STḶM·MIG: ASSESSING PROSODIC COMPREHENSION IN PRIMARY SCHOOL

ULRIKE BEHRENS* & SEBASTIAN WEIRICH**

** Universität Duisburg-Essen ** Institut für Qualitätssicherung im Bildungswesen, Berlin*

Abstract

The German speaking countries have tested L1 listening proficiency in large national assessment studies during the past decade. However, testing prosodic comprehension—that is, students' ability to understand prosodically encoded content—has remained a blind spot, primarily because test items focusing on this specific aspect of listening have been lacking. The project stḭm·mig aims to fill this gap by developing and evaluating test items that measure students' ability to understand prosodically encoded content in auditory texts. In this article, we explain the basic process of item construction, and we present sample items to illustrate the item design. Thanks to the collaboration of the Institute for Educational Quality Improvement (IQB) in Berlin, we were able to administer and evaluate the items in a large pilot study in 252 third-grade classrooms (*N* = 4,893 students). The main goal of this large-scale assessment was to evaluate the suitability of the reading and listening items so that they might be used for national, large-scale assessment studies. We tested the effects of the presentation modes (written vs. auditory) of the stimulus texts and test items in a multiple matrix sampling design. Our findings show that prosodic comprehension is a construct that is empirically distinguishable from both verbal comprehension and reading comprehension. However, more detailed analysis is needed to fully understand the structure of the prosodic comprehension construct.

Keywords: listening comprehension, prosody, third grade, listening test, prosodically encoded information

Behrens, U. & Weirich, S. (2019). Stḭm·mig: Assessing prosodic comprehension in primary school. Contribution to a special issue on Assessing Oracy, edited by Anne-Grete Kaldahl, Antonia Bachinger, and Gert Rijlaarsdam. L1-Educational Studies in Language and Literature, 19, 1-26. https://doi.org/ 10.17239/L1ESLL-2019.19.03.03

Corresponding author: Ulrike Behrens, Universität Duisburg-Essen, Universitätsstraße 2, D – 45141 Essen, email: ulrike.behrens@uni-due.de

1. INTRODUCTION

This article presents test items for assessing how well third-grade children understand information or meaning that is prosodically encoded in oral utterances (their *prosodic comprehension*). A Swiss-German project team (*stɱ·mig*[1]) developed the items using an educational assessment studies task format. The items then were included in the VERA-3 pilot study in Germany for validation. Although designed for primary school students, the items are to serve as a template for testing prosodic comprehension on other levels, as well.

The German-speaking countries of Austria, Germany, and German-speaking Switzerland implemented educational standards in the early 2000s, just as many other countries did. The standards specify competencies that students are expected to achieve during their school careers. Separate standards are in place for the primary and secondary levels. On each level, the subject of German is segmented into content areas that include (speaking and) listening, as well as writing and orthography, reading, and grammar (cf. D-EDK, BIFIE 2011, KMK 2005). In this context, national institutes or organizations (IQB in Germany, BIFIE in Austria, and the HarmoS-Konkordat in Switzerland) monitor the respective educational systems. The tests designed for assessment purposes are administered in group settings and include items on listening comprehension.

Thus far, most of these items have been based on the design of items assessing reading comprehension, following especially the example of the Programme for International Student Assessment (PISA) tests. As a result, listening items in these assessments focus largely on verbally encoded content (including factual information and affective meaning) that could be derived from a written text, as well. Especially teachers often have criticized this approach as too narrow because it ignores the fact that, in oral language, a lot of information, including information about the emotional state of the speaker, is also communicated by the way something is said.

In response to this critique, we have developed a new kind of listening item for measuring prosodic comprehension in primary school children. The project's strategy for item construction is derived from the teaching material, "Ohrwärts," developed by Bertschin, Käser-Leisibach, and Zingg Stamm (2014). Ohrwärts is designed to measure and promote listening abilities in nine- to ten-year-old children. It focuses on self-reflection, selective listening, listening strategies, and attentiveness to paraverbal information, such as stress or tone of voice. Included is an assessment of prosodic comprehension designed for classroom use. The assessment consists of a number of items related to an orally presented narrative text: the first chapter of "Drachenreiter" by Cornelia Funke. It was tested on a sample of 246 children. Based

[1] *In addition to the authors, the* stɱ·mig *project team consists of Michael Krelle (Germany), Ursula Käser-Leisibach (Switzerland), and Claudia Zingg Stamm (Switzerland). The project title is a play on (German) words:* stimmig *means* coherent*, and it shares its free morpheme with* Stimme *(=* voice*).*

on the response frequencies in this sample, "Ohrwärts" describes three levels of proficiency, which are used to give teachers reference values (cf. Bertschin et al., 2014, p. 97).

Stɨm·mig continues the work of the Ohrwärts material, seeking to develop items that would fit the needs of large-scale assessments for measuring prosodic comprehension. We developed a large number of test items that ask for verbal content on the one hand and prosodically encoded content on the other. In both cases, "content" refers to factual information, as well as to emotional meaning. In addition to narrative texts, we included radio programs for children on different topics. The items were administered in a large-scale pilot study (210 classes; *N* = 4,893 students). The results of the items on the verbal and prosodic comprehension were compared both to each other and to the performance data of a reading test, which was also part of the study. Furthermore, each item was presented to the students either in a *written+spoken* version or in a *written only* version (for verbal content) and in a *spoken only* version (for prosody). In comparing the results, we looked for evidence of the validity of the newly developed items as a measure of children's ability to understand prosodically encoded content.

In this article, we first introduce the concept of prosodic comprehension. We then explain the educational context—national assessment studies—for which the items were designed. These assessments are closely related to the educational standards in the German-speaking countries of Germany, Switzerland, and Austria. However, a gap still exists between the competencies required by the curriculum and those measured by the existing tests. To help fill this gap, the items developed in *stɨm·mig* follow certain construction principles that we describe in the next section.
For structural validation of the construct of understanding prosodically encoded content ("prosodic comprehension"), we compare different subsamples of items to determine whether prosodic comprehension can be distinguished empirically from related, but not identical, constructs, such as reading comprehension or understanding verbal content from audio texts. Also, we consider reliability coefficients and tests measuring local independence of item responses to determine whether the items measure a homogenous construct. In the results section of this article, we address two questions: First, is prosodic comprehension a competence that is different from verbal comprehension in listening? Second, does item presentation (written vs. auditory) affect the listening comprehension?

The main goal of the project is to determine whether understanding prosodically encoded content is a competence empirically distinguishable from understanding the verbal information in spoken language. We note that generating audio texts *without* prosodic features is impossible. Therefore, understanding verbal content is naturally supported—or in some cases distorted—by prosody. As a result, an element of prosodic comprehension is always present in items on verbally encoded content in audio texts. However, the items testing prosodic comprehension reduce the verbal support for comprehension as far as possible, whereas items testing reading comprehension provide no prosodic support for comprehension. We therefore expect

children's reading and "verbal" listening comprehension to be more closely related than their comprehension on "verbal" listening items compared to "prosodic" listening items. For additional evidence, we compare scores from both parts of the listening test to scores from the reading comprehension test, and we expect reading comprehension to be related more closely to the verbal than to the prosodic listening performance.

Our second question addresses assessment on the item level. In his work on *assessing listening*, Gary Buck states that it is "important to put the emphasis on assessing those language skills that are unique to listening, because they cannot be tested elsewhere" (Buck, 2001, p. 113). If prosodic comprehension can be conceptualized as a measurable aspect of listening, it is important to note that this applies not only to understanding the stimulus text but also to understanding the test items. Even if the text is presented acoustically, the questions testing comprehension of this aurally delivered item are usually displayed in written form in a booklet. In *stim·mig*, we varied the mode of item presentation (written vs. auditory) to evaluate whether item difficulty (measured as mean accuracy of item response) is affected by item presentation.

2. THEORY

2.1 Prosody in spoken language

For listeners, understanding the meaning of a spoken utterance requires not only understanding words and clauses, but also understanding the meaning of the paraverbal features. In an early work on listening assessment, Wilkinson, Stratta, and Dudley (1974, pp. 16ff.) refer to these two elements as the linguistic channel and the paralinguistic channel of human communication. In addition, they list six more channels: the visual channel, the proxemic channel, the kinesic channel, the tactile channel, the olfactory channel, and the taste channel. In authentic conversations, these channels of nonverbal communication (or "body language") have an important influence on understanding what is being said, and they add considerably to the complexity of analysis (cf. Fiehler, 2014). In this article, we set aside these other channels because of the solely auditory character of the presented material.

Looking at the linguistic and paralinguistic channels more closely, Wilkinson et al. (1974) distinguish between three codes or "levels" of language: "the words themselves, as found in a dictionary (the *lexis*), the way these words are related (the *grammar*), and the sounds we are required to make in order to utter these words (the *phonology*)" (Wilkinson et al., 1974, p. 14). We use the term "verbal" for content that is encoded in the lexis and grammar, and we refer to the paralinguistic channel as *prosody* or *prosodic features*, referring to "variations in *suprasegmental* parameters, such as duration, intensity, and *f*0 [the fundamental frequency; UB/SW] that contribute in various combinations to the production and perception of stress, rhythm and tempo, lexical tone, and intonation of an utterance" (Fletcher, 2010, p. 523).

Phonology (or prosody) can be viewed as having three functions:

1) First, some prosodic features are inseparable from the spoken words in a specific language or dialect: These features include the individual sounds that form a word and the accent on certain syllables (differences in pitch and volume within the word) that make a word understandable (Wilkinson, et al., 1974, p. 15). For example, the correct pronunciation of the word "cinema" requires that the emphasis be on the first syllable: *ci*nema. It would not fall on the penultimate or ultimate syllable: ci*ne*ma or cine*ma*. The same inseparability is true for the grammar level: Syntax in spoken language is marked by the sequence of words in an utterance, and by the intonation of the sequence, including stresses, pauses, and pitch (cf. Hirst and Di Christo, 1998).
2) Second, some features are inseparable from the person who speaks: The sound of a voice depends on sex, age, and physical and psychological conditions. Also, someone's place of birth might more or less strongly affect the pronunciation of words or speech, as well as conditions in the delivered context, such as being hoarse or breathless.
3) Third, some features are somewhat independent of the first two in that they can change the cognitive or affective meaning of a specific utterance (cf. Wilkinson et al., 1974, pp. 41–43) or add information through prosody. One important function is to highlight "the most important information from the speaker's perspective in a given context. The main accent within an utterance is realized on the focus exponent" (Richter & Mehlhorn, 2006, 349). For example, in the question, "Are you going to the *ci*nema today?" the questioner wants confirmation of the spatial information. If the questioner asks, "Are you going to the cinema to*day*?" he or she wants temporal information to be confirmed. In this example, prosody disambiguates the meaning on the verbal level. It also can clarify the emotional meaning of an utterance (cf. Burkhardt et al., 2005).[2] The verbal content is the same, as well as the pronunciation of the words.

Obviously, these three functions are not sharply distinguishable, and one might find examples where they overlap. For the purposes of our study, we focus on the third function: the prosody that signifies the precise cognitive or emotional meaning of an utterance. If the same verbal content can have different meanings, then proficient listeners must be able to recognize the correct intention not only from the context, but also from prosodic cues. In natural language, these cues are a composition of different paraverbal features, such as stress, loudness, tone of voice, and tempo. (See Imhof, 2003, p. 30; for vocal parameters, see Bose, 2010, pp. 35 ff.)

[2] *In the database developed by Burkhardt et al. (2005), seven realizations of the same sentence can be heard: neutral, angry, fearful, joyful, sad, bored, disgusted, each uttered by a man and a woman. (See database.syntheticspeech.de.)*

A vast body of linguistic research aims at differentiating between the manifold features and exploring their respective contribution to the functioning of language. (See, for example, the comprehensive research series by Kranich (2016) and Neuber (2002); see also Paeschke et al. (1999) for the emotions of boredom and disgust; Selting (1994) for emphasis; Richter and Mehlhorn (2006) for focus; and Schmiedel (2017) for irony.) Most of the extant work uses experiments and instrumental techniques to isolate properties of spoken language. We recognize that undertaking in-depth analyses of certain items to examine aspects that contribute to task or item difficulty might be beneficial. However, considerable challenges arise in matching phonological properties of utterances with their effect on the listener (cf., for example, Kranich 2016, p. 13).

In our work, we do not aim to distinguish between different properties of the sound signal. Rather, our focus requires an approach at the pragmatic level. In test development, the items generally are constructed on the basis of expert judgments. Some researchers have tried to determine influences of text and item properties on empirical difficulty. For example, in her analysis of such properties in listening tasks at the secondary level, Neumann (2012) examined 45 variables of the stimulus audio texts. Experts rated the variables on a two-step or three-step scale, and the ratings were correlated with the empirical response frequencies, interpreted as task difficulty. Only two of these variables were phonic and paralinguistic features: speech tempo and accent/dialect/articulation ("lautliche und paralinguistische Merkmale"; p. 37). Neumann found that only speech tempo showed a correlation with task difficulty (r= -0.27, not statistically significant). Because the test focused on verbal content and the assessment items were presented only in written form, prosody-related features on the *item level* were not even considered. The high number of possibly relevant features and the complexity of their interrelations make it difficult to interpret the findings.

Buck (2001, p. 133) suggests a number of "techniques for testing knowledge of the sound system," such as recognizing intonation patterns or stress that would require considerable abstraction. However, as listeners in our daily lives we do not have to be able to distinguish between the individual elements and portions of prosodic features. In fact, according to Pittam and Scherer (1993), listeners are more successful in differentiating between positive and negative tones or basic emotions than are objective acoustic analyses of vocal properties. Therefore, we do not ask students to describe features of language in an abstract manner. In contrast to Buck (2001), we try to capture the pragmatic abilities of proficient listeners to adequately understand the meaning of a whole utterance. Buck lists this kind of item as "conversation tasks", but doesn't differentiate between verbal and prosodic features (for example: "Test-takers listen to a statement followed by three possible continuations, and select the one that would continue the conversation best"; Buck, 2001, p. 135). Also, he does not provide an assessment instrument, himself, but provides a comprehensive survey of test requirements and material.

In contrast, Wilkinson et al. (1974) "decided to focus on intonation and to test understanding of changes in meaning brought about by its variation" (Wilkinson et al., 1974, p. 43), and they present different test batteries for three age groups (10–11 years old, 13–14 years old, 17–18 years old) including two subtests of phonology. However, the following example of an item from these tests shows how complicated it is to construct items that precisely target prosodic comprehension:

Figure 1. Sample item from test battery A, age 10+, with explanation (Wilkinson et al., 1974, p. 44)

(Sounds of Western film, gunshots, etc. on TV)
Mother What's happened in the Western, Jane?
Jane Pardon?
Mother Turn it down a little. What's happened in the Western, Jane?
Jane Well this man, he's the baddie, has got this girl and tied her up in the valley where the rattlesnakes are.
Mother Oh, charming

The narrator poses the question 'Does Jane's mother really think that the story is charming?' offering the answers:
A. *Yes.*
B. *No.*
C. *We can't tell from the conversation.*

The narrator explains:
She doesn't think it's charming; in fact, she probably thinks it's rather horrible. We can tell this by the way she says '*charming*' and not 'charming'. (…)

Although the correct answer can be derived from the ironic tone of the mother's voice in the listening task, it is just as easy to find it based on the given situational context without having to hear the exchange. In contrast, the next example (from battery B) cannot be solved based on the verbal information alone; the correct answer depends on the way the father says, "What's his name?":

Figure 2. Sample item from test battery B, age 13+ (Wilkinson et al., 1974, p. 45)

Jim Have you finished with the marmalade?
Molly Here you are
Jim Thanks.
Molly Well, whatever the delay it's not the new postman's fault.
Jim No, seems a very efficient chap. What's his name?
Narrator Question three. Has Janet's father ever heard the name before?

Possible answers are:
A. *Yes.*
B. *No.*
C. *You can't tell.*

The latter kind of item is what *stim·mig* seeks to develop for large-scale assessments.

2.2 Testing listening comprehension in national and international assessment studies

The assessment item development in *stim·mig* followed the approach generally used in national and international large-scale assessment studies, including PISA, PIRLS/IGLU, NAEP, and VERA-3.[3] In this approach, items are created "top-down" by finding audio texts that would require a certain proficiency level of comprehension for full understanding of the text, rather than being created "bottom-up" by manipulating specific features in a theoretically prescribed way. Test authors are especially concerned with finding age-appropriate texts so that the students listen to (or read) texts that are neither boring nor discouraging, but attractive, accessible and slightly challenging for them. The same concern applies to the items themselves: They should cover the relevant content of the text so that testers can assume that only the students who answered the items accurately have understood the text.

In these assessments, listening items normally can be solved on the basis of the verbal information alone; prosodic features might support comprehension but are not assessed separately.[4] Compared to a typical reading test, the only difference is the input or stimulus text, which is auditory rather than written. Otherwise, the test takers' challenge is the same: After having listened to a stimulus text, such as a story or a radio program, the students typically read a question or prompt and then choose the correct answer from several options or write a short-answer response. This part of the test could be completed based on the verbal information of the audio alone.[5]

2.3 Listening and reading comprehension

This similarity in test construction for both reading and listening is reflected in the similarity of the models of reading and listening performance. For example, the model underlying the German assessments uses five levels of competence both for listening comprehension and for reading comprehension. Students who have competences on the respective level are able to:

I. recognize single pieces of information in a prominent position in the text.
II. connect adjoining pieces of information and give information less prominently positioned.
III. connect scattered pieces of information and recognize the main idea of the text.
IV. recognize important relations in general and understand details in context.

---

[3] *oecd.org/pisa, timssandpirls.bc.edu/pirls2006, nces.ed.gov/nationsreportcard, iqb.hu-berlin.de/vera*

[4] *The same is true for one of the most popular English language tests: the IELTS by the British Council. (See the training test items at https://takeielts.britishcouncil.org/prepare-test/free-sample-tests/.)*

[5] *For sample items from the national assessment test in German primary schools, see https://www.iqb.hu-berlin.de/bt/BT2016/Bsp.*

V. evaluate and justify statements referring to main ideas of the text (cf. Bremerich-Vos et al., 2012, pp. 57ff.).

The pilot study for the evaluation of educational standards in Germany showed a latent correlation of .74 between listening and reading comprehension (cf. Behrens et al., 2009, p. 37). This correlation thus far has been explained based on an overlap of the tested constructs. That reading and listening share an underlying general comprehension skill is generally accepted (text comprehension; cf. Gernsbacher, Varner, & Faust, 1990; Sticht & James, 1984), regardless of the channel of perception (visual vs. auditory). Kürschner and Schnotz (2008) propose a more differentiated model derived from a survey of mostly experimental studies that compare the building of mental representations based on listening and reading. According to this model, reading and listening processes are rather similar in higher order processing. Modality effects between listening and reading are primarily caused by differences in lower level processing because of the (im-)permanence of the auditory signal vs. the visual signal. This model would be supported by substantial correlations still significantly below 1.

However, when empirical results of the reading and listening tests show these strong correlations, the results might at least partially be due to the methodological similarity of the tests and therefore represent an artefact of operationalization. To eliminate variance that arises on the verbal level, Behrens et al. (2009, p. 372) have suggested using identical texts as reading and listening stimuli and comparing the test results. Differences in performance could then easier be attributed to differences between listening and reading proficiency as a test construct. However, depending on the type of text used, the lexical and grammatical elements of oral and written language can differ considerably. Thus, the texts suitable for such an approach are limited to texts that can "naturally" be found in written and spoken form, such as stories that come as books and as audio books.

Furthermore, reading is not only involved in understanding the stimulus texts, but also the items. In paper-and-pencil tests, the items are usually displayed visually in booklets and have to be read. In a study on the effects of item presentation on listening test scores of L2 college students, Chang and Read (2013) found no significant differences between the oral and written item presentation. However, controlling for the differences in proficiency revealed an interaction: Poorer listeners scored significantly higher in the written mode, indicating that the cognitive load of the listening task was somewhat relieved when these test takers were allowed to read the items. To our knowledge, no similar study has been conducted that focuses on young children. Whether reading along would support or impede the comprehension of the questions in third graders who are still reading novices is an open question (cf. Schlücker, Hannken-Illjes, & Dehé, 2017, pp. 151ff.). As Rubin, Hafer, and Arata (2000) show in their summary of previous research, listening comprehension "enjoys an early developmental advantage over reading; this difference largely disappears for competent readers who have mastered decoding skills" (Rubin, Hafer, & Arata, 2000, p. 121).

This finding is supported by the findings of a pilot study for the evaluation of educational standards in Germany that showed a rather high listening proficiency among third graders, compared to results in the reading test. In fact, the listening items had to be revised to attain the targeted mean item difficulty of p=.50 (Behrens et al., 2009, p. 367). Developing items that would be solved by only a minority of very proficient students was particularly difficult. Furthermore, as in Chang and Read's study (2013), such an effect could also work differentially in the way that good readers profit from the additional support, whereas weaker readers struggle with monitoring the input (cf. Rost & Hartmann, 1992, pp. 346–48).

2.4 Listening in the curriculum

In many countries, large-scale assessments are directly connected to the adoption of national educational standards. However, a gap remains between the competencies described in the standards and those tested in the assessments. In the task of listening, educational standards in the three German-speaking countries assume a wide range of competencies. To illustrate, "listening comprehension" in the standards in Germany means "understanding content, inquiring purposefully, and expressing understanding and misunderstanding" (cf. KMK, 2005, p. 10). The Swiss curriculum 21 also includes the understanding of prosody, such as interpreting the tone of voice in the situational context. Understanding in monological listening situations means being able to detect important information; in dialogues, students are supposed to follow conversations and reflect on their own listening attitude and interests (cf. Kanton Zürich, 2017, p. 2ff.). In Austria's standards for the primary level, competence in oral language perception includes perceiving both linguistic and non-linguistic communication; using general knowledge for understanding; willingness to listen to others; and realizing particular requirements of different situations in oral communication (cf. BIFIE, 2011, p. 5f.). Standards serve as orientation for curriculum and classroom material, but they also establish the basis for the development of assessment instruments (cf. Krelle and Prengel, 2014). However, testing some of the standards, such as speaking, has been difficult because of methodological constraints.[6] Even the receptive aspects of oral language (listening) have been restricted to understanding the verbal content of audio texts, tested by developing items that strongly resemble the well-evaluated PISA-style reading tests. Therefore, national assessments so far have covered a rather narrow understanding of listening comprehension. Although politicians and practitioners in education are aware of the multimodality of oral communication, assessment tools do not yet reflect this multimodality. Consequently, we see a gap between the educational objectives and the abilities that actually are monitored. This gap can lead to a blind spot in the classroom because teachers might be tempted to focus on what is being tested rather than on the standards themselves.

[6] *So far, only Austria has sought to test speaking proficiency in students; for such testing at the primary level, see Breit, Bruneforth, & Schreiner, 2016, pp. 91–96.*

At the same time, the validity of educational monitoring on the national level (for listening comprehension) might also be questioned because the listening test does not include the important aspects of comprehending prosodically encoded content in auditory texts. The project *stim·mig* seeks to fill this gap by developing and evaluating test items that measure the student's ability to understand the meaning of utterances based on their prosodic features.

3. TASK DEVELOPMENT

As already noted, listening assessments in groups, as in a classroom setting, usually ask test takers to listen to a story or a radio program suitable for the respective age group. Afterward, they are presented with a number of content questions and have to write down the answer in a short-answer response (half-open item format) or choose an answer out of several options (multiple-choice). In VERA-3, multiple-choice items contain only one correct answer out of four options; sometimes, four to six questions in a true/false format are combined into one complex multiple-choice item; sometimes, matching items are used.[7] Developing a complete task therefore requires finding or producing a suitable audio text as stimulus and generating a number of test items—usually between five and fifteen, depending on the text's length and potential.

In *stim·mig*, we developed ten tasks based on six stories and four radio programs. Three additional collections of items are not based on stimulus texts. The first collection of stand-alone items ("Betonung," six items) focuses on understanding the meaning of short, context-free utterances based on intonation. Another collection ("Stimmklang", seven items) focuses on certain acoustic properties of voice, rather than on understanding meaning. These properties are inseparable from the person who speaks; for example, one item asks, "In which out of four sentences can you hear that someone is eating a banana?" The last collection ("Vorlesen", eleven items) determines children's ability to adequately rate readings by other children. For example, test takers listen to children reading aloud and are asked to decide which one of two readers performed better or which one of four feedbacks would fit the performance best (see Figure 3).

Figure 3. Sample item: The most adequate notion has to be chosen (see audio file#1)

Karl reads the end of a story. Which hint could you give him?

- ☐ He should pronounce more clearly.
- ☐ He should stress some of the words.
- ☐ He should not pause too often.
- ☐ He should read slower.

[7] *Sample items for listening in grade three are published at www.iqb.hu-berlin.de/vera/aufgaben/dep.*

In the following paragraphs we identify the most crucial features of the tasks based on stimulus texts. The description shows the general idea of the item design. (A more detailed explanation of the item constructions is not provided because of space constraints.) First, we briefly describe the audio texts used and then we explain the construction of the "prosodic" items. Table 1 provides an overview of the stimulus texts, their duration, and the number of items that focus on verbally and prosodically encoded content.

Table 1. Overview of stimulus texts and number of items

Genre Title	Short description	Duration	Number of "verbal" items	Number of "prosodic" items
	Stories			
Das goldene Herz	Clipping from *Das goldene Herz* by Ulf Stark	6:50	8	4
Bärbeiß	Chapter 2 of *Der Bärbeiß* by Annette Pehnt	8:30	7	7
Niklas und Karl	Clipping from *Nicht Chicago. Nicht hier* by Kirsten Boie	6:55	6	6
Tütenprinzessin	*Die Tütenprinzessin* by Robert Munsch	5:10	8	6
Aufregung im Schloss	Beginning of *Die wilde Sophie* by Lukas Hartmann	7:20	5	5
Drachenreiter	Chapter 1 of *Drachenreiter* by Cornelia Funke	10:05	5	5
	Radio programs			
Schluckauf	Program on the phenomenon of *hickup*	2:40	5	6
Geheimschrift	Program on *cryptographs*	5:56	9	2
Mit Tieren sprechen	Program in *animal language*	5:40	7	6
Bababa	Program reporting on a special kind of *music lesson*	4:08	4	4
	Other			
Betonung	short utterances; test takers decide on correct *meaning based on stress*	no stimulus text		6
Stimmklang	short utterances; test takers decide on *reason for vocal sound*	no stimulus text		7
Vorlesen	short readings by children; test takers rate *quality of reading*	no stimulus text		11

Stimulus texts

To represent a broad range of genre and topic, the *stim·mig* text corpus included both stories for young people and expository texts. Their topics were loosely connected to the field of language or voice.

If the stories were excerpted from longer texts, we ensured that they appeared as a complete narrative to provide an authentic and somewhat satisfying listening experience. In any case, the goal was to have three to five preferably very distinct main characters or speakers and a fair amount of dialogue. Conversations involved different, clearly distinguishable voices and/or emotional states of the characters.
However, the narratives and radio programs used for the test construction were not simply two homogenous groups of texts separated only by genre:

- All "expository" texts were pre-produced pieces and could not be changed. These broadcasts had the typical structure of a radio show, including a studio moderator, original "sound bites," different people talking, and atmospheric background noise. For the narratives, the analog form would have been the production of radio dramas, which was not possible because of financial constraints. Instead, a single professional speaker read all the narratives in an audio book style. These recordings were produced under the supervision of the project team.
- Especially the narratives, but essentially all the texts, covered a wide range of genre, topics, complexity of language, and characters. The narratives included literature for young people (e.g., "Nicht Chicago. Nicht hier.") and fantastic stories (e.g., "Der Bärbeiß"), and they featured "real" characters (e.g., "Das goldene Herz"), princesses (e.g., "Die Tütenprinzessin"), and dragons (e.g., "Drachenreiter") as protagonists.

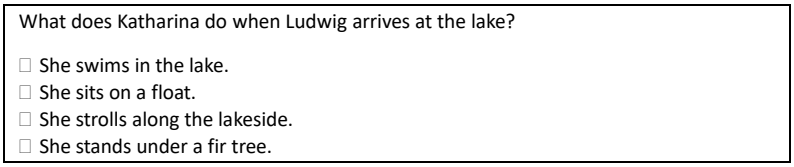
Because of this broad spectrum of texts, we did not investigate "genre effects" on listening performance. Instead, we assumed that the listening competence studied is independent of genre preferences in the test takers. Thus, the undeniable but uncontrollable influence of individual interests or previous knowledge is treated as randomized by means of a multiple matrix sampling design (see the methods section).

Item construction

To match the overall style of VERA-3 and other assessment studies, we included in every text-based task a set of items asking for verbally encoded content, similar to the sample in Figure 4. The complexity of these items is supposed to be dependent on the cognitive process necessary to identify the correct answer. Complexity was structured according to the descriptions of the five levels of the IQB proficiency model, as previously identified. Thus, because 72.2% of the students solved the item in Figure 4 correctly, it was located on level II ("connecting adjoining pieces of information and expressing information less prominently positioned"). The necessary information for the item was derived from the following text clipping:

... Er konnte kaum atmen, denn draußen auf dem Floß saß Katharina mit den Füßen im Wasser und sang das Lied, das sie bisher immer nur gesummt hatte.... [...He could barely breathe, for out on a float sat Katharina, her feet in the water, and she sang the song that, until now, she had only been humming....]

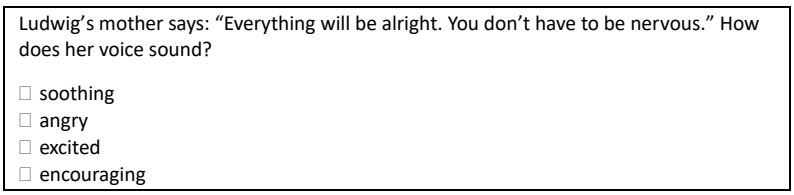
Figure 4. Sample item asking for verbally encoded content



In addition to these "verbal" items, a second set of items focused on understanding the prosodically encoded content of an utterance and therefore required the ability of prosodic comprehension. Development and evaluation of these prosodic items were the main objectives of the *stim·mig* project. Single audiotaped utterances were either cut out from the stimulus texts and replayed or were constructed according to the relevant texts (e.g., spoken in the voice of a specific character). The children were able to mark their answer while listening to the item.

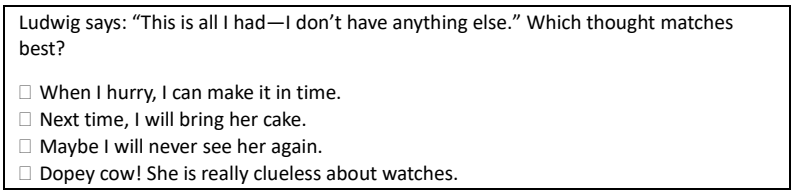
On the verbal level, the utterances in the items had varying emotional qualities. For example, in the item shown in Figure 5, Ludwig's mother might have tried to reassure her son using a soothing tone of voice or, using a more energetic expression, she might have tried to encourage her son. Also, she might have been irritated by Ludwig's nervousness and spoken angrily. In the text, though, the mother was actually rather nervous herself, and her prosodic tone was "excited."

Figure 5. Sample item: The emotional quality of an utterance has to be chosen (see audio file#2)



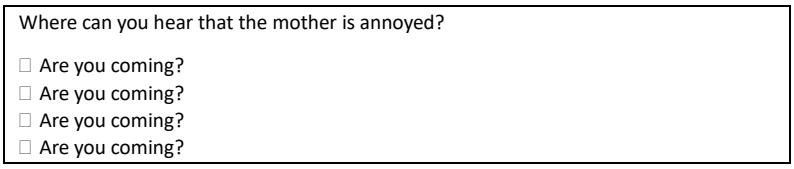
Other items did not explicitly name the emotional qualities but offered possible thoughts or alternative wording instead. Figure 6 shows a clip from the same stimulus text: Ludwig was sad and discouraged because Katharina kept rejecting his gifts, so he might have been thinking: "Maybe I will never see her again." But he also might have thought about rushing to get her something else, or he might have been thinking about another option, or he might have been mad at Katharina for her ignorance.

Figure 6. Sample item: A thought matching the emotional quality of an utterance has to be chosen (see audio file#3)



Another example is shown in Figure 7, in which a given emotional quality (annoyance) had to be recognized from four options that had identical verbal content.

Figure 7. Sample item: The utterance with a given emotional quality has to be chosen (see audio file#4)



Item and text presentation

As indicated, most paper and pencil listening tests, at least in the German-speaking countries, present the stimulus texts aurally, whereas the test instruction, the questions, and the answering options are displayed in writing, in a booklet. Accordingly, an undefined portion of reading ability is required for solving an item. Because third graders still must be considered reading novices, low performance scores might reflect weak reading skills rather than poor listening skills. To estimate the influence of reading ability, some of the items were presented in two modes:

- Some items asking for verbal content were displayed in a *written only* condition (i.e., items only appeared in the booklet and had to be read), and some verbal items were displayed in a *written+spoken* condition. (Test takers could hear questions and options and read along in their booklets.)
- Some prosodic items were presented in a *written+spoken* condition, while other prosodic items were presented in a *spoken only* condition. (Test takers found only checkboxes with the letters A, B, C, D in their booklet and had to listen to the item and the answering options.)

Finally, three of the narratives ("Niklas und Karl," "Die Tütenprinzessin," and "Aufregung im Schloss") also were administered as reading tasks. The verbal items were used for a comparison of reading and listening comprehension of the same texts. Obviously, it was not possible to have the same students answer one item in two

versions; rather, the comparison was based multiple matrix sampling design (see next paragraph).

For evaluation and validation, the items were included in the VERA-3 pilot study. The methodological details of the statistical analysis of the data are presented in the next section. We then address the following two questions based on these results:

1) Is prosodic comprehension a competence different from verbal comprehension in listening?
2) Does item presentation (written vs. auditory) affect the listening comprehension?

4. METHOD

4.1 Sample

We collected data by including our items in the VERA-3 pilot study ("Vergleichsarbeiten" in Germany). The main goal of this large-scale assessment was to evaluate whether newly developed reading and listening items (from an empirical point of view) are suitable to be used in national low-stakes, large-scale assessment studies. Hence, the focus of this study was on the evaluation of item characteristics (instead of, for example, students' competencies).

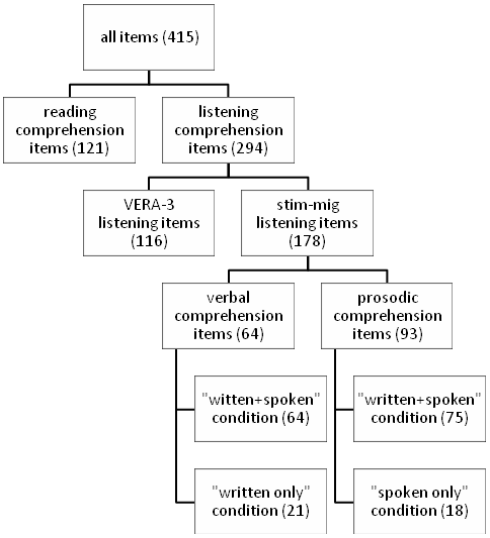
The total sample size was N = 4,893 third-grade students (49.9% female, 48.5% male, 1.6% missing information, mean age = 8.9 years), nested in 252 classes; 70% of the children were native German speakers (i.e., they responded that they speak exclusively German at home). Another 21.1% of the children said they speak both German and another language at home, and 3.8% said they exclusively speak a language other than German at home (5.1% missing information).

The sample was not representative for Germany, since only 8 of the 16 federal states took part in the study. Because the study primarily focused on inferences at the item level, the lack of a representative sample on the individual level is not considered critical.

Based on a multiple matrix sampling design (Gonzales & Rutkowski, 2010), each student worked on only a subset of tasks and items. Overall, the test comprised 415 dichotomously scored items: 121 items in the reading test and 294 items in the listening test. Of these 294 items, 116 items were part of the VERA-3 pilot study, and 178 items were developed specifically for the *stɪm·mig* study. As shown in Figure 8, the 178 listening comprehension items were divided into 93 prosodic and 85 verbal items. The 93 prosodic items were presented in the two conditions, *written+spoken* (75 items) and *spoken only* (18 items). The 85 verbal items were presented in the two conditions, *written+spoken* (21 items) and *written only* (64 items).[8]

[8] *Please note that the total number of items, 178, does not correspond to the sum of items listed in Table 1. The reason is that the listening items of three tasks were administered in two conditions, respectively, which counts as two single items. For example, "Tütenprinzessin"*

Figure 8. Subsamples of items in the study



The items were grouped into disjunct blocks, with a scheduled processing time of 20 minutes for each block. Each of the 56 different booklets consisted of four blocks, so that the overall testing time per booklet was 80 minutes.

4.2 Statistical models

To answer our study's first research question, a two-dimensional item response theory (IRT) model (Adams, Wilson, & Wang, 1997) was specified to evaluate whether prosodic comprehension can be empirically distinguished from verbal comprehension. To assess whether the construct is homogenously measured by the items, two competing models—the one-parameter (1pl) logistic model (Rasch model) and the two-parameter (2pl) logistic model (Birnbaum, 1968)—are specified, using the R package *TAM* (R Core Team, 2015; Robitzsch, Kiefer, & Wu, 2018). Both models are compared using the χ^2-test and the Bayesian Information Criterion (BIC) (Schwarz, 1978).

To answer the second research question, we used the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2015) to specify two uni-dimensional generalized linear mixed models (GLMM). One model was specified to compare the two different item presentation conditions *written+spoken* and *written only* for verbal items, and the other model was specified to compare the two conditions *written+spoken* and *spoken only* for prosodic items. In both models, effects of persons

included 6 prosodic items presented in a written+spoken *condition and a spoken only condition, which results in a total of 12 single items used in the calculation.*

and items were specified as random effects, whereas the effect of item presentation was specified as a fixed effect.

5. RESULTS

5.1 Preliminary analyses

Prior to the analyses that address research questions 1 and 2, descriptive analyses lead to the exclusion of 16 items, which showed poor discrimination values—i.e. the biserial correlation was below 0.1. Further inspection yielded that for some of these items the multiple-choice options were not distinguishable clearly enough. From a technical point of view, the remaining 162 items show satisfying fit indices (1pl item infit between 0.89 and 1.10; probability of solving an item between 13% and 96.5%). Hence, this subsample was considered to be sufficient to answer research questions 1 and 2.

5.2 Research question 1: Is prosodic comprehension a competence different from verbal comprehension in listening?

The latent correlation between prosodic comprehension and verbal comprehension was .755 for the 1pl model and .799 for the 2pl model. For both parametrizations, the χ^2-test shows a significantly better fit for the two-dimensional model: 1pl: $\chi^2 = 206.48$, $df = 2$, $p < 0.001$, $\Delta BIC = 190.0$; 2pl: $\chi^2 = 90.88$, $df = 1$, $p < 0.001$, $\Delta BIC = 82.7$. Hence, prosodic comprehension and verbal comprehension as measured by the test material described can be seen as two different but related facets of listening comprehension.

In addition, Table 2 shows correlations of both listening dimensions with reading competence, according to a three-dimensional GLMM.

Table 2. Latent correlations of verbal comprehension, prosodic comprehension, and reading comprehension

	Reading comprehension	Listening comprehension: verbal
Listening comprehension: verbal	.73	
Listening comprehension: prosodic	.61	.79

Comparing verbal and prosodic items in the 2pl model, we found substantially greater variation in item discrimination parameters for the prosodic items. Hence, the prosodic items are more heterogeneous in their discrimination between students. The mean discrimination is 0.756 (*SD* = 0.560) for the prosodic items and 0.965 (*SD* = 0.481) for the verbal items. Verbal items show a better discrimination overall, and they vary less in their discrimination than prosodic items. Correspondingly, the EAP reliability for verbal items (1pl: 0.599; 2pl: 0.615) is higher than for prosodic items (1pl: 0.525; 2pl: 0.568).

The question of whether 1pl or 2pl modeling is appropriate yields mixed results. The χ^2-test favors the 2pl model, whereas the BIC leads to the conclusion that the better model fit of the 2pl model does not justify the large number of additional parameters (comparison of uni-dimensional 1pl vs. 2pl for prosodic items only: $\chi^2 = 328.65$, $df = 76$, $p < 0.001$, $\Delta BIC = -296.9$).

So far, prosodic items have been modeled only uni-dimensionally. Without a concrete hypothesis about a possible multi-dimensional structure within the prosodic dimension, the appropriateness of uni-dimensional modelling can only be confirmed indirectly—for example, via the Q3 statistic (Yen, 1984, 1993). Substantial local dependencies often occur if the assumption of θ as an uni-dimensional latent trait is violated (Lord and Novick, 1968); thus, the Q3 statistic might give a hint as to whether this assumption is justified. Based on the 1pl model, 74 of 1,239 item pairs (6%) exceed the strictest threshold of $|r| > .2$, and 32 of 1,239 item pairs (2.6%) exceed the threshold of $|r| > .25$.

5.3 Research question 2: Does item presentation (written vs. auditory) affect listening comprehension?

Two GLMMs were specified separately for the verbal items and the prosodic items. Both models were tested based on a subsample of 2.026 students and show small but significant effects of the presentation mode (see Table 3): The first GLMM yields that verbal items in the *written+spoken* condition are significantly easier (i.e., more students answered them correctly) compared to the verbal items in the *written only* condition. However, the effect of .13 ($z = 2.83$, $p < .01$) logits is small. The second GLMM shows that prosodic items in the *written+spoken* condition are significantly more difficult compared to the items in the *spoken only* condition. Again, the effect of .12 ($z = 2.45$, $p < .05$) logits is small. Still, the previously hypothesized assumption is confirmed: The more listening and the less reading is involved in solving an item, the easier the items prove to be. Verbal items are easier if children hear them while reading along, whereas prosodic items are easier if the children only listen to them.

Table 3. Effects of item presentation

Parameter	verbal items			prosodic items		
fixed effects	Est.	*SE*	*p*	Est.	*SE*	*p*
Intercept	0.213	0.262	0.416	0.358	0.289	0.215
"written+spoken" (⍰)	0.134	0.048	< .01	-0.120	0.049	< .05
random effects	Var	SD		Var	SD	
persons	0.756	0.869		0.368	0.607	
items	1.415	1.190		1.069	1.034	
fit indices						
number of persons		2,026			2,026	
number of items		21			13	
number of item responses		19,159			11,741	
R^2		0.08%			0.08%	

Note. Only a subsample of items and persons was used for these analyses.

6. DISCUSSION

The project *stim·mig* sought to develop a new type of listening test item that measures the capability of understanding prosodically encoded content in large-scale assessments. The main goal was to produce, evaluate, and validate the items by administering them in a large-scale pilot study (i.e., VERA-3). Because no test procedure is available yet for the purpose of large-scale assessment, the results of the large-scale evaluation are used to prove their general practicability and the validity of the measurement. Based on theoretical considerations, we argued that prosodic comprehension is a competence different from verbal comprehension in listening. This notion was confirmed by comparing the results of the prosodic listening items with other item subsamples. Correlations with reading test items helped to further examine this construct. Evidence for the validity of the items was therefore drawn from test takers' performance, compared with test constructs that, from the theoretical point of view, are expected to be different, but related. For this purpose, correlations on different levels were calculated.

Not surprisingly, results showed that the two dimensions of listening proficiency have the closest connection (r = .79). We then compared the VERA-3 reading test results to the results on listening for verbally encoded content. Likewise, and in line with theoretical assumptions, the correlation between verbal listening and reading

comprehension of r = .73 (see Table 2) indicates that the two constructs are different but closely related.

Two interpretations are possible: On the one hand, reading and listening comprehension share much of the common construct of general language comprehension. This commonality is especially true for higher order processing of information, which Kürschner and Schnotz (2008) assume to be less affected by the channel of reception (visual vs. auditory). On the other hand, considerable similarity exists in the construction of verbal listening and reading items; the similarity therefore might be, at least in part, an artefact of item construction.

This impact of item construction becomes clearer when taking the results of the prosodic items into account: In comparing the two dimensions of listening (verbal and prosodic) to each other, and either of them with the VERA-3 reading test, the comparison reveals that the correlation between prosodic listening and reading is considerably lower (r = .61) than the correlation between verbal listening and reading (r = .73). This gap indicates that prosodic items actually add something to the construct that is different from reading. So, if one assumes a common basis of text comprehension or language comprehension shared by all three dimensions, the differences can be interpreted as indicating specific aspects of reading, verbal listening, and prosodic listening. However, the *stim·mig* items tested for the first time in this study are different from conventional verbal listening items because understanding prosodically encoded content is necessary for solving them. We therefore conclude that prosodic listening comprehension is an empirically distinguishable construct, and that items of the type presented can measure this construct.

However, the prosodic items' variance and reliability scores are fairly low (0.756; SD = 0.560). These values indicate that the prosodic listening test in its current form is not free of construct-irrelevant variance (Messick, 1984, 1998). Based on the data at hand, we cannot determine whether this variance results from the construct measured or from features of the test takers. In other words, the values could be low because the range of the children's prosodic listening comprehension levels is narrower than the range of their verbal listening skills. Also, the values could be low because (at least some) prosodic items fail to assess students' competences accurately, therefore failing to differentiate between stronger and weaker prosodic comprehension. This question cannot be answered simply by using a more complex model (2pl, for example) because either the items' ability to discriminate between proficiency levels or the between-subject variance has to be fixed for model identification.

Finally, on the item level, we examined the effect of item presentation in the listening test. For the three test tasks, each of the 39 items was administered in two versions: *written only* vs. *written+spoken* (for verbal comprehension) and *written+spoken* vs. *spoken only* (for prosodic comprehension). Our findings show that the item presentation mode has a small effect in favor of auditory input: Although the opportunity to hear the items while reading them in the booklet facilitates verbal comprehension, the opportunity for both reading and hearing seems to cause

confusion in the case of prosodic items: On verbal items, the children score higher in the *written+spoken* condition than in the *written only* condition (β = .13), whereas prosodic items are easier in the *spoken only* condition compared to the *written+spoken* condition (β = –.12).

These effects could point in the same direction as the previous findings in that the listening-specific portion of prosody makes a difference when listening is being tested. However, in this case, other possible interpretations have to be considered, too. First, the tested third graders must be viewed as reading novices. When they are looking at a written text as it is being read to them, their need for self-regulation in the reading process is diminished. Also, they are likely very familiar with the "listen and read along" setting, whether from home (e.g., bedtime stories) or in school. Prosodic items, in contrast, might work in reverse: Students might be confused if they have to concentrate on aspects of the acoustic input that are not verbalized in the spoken text. Therefore, only having to listen might be easier for children if there is nothing (meaningful) to read anyway.

In addition, reading proficiency levels also might yield a differential effect: The lower the additional cognitive load caused by reading along, the smaller the effect of reading along should be. In very proficient readers, the effect might even be inverted so that reading along facilitates prosodic comprehension for the student. Further analysis can show whether the *stɪm·mig* data support this idea, but will still have to be treated with caution for the reasons indicated.

While further research in smaller, more controlled settings is needed to provide additional support for these findings, we also need to look more closely at the aspects of the items on prosodic comprehension that yielded varying results:

- In all cases, the necessary information for solving an item must be derived from prosodic features. Items that did not meet this requirement were discarded before the data collection or were excluded from the analysis of the pilot study's results. However, in some items, the context might be easier to use than in others. For example, in Figure 6, children who built up an adequate mental model of the situation and characters might have been able to draw on this mental representation to solve the item; the correct option in this case was in line with the gist of the stimulus text. However, most of the other prosodic items had to be solved independently from the text.
- Different levels of ambiguity in the verbal content might have altered the importance of the prosodic features for comprehension. Still, differences in this respect cannot be attributed solely to features of the stimulus text or item; instead, the students' ability to achieve comprehension by resolving ambiguity interacts with numerous characteristics of the individual test taker (such as interest in or previous knowledge about the topic, familiarity with a certain genre, attention control, and basic cognitive competence) that couldn't be taken into account in this study.
- Prosodic items aim at various paralinguistic phenomena. The main concern of the project was the third function of prosodic properties in spoken utterances.

These properties clarify or complete the cognitive or emotional meaning of an utterance that is either ambiguous in or missing from the verbal level. However, in some cases, the necessary information that leads students to the correct option does not correspond to this function. Rather, they focus on prosodic features that, as we mentioned, are inseparable from the person who speaks. Although we did not ask students to identify dialect, age, or sex from the tone of voice, in some items we asked for current conditions, such as being breathless or eating a banana. These items focused on the general ability of the students to draw conclusions regarding the general (social) situation instead of asking them to understand linguistic aspects. This requirement is valid from the perspective of actual classroom work, which often focuses awareness for and concentration on sound. However, items of this kind probably weaken the data's informative value.

- The same effect might hold for a set of items that ask test takers to rate readings by children of the same age. Although this task is a realistic one for third graders, and it actually requires attention to prosody, additional knowledge also is involved (e.g., on norms of reading fluency). In this context, even prosodic function of the first kind (which is inseparable from the words and grammar, as discussed) matters. Items that test another aspect of prosody should be treated separately in further analyses.

In summary, in considering the requirements for measuring listening in the area of educational monitoring, as well as in classroom assessments, the test instruments that have been used in the past should be broadened. Listening items preferably should involve as little reading as possible; recorded and orally presented items (oral stimulus texts as well as response options and test instructions) are to be preferred. Such items promise a more valid measure of listening proficiency than verbal items alone. They also keep especially younger children from the strain of (possibly slow or weak) reading and might show more accurate results for listening comprehension.

However, written items displayed in a test booklet admittedly are considerably easier to produce and administer. Also, they give test takers the chance to reread questions and to work at their own pace, rather than standardizing the delivery rhythm of questions and answers after having heard a text. Therefore, as long as large-scale assessment is regularly being administered in a paper-and-pencil format (rather than, for example, in a tablet-based format or as online assessment), the significant effort and resources invested in recording and producing audio items must be well considered.

As a result of the *stim·mig* project, a rich pool of prosodic items is now available, along with the items' statistical parameters, based on a large sample. In light of our results and discussion, we call for further analysis of subsamples and individual items, as well as additional studies that investigate particular aspects of our study in a more detailed way. Still, the items have proven useful for large-scale assessment and can be recommended as a model for further development and study design.

AUTHORS' NOTE

For more information, please feel free to contact any team member:
Ulrike Behrens (ulrike.behrens@uni-due.de), Ursula Käser-Leisibach (ursula.kaeser@fhnw.ch), Michael Krelle (michael.krelle@zlb.tu-chemnitz.de), Sebastian Weirich (sebastian.weirich@iqb.hu-berlin.de), Claudia Zingg Stamm (claudia.zingg@fhnw.ch).

REFERENCES

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 57-75). New York, NY: Springer. https://doi.org/10.1007/978-0-387-49839-3_4

Allen, J., & Le, H. (2008). An additional measure of overall effect size for logistic regression models. *Journal of Educational and Behavioral Statistics, 33*, 416-441. https://doi.org/10.3102/1076998607306081

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.0-6). URL http://CRAN.R-project.org/package=lme4

Behrens, U., Böhme, K., & Krelle, M. (2009). Zuhören—Operationalisierung und fachdidaktische Implikationen [Listening—Operationalization and didactical implications]. In A. Bremerich-Vos, G. Walther, D. Granzer, & O. Köller (Eds.), *Evaluation der Bildungsstandards Deutsch und Mathematik* (pp. 357–376). Weinheim, Germany: Beltz.

Bertschin, F., Käser-Leisibach, U., & Zingg Stamm, C. (2014). *Ohrwärts. Zuhören und literarisches Hörverstehen. Kompetenzerhebung mit Förderangeboten für 9- bis 10-Jährige* [Ohrwärts. Listening and literary comprehension. Assessment and training for 9- and 10-year-olds]. Solothurn, Switzerland: Solothurner Lehrmittelverlag.

BIFIE (Ed.) (2011). *Themenheft für den Kompetenzbereich 'Hören, Sprechen und Miteinander-Reden.* Special issue for the competence area 'listening, speaking, and oral communication']. Graz, Austria: Leykam

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bose, I. (2010). Stimmlich-artikulatorischer Ausdruck und Sprache [Vocal-articulatory expression and language]. In A. Deppermann & A. Linke (Eds.), *Sprache intermedial. Stimme und Schrift, Bild und Ton* (pp. 29-68). Berlin, Germany: de Gruyter, 2010. https://doi.org/10.1515/9783110223613.29

Breit, S., Bruneforth, M., & Schreiner, C. (Eds.) (2016). *Standardüberprüfung 2015 Deutsch, 4. Schulstufe. Bundesergebnisbericht* [Review of the educational standards German, 4th grade. Results report]. Retrieved from https://www.bifie.at/wp-content/uploads/2017/05/BiSt_UE_D4_2015_Bundesergebnisbericht.pdf

Bremerich-Vos, A., Böhme, K., Krelle, M., Weirich, S., & Köller, O. (2012). Kompetenzstufenmodelle für das Fach Deutsch [Competence models for the subject German]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (pp. 56–71). Münster, Germany: Waxmann.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *9th European Conference on Speech Communication and Technology*. 5. 1517-1520.

Chang, A. C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System, 41*(3), 575-586. https://doi.org/10.1016/j.system.2013.06.001.

De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 3-42). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3990-9_1

D-EDK (=Schweizerische Konferenz der kantonalen Erziehungsdirektoren) (Ed.) (2018). *Nationale Bildungsziele für die obligatorische Schule: in vier Fächern zu erreichende Grundkompetenze*n [National educational objectives: Basic competencies to be achieved in four subjects]. Retrieved from https://www.edudoc.ch/static/web/arbeiten/harmos/grundkomp_faktenblatt_d.pdf

Fiehler, R. (2014). Von der Mündlichkeit zur Multimodalität ... und darüber hinaus [From oracy to multimodality ... and beyond]. In E. Grundler & C. Spiegel (Eds.), *Konzeptionen des Mündlichen. Wissenschaftliche Perspektiven und didaktische Konsequenzen* (pp. 13-31). Bern, Switzerland: hep Verlag.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 523-602). Oxford, UK: Blackwell.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 430-445. https://doi.org/10.1037/0278-7393.16.3.430

Gonzalez, E., & Rutkowski, L. (2010). *Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments.* IEA-ETS Research Institute Monograph, 3, 125-156.

Hirst, D., & Di Christo, A. (1998). A survey of intonation systems. In D. Hirst & A. Di Christo (Eds.), *Intonation systems. A survey of twenty languages* (pp. 1-44). Cambridge, UK: Cambridge University Press.

Imhof, M. (2003). *Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung* [Listening. Psychological aspects of auditory information processing]. Göttingen, Germany: Vandenhoeck & Ruprecht.

Kanton Zürich (Ed.) (2017). *Lehrplan für die Volksschule auf der Grundlage des Lehrplans 21*, [Curriculum for the primary school based on the curriculum 21]. Retrieved from https://zh.lehrplan.ch/lehrplan_printout.php?k=1&ekalias=0&fb_id=1&f_id=11.

KMK (=Kultusministerkonferenz) (Ed.) (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich* [Educational standards in German in primary school] München: Wolters Kluwer Deutschland GmbH. Retrieved from: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Deutsch-Primar.pdf

Kranich, W. (2016). *Sprechwissenschaftliche Grundlagen der Prosodieperzeption* [Phonological basics of prosodic perception]. Berlin, Germany: Frank & Timme

Krelle, M., & Prengel, J. (2014). Zur Konzeption von Zuhören im Rahmen der Vergleichsarbeiten für die dritte Klasse im Fach Deutsch [On the conception of listening in the context of the comparision tests in German]. In E. Grundler & C. Spiegel (Eds.). *Konzeptionen des Mündlichen—wissenschaftliche Perspektiven und didaktische Konsequenzen* (pp. 210–228). Bern, Switzerland: hep Verlag.

Kürschner, C., & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen [The relationship between spoken and written language in the construction of mental representations]. *Psychologische Rundschau, 59*(3), 139-149. https://doi.org/10.1026/0033-3042.59.3.139

Lord, F. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237. https://doi.org/10.1002/j.2330-8516.1984.tb00046.x

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45*(1-3), 35-44. https://doi.org/10.1023/A:1006964925094

Molenberghs, G., & Verbeke, G. (2004). An introduction to generalized (non)linear mixed models. In P. De Boeck & M. Wilson (Eds.). *Explanatory item response models* (pp. 111-166). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3990-9_4

Neuber, B. (2002). *Prosodische Formen in Funktion. Leistungen der Suprasegmentalia für das Verstehen, Behalten und die Bedeutungs(re)konstruktion* [Forms of prosody in function. Merits of

suprasegmentalia for understanding, remembering and (re)construction of meaning]. Frankfurt a.M., Germany: Peter Lang.

Neumann, D. (2012). *Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverstehen im Fach Deutsch in der Sekundarstufe I* [Features of listening comprehension tasks influencing difficulty in the school subject German on the secondary level]. Kölner Beiträge zur Sprachdidaktik 8. Köln: Gilles & Francke.

Paeschke, A., Kienast, M., & Sendlmeier, W. (1999). F0 contours in emotional speech. In *Proceedings of the International Congress of Phonetic Sciences 1999*, San Francisco, vol. 2, 929-932

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.). *Handbook of emotions* (185-198). New York, NY: Guilford Press.

R Core Team (2015). R: *A language and environment for statistical computing* (Version 3.2.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Richter, N., & Mehlhorn, G. (2006). Focus on contrast and emphasis: Evidence from prosody. In V. Molnár & S. Winkler (Eds.). *The architecture of focus*. Berlin, Germany: De Gruyter, 347-372.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules* (Version R package version 2.11-93). Retrieved from https://CRAN.R-project.org/package=TAM

Rost, D. H., & Hartmann, A. (1992). Lesen, Hören, Verstehen [Reading, listening, understanding]. *Zeitschrift für Psychologie 200*, 345–361.

Rubin, D. L., Hafer, T., & Arata, K. (2000). Reading and listening to oral-based versus literate-based discourse. *Communication Education 49*(2), 121–133. https://doi.org/10.1080/03634520009379200

Schlücker, B., Hannken-Illjes, K., & Dehé, N. (2017). Zuhören vs. Lesen: Verständnis literarischer Texte bei Schüler_innen [Listening vs. reading: Comprehension of literary texts in students]. *Zeitschrift für Angewandte Linguistik, 67*, 149–177. https://doi.org/10.1515/zfal-2017-0021

Schmiedel, A. (2017). *Phonetik ironischer Sprechweise* [Phonetics of ironic speech]. Berlin, Germany: Frank und Timme

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-465. https://doi.org/10.1214/aos/1176344136

Selting, M. (1994). Emphatic style: with special focus on the prosodic signalling of heightened emotive involvement in conversation. *Journal of Pragmatics, 22*(3/4), 375-408. https://doi.org/10.1016/0378-2166(94)90116-3

Sticht, T. G., & James, J. H. (1984). Listening and reading. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.). *Handbook of reading research* (pp. 293-317). New York, NY: Longman.

Wilkinson, A., Stratta, L., & Dudley, P. (1974). *The quality of listening*. London, UK: Macmillan.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER *ConQuest version 2.0: Generalised item response modelling software* [Computer software]. Camberwell, Australia: ACER.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x