LISTENING COMPREHENSION TESTS IN GERMANY AND AUSTRIA

Research report and critical review

SEBASTIAN WEIRICH*, ANTONIA BACHINGER**, MATTHIAS TRENDTEL*** & MICHAEL KRELLE****

Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin **Federal Institute for Educational Research, Innovation and Development of the Austrian School Sector (BIFIE)*TU Dortmund University ****Chemnitz University of Technology (TUC)*

Abstract

This article focuses on listening comprehension in large-scale assessments in Austria and Germany. L1 listening comprehension tests for elementary school students are part of the national educational assessment in both countries. The aim of this paper is to give an overview of the similarities and differences of these listening comprehension tests in Austria and Germany. Thus, we describe the educational policies and the underlying construct for the assessments. We will show that although both assessments are based on the same theoretical frameworks, test developers and policy makers made some different decisions for example about test procedures, statistical models and performance level descriptions (PLDs). Moreover, we illustrate that the choice of an appropriate statistical model is driven by empirical as well as didactical needs, which are difficult to reconcile with each other. These insights are illustrated by exemplary tasks and empirical examples. We use data from large-scale assessments (LSAs) from both countries, the BIST-Ü pilot study conducted by the BIFIE in Austria (*N* = 2,798) and the VERA-study conducted by the IQB in Germany (*N* = 3,107). We then draw conclusions primarily focusing on improving future tests and possible joint studies.

Keywords: L1 listening comprehension, large-scale assessments, oracy, IRT-Models, German language assessment, competency models

Weirich, S., Bachinger, A., Trendtel, M., & Krelle, M. (2019). Listening comprehension tests in Germany and Austria—Research report and critical review. L1-Educational Studies in Language and Literature Contribution to a special issue on Assessing Oracy, edited by Anne-Grete Kaldahl, Antonia Bachinger, and Gert Rijlaarsdam., 19, 1-26. https://doi.org/10.17239/L1ESLL-2019.19.03.05

Corresponding author: Sebastian Weirich, Hannoversche Straße 19, 10115 Berlin, Germany, email: sebastian.weirich@iqb.hu-berlin.de

1. INTRODUCTION

L1 listening comprehension tests for 3rd and 4th grade primary school students are part of the national educational assessment in Germany ("Bildungstrend" and "Vergleichsarbeiten") and in Austria ("Bildungsstandardüberprüfung"). The aim of this paper is to give an overview of the similarities and differences of these German as a native language tests in Austria and Germany. For this purpose, we will outline the different steps that are taken to develop an assessment framework of listening comprehension.

For a valid comparative assessment of Austrian and German students, a common assessment framework is necessary—i.e. both tests must be equivalent in terms of theoretical foundation, operationalization, measurement model, and the "rules of interpretation". Hence, the tests should refer to equivalent criteria, which can be used for passed/failed decisions. International assessment frameworks— such as PISA (OECD, 2003) or TIMSS (Bonsen, Lintorf, Bos, & Frey, 2008; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009)—already attempt to compare reading comprehension or mathematical achievement between countries.

The research report presents information on two recurring studies which are conducted in each country separately, the "Bildungsstandardüberprüfung" (BIST-Ü) in Austria and the "Vergleichsarbeiten" (VERA) in Germany. Both tests pursue monitoring purposes and provide criteria-based feedback to schools, classes, teachers and students. The assessments are repeated regularly in order to record potential changes over the years. The paper concludes with an overview over the issues that will have to be addressed when attempting a comprehensive study for both countries.

2. EDUCATIONAL POLICIES

Because of the worse-than-expected results of the first PISA examination in 2000, educational standards were developed in Germany as well as in Austria. Educational standards describe and define quality criteria of education systems with reference to specific subjects (i.e. mathematics, biology, native language, and foreign languages) and specific domains (i.e. subject knowledge, reading comprehension, and listening comprehension). Educational standards thus pursue to contribute to quality assurance and development in the education system (Shephard, Hannaway, & Baker, 2009). There are several forms of standards that are associated with educational norms: The most common ones, outcome standards are described either as "content" or as "performance standards". "Content standards" pertain to the respective areas of a particular subject, e.g. the knowledge of authors in the field of literature. "Performance standards" refer to abilities and competencies as goals of academic teaching-learning processes.

In the following, the focus lies on "performance standards" for the subject German and the domain listening comprehension. The primary objectives for the introduction of educational standards are to raise the awareness of teachers for the need of competency-oriented teaching (Köller, 2010, p. 530) and to support a sustainable development of the school system (Klieme et al., 2003).

Klieme et al. (2003) recommend that results-based standards (or performance standards) should be used as a basis to (1) set targets as a reference and (2) to provide sufficient autonomy for pedagogical practice. Standards, therefore, define which achievements students should accomplish "on average" in the respective subjects (KMK, 2004, p. 14). Standardized tests can help monitor students' achievement of educational standards empirically via regularly and repeatedly executed standardized assessments.

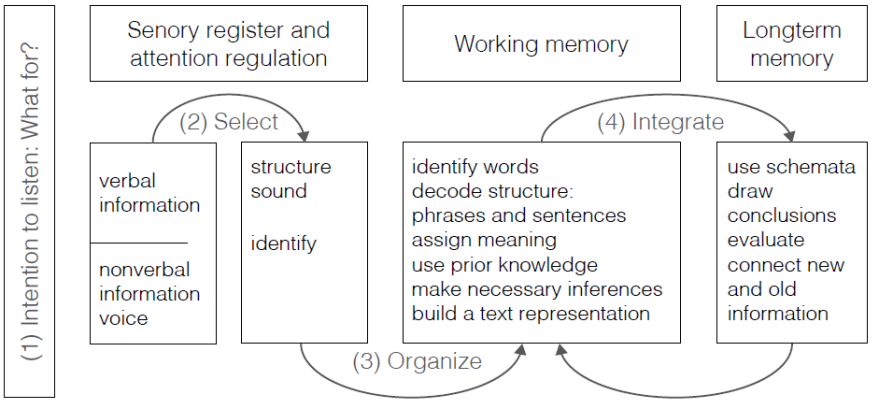
Such an undertaking requires: a) a theoretical framework of listening comprehension as a basis for test development, b) appropriate test procedures and instruments for educational measurement, c) statistical models to analyze the data in a way appropriate to the theoretical framework—thus, the statistical model should be suitable to represent and analyze the data in a way appropriate to the theoretical model, the data and research questions, and d) criterial standards to decide whether a specific score may be interpreted as "failing to fulfill the standard", "fulfilling the standard", or "exceeding the standard", for example.

It is important to mention that these four requirements are not independent of each other. Which statistical model is eligible depends on the test procedures and instruments. In turn, the criterial standards depend on the statistical model, because criterial standards are based on some specific properties of the underlying model. These interdependencies cause some constraints regarding the choice of an appropriate statistical model. These issues will be elaborated in the following sections, where these four requirements are reviewed and compared between Austria and Germany.

3. THEORETICAL FRAMEWORK OF LISTENING COMPREHENSION

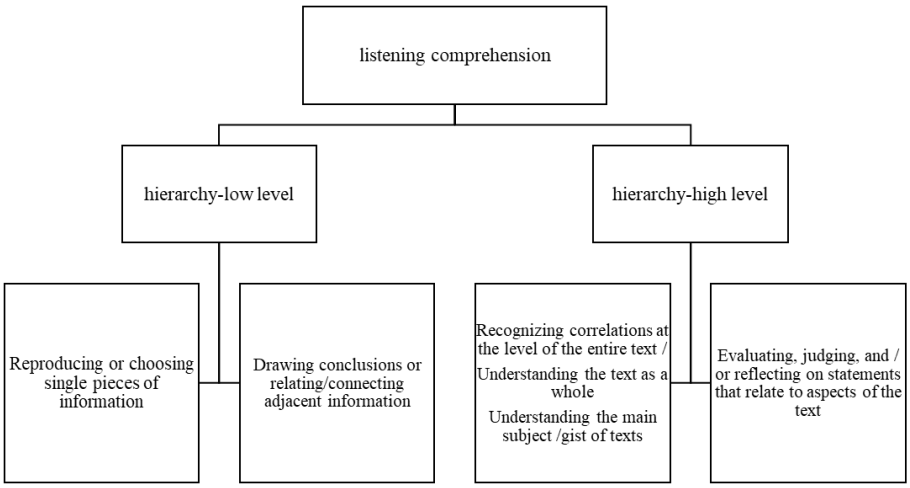
The listening comprehension tests in Germany and Austria draw upon the same conceptual model that defines listening as a multi-level process of information processing based on the listener's intention. The model draws the distinction between hearing and listening (Imhof, 2010; 2016, p. 1).

Figure 1. Listening as a four-step process of information processing (Imhof, 2016)



Listeners select acoustic signals from the continuous flow of sounds in order to organize information based on language and world knowledge. They integrate what they hear into their own representation of what they have heard. Working memory and long-term memory play a key role here: listeners check their comprehension based on their prior knowledge and new information concerning plausibility. If necessary, the listener extends or revises the information. In this respect, the listener performs an interpretive task (Imhof, 2016). To describe such achievements, various forms of mental representation are assumed to contribute to the construction of complex meaning. On the one hand, there is a hierarchy-low level of the cognitive system, which deals with the more or less analogous remembering of one or more formulations. On the other hand, there is a hierarchy-high level, in which representations are located, which refer to larger parts of the text or the entire text and its structures, e.g. in the form of a mental model (see figure 2 in Imhof, 2003, p. 65 ff.). In addition, judgments and evaluation of what is heard always play a role (not only in school contexts), because they are about literary objects, with or about which students can learn something about the respective culture or society.

Figure 2. Theoretical model of listening comprehension



In everyday life, listening comprehension is often integrated in complex communicative tasks including language production and processing of visual information. For example, when discussing, explaining or informing others, it is necessary to gather information about the listener in order to use it for their own goals and purposes. Thus, listening comprehension is often viewed as an aspect of conversation or oral communication skills. Obviously, these skills are acquired well ahead of school (and beyond the education system). Studies on language acquisition show evidence that auditory information processing is by far the earliest developed sense (Imhof, 2003). In dealing with other communication partners, listening evolves along with other social skills, such as empathy and teamwork or the ability to take on perspectives (Hagen, 2006, p. 18). In school, this complex bundle of competencies should be adapted to fit social requirements.

The theoretical concept described above forms the basis for the development of listening comprehension tests in Germany and Austria. Thus, test developers from both countries draw upon the same theoretical background.

4. TEST PROCEDURES AND INSTRUMENTS FOR EDUCATIONAL MEASUREMENT

In contrast to attitude tests, which appeal the examinees to answer honestly to some given statements (e.g. “I often feel uncomfortable in the presence of larger group of persons”), achievement tests require the examinees to try the best in finding the correct answer for some given tasks. Whereas attitude tests do not comprise “correct” or “wrong” answers, the tasks of achievement tests may be solved correctly or not. By convention, achievement tests are often binary coded; “1” indicates that the examinee’s solution is correct, and “0” indicates an incorrect solution.

Test developers are now faced with the challenge of finding some short listening texts (stimuli texts) and designing questions which can only be answered correctly if the examinee successfully accomplishes one or several elements of the four-step model of listening comprehension (see figure 2). Hence, the questions should not be answerable by general knowledge. The kinds of tasks used for measuring listening comprehension play a crucial role. Regarding the comprehensibility of the audio texts, important factors are, for example, the degree of concreteness or abstraction of the requested information. Furthermore, the type of assignment of the information (localization to integration) to the item and the type and amount of distractors both in the text and in the items play an important role (Bremerich-Vos & Böhme, 2009; Buck, 2001, pp. 154-155).

The items used in the listening tests must address rather easy as well as complex listening skills. Rather easy items ask to recognize some surface characteristics of the listening text (i.e., names or professions of main characters), whereas more difficult items allow for measurement of integrative understanding of the entire texts (Krelle & Prengel, 2014). As shown in figure 2, there are two levels of comprehension (hierarchy-low and high). These processes of comprehension are the basis for developing items for both levels. The hierarchy-low levels of the listening comprehension are represented in items that focus on retrieving explicitly stated information. Additionally, items solved by drawing directly from the text or by relating or connecting adjacent information are also considered to be on the lower level.

Depending on how elaborate inferences are, they are considered low or high level of comprehension (Cain & Oakhill, 1999; Zwaan & Radvansky, 1998). There are different items focusing on inferences on a local and a global level in the exemplary tasks. The tasks also include items about the structure of the text, the language and formulations and features of an audio text are considered to be high level. They usually refer to the whole text. The high level of the comprehension also includes recognizing the gist of the text, correlations in the text and inferring meaning about the whole text. Additionally, students have to interpret and relate what they heard about the texts' overall themes and ideas or to statements about the text. As we will show in section 7 (exemplary tasks) in detail, items in the German and Austrian test refer to the same processes of comprehension.

Additionally, the average difficulty of the test as a whole must fit to the average skill level of the examinees. One of the first empirical experiences in the development of listening tests for primary school students in 2007 showed that the items were empirically easier than expected (Behrens, Böhme, & Krelle, 2009). If the test is too easy, however, is it not possible to clearly differentiate between average and above-average students, because both subgroups tend to solve most items correctly. Consequently, subsequent listening texts with more complex content were selected as well as longer listening texts with a duration of up to 10 minutes.

One main difference between German and Austrian tests concerns the number of items per listening text. In the past, Austrian tests used rather short listening texts with a duration of up to 2 minutes and 1–3 items each. On the one hand, this leads

to a broad variability of listening texts within the single test booklets and reduces the local dependency of items, which belong to a common stimulus (Monseur, Baye, Lafontaine, & Quittre, 2011; Wainer & Thissen, 1996; Yen, 1984, 1993). On the other hand, short texts do not allow for as much character and plot development as longer texts do. This character and plot development make the texts interesting and challenging for children. Thus, using very short texts made it also quite challenging to develop difficult questions or to create items suitable for higher-performing students. Items where students need to reflect on the text or to make inferences require more complex and, therefore, longer listening tasks.

From our experience, complex, short texts (up to two minutes) can be found, but if they are not written explicitly for children, they can be unsuitable because of the level of abstraction, lexical density, complex subject matter, implied meaning, vocabulary etc. Choosing these short texts might lead to children not understanding the text at all. Additionally, the texts used in the assessment should reflect what students listen to in class and at home (Behrens, 2010).

The more recent Austrian and German tests use longer stimulus texts (Austria: 2–4 minutes for the text, 11 minutes for the task, 4–8 items; Germany: 6–9 minutes, 8–10 items) and, thus, eliminate some of the disadvantages mentioned above. However, some disadvantages of using complex listening texts may arise. In general, the students are not allowed to work on the items until after they are finished with listening to the text. Therefore, long stimulus texts not only require listening competencies but also memory capacities (Messick, 1984, 1995). Regardless of the length of the stimulus, all the items have to be read in the test booklet. However, one of the reasons for choosing longer texts is to use complex items that ask for some argumentation in favor or against a statement relating to the text. From our experience, these items are not only a good way to focus on highest level of listening comprehension according to the framework; these items are also relevant for teaching. To solve these kinds of items, students must cope with writing skills likewise. Thus, what is practically measured with tests incorporating longer stimulus texts is not only pure listening comprehension, but rather a mixture of listening comprehension, memorization, concentration, reading and writing skills.

Additionally, complex stimulus texts are prone to local dependency of items belonging to the same stimulus: Students who misunderstand relevant parts of the text will have a higher probability to fail at several items belonging to the input text. Local dependencies are primarily a statistical issue, as most common statistical models assume local independence of the item responses.

5. STATISTICAL MODELS TO OBTAIN COMPETENCE SCORES FROM EDUCATIONAL TESTS

From a conceptual point of view, listening comprehension can be classified as a latent capability (or competence) of an individual. The term “latent” refers to the fact that the listening competence as such cannot be observed directly—in contrast to each

person's body height or weight, for example. Hence, each examinee's listening comprehension competence must be derived from observable indicators or variables. Within the context of educational assessment, these indicators are the answers of examinees to test items. The measurement model (or statistical model) defines the relationship between observable variables and the unobservable (or latent) capability. Measurement models are derived from test theories.

Which measurement model of which test theory is suitable? The choice of a measurement model also depends on the response format—as the answers to test items are binary coded, logistic models derived from the item response theory (IRT) have proved to be useful and adequate in the past (Hambleton & Jones, 1993). The simplest logistic measurement model is the Rasch model (Adams & Wu, 2007; Fischer, 2006) which assumes an unobserved latent ability θ which is associated with the probability of solving an item correctly. Broadly speaking, the higher a person's θ, the higher the probability of solving an item. The Rasch model predicts this probability in the following way:

$$\text{logit}\left(P\left(X_i = 1\right)\right) = \theta_n - \beta_i \qquad (1)$$

β_i denotes the difficulty of a specific item *i*, and θ_n denotes the ability of a specific person *n* (for more details, see Embretson & Reise, 2000). The model can be used to empirically estimate the difficulties of items. In turn, these item parameters are used to formulate performance level descriptions (PLDs, see section 6).

Strictly speaking, the Rasch model consists of three components (De Boeck et al., 2011, p. 3): a) the model equation, b) the link function, and c) the random component. From equation 1, we see that the model equation does not predict a probability but the logit of a probability. The logit function is only one of several possible link functions (i.e., probit, loglog):

$$\text{logit}\left(P\left(X_i = 1\right)\right) = \ln\left(\frac{P\left(X_i = 1\right)}{1 - P\left(X_i = 1\right)}\right) \qquad (2)$$

The logit is defined as the natural logarithm of the odds ratio. An odds ratio is defined as the quotient of a probability and its inverse probability. Hence, the logit transformation is symmetrical and maps probabilities within the parameters of [0, 1] onto the latent continuous scale (De Boeck & Wilson, 2004).

The third component of the Rasch model is the random component, specifying the probability distribution of the dichotomous responses. The X_i responses follow a binomial distribution with

$$X_i \sim \text{binomial}\left(1, \pi_i\right), \text{ where } \pi_i = P\left(X_i = 1\right) \qquad (3)$$

Several properties of the Rasch model follow from equation 1 (Embretson & Reise, 2000):

1) Only two factors, β_i and θ_n determine the probability of solving an item *i*.
2) The items only vary with regard to their difficulty, not with regard to their discrimination.

3) As θ_n is defined as a unidimensional latent trait, the items are locally independent, i.e. after controlling for θ_n, there are no correlations between item responses. Lord and Novick (1968) show that both assumptions (unidimensional θ_n and local independence) are equivalent to each other.
4) The value of β_i marks "the point on the latent scale where the probability of a 1-response is .5" (Tuerlinckx & De Boeck, 2004, p. 299). This property is essential for the development of performance level descriptions (PLDs, see section 6).

The Rasch model can be seen as a special case of Generalized Linear Mixed Models (De Boeck & Wilson, 2004; Wilson & De Boeck, 2004). In its most common form, a random persons—fixed items model (De Boeck, 2008, p. 538) is used in practice. From a multilevel perspective, the Rasch model can also be described as a two-level model with responses at level 1 and items (as well as persons) at level 2 (Hedeker & Gibbons, 2006, chapter 9). In contrast to the common comprehension of multilevel models, the Rasch model differs in at least one important aspect: Common multilevel models often assume a hierarchical relationship, for example, students are nested within classes. Within the Rasch model, items as well as persons are nested within responses, but items and persons are (at least partially) crossed (Hecht, Weirich, Siegle, & Frey, 2015). Hence, there is no hierarchical relation between items and persons.

More complex IRT models—for example the two-parameter logistic (2PL) model (Birnbaum, 1968)—use a modified model equation, whereas the link function and the random component remain unchanged. The 2PL model, for example, adds a discrimination parameter to the model equation. In the Rasch model, the items differ only in their difficulty and are assumed to be equal in their discrimination. In the 2PL model, the items are modeled to differ additionally in their discrimination. The equation of the 2PL model can be written as

$$\operatorname{logit}\left(P\left(X_i = 1\right)\right) = \alpha_i\left(\theta_n - \beta_i\right) \qquad (4)$$

Hence, different models imply more or less restrictive assumptions concerning the person and/or item parameters. For example, both the Rasch model and the 2PL model assume that θ_n is univariate normally distributed with $\theta_n \sim N\left(\mu, \sigma_\theta^2\right)$. If students are nested in classes, this assumption might be violated—hence, both models can be seen as inappropriate. Mixture distribution Rasch models or multilevel IRT models might be an alternative then.

But which model is the most appropriate one? The short answer is: none of them. More complex models provide a better fit to the data (i.e., account for clustered data), but lose some of the properties which are essential for PLDs. Using the "inappropriate" Rasch model in spite of clustered data at least may lead to biased standard errors and a misinterpretation of group differences—for example if the competencies of boys and girls are compared (Lumley, 2004; Wolter, 1985).

A common approach in large-scale assessments to overcome this dilemma is to use a multi-step procedure (see, for example, the scaling procedures of the TIMSS

study, e.g. Foy, Galia, & Li, 2008) which incorporates several models: For item parameter estimation (step 1), the Rasch model is applied, and for person (or person group) estimates, multilevel models or replications methods (Rust & Rao, 1996) may be used in a second step to gain unbiased standard errors. The implementation of this multistep procedure is realized by the use of so-called "plausible values" (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). The PISA study, for example, also has adopted this method which was originally developed for the National Assessment of Educational Progress (NAEP, c.f. Allen, Donoghue, & Schoeps, 2001; Mislevy, Johnson, & Muraki, 1992). A detailed description of the "plausible values" method is beyond the scope of this paper—the basic idea is to use the measurement model as an imputation model for the inherently unobserved θ_n. The imputed θ_n-estimates may be analyzed using replication methods or multilevel models to account for clustered data.

In the German and Austrian listening comprehension tests, a comparable method is implemented: The Rasch model is used for item parameter estimation, whereas person parameters are estimated in a separate step.

6. DEFINING CRITERIAL STANDARDS VIA PERFORMANCE LEVEL DESCRIPTIONS (PLDS)

In order to obtain concrete interpretations from test scores which allow to classify groups of students into distinct classes (according to pass/fail decisions, for example), performance level descriptions divide the continuous θ-scale into disjunct sections (or levels). Students whose ability score lies within the interval of a specific level are assumed to have similar competencies. PLDs are developed from empirical data: the difficulty estimates of several items are ranked in ascending order whereupon a group of experts tries to find some "cut points", indicating where the items become more demanding in terms of meaning recognition or decoding. This process is known as standard setting (Cizek, 2001; Cizek & Bunch, 2007).

The standard setting procedure includes theoretical models as well as empirical data in order to describe PLDs. The functional principle of PLDs is "item based" (Hambleton & Jones, 1993). Within the classical test theory (CTT), students' true scores are usually interpreted in relation to a reference population (i.e., other students' true scores). For PLDs, *item* scores are used to interpret students' scores. Thus, IRT focuses on items when seeking to describe abilities of students.

Germany differentiates between five distinct levels (IQB, 2013), whereas Austria uses four levels (BIFIE, 2016). The top and bottom levels have no upper respectively lower borders. In Germany, the range of the middle levels has equal intervals (85 points[1]); however, the ranges of the middle levels are unequal varying from level to

[1] *The scale of competence values is quite arbitrary, like the scale of IQ values, which is defined to have a mean of 100 points. Competence scores have a mean of 500 points and a standard deviation of 100 points.*

level and differs from the range in Germany (e.g., Breit, Bruneforth, & Schreiner, 2016). It could be asked why the competence models differ even though they relate on very similar concepts of listening comprehension. This is not easy to answer as competence models are part of the specific educational policy of each country. Therefore, the features of competence models are also influenced by political needs and may vary between two countries.

Competencies at each level are described as "can-do" statements—hence, PLDs only describe competencies students are able to master. The description for the lowest level, for example, is not: "Students are not yet competent to make inferences in complex tests", but "Students are expected to recognize individual information from short listening texts". In Germany, "can-do" statements are also expressed for the lowest level, which might be questionable, because students without any correct answers are also assigned to the lowest level. For these students, no valid interpretations and no "can-do" statements can be derived from their test scores.

7. EXEMPLARY TASK

The following section introduces two exemplary tasks typically used in the German and the Austrian listening comprehension tests. The stimuli as well as some items along with parameter values are described in order to illustrate the typical kind of listening assessment tasks.

7.1 Germany

This 3rd grade listening comprehension task was used in 2016 in the German "Vergleichsarbeiten". The parameters stem from the pilot study, which was conducted in 2015—for more details about sample size etc., see Table 1 in section 8. The stimulus text refers to a short story by Gina Ruck-Pauquèt, called "The little zookeeper". The stimulus text contains 525 words and it takes students about four minutes to listen to it (Krelle et al., 2016). The gist of the text will be shown by the following summary:

> A zoo director is looking for a new zookeeper. The old one has quit his job because he had a problem with a snappish parrot. The zoo director invites a couple of candidates who boast about being the strongest, bravest, fastest—with one exception. A rather short man claims to understand what the animals are saying. The other candidates laugh at him and demand he proves his claim. The little man eavesdrops on the animals and realizes, for example, that the supposedly angry roaring lion has a thorn in his forepaw. After removing the thorn, the lion is calmed down. After some further proof of his skill, the parrot sits down on the short man's shoulder. Finally, the zoo director states, that the animals have voted for the little zookeeper.

A single adult speaker dramatizes the different roles in the text. The following multiple-choice question is classified as an easy item. It was solved by 78 % of the third-

grade students, which corresponds to an IRT-based item difficulty parameter of –2.1 logits and the easiest competence level 1.[2] The item fit was satisfactory (infit[3] = 0.92).

1) Why has the old zookeeper left the zoo?

- ☐ because of a conflict with the zoo director
- ☐ because he was afraid of one of the animals
- ☐ because of a conflict with one of the animals
- ☐ because he was afraid of the zoo director

This multiple-choice item asks students to identify the motive of someone's action. The relevant information to solve this item is placed at the very beginning of the text. The other options are easy to exclude. The listening text does not speak about conflicts or fears of the zoo director.

The following item asks students to match sentences to different animals (85 %, –2.7 logits, competence level 1, infit = 0.93). To solve this item, students have to focus on a specific part of the text, thus they have to make inferences on a local level. According to empirical data, 85 % of students solved this item. The students have to match the sentences with the pictures of animals. An elementary knowledge of the world and of language helps to solve this item. Students who know that the movement of seals is usually not described as "running" can easily exclude this option.

[2] *Each sentence should have a comparable grammatical structure, i.e. students should not be able to exclude options because of specific formulations. Of course, the specific formulations are not directly translatable. The following items might thus show some salience due to the translation.*

[3] *When items perfectly fit to the Rasch model, the infit value equals 1. Infit values between 0.85 and 1.15 are considered to be acceptable (Pohl & Carstensen, 2012).*

2) Which animal could have said this?
Connect the phrases to the according pictures. Attention: One picture remains.

The following constructed-response item, which asks for hypothetical reasons of someone's feelings, is a bit more difficult. 49 % of the students solved this item correctly (–0.5 logits, competence level 3, infit = 0.98):

3) Why might the little man be annoyed about the other men?

To solve the item, students must put themselves in the position of the main character. There are no options given, therefore the students must express their answers without drawing directly from the text.
The most difficult item within this listening task again is a constructed-response item. It was solved correctly by only 4 % of the students (2.5 logits, competence level 5, infit = 0.95). It requires finding some arguments in favor or against a specific statement:

4) Does the zoo director like his animals? Explain your answer.

Both decisions (yes/no) are possibly correct if they are justified properly in correspondence to the text, e.g.: Yes, because they can decide who is going to be the new zoo director. No, because he doesn't speak the language of the animals (like the little zookeeper).

7.2 Austria

The following text is an example of a typical listening text in the BIST-Ü assessment. The data originates from the pilot study conducted in spring 2018. For more details about sample size etc., see table 1 in section 8. The Austrian items cannot be assigned to PLDs yet, because the old PLDs have not been transformed for the new metric.

The text is a children's radio show about "Esperanto", the most widely spoken artificial language.

The text is non-fictional but includes some narrative elements like many radio shows for children of this age group. One adult, a female speaker, reads this short radio program. She addresses the listener(s) directly several times throughout the text. Thus, the text includes several formulations that are typical for scripted, spoken text. Additionally, central information is given more than once. The text includes 479 words, thus listening to it takes students a little more than four minutes. The following summary shows the gist of the text.

> The text introduces the topic by giving an example of the language. Then, it explains the concept of auxiliary languages in simple terms. The text also draws some comparisons with other languages like German, while especially focusing on the development of languages. After this introduction, the text focusses on the specific history of this language and its founder, Ludwik Lejzer Zamenhof. Thereafter, it gives some examples of Esperanto words and the basic structure of the language.
> The text compares Esperanto with other languages by describing it as a young plant still growing, as opposed to "naturally grown" languages. At this point, an advantage of learning "Esperanto" is presented by giving an example: being able to communicate across language barriers. This is then contrasted by mentioning other widely spoken languages like English, which already allow people to communicate without having to learn a new language.

There are six items attached to this text. Students have about five minutes to solve all items. In total, there are two multiple-choice items, two true/false items and two open-constructed items.

The first item of this task refers to the full text and its structure:

1) What can you find in this text?

- ☐ There is an exciting climax.
- ☐ People are described in detail.
- ☐ The text shows how languages are learned.
- ☐ There are many facts.

This multiple-choice item was solved by 30% of the students, which corresponds to a logit of 0.99. Thus, the item is considered to be difficult. The item fit was satisfactory (infit = 1.02). To solve this item, students must understand the gist and the tone of the whole text, however, not all given details. The three incorrect statements (distractors) are likely to be true in other texts listened to by students. Thus, students have to listen closely to the specific text and cannot draw from their general knowledge about radio shows.

The following item focuses on retrieving explicitly stated information.

2) Where do the rules for Esperanto come from?

- ☐ They were invented by a person.
- ☐ They developed over time.
- ☐ They were discovered in an old book.
- ☐ They are based on a secret language.

73 % of the students were able to solve this item, hence it is considered easy (–1.11 logits, infit = 1.01). Its difficulty was increased by taking up formulations like "secret languages" that are used in the text in different contexts.
There is a second item, which focusses on retrieving explicitly stated information. It is about Ludwik Lejzer Zamenhof (solved by 63 % of the students; –0.59 logits, infit = 1.03). All information needed for solving this item can be found in the respective paragraph.

To solve the following item, students have to make inferences: They have to link junks of information and draw conclusions. This includes implied information, which is not given explicitly. It only focusses on a part of the text:

3) Are these statements about Esperanto true?

	yes	no
Esperanto is older than Hebrew and Greek.	☐	☐
Esperanto should be as easy as possible to learn.	☐	☐
Esperanto is only spoken in few countries.	☐	☐
Esperanto should bring people closer together.	☐	☐

50 % of the students solved this item correctly (0.05 logits, infit = 0.99). For true/false-items like this, students have to make a decision for each sentence respectively. It is important that the different statements can be evaluated independently. The last two items are open-constructed items. Students have to assess, evaluate and / or reflect on statements relating to the text. The aim is to see if students can relate what they heard in the text to their prior knowledge. Additionally, these items aim to see if the students can reflect on the text, even if the information is not mentioned directly in the text.

4) Why might someone from England not consider Esperanto important?
Give reason for your answer based on what you hear.

There are no options given, thus students must express their answers in their own words. 49 % of the students answered this question correctly (0.06 logits, infit = 0.93). Answers that show that Englishmen speak English and probably have no need for a second world language are regarded as correct, for example: "because English is already spoken almost everywhere" or "because they speak English and are therefore understood everywhere". Wrong answers are those simply stating that Englishmen speak English without mentioning that this is a widely understood language.

The two stimuli have several similarities, for instance, the approximate length, the clear linear structure and that they are both read by a single speaker. Both use high-frequency vocabulary, which makes the texts easier (Buck, 2001, p. 159).

The aim of this section was to show two typical tasks from Austria and Germany to emphasize the similarities. Both item batteries as well as the examples shown in this section require similar comprehension processes. These different processes that stem from the theoretical model (section 3) will be discussed in the following paragraphs and ascribed to items from both countries: Some items focus on retrieving information that is presented explicitly in the text. Items like the first and the sixth item do not require inferring or interpreting but rather focus on information presented directly in the text. These pieces of information elicit particular focus. It is important—especially for items that focus on retrieving explicitly stated information—to have in mind that students are not able to go back to information if they missed it and that students listen to the text only once (Ruhm et al., 2016). Thus, only central information is asked in these kinds of items. This is an important difference to reading tests, in which students can usually go back to the test while answering questions.

Students also have to draw conclusions and make inferences, e.g. from characters' actions. Some of the items refer to the cohesion of the text, while others use information from outside the text. Students have to make inferences based on the text (Cain & Oakhill, 1999). These items concerning inferences might also include items about the structure of the text, the language and formulations and features of an audio text (item five). These items require students to connect pieces of information and recognize relationships both on a local and global level (Zwaan & Singer, 2003). The relationship between these pieces of information is not given explicitly in the text. Thus, students have to make straightforward inferences. For some items, students only have to connect two ideas or pieces of information, however, sometimes three or more ideas have to be connected. Some items demand listeners to focus only on parts of the text (local), like the German item two or the Austrian item three, whereas other items need listeners to discern the overall message of the whole text (global), like the third German and the first Austrian item. Sometimes, it is difficult for test developers to make a clear distinction between low-level and high-level conclusions and the respective items (Figure 2).

Students also have to interpret and relate what they heard to overall themes and ideas. The listeners have to focus on the local and global level of the text and relate it to themes and motives of the text. Additionally, they relate what they heard to their general understanding of the world. They use their own perspectives for this interpretive process to solve items like the fourth German item and the fourth Austrian item.

8. EMPIRICAL EXAMPLE

In the following section, we want to illustrate and to compare the evaluation of the listening comprehension test using data from Germany and Austria. Broadly speaking, we would like to demonstrate empirically what was discussed in theory in section 5: The most adequate model from a theoretical point of view is not the most adequate model from an empirical point of view—and vice versa. This contradiction can only be solved indirectly by using a workaround: the separation of the measurement model and the population model.

Table 1 lists some descriptive information about the number of items, tasks, persons and testlets for the German as well as for the Austrian assessment.

In order to reduce individual student workload, the listening comprehension items in both countries were used in a multiple matrix sampling design (Gonzalez & Rutkowski, 2010) in which a subset of items (i.e., a booklet) was randomly assigned to each examinee. According to a balanced incomplete block design (Frey & Bernhardt, 2012; Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010), several booklets were constructed. The entire test also includes some reading tasks which will not be considered here, and has a scheduled processing time of 80 minutes (Germany) and 240 minutes (Austria) overall.[4] After half of the test, a short intermission is provided to allow the students to relax. Due to the large size auf the Austrian assessment, the whole test procedure is separated into two parts, which are executed on two different days.

The original goal of both studies was to test whether newly developed listening tasks empirically fit to the measurement model on the one hand and to the already developed PLDs on the other hand. As mentioned in section 5, the Rasch model may be considered as a multilevel model with responses (level 1) nested in students (level 2) and items (level 2). Nevertheless, the model does not incorporate the hierarchical structure within the person level (students nested in classes). Ignoring the clustered structure may bias the estimation of standard errors which in turn leads to biased significance tests (Luke, 2004; Lumley, 2004). For example, the question whether males and females significantly differ in their listening comprehension skills, can lead to mistaken conclusions if the standard errors of the average male/female competence scores are underestimated. If the focus of the study lies only on the estimation of item parameters, this issue matters less. Sinharay and Haberman (2014) emphasize that it is important to consider the practical consequences of model misfit. If the model is wrong, the consequences might be substantial or not.

Applying the Rasch model using the R (R Core Team, 2015) package TAM (Robitzsch, Kiefer, & Wu, 2018), at first glance the results do not contravene the assumptions of the Rasch model (see Table 1): only 3 of 116 items (Germany) respectively 4

[4] *In Austria, the assessment also includes reading, writing, language in use and spelling. Speaking is also assessed but only with a small group of students on a different day.*

of 254 items (Austria) have an Infit value of above 1.15 which is considered to indicate item misfit (Pohl & Carstensen, 2012). Considering the Q3 statistics which indicates violation of local stochastic independence (Yen, 1984), the violation of local independence does not seem to be severe: Applying the most strict criterion (| *r* | > .2), 6% of the German and 3% of the Austrian item pairs exceed this threshold value. Comparing the Rasch model with the more liberal two-parameter logistic model (Birnbaum, 1968), the Bayesian Information Criterion (BIC, Schwarz, 1978) yields that the improved model fit of the 2PL model does not outweigh the additional discrimination parameters. This holds for Germany and Austria as well.

On the first view, the Rasch model seems justifiable also from an empirical point of view. To investigate whether we would have to consider the hierarchical structure at the person level, we specified a multilevel model (Doran, Bates, Bliese, & Dowling, 2007) within the framework of Generalized Linear Mixed Models (De Boeck & Wilson, 2004) using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014, 2015). Broadly speaking, within the right-hand side of equation 1, θ_n is separated into a between-class variance δ_c and a within-class variance θ_{pc}. Hence:

$$\text{logit}\left(P\left(X_i = 1\right)\right) = \theta_{pc} + \delta_c + \beta_i \quad [5] \qquad (5)$$

Table 1. Descriptive statistics for both assessments

	Germany	**Austria**
year of assessment	2015	2018
number of listening items	116	254
number of listening tasks	11	50
number of testlets	11	50
number of booklets	45	37
number of persons	3107	2798
number of classes	163	154
number of items per testlet (min.-max.)	6-14	3-8
total number of responses	57,260	55,316
time allocated for each task (min.-max.)	10-20 minutes	3-11 minutes
entire test length	80 minutes	240 minutes
average age of students	8.9 years	10.4 years
Applying the Rasch model		
number of items with Infit > 1.15	3 of 116	4 of 254
number of item pairs with |Q3|> 0.2	190 of 3261	141 of 4778
number of item pairs with |Q3|> 0.25	75 of 3261	56 of 4778
number of item pairs with |Q3|> 0.35	4 of 3261	9 of 4778
BIC Rasch	58465	61868
BIC 2PL	58591	63025

[5] *We adopt the commonly used „+"-notation of generalized linear mixed model.*

The multilevel model additionally parametrizes a random effect for the 163 (Germany) resp. 154 (Austria) classes. Table 2 lists the results for the Rasch model and the multilevel model, which was fitted to the German and Austrian data each. Comparing the multilevel model against the random persons—random items Rasch model (RPRI, see De Boeck, 2008, p. 538), the results of both countries clearly prefer the multilevel model. All three indices (AIC, BIC, and the likelihood ratio test—Germany: $\chi^2 = 321.4$; $df = 1$; $p < .001$; Austria: $\chi^2 = 271.9$; $df = 1$; $p < .001$) indicate that the fit of multilevel model to the data is significantly better than the fit of the Rasch model.

These results suggest that the Rasch model may be sufficient as long as only item parameters are in the focus of interest: A possible bias due to the clustered structure mainly affects θ. The results of the Q3 statistic suggest that the multilevel structure of the data *within the item level* is not as severe as it has to be explicitly considered by the model.

Referred to the practice of common large-scale assessments, the Rasch model is often used in spite of its empirical misfit. However, the possible solution seems obvious: Why not use a more appropriate statistical model, which accounts for clustered data? Why do we try to adapt the assessment to the statistical model instead of adapting the model to the data? The reason for this was shortly addressed in section 2 and section 5: Statistical models and the definition of criterial standards are not independent from each other. One cannot (substantially) change the statistical model and maintain the interpretation rules from performance level descriptions (PLDs). As Tuerlinckx and De Boeck (2004) emphasize: Item parameters resulting from response models which account for local dependence (for example, copula models, see Braeken, 2011) cannot be understood as item difficulties in the sense which is crucial for PLDs. The main benefit of PLDs is to describe students' performances by means of item difficulties. Hence, no reference to some kind of norm population is necessary for the interpretation of test scores. For this purpose, the scales of item and person parameters must be identical—the Rasch model meets these requirements which follow from its properties (see section 5). More complex models like the multilevel model however, do not share these characteristics.

To evaluate whether the Rasch model must be considered as an inappropriate model, it is self-evident to ask for practical consequences of misfit. For example, Glas and Suarez Falcon (2003) showed that the violation of local independence does not bias the estimation of item response curves in a serious manner. However, this is not necessarily true if the Rasch model assumption of $\theta_n \sim N\left(\mu, \sigma_\theta^2\right)$ is violated due to clustered data.

Table 2. Fixed and random effects for the RPRI Rasch model and the multilevel model

	Germany						Austria					
Parameter	**RPRI Rasch model**			**Multilevel model**			**RPRI Rasch model**			**Multilevel model**		
Fixed effects												
	Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*
Intercept	0.240	0.148	0.105	0.227	0.153	0.138	0.562	0.085	<.001	0.562	0.084	<.001
Random effects												
	Var	*SD*		Var	*SD*		Var	*SD*		Var	*SD*	
theta	1.060	1.030		0.805	0.897		0.661	0.813		0.508	0.712	
beta	2.484	1.576		2.473	1.572		1.706	1.306		1.710	1.308	
delta				0.264	0.514					0.162	0.403	
Model fit												
AIC	58288			57969			61147			60877		
BIC	58316			58005			61174			60913		
deviance	58282			57961			61141			60869		

Note. theta – within class variance, beta – item variance, delta – between class variance.

Hence, the Rasch model and the multilevel model—on its own—must be considered as inappropriate. As mentioned in section 5, the solution is to combine both imperfect models to reach a satisfying solution. The multi-step procedure allows us to separate the estimation of item parameters from the estimation of person parameters. By using a conditioning model, the inappropriate assumption $\theta_n \sim N\left(\mu, \sigma_\theta^2\right)$ is replaced by $\theta_n = \mathbf{Y_n}\boldsymbol{\beta} + E_n$ with $E_n \sim N\left(0, \sigma^2\right)$. Conditioning models are part of the "plausible values"-procedure but do not account per se for clustered data. Monseur and Adams (2009) emphasize that the multilevel structure has to be specified in the conditioning model to yield unbiased proficiency estimates as well as accurate within- and between-class variance.

9. OUTLOOK: WHAT CAN WE LEARN FROM THE EXPERIENCES IN BOTH ASSESSMENTS?

The tests in Austria and in Germany have both gone through a couple of adaptations to fix some problems, which became apparent in their first implementations. The assessments might be seen as "learning systems", which are continuously being refined—future adaptations, for example, will integrate methods of computer-based assessments into listening comprehension tests. The conclusions which result from

our experiences and from the comparison of the German and the Austrian assessment are:

- Length of stimulus texts should vary between one and five minutes. If the text is too short, higher performance levels cannot be measured adequately; if the text is too long, memory or concentration effects become a source of construct-irrelevant variance (Messick, 1984, 1995). This issue was solved differently by test developers in Austria and Germany considering the different pros and cons for this decision. In Germany, test developers chose longer texts to better represent texts used in schools and to be able to develop items that are more difficult. In Austria, shorter texts were used to minimize memory or concentration effects and to have more texts per test in limited test time. The differences in length are minor and we consider both solutions viable options.
- Developing test tasks is not an easy business. It also requires a piece of creativity. To name just a few examples: Depending on the complexity of the text, it is easier to develop tasks at several levels. In addition, tasks can be easily developed to low-hierarchy processes. However, it is more difficult to create tasks to high-level processes. The quality and scope of the tasks thus depends essentially on the selected texts.
- Items which ask to argue in favor or against an opinion, which might be derived from the listening text are both a curse and a blessing. From a methodological point of view, these items call for more than only listening comprehension—the construct measured by these items is essentially multi-dimensional. These items, on the other hand, address a central point of literacy, which is forming an argument on basis of a text. As was shown in the theoretical model (section 3), this is an important part of what defines listening comprehension. Test construction is often confronted with such dilemmas between statistical and didactical demands. By now, these items are used in both countries.
- In both countries, there is limited time for the assessment, thus, using longer texts reduces the number of different texts per test. If there is an issue with one of the texts, this increases the influence on the overall performance because a bigger percentage of all items administered in this group depend on that text. If a student knows the text/topic of the text, he or she might be able to solve more items. If a student does not like a specific genre or is distraught by the content, the performance might be lower than could be accepted. Of course, subjects that could possibly be led to this are avoided in the assessment; however, this is not always possible (i.e. text about animals could lead children to think about recently deceased pets etc.).
- Considering the empirical analyses, a single statistical model that has all the desirable qualities which are necessary for the evaluation of listening assessment, does not exist. Hence, we have to deal with various models in multiple steps that build on one another. The model, which is used for item parameter estimation in the first step, differs from the model, which is used for person parameter estimation in a subsequent step.

10. FUTURE RESEARCH

We are planning to empirically compare the German and the Austrian tests in a pilot study carried out by the BIFIE or the IQB. Students will work on tasks from both countries in one booklet. Fortunately, the regional differences in standard German are only minor between Austria and Germany, therefore the same texts without translations can be used in both countries. Several items from each country will be used to guarantee that both constructs are reflected in the tests. There will be at least three individual booklets tested.

Some of the preparatory work for this project has already started. The items will also have to be combined in different ways to limit the effects they have on each other. The tasks will be similar in length and administrative form to avoid undesirable context effects. We plan to have about 200 students solve every individual item. This results in a minimum sample of 600 students. It would be ideal to use the same test booklets in Germany to get data from German students on Austrian tasks and vice versa. Furthermore, we hope to extend our research project to Switzerland. However, it is not certain yet that this will be possible in the near future.

In addition, there are no comparisons with models and studies on listening comprehension in the native language in other countries. Furthermore, researchers in foreign language didactics have been studying aspects of listening for quite some time (Osada, 2004). Especially with regard to the requirements of the test, considerable differences in the constructs and in the tests are addressed in research (Behrens & Krelle, 2014). A comparison is still pending.

REFERENCES

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, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report*. Washington, DC: National Center for Educational Statistics.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung [Aspects of a competence model for listening skills and options how to test it]. In V. Bernius & M. Imhof (Eds.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis.* [Listening competencies in education and teaching. Contributions from science and practice] (pp. 31-50). Göttingen, Germany: Vandenhoek & Ruprecht.

Behrens, U., Böhme, K., & Krelle, M. (2009). Zuhören—Operationalisierung und fachdidaktische Implikationen [Listening comprehension—Operationalization and pedagogical implications]. In A. Bremerich-Vos, D. Granzer, & O. Köller (Eds.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* [Educational standards German and Mathematics. Performance assessements in elementary school] (pp. 357-375). Weinheim, Germany: Beltz.

Behrens, U., & Krelle, M. (2014). Hörverstehen—ein Forschungsüberblick [Listening comprehension—research overview]. *Didaktik Deutsch, 36*, 86-107.

BIFIE (2016). Konstrukt- und Kompetenzstufenbeschreibung in Deutsch/Lesen/Schreiben 4. Schulstufe. Die Kompetenzstufen für die Überprüfung der Bildungsstandards [Construct and competency description in German / Reading / Writing 4th grade. The competency levels for the review of educational standards]. Retrieved from https://www.bifie.at/wp-content/uploads/2017/05/BIST-UE_D4_Konstruktbeschreibung.pdf

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.

Bonsen, M., Lintorf, K., Bos, W., & Frey, K. A. (2008). TIMSS 2007 Grundschule—Eine Einführung in die Studie [TIMSS 2007 Elementary school – an Introduction to the study]. In W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter, & G. Walther (Eds.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich [Mathematical and scientific competences of primary school children in Germany in an international comparison] (*pp. 19–48). Münster, Germany: Waxmann.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika, 76*(1), 57-76. https://doi.org/10.1007/s11336-010-9190-4

Breit, S., Bruneforth, M., & Schreiner, C. (2016). *Standardüberprüfung 2015 Deutsch, 4. Schulstufe. Bundesergebnisbericht* [Educational Standard Assessment 2015. German, 4th grade. National report]. Salzburg, Austria: BIFIE.

Bremerich-Vos, A., & Böhme, K. (2009). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. [Reading Competency Diagnostics - the development of a competency level model for the field of reading]. In D. Granzer, O. Köller, & A. Bremerich-Vos (Eds.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* [Educational standards German and Mathematics. Performance assessements in Elementary school] (pp. 228-261). Weinheim, Germany: Beltz.

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

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing: An Interdisciplinary Journal, 11*(5-6), 489-503. https://doi.org/10.1023/A:1008084120205

Cizek, G. J. (2001). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28. https://doi.org/10.18637/jss.v039.i12

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

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 Package. *Journal of Statistical Software, 20*(2), 1-18. https://doi.org/10.18637/jss.v020.i02

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fischer, G. H. (2006). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 515-585). Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/S0169-7161(06)26016-4

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science asssessment. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 Technical Report* (pp. 225-280). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling, 54*(4), 397-417.

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

Glas, C. A. W., & Suarez Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87-106. https://doi.org/10.1177/0146621602250530

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.

Hagen, M. (2006). *Förderung des Hörens und Zuhörens in der Schule* (Vol. 6) [Fostering listening skills in school]. Göttingen, Germany: Vandenhoeck & Ruprecht.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement, 75*, 1021-1044. https://doi.org/10.1177/0013164415573311

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Imhof, M. (2003). *Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung [Listening – Psychological Aspects of auditory information processing]* (Vol. 4). Göttingen, Germany: Vandenhoeck & Ruprecht.

Imhof, M. (2010). Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht [Learning to listen and teaching listening. Psychological basis for describing and promoting listening comprehension competencies in education and teaching]. In V. Bernius & M. Imhof (Eds.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis [Listening competencies in education and teaching. Contributions from science and practice]* (pp. 15-30). Göttingen, Germany: Vandenhoeck & Ruprecht.

Imhof, M. (2016). Listening is easy!? Looking at critical factors for listening performance. In K. M. Carragee & A. Moennich (Eds.), *Communication as Performance and the Performativity of Communication. Proceedings of the 2014 International Colloquium on Communication* (pp. 79-88). Blacksburg, VA: Virgina Tech.

IQB. (2013). Kompetenzstufenmodell zu den Bildungsstandards für das Fach Deutsch im Kompetenzbereich „Sprechen und Zuhören"—Primarbereich. Beschluss der Kultusministerkonferenz (KMK) vom 04.03.2010. Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf in der Version vom 13. Februar 2013. [Competency level model for the educational standards for the subject German in the competency area "Speaking and Listening". Decision of the Conference of the Ministers of Education (KMK) of 04.03.2010. Based on state comparison 2011 revised draft in the version of 13 February 2013]. Retrieved from https://www.iqb.hu-berlin.de/bista/ksm/KSM_GS_Deutsch_Z_2.pdf

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., . . . Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* [On the development of national educational standards. An expertise]. Berlin, Germany: BMBF.

KMK. (2004). Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung [Educational standards of the Conference of the Ministers of Education. Explanations to the concept and developments]. München, Germany: Luchterhand.

Köller, O. (2010). Bildungsstandards [Educational standards]. In R. Tippelt & B. Schmidt (Eds.), *Bildungsforschung* [Educational research] (pp. 529-548). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92015-3_28

Krelle, M., Dünschede, S., Bremerich-Vos, A., Bittins, P., Dörnhaus, S., Engelbert, M., . . . Weirich, S. (2016). *Vergleichsarbeiten 2015. 3. Jahrgangsstufe (VERA-3) Deutsch—Didaktischer Aufgabenkommentar „Zuhören"* [Vergleichsarbeiten 2015. 3rd grade (VERA-3) German—pedagogic commentary on tasks for "Listening comprehension"]. Berlin, Germany: IQB.

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 comprehension for the Vergleichsarbeiten for 3rd grade (VERA-3) German]. In E. Grundler & C. Spiegel (Eds.), *Konzeptionen des Mündlichen [Concepts for Oracy]* (pp. 208-226). Bern, Switzerland: hep-Verlag.

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

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oakes, CA: Sage. https://doi.org/10.4135/9781412985147

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(1), 1-19. https://doi.org/10.18637/jss.v009.i08

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237. https://doi.org/10.1111/j.1745-3984.1984.tb01030.x

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037//0003-066X.50.9.741

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196. https://doi.org/10.1007/BF02294457

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131-154. https://doi.org/10.3102/10769986017002131

Monseur, C., & Adams, R. (2009). Plausible values: how to deal with their limitations. *Journal of Applied Measurement, 10*(3), 320-334.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 4*, 131-155.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

OECD (Ed.) (2003). *The PISA 2003 Assessment framework mathematics, reading, science and problem solving knowledge and skills*. Paris, France: OECD.

Osada, N. (2004). Listening comprehension research: A brief review of the past thirty years. *Dialogue, 3*(1), 53-66.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report—scaling the data of the competence tests*. Bamberg, Germany: German National Educational Panel Study (NEPS).

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/

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

Ruhm, R., Leitner-Jones, C., Kulmhofer, A., Kiefer, T., Mlakar, H., & Itzlinger-Bruneforth, U. (2016). Playing the recording once or twice: effects on listening test performance. *The International Journal of Listening, 30*, 82-98. https://doi.org/10.1080/10904018.2015.1104252

Rust, K., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research, 5*, 283-310. https://doi.org/10.1177/096228029600500305

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

Shephard, L., Hannaway, J., & Baker, E. (2009). Standards, assessments, and accountability. Education policy white paper. *National Academy of Education (NJ1)*, 16.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23-35. https://doi.org/10.1111/emip.12024

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 289-316). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3990-9_10

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, 2*, 9-36.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29. https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 43-74). New York, NY: Springer.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York, NY: Springer.

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

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185. https://doi.org/10.1037//0033-2909.123.2.162

Zwaan, R. A., & Singer, M. (2003). Text comprehension. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discours processes* (pp. 83-122). Mahwah, NJ: Lawrence Erlbaum Associates.