilla provides a cloud-based server and server management that also fulfil the requirements of the General Data Protection Regulation [4]. For example, the basic functionality of an audiological measurement, the presentation and modification of audio signals, and the recording of the participant's reaction to these signals, is possible via a browser. The experiments can be made accessible to potential participants, for example via a link, or the participants can be recruited via recruitment portals such as Prolific or MTurk (for an overview, see [2], [5]). Once the experiments have been completed, the data collected can be accessed via a database, downloaded, and further analysed.

Various aspects limit the selection of the speech test and its programming for an online experiment. In principle, there are no options for influencing the measurement equipment, such as the headphones and sound card used. With an antiphase test [6], the general use of headphones can be ensured, but an exact calibration of the measurement setup is not possible. Therefore, the online experiment can only ensure audibility and determine relative level changes as results. The influence of possible background noise can only be minimised by presenting the signals at a level clearly above threshold. In view of these aspects, a test in background noise that determines a speech-recognition threshold (SRT) is therefore suitable as an online speech test. The SRT describes the signal-to-noise ratio (SNR) at which 50% of the speech is correctly recognised. In an online test, it is also easier to offer closed response options, i.e., a selection of all possible responses, than to allow free re-

sponses. The response entered in the closed test can be converted directly into recognition scores and used further in the progress of the measurement, e.g., for adaptive control. Alternatives in an open test, such as audio recordings or text input during the online experiment, would first have to be interpreted as correct or incorrect by speech or text recognisers to convert them into recognition scores. Errors can occur during interpretation, as it is not possible for the examiner to ask questions as in the laboratory. To enable a closed implementation for an online experiment, clearly limited speech material is required. At the same time, it must be noted that the limited speech material itself cannot be learnt, or the learning effect can be monitored.

One speech test that fulfils these criteria is the Oldenburg Sentence Test (OLSA, [7], [8], [9]). Its sentences have a standardised syntactic structure: name, verb, number, adjective, object. Ten words are available in each word group, which can be combined in different ways. This results in sentences that are grammatically correct, but do not always make sense (e.g., "Peter buys ten wet armchairs."). The associated noise was generated by randomly superpositioning the speech material, and therefore has the same long-term spectrum as the speech. The OLSA sentences in the noise are presented to the participants, and their task is to repeat the recognised words or to select them in the input mask. Depending on the words recognised, the SNR is adaptively changed to determine the SRT. The limited speech material makes it possible for the OLSA to be carried out as a closed test, and the response behaviour of the participants can be documented word by word. Due to the small number of words per word group, a learning effect can occur [7], [8]. Therefore, the speech material must be trained with two lists before the measurement [10].

For the original version of the OLSA, the sentences were recorded by a male speaker [7], [8], [9]. As a further variant, the speech material was recorded with a female speaker [11], [12]. Nuesse et al. [13] synthesised the speech material with a female voice using a text-to-speech system, and compared it with the recordings of the natural female speaker in the laboratory. Forty-eight participants (mean age: 21.8 years) took part in these measurements. They had normal hearing, spoke German as their native language, and were unfamiliar with the OLSA. Nuesse et al. [13] showed that a similar level of speech recognition was achieved with the synthesised female speaker as with the natural female speaker. The synthesised speech material is freely available for research purposes [14]. It was therefore used for the online experiment.

There are advantages and disadvantages to recruiting for and conducting online studies. Reips [15] describes the possibility of recruiting a larger number of participants in a shorter time. The sample generally shows greater diversity, and is therefore more similar to the general population, compared to participants recruited for laboratory measurements. This diversity is also reflected in the measuring equipment used. As the participants carry out the measurements independently, the measurements are cost-effective and free from errors that could be generated by the investigator. However, the lack of personal contact also means that the participants' motivation and response behaviour cannot be observed. Multiple participation or the participation of bots is possible. In addition, with increased duration of the online study, the participants' willingness to participate and their concentration decreases (e.g., [16], [2]). A study length of more than 30 minutes is feasible, but can have a demotivating effect [2]. In principle also, higher dropout rates must be expected than in laboratory studies [15].

To answer the fundamental question as to whether speech tests can be used online at home, an online experiment with the OLSA was created and conducted via the Gorilla platform. Aspects that describe the boundary conditions of the experiment, such as the number of participants, dropout rates, and completion times, are reported. Based on the hypothesis that there are no differences, the SRT values and slopes determined in the online experiment were compared with those measured in the laboratory by Nuesse et al. [13].

2 Methods

2.1 Measuring environment and equipment

The participants were able to access the online experiment via a link provided by Gorilla at any time from any location with a tablet or computer (laptop and desktop) that allowed internet access. Gorilla offers various options for restricting the framework of an experiment. For example, the medium through which an online experiment can be carried out by the participant can be selected. Tablets or computers were permitted for this experiment. To support a stable and reliable measurement environment, mobile phones were excluded, as it was assumed that they would be used more frequently in rapidly changing situations than computers or tablets. Browser type, location, and connection speed were not restricted. In addition, the participants were asked to wear headphones instead of using loudspeakers during the experiment. This was also intended to support the stability and reliability of the measurement situation.

2.2 Creation of the online experiment

2.2.1 Setup of the experiment

Online experiments on the Gorilla platform [5] have a modular structure. Various elements, such as Questionnaires and Tasks, are generated separately for application by the participants. The components of the Questionnaires are suitable, for example, for constructing forms or integrating scales and other types of questionnaires. Tasks are created in the Task Builder. This allows the preparation of various graphical interfaces in which participants' tasks are integrated. The Spreadsheet in the

Task Builder enables defining the visualisation sequence of the graphical interfaces, as well as the results structure and stimuli (e.g., images, audio files or videos).
The general procedure of the experiment used here is shown in Figure 1. The experiment was started with a Questionnaire, the first page of which welcomed the participants, and informed them about the procedure and the basic requirements. In addition, the approximate measurement duration of 30 minutes was indicated, the use of headphones was recommended, and a link was provided to access detailed information about participation. On further pages, contact information was displayed, consent to voluntariness and anonymity was obtained, information on data protection was provided, and consent to participation was requested. In order to pseudonymise the results, a Task was used to create a personal code word.
In the further progress of the experiment, the participants were asked to complete a questionnaire about their personal medical history. General data, such as age, gender, native language, and information regarding general health, a possible hearing problem, and familiarity with the OLSA with a female speaker were recorded here. The health status was requested to make a subsequent selection of the data and to exclude any participants who were ill, and for whom a possible influence on the measurement result due to illness was unclear. The familiarity with the female OLSA was explicitly tested to make a subsequent selection of the data, and at the same time not to exclude too many data sets due to the participants' general knowledge of the OLSA. If the participants stated that they had identified a hearing problem, they were directed to a further questionnaire. This questionnaire asked about asymmetrical hearing (question: Do you hear equally well with both ears?, response options: Yes, No, Don't know) and the degree of hearing loss (question: Please describe the degree of your hearing loss (in the better ear), response options: very mild, mild, moderate, severe, very severe). Only people between the ages of 18 and 30 years, whose native language was German and who had no or very mild hearing loss were allowed to take part in the study. If the information provided did not meet the specified criteria for the study, the participants were informed of their exclusion and the experiment was automatically terminated. Otherwise, they were able to continue with the experiment.
This was followed by a Task to set a comfortable presentation level for the stimuli (speech and noise). The level was set using the main volume control for the audio output of the user's own computer or tablet. First, the main volume was set to 0 and then the level of the presented noise was adjusted by the participant to a comfortable presentation level using the main volume control. This setting should then be readjusted by the participant if the subsequently presented speech signal was not understood. Finally, the participant confirmed the adjusted presentation level. This setting of the main volume control of the device (tablet or computer) was the basis for the further experiment. It could not be read out or changed in Gorilla. Clipping of the signals in the online experiment was not to be expected, as only participants with normal hearing were examined, and the level of pleasant loudness was adjusted in Gorilla by changing the SNR in the OLSA only.
After setting the presentation level, a Task checked the use of headphones with the help of an antiphase test [6] available in Gorilla. Three pure tones were presented in stereo, one of which had a 180° phase shift on one channel. The participants had to select the quietest signal. Due to phase cancellation, this task should be difficult when using loudspeakers. Woods et al. [6] were thus able to reliably distinguish between headphone and loudspeaker users. After performing the antiphase test, the experiment was terminated for participants who did not use headphones. The last step of the experiment was the Task that included the OLSA (see section 2.2.2).
In Gorilla, the processing time for this online experiment was limited to 2 hours, after which it was automatically terminated. This limitation was necessary to complete the data collection and to cancel experiments that had been started but never continued. After completion of each Task and each Questionnaire, the corresponding data was automatically saved by Gorilla in tables. However, the measurement data could only be retrieved for experiments that had been fully completed by the participants. A fee had to be paid to Gorilla for each completed experiment [5].

2.2.2 OLSA in Gorilla

For the measurement of OLSA in Gorilla, the speech material generated with a synthesised female voice was used together with the corresponding noise [13]. Thirty sentences were part of each list. A total of 45 lists were implemented in Gorilla as mp3 files (bit rate when converting from wav to mp3 format with a maximum possible quality: 320 kbits/s, conversion carried out using the website: https://online-audio-converter.com/de/). These correspond to the lists from Nuesse et al. [13]. The participants had the task of selecting the words they understood from a response matrix (10x5) displayed on the screen. During the measurement, a fixed speech level was presented and the noise level was adaptively adjusted, depending on the participant's response. The speech level was kept constant once the participant achieved intelligible speech in the presentation lesvel setting. Adaptive adjustment is also used in the original OLSA [7], [8], [11] and was developed by Brand and Kollmeier [17]. The SNR of the first sentence of a list was always 0 dB. The participants performed the OLSA three times in succession. The first two lists were training lists [7], [8], [9], [10].
Four interfaces were created in the Task Structure for the OLSA. The Fullscreen function was built into the first interface. This automatically displayed the user's screen in full-screen mode at the beginning, with the intention of making it easier for the participants to concentrate on the speech test. A text window followed, with detailed in-

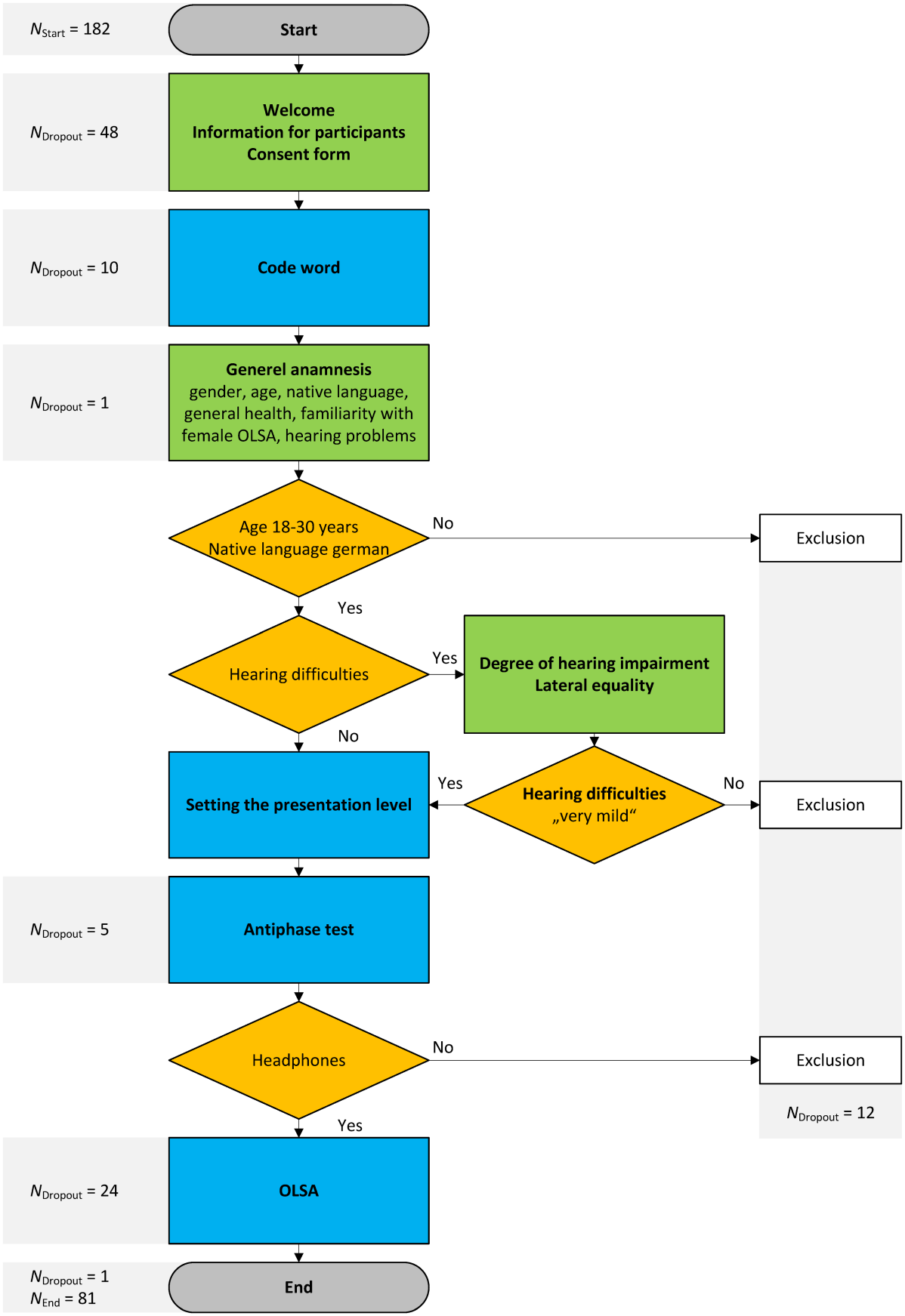

Figure 1: Flowchart of the online experiment created in Gorilla to measure the OLSA. Green processes were implemented as Questionnaires in Gorilla. Blue processes were designed as Tasks. Orange sections were decisions. The number (*N*) displays the number of experiments started, of experiments cancelled within a section, of exclusions, and of experiments completed at the end.

structions about the OLSA. This included the aim of the measurement, the exact procedure, a description of the stimuli, the duration and number of sentences or sentence lists, and the participant's exact task. It was also pointed out that the speech test intentionally takes place under difficult conditions, and that pauses are possible at any time.
After starting the measurement, an interface followed in which the start of a sentence list or the announcement of the next measurement run was embedded. When the "Next measurement"-button was pressed, the response interface was activated. It initially showed a white screen for the duration of the sentence presentation, so that the participant's attention was focussed on recognising the sentence and the display of the response matrix and a progress bar did not distract. Only after the sentence presentation was the response matrix for entering the recognised parts of the sentence displayed and a progress bar shown. This interface was linked to a Spreadsheet that enabled the display of the response options, and also took up the sequence and frequency of the previously created interfaces. The spreadsheet was also used to determine the result structure for each participant. Among other things, the SNR values and the number of correctly recognised words per sentence were included in the table. Then the next sentence was started with the "Next"-button. The participants were not given any feedback on their recognition of the words. The last interface of the OLSA contained a comment field for the people who took part in the experiment, and the contact information of the lead investigators.
The OLSA task was extended using the Script-Editor in JavaScript or HTML, so that the OLSA stimuli could be presented. The fact that the lists and the sentences within a list were randomised was taken into account. In addition, the comparison of the participant's input with each OLSA sentence presented, and the level control used for the adaptively adjusted noise level according to Brand and Kollmeier [17] was implemented.
At the end of the measurement phase, all the results stored on the Gorilla platform were downloaded by the lead investigators for each participant who had completed the measurements. The number of correctly recognised words and the corresponding SNR values for each sentence presented were thus available for the OLSA. To estimate the SRT and slope values for each test list presented, a Matlab implementation of the maximum-likelihood estimation according to Brand and Kollmeier [17] was applied, which is also used in the original OLSA. As in the original OLSA, the results of the third and therefore last measurement were used for the evaluation, as the training effect is considered negligible for this measurement [10].

2.3 Study

2.3.1 Participation options

People participated in the study via the link provided by Gorilla. The link was mainly distributed within universities to match the participants of Nuesse et al. [13] between 18 and 30 years old.

2.3.2 Participation

In the period from 25/05/2021 to 13/06/2021, the link was accessed 182 times. The experiment was completed 81 times and cancelled 101 times. Of these, 89 were cancelled because the time limit of 2 hours had been exceeded. The time limit was exceeded when reading the general information about participation and the consent forms (N=48), when creating the code word (N=10), during the general medical history questionnaire (N=1), during the antiphase test [6] to check headphone use (N=5), when carrying out the OLSA (N=24), and after completion of the OLSA measurements during the final information (N=1). Twelve further exclusions were generated because the participants did not fulfil the specified requirements (age, native language, and hearing ability).

2.3.3 Participants

The age range of the 81 participants (47 female, 33 male, 1 diverse) was between 18 and 30 years (mean: 23.7 years). Of these, 91.0% stated that they had no hearing problems and 9.0% had very minor hearing problems. Their general state of health was described by 28.4% as "excellent", by 54.3% as "very good" and by 17.3% as "good".
For the comparison of the results, the selection criteria were adapted to the criteria of Nuesse et al. [13]. Results from 69 participants (41 female, 27 male, 1 diverse) who stated that they did not know the female OLSA were included in this comparison. They were on average 23.6 years old, and 29.0% described their general state of health as "excellent", 53.6% as "very good" and 17.4% as "good". 89.9% of the participants stated that they had no, and 10.1% had very minor hearing problems.

2.3.4 Statistical analysis

The statistical analysis was carried out using Matlab from Mathworks (version R2021a), and a significance level of α=0.05 was used. A Shapiro-Wilk test was used to test the SRT and slope values for normal distribution. As the data were not normally distributed, significant differences were analysed using the Mann-Whitney U test. An F-test was used to examine differences in the distribution of the data.

3 Results

3.1 Measurement duration

The date and time of the first and last entry in the participants' result tables (N=81) were used to analyse the measurement duration (see Figure 2). The participants needed a median of 32 min for the measurements, the shortest took 20 min, the longest 63 min.

3.2 SRT and slope

Figure 3a shows the SRT values for all participants, for the participants for whom the female OLSA was unknown, and the participants' results from Nuesse et al. [13]. The median SRT were –9.1, –8.9, and –8.7 dB SNR, respectively. The SRT values of the participants who were unfamiliar with the female OLSA and those of Nuesse et al. [13] show only small median differences (see Figure 3a). The Mann-Whitney U-test revealed that the SRT values were not significantly different (U=3732, p=0.060). However, the variance of the SRT values collected online was greater than that of Nuesse et al. [13]. An F-test confirmed this (p=0.002; F(68,47)=2.450).
Figure 3b shows the slope of the discrimination functions measured from all participants, those to whom the female OLSA was unknown, and the results of Nuesse et al. [13]. The slopes were 16.7, 17.6, and 13.2 percentage points/dB, respectively. The slopes of the participants who were unfamiliar with the female OLSA and those of Nuesse et al. [12] differ significantly (U=4995, p<0.001). As also confirmed by an F-test, the variance of the slopes established online was also greater than in Nuesse et al. [13] (p< 0.001; F(68,47)=41.329).

4 Discussion

4.1 Recruitment

To recruit the participants, the link was published via private contacts, WhatsApp groups, email distribution lists, a digital noticeboard, and through social media. Despite this wide reach of publication, the response and willingness to participate of 182 link views in a period of 20 days seems rather low. Compared to the distribution of the link at universities, other recruitment alternatives might enable a higher number of participants in a shorter time. Offers from participant/recruitment databases such as Prolific or MTurk (for an overview see [2]) could be of help here, but the relatively low proportion of German native speakers in these databases must be taken into account. They can be integrated into Gorilla, but are usually fee-based. The databases offer more targeted contact with participants, who can be filtered according to criteria. The participants are compensated for their efforts via these platforms. An expense allowance could also increase the willingness to participate. Nevertheless, this is often difficult to implement under the given data protection regulations and possible requirements of public organisations.
However, if this willingness to participate is compared with a laboratory study, it is apparent that many usable results (N=81) were recorded in a relatively short time (20 days). In comparison to the laboratory situation, the results were obtained without the need of a laboratory room with measuring equipment and the support of an investigator. An online study is therefore more cost-effective and time-saving than a laboratory measurement [2]. The link to participate in the online experiment was mainly distributed within universities, so that the participants were similar to those of Nuesse et al. [13]. It can therefore be assumed that mainly students took part in the study. However, in comparison to a laboratory study, in which only participants take part who could reach the study location, students from a university with different study locations could be approached. This shows that various selection factors can limit the diversity of the participants, but can also increase it.

4.2 Dropout rates and measurement duration

A total of 182 persons used the participation link, which indicates that the target group for the study was generally reached. When looking at the number of link views, the dropout rate (N=101) seems relatively high. However, according to Reips [15], a high dropout rate is to be expected. Most dropouts were not due to the exclusion criteria (N=12), but because the maximum measurement time of two hours was exceeded (N=89). This means that the participants had started the study but not completed it. For many of the participants (N=59), the time out was already reached during the introduction to the experiment (participant information, declaration of consent, code word creation, and short questionnaire). During the reading of the initial information, participants presumably did not feel further addressed and did not continue their participation. It is also possible that the participants found their current life situation unsuitable for the experiment, and therefore decided to continue the measurement later. In principle, it was possible for the participants to execute or discontinue the study several times. However, this could not be taken into account when recording the number of participants and dropouts and when recording the data. This means that only started and completed experiments were counted. It is not possible to determine from the data collected whether a participant cancelled an experiment and then completed it later.
During the course of the main tests (antiphase test and OLSA), other participants exceeded the time limit (N=29). One possible reason for this may have been the long duration of the experiment. Only about half of the participants completed the experiment within the recommended duration of 30 min [2]. Other participants needed longer. To avoid training effects, the OLSA always requires two training lists per session at the beginning, and only the third list can be counted as a result [8], [10]. This re-

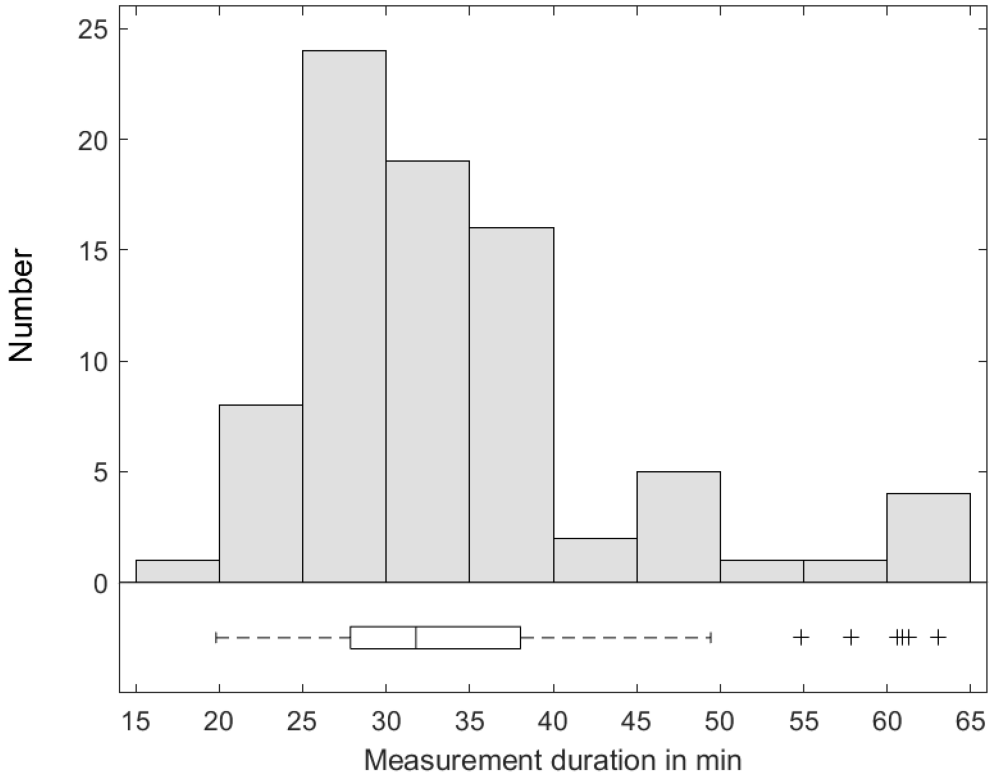

Figure 2: Duration of the measurements for all 81 participants; both the frequency of the duration grouped in 5-minute sections and a box plot describing the data are shown.

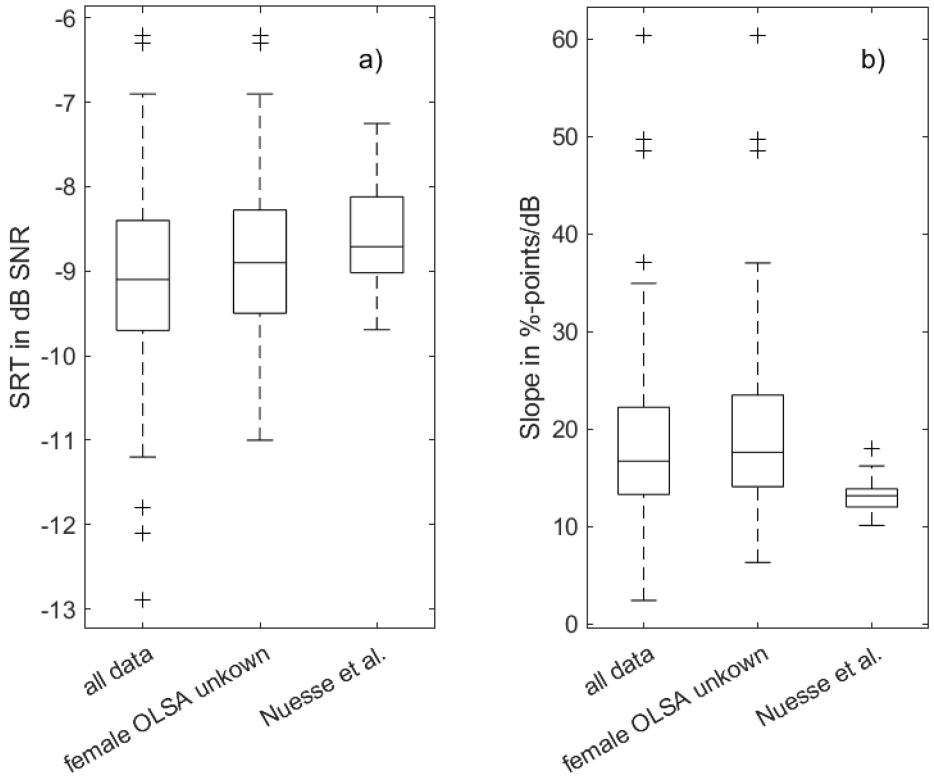

Figure 3: a) SRT and b) slope values for all participants, for the participants for whom the female OLSA was unknown, and the results of Nuesse et al. [13]

petitive execution and the associated monotony of the process may also have caused dropouts. In general, the application of OLSA within the recommended measurement duration is rather difficult. If different test situations are to be compared with each other, the recommended duration is quickly reached by training the test. In that case, speech tests such as the Göttingen sentence test [18] or the digit triple test [19], which do not require training, could be useful alternatives.

4.3 Implementation of the OLSA in Gorilla

Gorilla provides a simple interface for designing an online experiment via the Task Builder. However, these modules were not always sufficient to achieve the complex functionality of an OLSA online, and had to be supplemented by programming in JavaScript. This was the case, for example, for reading in the participants' responses and comparing them with the correct responses. The adaptive level control also had to be programmed in Gorilla. The OLSA implementation chosen here only saved the recognition score and the SNR presented for each sentence as a result in Gorilla. A calculation of the SRT using the maximum likelihood method [17] was performed after downloading the results in Matlab. Although in principle it would have been possible to determine the SRT directly in the online experiment, this was not done due to the effort involved. This meant that the SRT could not be used as a reference in the further progress of an online experiment, e.g., to present a specific SNR or to select participants.

In addition, the audio files of the sentence material could not be stored in Gorilla as lossless wav files, but had to be converted to mp3 format. According to Brandenburg [20], the mp3 format performs data reduction on the basis of psychoacoustics. This means that signal components that are not perceptible are reduced. The bit rate used is decisive for the audio quality. The higher it is, the better the audio quality. Brandenburg describes for mp3 that an efficient compression is achieved at a bit rate of, e.g., 128 kbit/s for a stereo signal at 48 kHz, where the audio quality is high and the data size is small. In comparison, higher bit rates only lead to minor improvements in audio quality for larger files. As a bit rate of 320 kbit/s was used in this study, no differences in recognition between the data formats are to be expected for the young participants with normal hearing, as tested here. Furthermore, the signals were presented with additional noise, the energetic masking of which is presumably decisive for the recognition of the speech.

4.4 Realisation of the online experiment

The fundamental hurdles of an online experiment are also evident in this study. The equipment used by the participants and its adjustment remain unknown. It is also not possible to check, observe, or directly influence the correctness of the responses, the understanding of the task, the motivation and the attention of the participants, and, for example, to repeat questions, explain, or suggest pauses. Hearing abilities can only be assessed using a questionnaire and cannot be tested using a pure-tone audiogram in an acoustic booth with calibrated equipment. This study involved participants who were subjectively normal-hearing or reported a very mild hearing loss. When comparing subjective assessments with hearing loss assessed according to World Health Organisation criteria, v. Gablenz et al. [21] found that young participants tended to overestimate, while older participants tended to underestimate their hearing problems. The current study and Nuesse et al. [13] involved young participants. In Nuesse et al. [13], the hearing thresholds of the participants were 15 dB HL at a maximum of two frequencies, which can be considered normal hearing according to the World Health Organisation [22]. The SRT values determined here are in agreement with Nuesse et al. [13]. According to Wagener and Brand [23] the SRT values in the OLSA for participants with normal and impaired hearing do not change when the presentation level changes. People with a hearing impairment always show a higher (worse) SRT value at different presentation levels than people with normal hearing. Thus, the results obtained with Nuesse et al. [13] illustrate that the hearing status of the participants for speech recognition was very similar. The self-assessment of hearing ability in this study therefore appears to have been carried out reliably by the participants most of the time.

4.5 SRT and slope

The comparison of the results obtained with the study by Nuesse et al. [13] shows that the OLSA can also be used in an online study to determine the SRT. The SRT values of both studies differ only slightly, and not significantly, from each other. The small median differences in the SRT values of 0.2 dB may have been caused by the task. In Nuesse et al. [13], the OLSA was performed as an open variant. In contrast, the participants in this study did not repeat the recognised words aloud, but used a closed response setting and had to select the words in a response matrix. Holube et al. [24] found a similar improvement in SRT values of 0.2 dB when comparing the closed and open variants. The closed variant allows the participants to draw conclusions about possible responses that are missing in an open variant.

The studies also differed in how the speech material was presented. In Nuesse et al. [13], three fixed SNR values (−11.0, −8.5, and −6.0 dB SNR) were presented and then a discrimination function was fitted. In contrast, the data obtained via Gorilla is based on an adaptive change in SNR during each OLSA measurement, so that a change in the SNR value occurred for each individual sentence. This was carried out according to the adaptive control of Brand and Kollmeier [17]. To present SNR values close to the SRT as quickly as possible, the adaptive method used a large step size at the beginning of a list. The step

size was then quickly reduced to fluctuate closely around the SRT. The recognition scores and SNR values assessed in this way were then used in the maximum likelihood method [17] to estimate a discrimination function and its SRT and slope. If the recognition scores show a steadily increasing or decreasing trend depending on the SNR within a list, and do not fluctuate around the SRT as is desired with the adaptive method, this can lead to an over- or underestimation of the SRT and the slope. This might also have been the cause of the high variance and the clear outliers in this study. Presumably, the participants' attention was diverted at an early stage within a list. The resulting incorrect responses led to a rapid reduction in the step size, although SNR values close to the SRT were not yet reached, and the curves described above emerged. As already noted, the possibilities of distraction in the online experiment are large and difficult to control, and the participant's motivation cannot be checked.

Despite the increased variance of the data and the presumed lower level of attention and greater diversity of the participants and especially the conditions of participation (e.g., headphones, sound cards, computer, browser, measurement environment), it was still possible to achieve comparable SRT values to those measured in the laboratory. However, the diversity also increases the possibility of generalising the results to more realistic measurement situations in the home environment.

5 Conclusion

It is clearly possible to carry out a speech test and measure an SRT in an online study. However, when realising the experiment, the measurement setup and the measurement environment are unknown, and this must be taken into account. This has consequences for the selection of the test, as only relative level changes can be recorded as results, and the measurements should be carried out with additional noise, so that background noise can be masked. In principle, more extensive recruitment is to be expected in an online study, as there are higher dropout rates. Nevertheless, online studies are more cost-effective and time-saving compared to laboratory measurements. The diversity of the participants, the equipment, and the measurement environment is also reflected in the variance of the data. In addition, the motivation of the participants must be ensured as consistently as possible, as it cannot be verified. This could be achieved through shorter measurement durations and varied tasks.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

This study was conducted as part of university teaching at the Jade University of Applied Sciences in Oldenburg. It was part of the module Practical course 2 in the 6th semester of the course of study Hearing Technology and Audiology. Payment for the data sets received by Gorilla was made from a Jade University fund for these modules.

Acknowledgement

We would like to thank the participants for taking part in the online experiment and Theresa Nüsse and Bianka Wiercinski for providing the data from their publication [13]. Language support was provided by Scientific and Technical English Language Services (desmosa@gmx.de).

Author's ORCID

Anne Schlüter: 0009-0006-1062-2702

References

1. Bundesministerium für Gesundheit. Chronik zum Coronavirus SARS-CoV-2. Berlin: BMG; 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html
2. Sauter M, Draschkow D, Mack W. Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. Brain Sci. 2020 Apr;10(4):. DOI: 10.3390/brainsci10040251
3. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: An online behavioral experiment builder. Behav Res Methods. 2020 Feb;52(1):388-407. DOI: 10.3758/s13428-019-01237-x
4. Gorilla. Gorilla Experiment Builder: Gorilla Support, Due Diligence. 2023 [updated 2023 Feb 20; cited 2023 Feb 20]. Available from: https://support.gorilla.sc/support/due-diligence#overview
5. Gorilla. Gorilla Experiment Builder: Gorilla's Support Documentation. 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://support.gorilla.sc/support/
6. Woods KJP, Siegel MH, Traer J, McDermott JH. Headphone screening to facilitate web-based auditory experiments. Atten Percept Psychophys. 2017 Oct;79(7):2064-72. DOI: 10.3758/s13414-017-1361-2
7. Wagener K, Brand T, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(2):44-56.
8. Wagener K, Brand T, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(3):86-95.
9. Wagener K, Kühnel V, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Zeitschrift für Audiologie (Audiological Acoustics). 1999;38(1):4-15.

10. Schlueter A, Lemke U, Kollmeier B, Holube I. Normal and Time-Compressed Speech: How Does Learning Affect Speech Recognition Thresholds in Noise? Trends in Hearing. 2016;20(1). DOI: 10.1177/2331216516669889

11. Wagener KC, Hochmuth S, Ahrlich M, Zokoll MA, Kollmeier B. Der weibliche Oldenburger Satztest. In: Deutsche Gesellschaft für Audiologie, ed. Abstracts der 17 Jahrestagung der Deutschen Gesellschaft für Audiologie; 2014. Abrufbar unter/Available from: https://www.dga-ev.com/fileadmin/daten/downloads/bisherige_Jahrestagung/dga2014_programm_final.pdf

12. Ahrlich M. Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit [Bachelorarbeit]. Oldenburg: Carl von Ossietzky Universität Oldenburg; 2013.

13. Nuesse T, Wiercinski B, Brand T, Holube I. Measuring Speech Recognition With a Matrix Test Using Synthetic Speech. Trends in Hearing. 2019;23. DOI: 10.1177/2331216519862982

14. Nuesse T, Wiercinski B, Holube I. Synthetic German matrix speech test material created with a text-to-speech system. Zenodo; 2021 [cited 2022 Jul 1]. DOI: 10.5281/zenodo.4501212

15. Reips UD. The Web Experiment Method. In: Birnbaum MH, editor. Psychological experiments on the internet. San Diego: Academic Pr; 2000. p. 89-117. DOI: 10.1016/B978-012099980-4/50005-8

16. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspect Psychol Sci. 2011 Jan;6(1):3-5. DOI: 10.1177/1745691610393980

17. Brand T, Kollmeier B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J Acoust Soc Am. 2002 Jun;111(6):2801-10. DOI: 10.1121/1.1479152

18. Kollmeier B, Wesselkamp M. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. J Acoust Soc Am. 1997 Oct;102(4):2412-21. DOI: 10.1121/1.419624

19. Van den Borre E, Denys S, van Wieringen A, Wouters J. The digit triplet test: a scoping review. Int J Audiol. 2021 Dec;60(12):946-63. DOI: 10.1080/14992027.2021.1902579

20. Brandenburg K. MP3 and AAC Explained. In: Audio Engineering Society, editor. Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding; 1999.

21. von Gablenz P, Otto-Sobotka F, Holube I. Adjusting Expectations: Hearing Abilities in a Population-Based Sample Using an SSQ Short Form. Trends Hear. 2018;22. DOI: 10.1177/2331216518784837

22. World Health Organization. Deafness and hearing loss. 2022 [updated 2022 May 18; cited 2022 May 18]. Available from: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

23. Wagener KC, Brand T. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. Int J Audiol. 2005 Mar;44(3):144-56. DOI: 10.1080/14992020500057517

24. Holube I, Blab S, Fürsen K, Gürtler S, Taesler S. Einfluss des Maskierers und der Testmethode auf die Sprachverständlichkeit von jüngeren und älteren Normalhörenden. Zeitschrift für Audiologie (Audiological Acoustics). 2009;48(3):120-7.

Corresponding author:
Anne Schlüter
Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany
anne.schlueter@jade-hs.de

Please cite as
Schlüter A, Baumann ES, Ben Ghorbal F, Hauenschild N, Kleinow A, Mazur V, Thomas J, Holube I. Sprachtests im Online-Experiment am Beispiel des Oldenburger Satztests. GMS Z Audiol (Audiol Acoust). 2024;6:Doc05.
DOI: 10.3205/zaud000040, URN: urn:nbn:de:0183-zaud0000401

This article is freely available from
https://doi.org/10.3205/zaud000040

Published: *2024-04-16*