




# Difficulty Indices for 34 Texts in the Heard Text Recall (HTR) Paradigm: Measuring Text Comprehension and Memory

Chinthusa Mohanathanas<sup>1</sup> , Isabel S. Schiller<sup>1</sup>  & Sabine J. Schlittmeier<sup>1</sup> 

<sup>1</sup> Work and Engineering Psychology, Institute of Psychology, RWTH Aachen University, Germany

Corresponding Author: [sabine.schlittmeier@psych.rwth-aachen.de](mailto:sabine.schlittmeier@psych.rwth-aachen.de)

**Abstract:** This publication presents difficulty indices for 34 texts and corresponding questions (Set\_2) of the Heard Text Recall (HTR) paradigm as introduced by Schlittmeier et al. (2023). In the HTR, participants listen to short texts and answer nine open-ended questions after each text presentation, allowing for measuring comprehension and memory across various experimental conditions in a within-subject design. The questions are categorized into three types, asking for either names, family relationships, or other factual details. Data from 13 experiments involving 429 participants were analyzed to determine difficulty indices for each text (with each text tested a different number of times), each question type, and individual questions. These difficulty indices enable researchers to ensure a balanced level of HTR-task difficulty, which is particularly important for drawing meaningful comparisons between different experimental conditions. This publication serves as an initial reference, and we encourage researchers to contribute further datasets.

## 1 Introduction

The Heard Text Recall (HTR) paradigm (Schlittmeier et al., 2023) provides a cognitive and experimental psychological approach for assessing a listener's text comprehension and memory for coherent, running speech. In the HTR, participants are required to listen to brief texts and respond to nine open-ended questions following each text. The HTR enables a systematic within-subject examination of factors affecting comprehension and memory across varied experimental conditions—for example, background noise versus quiet listening conditions, a speaker's hoarse versus normal voice quality, and audio-only versus audiovisual presentation. Information on HTR-text difficulty is essential for designing experiments that minimize confounding variables and enhance the validity of results. By grouping texts of comparable difficulty, researchers can create balanced experimental conditions, ensuring that participants are exposed to texts that are equally challenging across all conditions. This enables a valid attribution of differences in experimental outcomes between conditions to the experimental manipulations, rather than to inherent variations in text difficulty, strengthening the overall reliability of the study's conclusions.

To achieve such balance in experimental design, standardized text materials with known difficulty indices are essential. The initial work by Schlittmeier et al. (2023) provided two sets (Set\_1 and Set\_2) of German texts with open-ended recall questions, standardized in terms of topic (family stories), length (120–131 words), and syntactic structure (e.g., each text includes 10 sentences with no more than one subordinate clause per sentence). Despite these parallelization efforts, texts vary in difficulty. Difficulty indices for Set\_1 were already provided in the initial publication (Schlittmeier et al., 2023).

In this publication, we provide difficulty indices for Set\_2, which consists of 34 texts, each accompanied by nine open-ended questions. Background on the Heard Text Recall (HTR) and Read Text Recall (RTR) paradigms is provided for context (see Schlittmeier et al., 2023, for more detailed information). Each text describes a family, detailing family members, their relationships, and additional aspects like residence, hobbies, and occupations. To minimize learning effects, each text is based on a unique family structure with distinct names and details. Family narratives were selected as they allow for the creation of multiple, largely parallel texts—an approach that would be challenging with scientific or other topics. Additionally, standard family structures (grandparents, parents, children) offer minimal variation in partici-



pants' prior knowledge and interest. In the HTR paradigm, correct answers to open-ended questions often require the integration and processing of information across multiple sentences. This is primarily achieved through cross-sentence questions (termed indirect questions in the original database), which cannot be answered based solely on information from a single sentence. Further, the questions require free recall responses (1-2 words, avoiding yes or no answers). The open-ended questions ask for relationships and other factual details.

The difficulty indices presented here for the 34 texts of Set\_2 are based on data from 13 experiments (see Ehret et al., 2023, 2024, 2025, under review; Mohanathasan et al., 2024, 2025, under review; Schiller et al., 2023, 2024; unpublished data from bachelor thesis, 2022). They assess difficulty at multiple levels: overall text difficulty (i.e., based on all nine questions), by question type (names, relationships, factual details), and at the level of individual questions.

Please note that we consider the present difficulty indices as an initial step, with updates planned as additional experiments are currently running and more data will become available. This approach allows the difficulty indices to become increasingly accurate and more informative for diverse research needs, like adaptive testing (cp. Section 3.2).

## 2 Calculating Difficulty Indices

The statistical analysis for calculating difficulty indices was conducted using R (version 4.1.2; R Core Team, 2023). Data were sourced from experiments utilizing the texts and questions in Set\_2, with all texts presented auditorily and thus exclusively within the HTR paradigm. A total of 13 experiments contributed data to the difficulty analysis (see Ehret et al., 2023, 2024a, 2024b, 2025, under review; Mohanathasan et al., 2024, 2025, under review; Schiller et al., 2023, 2024; Mofti, 2022), with each experiment including a subset of the 34 texts from Set\_2. The experiments used either audio recordings from the AuViST database (Ermert et al., 2023) or researcher-generated recordings (Schiller et al., 2024).

Table 1 in the appendix presents detailed information on participant numbers, descriptive results, texts used as well as descriptions and references for each experiment. In all experiments used for the difficulty analysis, each text was followed by a consistent set of nine open-ended questions. Each correct answer was coded as 1, while incorrect answers or missing responses were coded as 0. The difficulty index represents the percentage of correct answers and was calculated as

$$\text{difficulty index} = \frac{n_r}{n} \times 100 \quad (1)$$

where  $n_r$  represents the number of correct answers and  $n$  is the total number of responses at the respective level of analysis (cf. Kelava & Moosbrugger, 2020).

Since responses in the HTR paradigm are collected at multiple levels, difficulty indices were calculated separately for

- (a) each text (based on all nine questions combined),
- (b) each question group (1 = names, 2 = relationships, 3 = other facts), and
- (c) each individual question.

As different experiments used different subsets of texts (see Table 1 in the appendix), the total number of responses  $n$  varies across texts and questions. The resulting difficulty indices express the difficulty level as a percentage, with higher values indicating an easier task. As a proportion of correct responses, difficulty indices range from 0% (no correct answers) to 100% (all answers correct).

## 3 Selecting Texts Based on Difficulty Indices

A key question in using difficulty indices is determining the most appropriate difficulty range for a performance test designed to compare experimental conditions, such as the HTR paradigm.

Following classical test theory, we can expect an inverted U-shaped relationship between the difficulty of test items, such as the texts or the questions in the HTR, and their ability to differentiate between conditions. In traditional ability tests (e.g., school exams), medium-difficulty items often best differentiate between individuals with high and low performance levels (cf. e.g., Schmidt-Atzert, Krumm & Amelang, 2021). However, the goal of the HTR paradigm is not to assess individual ability, but to compare cognitive performance across different experimental conditions. This means that although one might assume that selecting texts (or question groups, or individual questions in the HTR) of medium difficulty (e.g., 50%  $\pm$  10%) is ideal, this is not always the case (see, e.g., Priebe, 2024), and specifically not in the HTR paradigm. Here, several considerations justify a shifted selection strategy that favors a range of difficulty indices from 50–70% or even 60–80%.

The most important reason for recommending such a shift toward slightly easier texts is that even our student samples, who generally perform well in cognitive tests, report finding the HTR both challenging and mentally fatiguing. If texts are too difficult (i.e., if correct recall rates, and thus difficulty indices, are very low), this can lead to

frustration, disengagement, guessing, and dropout effects as an experimental session continues, particularly under challenging testing conditions, like listening in noise. However, texts that are too easy pose problems regarding reliability. If a text is so simple that even under a challenging experimental condition (e.g., background noise), participants can still maintain (near-to) perfect performance, this leaves no room to detect differences between experimental conditions (ceiling effect). Thus, presenting challenging but not excessively difficult or easy texts is an effective way to enhance and sustain participant engagement, ensuring the reliable measurement of actual performance (rather than, for example, frustration levels), and allowing for differences between conditions to be detected and meaningfully interpreted. It should be noted that very easy texts are particularly well suited for practice or concealed exercise trials (cf. Schlittmeier et al., 2023), where the primary goal is to familiarize participants with the task demands and response format.

#### 4 Exemplary applications of difficulty indices in experiments employing the HTR paradigm

The following section presents exemplary considerations and uses cases for applying difficulty indices to refine experimental design and improve text assignment across experimental conditions, participant groups, or individual participants within the HTR paradigm.

##### 1. Selecting texts for more than two within-subject conditions

In the initial publication (Schlittmeier et al., 2023), a within-subject design was introduced as effective for capturing text comprehension and recall performance across conditions. This design included a practice block followed by two experimental blocks (A and B), each containing a concealed exercise text and six test texts (see Figure 1 in Schlittmeier et al., 2023). The inclusion of Set\_2 and its difficulty indices enables precise matching of text difficulty between blocks, allowing for refined comparisons. As previously recommended, the text sequence within each block should be balanced, and assignments to conditions counterbalanced across participants. With the expanded Set\_2 stimulus material, researchers can now implement additional conditions within a within-subject design.

Implementing more than two conditions, however, requires careful planning to manage participant fatigue and ensure optimal measurement accuracy. Strategies

to address participant fatigue include reducing the number of texts per block or condition, omitting the concealed exercise text from each block, or scheduling participants for multiple testing sessions. However, reducing the number of texts per condition may decrease reliability due to fewer measurement points. Thus, balancing reliability (which improves with more texts) against potential fatigue is essential. Using multiple texts per condition is therefore recommended, as averaging performance across texts per condition yields more stable measures and reduces fluctuations that might obscure experimental effects.

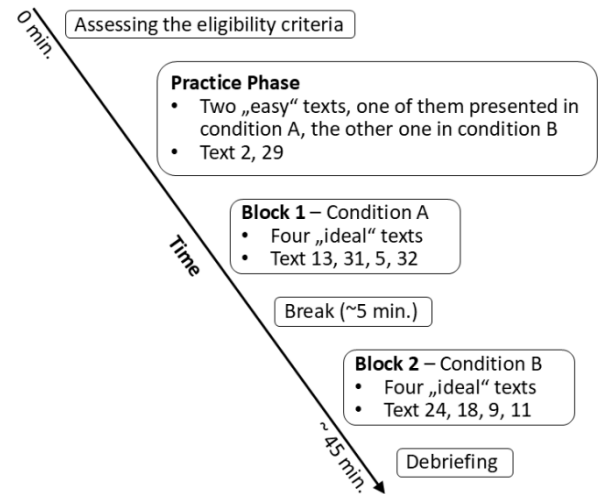


Figure 1: Example of an experimental design assessing the impact of two listening conditions (A and B) on memory performance. A practice phase familiarizes participants with the task and the two listening conditions. Here, two texts with difficulty indices indicating “easy” texts were chosen. The two subsequent experimental blocks feature texts of moderate difficulty, i.e. “ideal” for detecting condition effects. Text order and block order are balanced across participants.

##### 2. Selecting texts for parallel or adaptive testing

When a between-subject design is required, such as in studies comparing independent groups, difficulty indices enable the compilation of parallel test versions with comparable difficulty levels. In such designs, this approach also reduces the risk that observed differences are due to variations in text difficulty rather than to experimental manipulations. Additionally, difficulty indices support adaptive testing. At the group level, lower-performing groups can receive easier texts to achieve comparable performance levels and accurately assess treatment effects. At the individual level, the assessment of comprehension and recall capacities can start with more difficult texts for higher-performing participants, reducing the number of texts needed.

## 5 Considerations and current data limitations

The following considerations are critical for interpreting the difficulty indices provided in this publication. They address limitations related to baseline conditions, data density, and potential modality effects, each of which should be considered when evaluating the reliability and generalizability of these indices.

### 1. Inclusion of diverse experimental conditions

Difficulty indices were derived from data collected under a variety of auditory conditions and tasks (see Table 1 in the appendix), including both typical and altered speakers' voices (e.g., hoarseness) and, in some experiments, concurrent secondary tasks to assess listening effort. These variations mean that some texts were tested under more demanding conditions, which may result in overestimating their difficulty compared to simpler experimental settings. However, we chose to include all available data to provide a comprehensive resource for varied experimental conditions. Consequently, each data point in the difficulty analysis is annotated with the specific experimental condition under which it was collected, enabling researchers to perform custom difficulty analyses as needed. While the difficulty indices provided here serve as general guidelines, they should be interpreted with the understanding that systematic shifts in difficulty indices may occur for specific texts depending on their testing conditions.

### 2. Data density across texts

A limitation of the present difficulty indices is the variability in data density across texts. Differences in the number of data points per text, resulting from varied experimental usage, affect the stability and reliability of these values. Texts with fewer data points may yield less stable difficulty estimates, while frequently used texts provide more robust indices. To support informed text selection, the difficulty analysis includes the number of data points per text ( $n$ ), enabling researchers to assess the stability of difficulty estimates and make informed selection when precise estimations are critical.

### 3. Generalizability of difficulty indices

It is important to note that the present difficulty indices were derived from student samples, who typically perform well in cognitive tasks and may not represent other target populations. Consequently, these indices may not be fully generalizable to groups with different levels of familiarity with cognitive testing or varying cognitive or auditory-perceptive abilities.

### 4. Modality effects and generalizing from HTR paradigm

The difficulty indices presented in this publication are based exclusively on experiments where texts were presented auditorily, utilizing the Heard Text Recall (HTR) paradigm. However, comprehension and recall outcomes – and thus difficulty indices – may differ if texts are presented in the Read Text Recall (RTR) paradigm. Listening requires participants to process the text sequentially, within a time-limited, non-repeatable, and not self-paced format, relying on a single exposure to the material. In contrast, reading allows participants to adjust their pace and direct their gaze back to specific information as needed. Given these modality-specific processing demands, caution is advised when applying the present difficulty indices to studies employing visual text presentation, since comprehension and recall outcomes may differ by modality.

## 6 Outlook

Addressing the identified limitations in difficulty estimation requires an ongoing, collaborative effort to expand and refine the dataset. To enhance data quality and reliability, we plan to increase data density across all texts by systematically incorporating new experimental results. This will enhance the stability and reliability of difficulty indices, particularly for those texts that currently rely on limited data points.

Furthermore, expanding difficulty analysis to include text recall data from both auditory (HTR) and visual (RTR) modalities would allow for a comprehensive comparison of modality-specific difficulty indices. Such additions would support modality-based adjustments in text or question selection, if necessary, making the difficulty analysis a more versatile resource for diverse experimental designs.

To achieve these improvements, we invite researchers to contribute their data on text recall and comprehension, collected under various experimental conditions. Submitted datasets should adhere to a standardized format and include detailed information (see supplementary templates for guidance). By pooling data from different research groups, we can enhance the robustness and comprehensiveness of the database, ensuring reliable difficulty indices for specific task demands, modality effects, and baseline conditions. This collective effort aims to provide difficulty indices across all texts and a wide variety of experimental conditions, facilitating effective use of Set\_2 of the HTR paradigm (Schlittmeier et al., 2023) in a broad range of research contexts.

We encourage researchers with relevant datasets to contact the corresponding author to contribute to the next DIF update.

## 7 Files

The difficulty analysis for Set\_2 of Schlittmeier et al. (2023) can be downloaded from <https://doi.org/10.18154/RWTH-2025-05207> and includes:

- A .txt file containing citation instructions (citation-instruction.txt).
- An .xlsx file showing difficulty indices, with difficulty levels highlighted: (a) for each text (Sheet 1), (b) for each text and question group (Sheet 2), and (c) for each text and individual question (Sheet 3; Set-2\_difficulty-indices.xlsx).
- An .xlsx file containing all collected and analyzed data (data.xlsx) and a .txt file explaining the data columns (data\_read\_me.txt).
- A .pdf file containing detailed information on the experiment, including participant numbers, descriptive results, texts used, and a description with references for each experiment (Table 1.pdf).
- The R script used for analysis, allowing for further examination (difficulty\_index.R).

### Acknowledgments

The authors would like to thank Jonathan Ehret for sharing detailed information about his experiments and providing the respective data. We also thank all student assistants who supported data collection for the studies included in this difficulty analysis.

### Funding

The contributions of all authors to this work were supported by a grant from the HEAD-Genuit-Stiftung (HEAD-Genuit-Foundation: P-16/10-W).

### ORCID

Chinthusa Mohanathanasan

 <https://orcid.org/0000-0001-6916-1425>

Isabel S. Schiller

 <https://orcid.org/0000-0003-2387-7625>

Sabine J. Schlittmeier

 <https://orcid.org/0000-0001-9051-4547>

### References

Ehret, J., Bönsch, A., Nossol, P., Ermert, C.A., Mohanathanasan, C., Schlittmeier, S.J., Fels, J., & Kuhlen, T.W. (2023). Who's next? Integrating Non-Verbal

Turn-Taking Cues for Embodied Conversational Agents. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 1–8.

<https://doi.org/10.1145/3570945.3607312>

Ehret, J., Bönsch, A., Schiller, I. S., Breuer, C., Aspöck, L., Fels, J., Schlittmeier, S. J., & Kuhlen, T. W. (2024b). Audiovisual coherence: Is embodiment of background noise sources a necessity? *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 61–67.

<https://doi.org/10.1109/VRW62533.2024.00017>

Ehret, J., Dasbach, V., Hartmann, N., Fels, J., Kuhlen, T.W. & Bönsch, A., (2025). Exploring Gaze Dynamics: Initial Findings on the Role of Listening Bystanders in Conversational Interactions. *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 748–752.

<https://doi.org/10.1109/VRW66409.2025.00151>

Ehret, J., Schüppen, J., Mohanathanasan, C., Ermert, C. A., Fels, J., Schlittmeier, S. J., Kuhlen, T. W., & Bönsch, A. (under review). Objectifying Social Presence: Evaluating Degraded Speech Performance in ECAs Using the Heard Text Recall Paradigm. *IEEE Transactions on Visualization and Computer Graphics*.

Ermert, C. A., Mohanathanasan, C., Ehret, J., Schlittmeier, S. J., Kuhlen, T. & Fels, J., (2023) AuViST – An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm. *RWTH Publications*. <https://doi.org/10.18154/RWTH-2023-05543>

Kelava, A., & Moosbrugger, H. (2020). Deskriptivstatistische Itemanalyse und Testwertbestimmung [Descriptive item analysis and test score determination.]. *Testtheorie und Fragebogenkonstruktion*, 143–158.

Krumm, S., Schmidt-Atzert, L., & Amelang, M. (2021). Grundlagen diagnostischer Verfahren [Fundamentals of diagnostic procedures]. In L. Schmidt-Atzert, S. Krumm, & M. Amelang (Hrsg.), *Psychologische Diagnostik* (6. Aufl., S. 39–208). Springer. <https://doi.org/10.1007/978-3-662-61643-7>.

Mofti, M. (2022). Vergleich zweier Aufgaben zur Erfassung der Erinnerungsleistung unter akustisch aversiven Zuhörbedingungen [Comparison of two tasks for assessing memory performance under acoustically aversive listening conditions]. *Unpublished Bachelor Thesis*. RWTH Aachen University.

Mohanathanasan, C., Fels, J. & Schlittmeier, S.J. (2024) Listening to two-talker conversations in quiet settings: the role of listeners' cognitive processing

- capabilities for memory and listening effort. *Scientific Report*, 14, 22764.  
<https://doi.org/10.1038/s41598-024-74085-1>
- Mohanathasan, C., Ermert, C. A., Fels, J., Kuhlen, T. & Schlittmeier, S. J. (2025). Exploring short-term memory and listening effort in two-talker conversations: The influence of soft and moderate background noise. *PLoS ONE* 20(2): e0318821.  
<https://doi.org/10.1371/journal.pone.0318821>
- Mohanathasan, C., Koleva, P.B., Ehret, J., Bönsch, A., Fels, J., Kuhlen, T. & Schlittmeier, S. J. (under review). Beyond words: The impact of static and animated faces as visual cues on memory performance and listening effort during two-talker conversations.
- Priebe, J. A. (2024). Testkonstruktion – das Herz der psychologischen Diagnostik: Von Gütekriterien, lokaler stochastischer Unabhängigkeit, Normen und dem IQ [Test construction – The heart of psychological assessment: On quality criteria, local stochastic independence, norms, and IQ]. Springer. <https://doi.org/10.1007/978-3-662-67547-2>.
- R Core Team. (2023). *R: a language and environment for statistical computing.*: Bd. R Foundation for Statistical Computing.
- Schiller, I. S., Breuer, C., Aspöck, L., Ehret, J., Bönsch, A., Kuhlen, T. W., ... & Schlittmeier, S. J. (2024). A lecturer's voice quality and its effect on memory, listening effort, and perception in a VR environment. *Scientific Reports*, 14(1), 12407.  
<https://doi.org/10.1038/s41598-024-63097-6>
- Schiller, I. S., Aspöck, L., & Schlittmeier, S.J. (2023). The impact of a speaker's voice quality on auditory perception and cognition: a behavioral and subjective approach. *Frontiers in Psychology*, 14, 1243249.  
<https://doi.org/10.3389/fpsyg.2023.1243249>
- Schlittmeier, S.J., Mohanathasan, C., Schiller, S. & Liebl, A. (2023). Measuring text comprehension and memory: A comprehensive database for Heard Text Recall (HTR) and Read Text Recall (RTR) paradigms, with optional note-taking and graphical displays. *RWTH Publications*.  
<https://doi.org/10.18154/RWTH-2023-05285>