# International assessment of low reading proficiency in the adult population: A question of components or lower rungs?

Grotlüschen, Anke; Nienkemper, Barbara; Duncker#Euringer, Caroline

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

**Empfohlene Zitierung / Suggested Citation:**
Grotlüschen, A., Nienkemper, B., & Duncker#Euringer, C. (2020). International assessment of low reading proficiency in the adult population: A question of components or lower rungs? *International Review of Education*, *66*(2-3), 267-288. https://doi.org/10.1007/s11159-020-09830-5

International Review of Education (2020) 66:235–265
https://doi.org/10.1007/s11159-020-09829-y

ORIGINAL PAPER

# International assessment of low reading proficiency in the adult population: A question of components or lower rungs?

**Anke Grotlüschen**[1] · **Barbara Nienkemper**[2] · **Caroline Duncker-Euringer**[3]

Published online: 2 April 2020


## Abstract

Among the United Nations' 17 Sustainable Development Goals (SDGs) launched in 2015, the fourth goal (SDG 4) is dedicated to education, and one of the ten targets within that goal specifically addresses adult literacy and numeracy skills. Efforts to reach this target involve monitoring, which in turn involves assessment. The most powerful instrument for assessing literacy proficiency is the Programme for the International Assessment of Adult Competencies (PIAAC), conducted by the Organisation for Economic Co-operation and Development (OECD). It has five hierarchically organised proficiency levels for literacy. A sixth category, labelled "below Level 1", lumps together low proficiencies at the bottom end of the proficiency continuum. To boost effective action in addressing SDG 4, the UNESCO Institute for Statistics (UIS) recently launched the Global Alliance to Monitor Learning (GAML), which aims to support national assessment strategies and to develop internationally comparable indicators and methodological measurement tools. While PIAAC Levels 1–5 are already broadly suitable for international comparison, the "below Level 1" category has so far only been assessed by individual countries (e.g. Canada, the United States, the United Kingdom and Germany) using instruments developed nationally. Focusing on the reading aspect of literacy, the authors of this article investigate how these nationally developed low proficiency assessment instruments might be adjusted to facilitate international comparability.

✉ Anke Grotlüschen
anke.grotlueschen@uni-hamburg.de

Barbara Nienkemper
Barbara.Nienkemper@gmx.de

Caroline Duncker-Euringer
caroline.euringer@polizei-studium.org

[1] Universität Hamburg, Hamburg, Germany

[2] Grundbildungszentrum, Hamburger Volkshochschule, , Hamburg, Germany

[3] Akademie Der Polizei Hamburg, Hamburg, Germany

**Keywords** Lower-rungs approach · Reading components approach · Reading proficiency levels · Assessment · Monitoring · Fourth sustainable development goal (SDG 4) · Global alliance to monitor learning (GAML)

**Résumé**
Évaluation internationale de la mauvaise maîtrise de la lecture au sein de la population adulte: sur la base des composantes de la lecture ou du niveau au sein de la société ? – Parmi les dix-sept Objectifs de développement durable des Nations unies (ODD) lancés en 2015, le quatrième (l'ODD 4) porte sur l'éducation, et l'une de ses dix cibles est axée spécifiquement sur la littératie et la numératie chez les adultes. Le suivi de cette cible, qui implique une évaluation, fait partie des efforts à entreprendre pour l'atteindre. Le Programme d'évaluation internationale des compétences des adultes (PIAAC-Programme for the International Assessment of Adult Competencies), dirigé par l'Organisation de coopération et de développement économique (OCDE), est l'instrument le plus puissant d'évaluation du niveau de littératie et de numératie. Il est structuré en cinq niveaux hiérarchiques de maîtrise de la lecture, de l'écriture et du calcul. Une sixième catégorie, sous l'intitulé « niveau inférieur au niveau 1», résume les faibles niveaux en bas de l'échelle des compétences. Pour stimuler l'efficacité des efforts entrepris pour répondre à l'ODD 4, l'Institut de statistique de l'UNESCO (ISU) a récemment lancé l'Alliance mondiale de suivi de l'apprentissage (GAML-Global Alliance to Monitor Learning) qui vise à soutenir des stratégies nationales d'évaluation et à développer des indicateurs internationalement comparables et des outils méthodologiques de mesure. Tandis que les niveaux 1 à 5 sur l'échelle d'évaluation du PIAAC conviennent déjà globalement pour une comparaison internationale, le « niveau inférieur au niveau 1» a jusqu'à présent seulement été évalué par certains pays (p. ex. le Canada, les États-Unis, le Royaume-Uni et l'Allemagne) au moyen d'outils développés au plan national. Les auteures de cet article se sont penchées sur les compétences en lecture pour examiner dans quelle mesure ces outils nationaux d'évaluation des niveaux faibles développés pourraient être adaptés afin de faciliter une comparaison à l'échelle internationale.

## Introduction

### Relevance

Among the 17 Sustainable Development Goals (SDGs) launched by the United Nations (UN) in 2015 (UN 2015; 2016), the fourth goal (SDG 4) is dedicated to education. Extending the scope beyond the previous agenda's focus on primary education,[1] it aims to "promote lifelong learning opportunities for all". This has led to "hopes for a stronger role" of adult learning and education "in global education

[1] The 2030 Agenda (2015–2030) with its 17 SDGs was preceded by the Education for All agenda (1990–2015) with its 8 Millennium Development Goals (MDGs). For more information, visit https://www.un.org/millenniumgoals/ [accessed 12 January 2020].

agendas and policies" (Elfert 2019, p. 537). While UN Agendas fall into the category of *soft law*,[2] they reflect a need for action, and by endorsing them, UN Member States have made commitments towards trying to achieve the targets.

One of the core instruments of soft law is monitoring (Grek 2019), and it often relies on assessment (Hamilton et al. 2015). Monitoring countries' progress towards achieving the targets of the SDGs on an international scale makes it necessary to discuss methods of assessment, especially for adult literacy and numeracy. One of the ten targets within SDG 4 directly addresses adult literacy and numeracy skills:

> By 2030, ensure that all youth and a substantial portion of adults, both men and women, achieve literacy and numeracy (SDG target 4.6; UN 2016).

To boost effective action in addressing SDG 4, the UNESCO Institute for Statistics (UIS) recently launched the Global Alliance to Monitor Learning (GAML), which

> is designed to improve learning outcomes by supporting national strategies for learning assessments and developing internationally-comparable indicators and methodological tools to measure progress towards key targets of … SDG 4 (UIS 2017).

This initiative covers all ten targets of SDG 4, with thematic task forces established to address each of them. Since 2017, the task force for SDG target 4.6 has held several expert meetings in order to collect and evaluate existing tests and findings and discuss adequate testing instruments.

The dilemma is how to build on earlier – mostly Western – research on the one hand, and how, on the other hand, to avoid a monopolistic spread of definitions and test instruments throughout the world (Addey 2018). Another challenge is that the most powerful instrument, the Programme for the International Assessment of Adult Competencies (PIAAC) conducted by the Organisation for Economic Co-operation and Development (OECD),[3] is too expensive for most UN Member States. The OECD asks participating countries to organise the data collection and test analysis themselves. This requires sample sizes of around 5,000 test takers per country. Respondents' completion of the test and questionnaire takes approximately two hours and also includes a computer-aided personal interview which is usually carried out by a survey company that charges several million Euro for the data collection.

Moreover, the five proficiency levels for literacy do not cover the most basic levels of literacy, i.e. from total illiteracy onwards (there is simply a sixth category

[2] The term "soft law" refers to officially ratified but not legally binding instruments like resolutions and declarations of international entities such as, for example, the UN, and the European Union (EU) with the Council of Europe and the European Commission,

[3] The first cycle of PIAAC was conducted in three rounds. Many of the countries participating in Round 1 (2011–2012) will be included again in the first round (2021–2022) of the second cycle. For more information, see https://www.oecd.org/skills/piaac/about/#d.en.481111 [accessed 24 February 2020].

labelled "below Level 1").[4] Since GAML is monitoring improvement by 2030, at least two reports will be needed from each country before 2030: The first assessment would serve as a starting point which the second assessment can then be compared against, hopefully demonstrating improvement in adult literacy and numeracy. So the timeframe for coming up with suitable assessment methods and tools to begin the first round of assessments as soon as possible is tight. What is especially urgently needed are tests that cover the most basic levels of literacy in a more differentiated way than "below PIAAC Level 1". Moreover, the question arises whether existing instruments that cover lower levels of literacy can be integrated into a common scale with instruments that cover higher levels of literacy, e.g. the PIAAC scale.

## State of the art

In terms of existing instruments, there are two competing approaches, which we discuss in detail in the course of this article. One is the *lower-rungs approach* (Brooks, Davies et al. 2001a, b), and the other is the *reading components approach* (Sabatini and Bruce 2009; Strucker et al. 2007). In a nutshell, the lower-rungs approach takes a differentiated look at the lowest level of literacy, and the reading components approach indicates adults' proficiency in decoding, word recognition and word meaning (vocabulary). Both approaches have strengths and weaknesses.

Test items of the lower-rungs type have the advantage of correlating with, and complementing, higher levels on international literacy proficiency scales such as those used by PIAAC. But they have not, in fact, been translated into languages other than English and German.

By contrast, the reading components test items are not hierarchically organised and therefore are not aligned with the PIAAC scale, but they do exist in several languages. Moreover, they have been administered internationally as an add-on to the OECD's PIAAC programme, under UNESCO's Literacy Assessment and Monitoring Programme (LAMP)[5] as well as the World Bank's Skills Towards Employability and Productivity (STEP) skills measurement programme.[6] While both of these programmes were run in middle-income countries or regions, their suitability for low-income countries is unlikely. Another complicating factor is that the Reading

[4] PIAAC literacy proficiency "below Level 1" is described as follows "Individuals at this level can read brief texts on familiar topics and locate a single piece of specific information identical in form to information in the question or directive. They are not required to understand the structure of sentences or paragraphs and only basic vocabulary knowledge is required. Tasks below Level 1 do not make use of any features specific to digital texts" (OECD 2013, p. 67). For descriptions of Levels 1 to 5, see OECD 2013, pp. 66–67.

[5] Initiated in 2003 by the UNESCO Institute for Statistics (UIS), LAMP was "the first international [testing] experience concerning youth and adult literacy comprising non-European languages" (Guadalupe and Cardoso 2011, p. 213)). For more information, see https://www.uis.unesco.org/literacy/Pages/lamp-literacy-assessment.aspx [accessed 14 February 2020].

[6] Launched in 2012, the World Bank's STEP skills measurement programme was the first-ever initiative to measure skills in low and middle-income countries. For more information, see https://microdata.worldbank.org/index.php/catalog/step/about [accessed 14 February 2020].

Components test items originate from many sources and there are different versions of test sets – with different ownership.

### Purpose and structural organisation of this article

Our aim in this article is to explore and clarify whether the Reading Components, as they are used in their international version (e.g. as a PIAAC add-on), can be understood as hierarchical and therefore be organised on a proficiency scale which can be aligned with and connected to international literacy scales like the one applied by PIAAC. If this is possible, the reading component items would perform like lower-rungs items and then enhance the bottom end of the scale where the most basic skills are situated. This would solve the problem of where to find test items for a range of countries (including low-income ones), as the international Reading Components are already widely used, well-accepted and available in many languages, and have also already been pretested in the countries that participated in LAMP and STEP as well as those who bought the add-on module under PIAAC.

We begin with a review, looking back into the development of each of the two competing approaches (lower rungs versus components). This is necessary to avoid confusion between earlier and more recent versions. We also present the theoretical background, the development of test items as well as pretest and main test results for both approaches, and sum up the differences in a table. We then discuss both approaches with regard to their strengths and weaknesses for monitoring SDG target 4.6 globally. This discussion leads to our three research questions, the overarching purpose of which is to find out whether one of the item sets (the Reading Components test set) could be disconnected from its theoretical background (the components approach) and re-organised in a hierarchical way (as rungs on a ladder). This would meet the requirements specified by the GAML initiative for effective assessment methods to monitor a wide range of countries' progress in achieving SDG target 4.6. In our methodology, we describe and report on the relevant statistical tests which we carried out using *item response theory* (IRT)[7] and the German PIAAC Reading Components subset of data. After presenting the results, addressing each of our three research questions, we evaluate the outcomes and conclude our article with recommendations for further re-analysis and refinement.

## Review: assessing the most basic levels of literacy

International large-scale assessments currently measure literacy with *unidimensional* and *continuous* competence models. What this means is that individual proficiencies are hierarchically described as being situated on a scale rising from low

[7] Item response theory is used in *psychometrics* (the measuring of mental capacities and processes). Initially applied in mainly educational contexts, it enables the development and evaluation of surveys carried out using questionnaires and other proficiency assessment instruments which feature test items. For an overview, see Carlson and von Davier (2013).

to high levels of competence. In terms of the main results, PIAAC and earlier international assessments[8] have defined four or five proficiency levels and documented the percentage of adults scoring at each of these levels for each of the participating countries (OECD 2013; OECD and Statistics Canada 2000, 2005) and an average for all of them together. For example, in 2012, on OECD average, 15.5 per cent of the participating international population (ages 16–65) scored at literacy Level 1 or below (OECD 2013, p. 257).[9]

In the underlying theoretical model, literacy is defined as

> the ability to understand, evaluate, use and engage with written texts to participate in society, achieve one's goals, and develop one's knowledge and potential (OECD 2013, p. 61).

In addition to the literacy scale, a "Reading Components" assessment was included in PIAAC's international "Survey of Adult Skills" (OECD 2013, pp. 59, 67). According to John Sabatini, the intention was to use the information collected through this additional "battery of reading component tasks" to "draw implications for policy, as well as for learning and instruction, for adults who score *at or below Level 1* in literacy proficiency" (Sabatini 2015, p. 2; emphases added).

There are also approaches to assessing basic reading and writing skills with continuous models, so-called "lower-rungs approaches" (Brooks, Giles et al. 2001), which think of the continuum as a ladder and take into account even barely measurable low proficiency levels. However, when complementing (rather than extending) PIAAC with the above-mentioned "battery of reading component tasks" (Sabatini 2015, p. 2), the OECD preferred a non-continuous model of three reading components and did not integrate these into the six-level literacy scale (Levels 1–5 and the "below Level 1" category). The three reading components the PIAAC add-on module tests participants on are (1) *word recognition,* (2) *sentence processing* and (3) *passage fluency* (Sabatini and Bruce 2009). It remains unclear why there have been no attempts up to now to find out whether it would be possible to link either these three components or the total set of component items to the PIAAC scale. Perhaps one reason is the theoretical quality of the three components. Since these components were developed independently of PIAAC, they are different from what is being tested on the overall literacy scale now (Strucker et al. 2007). However, the preparations for PIAAC did polish the reading component subtests in a way that made them suitable for international comparison (Sabatini and Bruce 2009). We assume that the theoretical differences may have decreased during this process while the similarity to the overall PIAAC literacy scale increased.

[8] The OECD conducted the first round of PIAAC in 2008–2013. Earlier comparative international adult assessments, also run by the OECD, include the International Adult Literacy Survey (IALS), conducted from 1994 to 1998, and the Adult Literacy and Life Skills Survey (ALL), conducted in two rounds between 2003 and 2008.

[9] The corresponding national figure for Italy, for example, was 27.7; for Germany it was 17.5; and for Japan it was 4.9 (OECD 2013, p. 257).

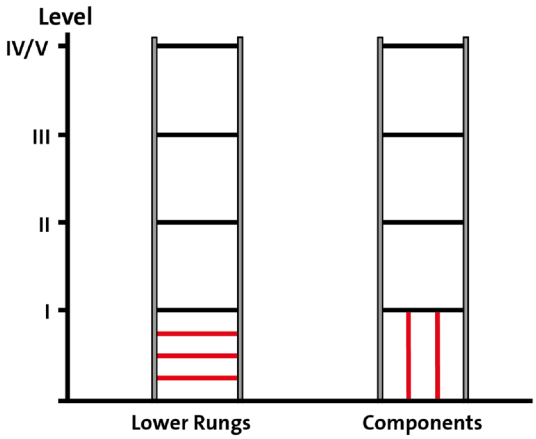


**Fig. 1** Lower-rungs vs. components approach

Unlike the PIAAC literacy scale, which builds on *item response theory* (briefly explained in footnote 7), the Reading Components in the add-on module are tested using *classical test theory*[10] methods (Yamamoto et al. 2013, p. 16; Zabal et al. 2014, p. 106). Again, it is not clear why this is so. It remains open to investigation whether it would be possible to run the reading component tests under an item response model as well, and also whether they would meet the necessary quality controls. If the answer to both of these questions turned out to be yes, the reading component tests would lose their full status as component tests, but they would gain the highly relevant quality of being statistically linkable to established international literacy scales.

Figure 1 illustrates the theoretical assumption about the main difference between rungs and components. While both are located inside the lowest level of literacy, labelled Level I in the graph, only the *lower rungs* claim to be hierarchical and part of the overall literacy scale. The *components* claim to be different elements of the reading process and thus non-hierarchical and non-comparable to the literacy scale. Both approaches are explained further below.

Both approaches compete with each other in assessing proficiencies of adults with low literacy skills. While the reading components approach was very fruitful in the early 2000s in Canada and the United States (US), the development of testing materials in the United Kingdom (UK) and in Germany focused on the lower-rungs approach. Among early versions of component approaches, the components were clearly differentiated and were linked to reading. When PIAAC chose to take the reading components approach on board in an add-on module, it became necessary to translate the test instruments already existing in individual countries (such as Germany, for example), and to reduce them to make them suitable for application in and comparison among a wide range of countries. It can be assumed that the components approach consequently became more similar to a (lower-) rungs approach than

[10] Classical test theory is sometimes also referred to as "true score theory", where the *true score* is an error-free measurement. Classical test theory was the predecessor of item response theory.

expected. The question is whether the reduction made to meet the needs of international comparability subsequently led the components to become hierarchical parts of one *latent variable*[11] (i.e. reading). We return to this question in a later section of this article.

## The lower-rungs approach and its implementation in the Level-One Survey (LEO) in Germany

A lower-rungs approach can be applied to describe and examine low skills in literacy. This means it enables differentiating the lowest level of the literacy scale more finely – in other words, "creating the lower rungs of the ladder" (Brooks, Davies et al. 2001a, b, p. 55). By including proficiencies "below Level 1", the lower-rungs approach extends the lower end of the established ranking of proficiency levels, which is based on a hierarchical and unidimensional model of literacy.

For example, the New Standards Level, developed in the UK in 2000 by the Basic Skills Agency (BSA) and the Qualifications and Curriculum Authority (QCA), comprised one "Entry Level", subdivided into Entry Levels 1–3 (E1, E2, E3), describing reading skills that are comparable to the range below IALS Literacy Level 1[12] (Brooks, Davies et al. 2001a, b; QCA 2005). These levels were applied in the Skills for Life survey conducted by the UK Department for Business, Innovation and Skills in 2011 (BIS 2012).

Another example is the Level-One Survey (LEO), which implemented four so-called Alpha Levels (α1 [letters], α2 [words], α3 [sentences], α4 [whole texts]) in Germany. They are based on theories about the acquisition of written language,[13] international large-scale assessments, national and international educational standards, and concepts of the practice of adult basic education (Dessinger 2011; Kretschmann 2011). Furthermore, the Alpha Levels were theoretically anchored within the IALS literacy scale (i.e. below IALS Level 1) by the level definitions and the "can-do" descriptions and characteristics for determining the level of difficulty (Grotlüschen 2011). Examples are provided in Fig. 2, which shows the "can do" descriptors of Alpha Level 3 in reading, and Fig. 3, which shows the "can do" descriptors of Alpha Level 4 in writing.

Furthermore, the Alpha Levels have had an influence on the development of instruments and tools for assessing adult literacy proficiency in Germany. The curriculum framework for literacy and adult basic education (DVV 2014), which contains guidelines for teaching and testing reading, writing and calculating in adult basic education, was developed following Alpha Levels 1–4 (ibid.).

[11] In item response theory, a latent variable is something which is not directly observable, but only inferred from other, directly observable variables.

[12] IALS measured literacy proficiency on a scale of 0 to 500 points. Adults who had achieved 0–225 points on that scale in the assessment scored at Level 1.

[13] For reading, see Bamberger and Vanecek (1984) and Coltheart et al. (2001); for writing, see Brügelmann (2000); Frith (1985); Kretschmann (2005); Reuter-Liehr (2008) and Spitta (1997).

## The reading components approach and its implementation in PIAAC

Representing basic "building blocks" of reading, component reading tasks also examine very foundational reading abilities, albeit not in a hierarchical order. Before the OECD added a reading components assessment module to the international assessment of PIAAC in 2012, the Statistics Canada research institute decided to implement a components approach in the Canadian part of the OECD's ALL Survey in 2003.

### Early Canadian and US-American national testing components

The reading components identified in Canada offered some additional information that differentiated among types of struggling readers. The advantage of a components approach was seen in its potential to offer insights into the different ways in which weak readers lag behind. Possible difficulties are insufficient vocabulary, difficulties with basic word decoding, inadequate strategies for dealing with new or complex texts, or general comprehension problems. Statistics Canada's expectation was that these differentiations would provide useful information to programme providers and policymakers (Murray 2001). Table 1 shows the components and tests which were discussed and subsequently recommended as being suitable for a household survey investigating adults' reading proficiency – in this case the Canadian ALL Survey, conducted in 2003.

The *Adult Reading Components Study* (ARCS; Strucker and Davidson 2003) carried out in the United States by the National Center for the Study of Adult Learning and Literacy (NCSALL) served Statistics Canada as a model for clustering adult learners into groups of reading skills levels. John Strucker and Rosalind Davidson tested 955 randomly selected learners from adult basic education (ABE) and English for speakers of other languages (ESOL) classes to assess their phonological awareness, rapid naming, word recognition, oral reading, spelling, vocabulary and background knowledge. Using a *cluster analysis*[14] methodology, they discerned ten clusters of reading skills levels in their sample which they deem relevant for effective teaching and learning (Strucker and Davidson 2003, p. 126).

Further components research was conducted jointly by John Strucker (NCSALL) as well as Kentaro Yamamoto and Irwin Kirsch from the Educational Testing Service (ETS), also in the United States. They took a sample of 1,034 adults and ran, among other things, a *latent class analysis* (LCA)[15] based upon participants' scores on:

[14] A cluster analysis methodology groups a set of data objects into clusters to analyse data distribution.

[15] A latent class analysis classifies individual test respondents into mutually exclusive types, or latent classes, based on their pattern of answers.

**Alpha Level 3: Reading**

**Central requirement: able to read at sentence level, predominantly construing reading and lexical reading of standard words**

*"Can do" descriptors*

- Can read single words within the context of a sentence
- Can read orthographically complex words
- Can make sentence–image connections
- Can read and understand sentences of increasing length
- Can read and understand subject-verb-object (SVO) sentences and SVO sentences with insertions
- Can follow simple instructions, especially when they include images
- Can read the TV programme including time designations

*Task characteristics*

Text length: 1–2 sentences
Sentence length: 5, 6, 7 or more words; main clause; main clause and subordinated clause
Sentence construction: SVO sentences; SVO sentences with insertions; Germanised foreign words and common Anglicisms
Typography: Arial, Times New Roman
Line spacing: double; left justification (line breaks after units of meaning)
Difficulty of words: increasing difficulty of words; up to level 3 (model of word difficulties)

**Fig. 2** "Can do" descriptors and task characteristics of Alpha Level 3 (reading) (translated from Kretschmann 2011, p. 53)

- oral vocabulary (PPVT);
- real word reading (TOWRE A);
- pseudo-word reading (TOWRE B);
- spelling; and
- short-term memory (digit span).

The result was a distinction of five classes of readers:

(1) proficient ABE, adult secondary education (ASE), and household sample readers with very strong decoding and vocabulary skills;
(2) ABE and ASE students with strong decoding skills that tend to undermine their vocabulary skills;
(3) advanced ESOL students with strong decoding but noticeably weaker English vocabulary skills;
(4) intermediate ESOL students with moderate weaknesses in decoding and vocabulary skills in English; and
(5) low intermediate ESOL students and reading disabled ABE native speakers with marked needs in decoding and vocabulary (Strucker et al. 2007).

**Alpha Level 4: Writing**

**Central requirement: able to write entire texts, albeit not flawlessly**

*"Can do" descriptors*

- Can use final-obstruent devoicing with adjectives*
- Can write the same consecutive letters in compound words
- Can use differences in lengths of vowels or consonants
- Can write the prefix "ver" correctly
- Can use *s* sounds correctly
- Can use abbreviations correctly
- Can capitalise definite abstract nouns
- Can recognise and write interfixes**
- Can write compound words together/separately
- Can use commas in lists
- Can write sentences at least phonetically

*Task characteristics*

Written element: words, sentences
Symbol length: number of letters up to 19; five-syllable words max.
Sentence length: 11 words max. (task context: text length 13 sentences max.)
Phoneme stage 4 (Reuter-Liehr 2008): consonant cluster with stop consonants: ck, tz
Phoneme stage 5 (ibid.): Elongation of vowels ie, ah, eh, üh, ih
Phoneme stage 6 (ibid.): ß (ss) at the start of a syllable
Use: up to CEFR*** B1: Words with a high degree of abstraction
Strategy: alphabetic, orthographic and morphemic

**Fig. 3** "Can do" descriptors and task characteristics of Alpha Level 4 (writing) (translated from Grotlüschen et al. 2010, p. 38). *Note: *Final-obstruent devoicing* means that spoken words like "*Hun-d*" [dog] sound as if they were spelled with a hard consonant at the end (*Hun-t* [do-k]), making it difficult to draw conclusions from the sound towards the spelling. ** *Interfixes* are the spoken gaps between syllables in compound words, e.g. "*Bus_halte_stelle"* [bus stop]. *** CEFR is the Common European Framework of Reference which has standardised proficiency levels from A1 (lowest), A2 and B1, B2 to C1, C2 (highest)

Further results of latent class analysis with component assessment data from the Canadian International Survey of Reading Skills (ISRS)[16] were published by the Canadian Council on Learning (Murray et al. 2008). The report distinguishes six groups (A1, A2, B1, B2, C and D) based on mother tongue, immigrant status and

[16] The International Survey of Reading Skills (ISRS) wss conducted by Statistics Canada in 2004 and 2005. For more information, see https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5070 [accessed 28 February 2020].

**Table 1** Table of recommended components and tests*

| | | |
|---|---|---|
| 1 | Word recognition | Test of Word Reading Efficiency (TOWRE) |
| 2 | Vocabulary knowledge | Peabody Picture Vocabulary Test (PPVT) |
| 3 | Listening comprehension | Ordinate Corporation PhonePass |
| 4 | Processing and memory | Digit span test |
| 5 | Processing and memory | Rapid automatised naming (RAN) of letters |

*Source:* Statistics Canada (Murray 2001)

**Note:* Many of the tests listed in this table have been revised and updated over the course of time. For more information about their original versions, see Torgesen et al. 1999 (TOWRE); Dunn 1959 (PPVT); Ordinate Corporation 1999 (Ordinate Corporation PhonePass); Wechsler 1997 (Digit span test); and Denckla and Cutting (1999). (RAN)

other key characteristics including age, gender, education and employment status (ibid.).

### International components suitable for comparative analyses

The developers of the reading components assessment in PIAAC 2012 applied none of the above-named tests, because they needed instruments that would enable international comparison. Whereas the developers' conceptual framework suggested five components, only three of these reading components made it into the final assessment set. Since languages vary in terms of their writing systems (alphabetic [e.g. English], syllabic [e.g. Japanese] or logographic [e.g. Chinese]), the PIAAC Reading Components test excluded tasks for alphanumeric perception and efficiency as well as tasks for word recognition and decoding (Sabatini 2015; Sabatini and Bruce 2009). Below, we explain the three remaining components and their task-sets.[17]

*Print vocabulary (word meaning).* To ensure cross-country comparability, the language chosen for this component in the PIAAC's add-on module was the local language being used in the respondents' neighbourhood, in the market and in popular media. The print vocabulary tasks are based on the assumption that adults know the meaning of everyday words from pictures and from listening. The 34 print vocabulary tasks assess whether a person also knows their meaning from print. For this purpose, the respondent is given a four-item multiple choice list and asked to circle the correct word that represents the meaning of an image. Thus the print vocabulary task-set seeks to determine whether individuals can identify everyday words of their local language in print.

*Sentence processing.* To ensure this component's cross-country comparability, the tasks in this set were created without varying the grammatical/syntactic complexity of the sentences. Variation was, however, taken into account in the length of sentences within a basic grammatical structure, and also in the logical relationships that comprise meaning. These variations were designed with increasing difficulty and therefore indicate the individual's proficiency at constructing basic meaning from

[17] Examples of the English-language reading component items are published in Sabatini (2015).

print (Sabatini 2015, pp. 7, 11). The 22 sentence-processing tasks ask an individual to judge "whether the sentence makes sense in relation to common knowledge about the world […] or based on the internal logic of the sentence" (ibid., p. 12). Therefore, a "yes" or "no" answer represents a 50 per cent guess probability. Thus the sentence processing tasks assess the individual's proficiency in applying his or her language skills in the context of printed text.

*Passage comprehension.* The passage comprehension task-set measures fluent, efficient reading performance. The 44 passage comprehension tasks are embedded in four short basic text passages designed for adult readers. In each task, respondents are asked to choose between a word that correctly fits a sentence in a passage and a second option that a skilled reader would recognise as being obviously wrong. Although reading fluency and efficiency are usually assessed by giving participants only a fixed amount of time to do the task, PIAAC 2012 allowed them as much time as they needed. The individual total time required to complete it was recorded, and average reading rates were compared afterwards. The purpose of this was to prevent biases, caused by cross-country comparison, because differences between languages, writing systems and cultural variables were expected to affect average reading rates (Sabatini and Bruce 2009, p. 13). In their conceptual framework, John Sabatini and Kelly Bruce explain that "the time to complete will add very little additional information" about the skills of "the very low-skilled beginning reader", but low-ability adults with high accuracy scores within the passage comprehension tasks can be identified by this measurement, because they need more time to complete than the subsample of skilled readers in each country (Sabatini and Bruce 2009, p. 13).

Table 2 sums up the differences between the lower-rungs approach and the reading components approach and their development for PIAAC.

## Research questions

Having elaborated the differences between the lower-rungs approach and the reading components approach in the previous sections of this article, we now discuss both approaches with regard to their potential suitability for monitoring SDG target 4.6 globally, which then leads to our presentation of our own research.

To assess lower reading skills, PIAAC 2012 opted for a components approach rather than a lower-rungs approach. There are two possible reasons for this. First, the design of the survey suggests there was no plan to link the Reading Components to the continuous literacy scale. The Reading Components assessment was implemented as a new domain and as an *optional* element of the assessment in Round 1 (2011–2012) of PIAAC's first cycle. Furthermore, it was provided in pencil-and-paper format, whereas the main assessment was designed in a computer-based format (Kirsch and Thorn 2013). This certainly limits the comparability of both measures.

Second, Sabatini and Bruce state that even in theory, the components "do not strictly develop hierarchically" during the acquisition of reading skills (Sabatini and Bruce 2009, p. 7). Therefore, it might be inadequate to treat them as having a clearly hierarchical order.

**Table 2** Summary of differences between the lower-rungs approach and the reading components approach before and during PIAAC

| | Lower-rungs approach (Brooks, Davies et al. 2001a, b) | Reading components approach as administered in Canada and the United States (Murray 2001; Strucker et al. 2007) | | Reading components approach as administered in PIAAC and STEP (Sabatini and Bruce 2009) |
|---|---|---|---|---|
| Theoretical assumption about what is being tested<br>and<br>Instruments which the test builds on | Literacy<br>as defined in IALS, with can-do-descriptions for the lower rungs<br>Skills for Life Survey (UK):<br>Development of 25 new test items that are easier than the lower IALS levels | (1)Word recognition | Test of Word Reading Efficiency (TOWRE) | Reduction according to international requirements<br>Print vocabulary<br>Sentence processing<br>Passage fluency |
| | | (2)Vocabulary knowledge | Peabody Picture Vocabulary Test (PPVT) | |
| | | (3)Listening comprehension | Ordinate Corporation Phone-Pass | |
| | | (4)Processing and memory | Digit span test | |
| | | (5)Processing and memory | Rapid automatised naming (RAN) of letters | |
| | LEO (Germany):<br>Development of 71 new test items that are easier than the lower PIAAC levels, testing both reading and writing | Subtests<br>(Strucker et al. 2007):<br>oral vocabulary (PPVT),<br>real word reading (TOWRE A)<br>pseudo-word reading (TOWRE B)<br>spelling<br>short-term memory (digit span) | | |
| Statistical testing model | Item response theory | Classical test theory | | Classical test theory (so far), but the dataset can be re-run with item response theory, as demonstrated in this article |
| Linkable to existing international literacy scale | Yes (IALS, PIAAC) | No | | No |

**Table 2** (continued)

| | Lower-rungs approach (Brooks, Davies et al. 2001a, b) | Reading components approach as administered in Canada and the United States (Murray 2001; Strucker et al. 2007) | Reading components approach as administered in PIAAC and STEP (Sabatini and Bruce 2009) |
|---|---|---|---|
| Surveys which apply this model | Skills for Life (UK)<br>Level-One Survey (LEO) in Germany | The OECD's Adult Literacy and Lifeskills Survey (ALL) in Canada; in English and French<br>UNESCO's Literacy Assessment and Monitoring Programme (LAMP) in several UNESCO Member States | PIAAC (Round 1, 2011–2012) as a voluntary add-on, pretested in several languages and countries<br>The World Bank's Skills Towards Employability and Productivity (STEP) skills measurement programme in several middle-income countries in several world regions<br>The OECD's Programme for International Student Assessment (PISA) for Development survey (in preparation) |
| Applicable for the Global Alliance to Monitor Learning (GAML) for monitoring SDG target 4.6 (adult literacy and numeracy) | No, because translation and pretesting are too expensive | No, because translation and pretesting are too expensive | Not clear, because not linked to international scale; but linkage seems potentially feasible and the necessary checks are provided in in this article. For linkage, the items are treated as if they were lower-rungs items and as if they would test only one skill (literacy), not three components |

However, the published results of the PIAAC Reading Components assessment (OECD 2013) as well as the progression of the components (from words to sentences to text passages) could point to a hierarchy among the three different types of the assessed reading component tasks.

## A hierarchy of difficulty?

The published average proportions of the correctly answered reading component items show differences among the three dimensions (print vocabulary, sentence processing and passage comprehension). The highest average proportions of correct answers were reached for the print vocabulary dimension, whereas the lowest were reached for the sentence processing items. This result is stated independently of the individual literacy level. Furthermore, this is not only true for the German data, but also for the OECD average (OECD 2013, pp. 416–418).

Table 3 shows the average proportion of correctly answered reading component items by literacy proficiency level for the German sample. From this table it is reasonable to assume that the print vocabulary items are the easiest, and the sentence processing items are more difficult than the passage comprehension items.

Also, Sabatini states for the US reading components:

> One may have noticed that sentence and passage reading means were closely aligned across the higher levels of literacy proficiency, with passage means sometimes higher than sentence means toward the higher proficiency levels. This is because the most difficult sentence items are typically more difficult than any of the passage items. Thus, even adults who are relatively more proficient may still make errors on these challenging sentence items while likely finding all passage items relatively easy to answer (Sabatini 2015, p. 16).

Table 4 shows the average time spent completing a reading component item, in seconds, by PIAAC literacy proficiency level for the German sample. Here, too, print vocabulary emerges as the easiest dimension, because the average time spent on completing these tasks is comparably the shortest for all literacy levels. But responding to the passage comprehension items takes a little longer than answering

**Table 3** Average proportion of reading component items answered correctly, in per cent, by PIAAC literacy proficiency level (German sample)

| | Below Level 1 | Level 1 | Level 2 | Level 3 and above |
|---|---|---|---|---|
| Print vocabulary ($n=817$) | 93.6 | 97.2 | 98.5 | 99.3 |
| Sentence processing ($n=809$) | 75.8 | 87.3 | 94.0 | 97.5 |
| Passage comprehension ($n=785$) | 81.0 | 90.9 | 96.5 | 99.1 |

*Source:* PIAAC and PIAAC Reading Component datasets, own calculations

*Note:* Deviations from the OECD table (OECD 2013, pp. 416–417) are due to our exclusion of respondents who did not reach the end of the Reading Components assessment despite their not being given a time limit.

**Table 4** Average time spent completing a reading component item, in seconds, by PIAAC literacy proficiency level (German sample)

| | Below Level 1 | Level 1 | Level 2 | Level 3 and above |
|---|---|---|---|---|
| Print vocabulary ($n=817$) | 7.2 | 5.7 | 4.7 | 3.8 |
| Sentence processing ($n=809$) | 16.7 | 11.7 | 9.1 | 7.2 |
| Passage comprehension ($n=785$) | 17.3 | 13.4 | 9.9 | 7.6 |

*Source:* PIAAC and PIAAC Reading Component datasets, own calculations

the sentence processing items. Therefore, in terms of time spent on completing the tasks, it is reasonable to assume that the passage comprehension items are more difficult than the sentence processing items.

This pattern is also the same for the OECD average across all participating countries (in Round 1 of PIAAC's first cycle) for time spent completing the reading component items (OECD 2013, pp. 417–418).

Considering these results, the research questions (RQ) we decided to investigate in our own research, presented in this article, were:

RQ1 *Is it possible to describe the PIAAC reading component items (in the German PIAAC questionnaire) hierarchically by their difficulty?*
RQ2 *Provided that it is possible, what kind of hierarchical relationship exists among the three components and across all items?*
RQ3 *If the Rasch model proves unsatisfactory, does a 2PL Birnbaum model fit the reading component data better?*[18]

## Methodology

In addressing our research questions, we applied methods of item response theory (IRT) to the German sample of the PIAAC Reading Components data. IRT provides probabilistically combined results regarding respondents' *trait level* (competences) and *item properties* (difficulties) based on the probability of a correct response to a test item (Embretson and Reise 2000).

The simplest item response model, the so-called *Rasch model,*[19] assumes that the probability of a specified response depends on two variables: the respondent's trait

[18] In a nutshell, both models (further explained in the methodology section) serve to statistically estimate the probability of survey respondents' correct answers for test items of varying difficulty. While the *Rasch model* is a so-called one-parameter logistic (IPL) model, a *Birnbaum model* has more than one parameter. The one we used for our research was a two-parameter logistic (2PL) Birnbaum model.

[19] Named after Danish mathematician, statistician and psychometrician Georg Rasch, this model estimates test reliability in proficiency assessments where there are only two options for answering test items, either correctly or incorrectly. It considers the probability of a respondent with a certain aptitude choosing a correct answer, factoring in the difficulty of tackling that item.

level and the difficulty of the test item (Embretson and Reise 2000, pp. 48–51). If a respondent's trait level exceeds the difficulty of the item, then there is a strong possibility that this person will respond correctly to the item. If the difficulty of the item exceeds the respondent's trait level, there is a strong possibility that this person will not respond correctly to the item. In other words, the more difficult an item is, the less likely it is that a person with a particular trait level will respond correctly to this item (Embretson and Reise 2000, p. 49).

In our research, we focused in particular on the item difficulties of all three reading component items, re-analysing them in terms of their hierarchical relationship. For this purpose, we chose the one-parameter logistic Rasch model, because it is particularly suitable for estimating and scaling test items on a common scale, ordered by their difficulties. In case of model conformity, the Rasch model has the property of specific objectivity. This means that differences in terms of item difficulties can be stated independently of the sample's skills distribution (Embretson and Reise 2000; Moosbrugger 2012, p. 49).

A necessary precondition of IRT analyses is the assumption of item homogeneity and local independence, meaning that all item responses depend on the same latent variable and that, given the model parameters, no further relationships exist in the data (Embretson and Reise 2000, p. 60). One important advantage of the Rasch model is that it provides appropriate and strict model fit criteria to evaluate item homogeneity and item quality.

We carried out the estimation of a one-dimensional Rasch model using ConQuest software. Sabatini states that the translation of reading component items across languages may result in different item level difficulty estimates (Sabatini 2015, p. 11). Therefore, the analysis we present here refers to the reading component data from only one country (Germany). Our input file contained the full response data of the German sample in the PIAAC Reading Components assessment based on the reduced version of the *German PIAAC Scientific Use File* (SUF; Rammstedt et al. 2015).

The Reading Components sample for Germany comprises 822 cases, whereas the whole German PIAAC sample comprises 5,465 cases. Therefore, the sample is not representative for the German adult population. Furthermore, Claudia Tamassia et al. note in the OECD's *Technical Report of the Survey of Adult Skills (PIAAC)* that the criterion for routing respondents into the paper-based reading components assessment was not only lower literacy and numeracy skills, but also a lack of experience in handling a computer.

> [The] paper-based assessment was administered to respondents who either reported they had no computer experience; failed the test of basic computer skills required to take the assessment; or refused to take the assessment on the computer (Tamassia et al. 2013, p. 2).

As a consequence of this routing process, the German sample contains relevant proportions of respondents with higher literacy skills, who solved the reading

components tasks with ease, while a relevant proportion of adults with lower literacy skills also remained in the sample. Across the entire 23-country PIAAC sample, an above-average proportion of 31 per cent of the adults who took the Reading Components assessment (compared to 15.5 per cent total) scored at or below Level 1 (Sabatini 2015, p. 9).

Comparison of the sociodemographic bias of the German sample against that of the German adult population as a whole can be described by, for example, a higher mean age and a higher proportion of adults who speak German as a second language.[20]

The dataset we analysed comprised responses for a total of 100 PIAAC reading component items. These were 34 print vocabulary items (numbered 1–34), 22 sentence processing items (numbered 35–56) and 44 passage comprehension items (numbered 57–100). For our IRT analysis, we recoded the response data into dichotomous (0/1: incorrect/correct) data. Missing values were treated as follows: in cases where questions had been skipped (refused or not done), we recoded missing values into incorrect responses; in cases where the whole reading components assessment was broken off, we recoded the first missing value into an incorrect response and left all further missing values as missing. Afterwards, we estimated, mapped and analysed the item parameters and evaluated the quality of the items. We checked the Rasch model fit by *weighted mean squares* (MNSQ). A perfect item fit in terms of mean squares would be 1.0 (Wu et al. 2007, p. 54). For this research study, we chose MNSQ $\geq$ 1.33 as criterion for a bad item fit (Wilson 2005, p. 129; Grotlüschen et al. 2012, p. 63). Furthermore, we illustrated and described the distribution of item difficulties by a *Wright map* (see results section).[21]

Subsequently, we compared the results to the outcomes of a two-parameter logistic (2PL) *Birnbaum model,*[22] which considers varying item discriminations. Since items differ in their discriminating power, trait level estimates depend on the specific patterns of success and failure in the item set. In contrast to the Rasch model, items do not have equal weight in estimating trait levels (Embretson and Reise 2000, p. 53). We estimated the 2PL model using Mplus 7 softwareJavaScript:TypeChar(73).

[20] Beatrice Rammstedt and Britta Gauly of the Leibniz Institute for the Social Sciences (GESIS, in Mannheim, Germany) are currently preparing their analysis of the sociodemographic data of the German reading components sample for publication.

[21] Named after American psychometrician Benjamin D. Wright. a Wright map is an item map, which "is organized as two vertical histograms. *The left side shows candidates and the right side shows items.* The left side of the map shows the distribution of the measured ability of the candidates from most able at the top to least able at the bottom. The items on the right side of the map are distributed from the *most difficult* at the top to the *least difficult* at the bottom" (Lunz 2010; emphases in original).

[22] Named after Polish-American mathematician and statistician Zygmunt Wilhelm Birnbaum, this model extends the one-parameter (1PL) Rasch model by one or two more parameters (resulting in a 2PL or a 3PL model), factoring in the possibility that respondents' answers might be the result of guessing.

## Results

### RQ1: Is it possible to describe the PIAAC reading component items (in the German PIAAC questionnaire) hierarchically by their difficulty?

As a main result of our analysis, we found that the applied Rasch model confirmed the possibility of representing the 100 reading component items on a hierarchical scale (i.e. the overall answer to our first research question seemed to be yes). The mean squares of most items ($n = 92$) met the model fit criterion (MNSQ $\leq$ 1.33). Only eight items, in this analysis, did not meet this criterion. These were one item (item 17) from the print vocabulary item set, and seven items (items 39, 40, 44, 45, 50, 51 and 56) from the sentence processing item set.[23] Their mean squares range from 1.33 (item 45) to 1.47 (item 40). On the one hand, these items are characterised by very low discriminations. This could mean that respondents with higher abilities are not more likely to solve them than respondents with lower abilities. On the other hand, the unsatisfactory item fits could also indicate that these items do not fit a one-dimensional construct of the kind we applied here (Rost 2004, p. 98; Kelava and Moosbrugger 2012, p. 86).

Seven out of the eight unsatisfactory items belong to the sentence processing item set, indicating that roughly one-third of the sentence processing scale either does not discriminate well, or might be testing something other than sentence processing. Respondents are asked to check the sentences in terms of whether they make sense; it is possible that people only check whether they are grammatically correct without deciding whether or not they are reasonable.

Nevertheless, our Rasch analysis of the Reading Components data did result in a *coefficient alpha*[24] of 0.95. This indicates an overall internal consistency, although this is also not a measure for item homogeneity (Schermelleh-Engel and Werner 2012, p. 132).

### RQ2: Provided that a hierarchical description of the PIAAC reading component items (in the German questionnaire) is possible, what kind of hierarchical relationship exists among the three components and across all items?

With the overall answer to our first research question being yes, we then addressed our second research question. The Wright map in Fig. 4 shows the results for our analysis of the 100 Reading Components items in the German PIAAC sample when applying a one-parameter logistic Rasch model. This map of latent distributions and

[23] Unfortunately, we are unable to provide descriptions of single items, because the reading component tasks are treated as strictly confidential by the OECD and the Educational Testing Service (ETS).

[24] A coefficient alpha (or Cronbach's alpha) measures the reliability of the average of questionnaire items.

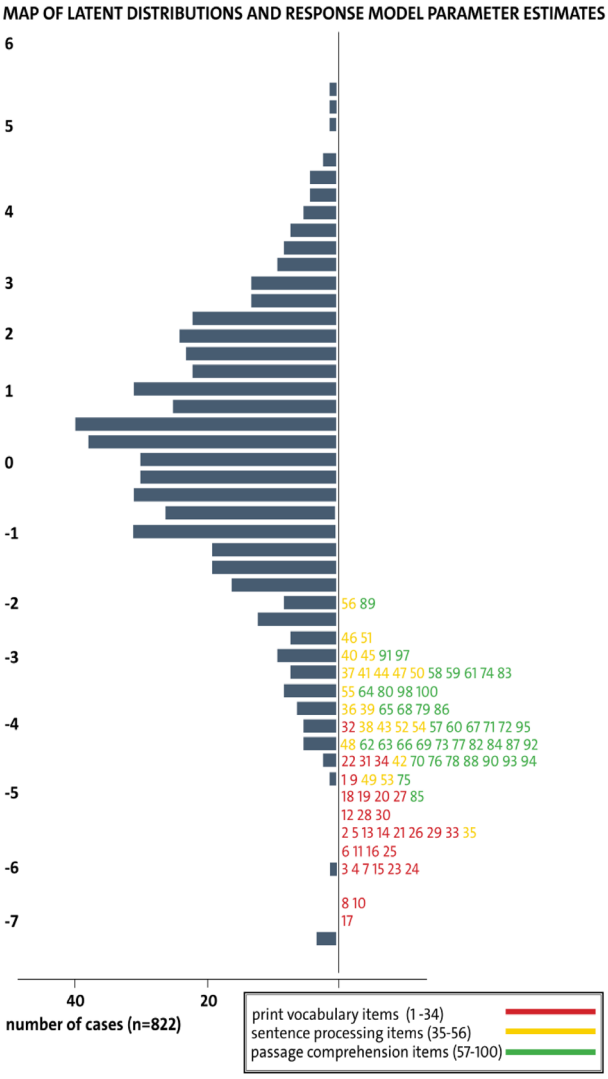


**Fig. 4** Map of latent distribution and response model parameter estimates for the German PIAAC 2012 sample. *Note:* The vertical axis designates level of difficulty

response model parameter estimates displays a joint hierarchical scale. The horizontal axis designates the number of cases/respondents; the vertical axis designates the level of difficulty. The scale is adjusted in a way that uses "zero" as the average competence of the sample. This gives the difficult items a positive value and the easier items a negative value.

The left-hand panel shows a representation of the latent reading competencies distribution, and the right-hand panel indicates the difficulty of the test items. Each number represents one item and the items are plotted according to their difficulties. Here, the difficulties range from –6,86 to –2,14. Item 56 and item 89 have the highest item difficulties, so they are plotted at the top of the figure, while item 17 has the lowest item difficulty, so it is plotted at the bottom of the figure. According to Rasch's model, a person with a latent ability estimate that corresponds to the level at which the item was plotted would have a 50 per cent chance of success on that item (Wu et al. 2007).

As expected, the item difficulties are located clearly below the average of the competence distribution. This means that the majority of the sample responding to the Reading Components add-on was able to solve most of the items correctly.

When comparing the item difficulties of the three components, it is evident in Fig. 4 that most of the print vocabulary items (numbers 1–34, shown in red) are relatively easy, as expected. Furthermore it is noticeable that the sentence processing items (numbers 35–56, shown in yellow) and the passage comprehension items (numbers 57–100, shown in green) have higher difficulties, but mix with each other relating to their difficulties. Therefore, the implicit assumption of a components-related hierarchy of the three scales cannot be confirmed. However, while print vocabulary, sentence processing and passage comprehension are not *clearly* ordered like lower rungs, the general trend is that words (print vocabulary) are easier than sentences (sentence processing); which are easier than short texts (passage comprehension). Thus the test items do form a hierarchy. Even under the rather strict assumption of Rasch homogeneity, all but 8 items meet the model fit requirements.

The *print vocabulary* items numbered 17, 8 and 10 have the lowest item difficulties, ranging from − 6,86 to − 6,51. It is worth noting that the correct answers for these three easier items are monosyllabic, which might explain their position on the Wright map.[25] Further up in the map, the most difficult print vocabulary items do already mix with items from both the sentence processing and the passage comprehension components.

Within the *sentence processing* component, item 35 is the one with the lowest difficulty. This seems to be reasonable, because the sentence consists of one definite article, one subject and one verb in simple past form. Items 49 and 53 also have low difficulties, but their sentence structures are far more complex. For example item 49 involves an interjectional relative clause, and the length of the text comprises 14 words. By contrast, items 36, 37 and 41 have higher item difficulties, although they are main clauses and their lengths range from four to eight words. This order of difficulties disagrees with Sabatini's theoretical description, which states that the sentence processing items in the test booklet would rise in terms of their difficulties (Sabatini and Bruce 2009, p. 11).

The difficulties of the *passage comprehension* items concentrate on the range between –4,95 and –2,18. Each of the passage comprehension tasks requires the respondents to choose between two words within a short text. According to the

[25] As mentioned earlier, for legal reasons we are unfortunately unable to provide more differentiated descriptions and interpretations of single items.

results of the Rasch analysis, items 85 and 75 have the lowest item difficulties, whereas item 89 is the most difficult one. The varying difficulties of passage comprehension items could depend on the length and familiarity of the words, the abstractness of the word meaning, and how obviously they seem to be correct answers in the context of the text passage.

### RQ3: If the Rasch model proves unsatisfactory, does a 2PL Birnbaum model fit the reading component data better?

According to Kentaro Yamamoto et al., for PIAAC, "a common set of item parameter estimates of the two-parameter logistic (2PL) model and the general partial credit model (GPCM)[26] was estimated and found to fit quite well to all countries" (Yamamoto et al. 2013, p. 16), i.e. not with a simple Rasch model. Indeed, the Rasch model assumption of homogeneous item discrimination is often non-realistic and artificial. More sophisticated models can cope with inhomogeneity of discrimination.

As already mentioned, we found that eight reading component items showed poor Rasch model fit, that is, they did not discriminate the same way as the others or they did not test the same latent variable. For these reasons we estimated a two-parameter logistic (2PL) Birnbaum-model in order to check the item difficulties taking different discrimination characteristics into account. The item discriminations ranged from 0.73 to 11.91.

All in all, we found that the two-parameter logistic Birnbaum model fit the Reading Components data better than the one-parameter logistic Rasch model. In comparison (see Table 5), the 2PL model fits show lower Akaike and Bayesian information criteria (AIC and BIC)[27] and sample-size adjusted BIC and should therefore be preferred (de Ayala 2009, pp. 141–142).

## Discussion

### Component items do also function as hierarchical test items and therefore meet GAML requirements

To sum up, we found that the first research question (Is it possible to describe the PIAAC reading component items [in the German PIAAC questionnaire] hierarchically by their difficulty) can be answered positively. Two different approaches (applying the Rasch model and the Birnbaum model) show that the component approach at least partly contains hierarchical item difficulties.

Our second research question investigated the kind of hierarchical relationships existing among the three components and across all items. We found that

[26] A *general partial credit model* (GPCM) allows for partially correct solutions, while the Rasch (1PL) model only allows right or wrong responses.

[27] Akaike and Bayesian information criteria (AIC and BIC) are used in model selection to avoid overfitting. Models with lower AIC and BIC values are preferable to those with higher ones.

**Table 5** Model fit of 1PL and 2PL (both calculated using Mplus software)

| | 1PL Model (Rasch) | 2PL Model (Birnbaum) |
|---|---|---|
| Number of free parameters | 101 | 200 |
| Log likelihood | | |
| H0 value | − 11,126.158 | − 10,541.422 |
| H0 scaling correction factor for multiple linear regression (MLR) | 0.9914 | 1.0519 |
| Akaike information criterion (AIC) | 22,454.316 | 21,482.844 |
| Bayesian information criterion (BIC) | 22,930.202 | 22,425.192 |
| Sample-size adjusted BIC ($n^* = (n+2)/24$) | 22,609.464 | 21,790.068 |

while the print vocabulary scale is easier than the two others, the latter have internal hierarchies but mix with each other in terms of difficulty. Our first method, which applied the Rasch model, showed unsatisfactory item fits for 8 out of 100 items, with 7 of them belonging to the 21-item sentence processing subscale.

Our third research question investigated whether a two-parameter logistic model would lead to better fit values. The results indicate that the model fit was indeed better and that the reading components approach as used in PIAAC can also be interpreted as a hierarchical scale modelling a latent variable that could be called "reading".

Our findings indicate an overall hierarchy of the Reading Component items, although two of the dimensions, namely sentence processing and passage comprehension, cannot be clearly separated in terms of the rise in difficulty. Reading comprehension in a single sentence as distinct from the comprehension of a multi-sentence text section is not tested selectively in the PIAAC assessment tasks. This is certainly a consequence of choosing especially those reading tasks for the assessment that are less language-specific in order to improve the international comparability.

Moreover, our findings demonstrate that the Reading Component test set under PIAAC 2012 also works as a hierarchy which would indeed be linkable to an international literacy scale. The test items are available in many languages. This already enables usage of component items in a wide range of countries. Many of the subsets have been applied under PIAAC, STEP, LAMP or even IALS, ranging across several supra-national organisations and thus indicating that the items are widely accepted (which would probably be more difficult if the items were purely owned by the OECD or ETS).

One conclusion of our research therefore is that it is technically possible to use the full set of PIAAC reading component test items to meet the requirements of the Global Alliance to Monitor Learning (GAML) initiative's efforts to address all ten targets of SDG 4. Participating UN Member States can add the tests to national micro-censuses or similar surveys. Findings can be displayed in a hierarchy that is comparable across countries because of its linkability to an anchor literacy scale (e.g. PIAAC).

### Tests that were developed under the Reading Components scheme become disconnected from their origins when they are made internationally comparable

To break down the reading proficiency within the lowest literacy level ("below PIAAC Level 1") into more differentiated categories, a lower-rungs approach was developed in Europe (in the UK and Germany) and a reading components approach was developed in the United States and Canada. Both have advantages and disadvantages. The most recent and widespread version of the Reading Components is the one used in PIAAC and STEP. It differs from earlier versions, because it was adjusted for the purpose of being applicable in different countries, settings, languages and scripts. While these adjustments and test development efforts polished the test (Sabatini and Bruce 2009), one unavoidable side effect was the blurring of some of the clear differences which had been discernible among earlier components (Strucker et al. 2007).

Earlier component versions differed much more from each other and were more closely linked to different aspects of reading. One aspect, for example, was the strategy of letter-by-letter-decoding of unknown words, mostly tested by using nonsense words (i.e. in the TOWRE test). Another aspect was the existence of *lexical memory*[28] entries according to a lexical strategy of reading where fast word recognition is required. This can be tested with word recognition tests (TOWRE, PPVT). These two aspects can be interpreted by using Coltheart's *dual-route theory of reading* (Coltheart et al. 2001)[29] and they show up in readers with different kinds of dyslexia, requiring different treatments. Both are different from tests on language and vocabulary or tests on grammar, which indicate low language proficiencies – and thus require provision of language lessons rather than making efforts to improve learners' decoding or memorising skills. Another aspect has been the test of short-term memory, attention or concentration. Many foreign-born readers may have excellent short-term memories, while locally born struggling readers may not because of generally low cognitive skills. The latter may indicate learning disabilities but may also need psychological treatment. Earlier reading component approaches also tested listening and differentiation skills, phonemics or *phonemic awareness*.[30] In cases of low test results, training would focus on syllables and rhymes, precise pronunciation and listening skills. Less important for reading but a good indicator for literacy proficiency are spelling skills which require a good command of writing skills as well. Overall, the earlier versions of reading components provided in-depth knowledge about adequate pedagogical treatment. The problem is that these tests do not work for comparative studies of surveys conducted using different language and letter systems. Most of the nationally developed reading components correlate very closely with the phonemic characteristics of particular languages and their written equivalents (Sabatini and Bruce 2009). For these reasons, it is rather difficult to develop

[28] Lexical memory refers to being able to remember particular written words as pictures.

[29] The "dual-route theory of reading [concerns] the 2 tasks most commonly used to study reading: lexical decision and reading aloud" (Coltheart et al. 2001, p. 204).

[30] Phonemic awareness refers to the ability to discern distinctly separate units of sound in a particular language which determine the meaning of a word. Example: being able to distinguish between d and t in the words bad and bat.

test items that still keep a close relation with the theoretical explanations and are internationally comparable.

In sum, useful information from earlier component versions (covering the dual-route theory of reading, short-term memory or learning disabilities; language or grammar and vocabulary; phonemic awareness, grapheme-morpheme-correspondence or spelling) has been lost in the efforts of trying to make items internationally comparable. Thus, as we already assumed before embarking on our research, an internationally comparative approach at this level indeed proves to be extremely difficult and, in cases where it does work, loses the components character, shifting slightly towards a lower-rungs approach.

While on the one hand components lose their strong connection to the original theoretical background, lower rungs in recent years have tended to be described in more detail, providing rich didactical insights and knowledge. The can-do descriptions provide a good example of better theoretical knowledge (see also Durda et al., in this issue).

#### Limitations: custody of an international literacy scale – who owns it?

There is no such thing as the one and only common literacy scale, even though the items used in the PIAAC add-on module have proven to test literacy in a hierarchical order. Further research with open and large datasets would be necessary to link them to the overall PIAAC scale or any other international literacy scale. For the moment, the OECD holds custody of its PIAAC scale, and UNESCO's LAMP component datasets are not large enough to run the necessary analyses. The dilemma remains the same. GAML has to avoid implementing a single scale and definition with a single test in possibly all UN Member States, because researchers claim that this would lead to a monopoly (Addey 2018) and re-colonisation of the so-called Global South (Grotlüschen 2018). The current solution (UIL 2019) is to propose two reporting levels according to Member States' income category.

Moreover, the tests in the PIAAC add-on module were developed for industrialised countries. They still have a blind spot at a certain point that lies between virtually no reading skills and the easiest test item. This section may be highly relevant for low-income countries.

## Recommendation: re-analyse LAMP and STEP items and refine the theoretical approach to assessing reading proficiency in an internationally comparable manner

At this point it seems necessary to re-run analyses of reading components data from several other surveys like LAMP and STEP in order to find out whether they deliver similar hierarchies and, if not, whether eliminating some items might improve the scales. Another necessity is to discuss a common anchoring scale. This would enable countries to develop further and perhaps even easier test items, co-run them in their national surveys and link them to the existing set.

More theoretical work is needed for the development and interpretation of tests at the very lowest levels of literacy (see Durda et al., in this issue). Lower rungs can be described according to what proficiency the items require or according to can-do-descriptions, i.e. Alpha Levels with 7–10 can-do descriptions on each level for both reading and writing. This would provide detailed knowledge via the descriptions of lower rungs. For surveys to be run in Germany, an adult education curriculum with formative assessment tools has already been developed based on the lower rungs level descriptions.

Hence, to improve learning outcomes within the GAML initiative, instead of trying to find a language-independent set of test items, it would be appropriate to reconsider the advantages of a lower-rungs approach for the international assessment of reading skills, either to supplement the components approach or to leave the language-related area of "below Level 1" research to UN Member States.

**Acknowledgements** Open Access funding provided by Projekt DEAL. The research presented in this article was conducted within a project called "Reading Components and lower competencies – a better understanding of adults with lower competence" which is being funded by the German Federal Ministry of Education and Research under the funding code PIAACRC2. The data we used were provided by the Leibniz Institute for the Social Sciences GESIS in Mannheim, Germany. In this article we publish and discuss new empirical results of our own analysis of the German PIAAC datafile ("Scientific Use File: Germany") for secondary analysis. We also contribute to the development of measurement instruments for the assessment of lower literacy skills. Carol Bennett provided professional language editing services. Sole responsibility for content lies with the authors of this article, all of whom confirm that they have approved the manuscript for submission. We confirm that the content of the article has not been published or submitted for publication elsewhere. The authors declare that they have no competing interests.



## References

Addey, C. (2018). Assembling literacy as global: The danger of a single story. In M. Milana, J. Holford, S. Webb, P. Jarvis, & R. Waller (Eds.), *The Palgrave international handbook of adult and lifelong education and learning* (pp. 315–335). London: Palgrave Macmillan UK.

Bamberger, R., & Vanecek, E. (1984). *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache* [Reading – understanding – learning – writing: The difficulty levels of texts in German]. Frankfurt am Main: Diesterweg.

BIS (Department for Business Innovation and Skills) (2012). *The 2011 Skills for Life Survey: A survey of literacy, numeracy and ICT levels in England*. London: BIS. Retrieved 21 February 2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/36000/12-p168-2011-skills-for-life-survey.pdf.

Brooks, G., Giles, K., Harman, J., Kendall, S., Rees, F., & Whittaker, S. (2001). *Assembling the fragments. A review of research on adult basic skills.* Research report no. 220. London: Department for Education and Employment (DfEE).

Brooks, G., Davies, R., Duckett, L., Hutchinson, D., Kendall, S., & Wilkin, A. (2001). *Progress in adult literacy. Do learners learn?*. London: Basic Skills Agency.

Brügelmann, H. (2000). *Kinder auf dem Weg zur Schrift. Eine Fibel für Lehrer und Laien*. [Children on the way to writing: A primer for teachers and laypeople]. Bottighofen: Libelle.

Carlson, J. E., & von Davier, M. (2013). *Item response theory*. ETS R&D Scientific and Policy Contributions Series, ETS SPC-13-05. Princeton, NJ: Educational Testing Service (ETS). Retrieved 6 February 2020 from https://www.ets.org/Media/Research/pdf/RR-13-28.pdf.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Denckla, M. B., & Cutting, L. E. (1999). History and significance of rapid automatized naming. *Annals of Dyslexia, 49*(1), 29–42.

Dessinger, Y. (2011). Kompetenzmodelle des Schriftspracherwerbs [Competence models of written language acquisition]. In A. Grotlüschen, R. Kretschmann, E. Quante-Brandt, & K. D. Wolf (Eds.), *Literalitätsentwicklung von Arbeitskräften [Literacy development of the labour force]* (pp. 68–85). Münster: Waxmann.

Dunn, L. M. (1959). *Peabody picture vocabulary test*. Circle Pines, MS: American Guidance Service.

DVV (Deutscher Volkshochschul-Verband) (2014). *Rahmencurriculum und Kurskonzept für die abschlussorientierte Grundbildung* [Framework curriculum and course concept for graduation-oriented basic education]. Bonn: DVV. Retrieved 13 July 2016 from https://grundbildung.de/projekte/rahmencurriculum/material.html.

Elfert, M. (2019). Lifelong learning in Sustainable Development Goal 4: What does it mean for UNESCO's rights-based approach to adult learning and education? *International Review of Education, 65*(4), 537–556.

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

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J.C. Marshall, & M. Coltheart (Eds), *Surface dyslexia. Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). London: Lawrence Erlbaum. Retrieved 2 August 2016 from https://www.icn.ucl.ac.uk/dev_group/ufrith/documents/Frith,%20Beneath%20the%20surface%20of%20developmental%20dyslexia%20copy.pdf.

Grek, S. (2019). METRO – International organisations and the rise of a global metrological field - ERC. *Impact, 2019*(1), 41–43.

Grotlüschen, A. (2011). Zur Auflösung von Mythen. Eine theoretische Verortung des Forschungsansatzes lea. – Literalitätsentwicklung von Arbeitskräften [Dispelling myths: A theoretical location of the *lea.* research approach. – Literacy development of the labour force]. In A. Grotlüschen, R. Kretschmann, E. Quante-Brandt & K. D. Wolf (Eds), *Literalitätsentwicklung von Arbeitskräften* [Literacy development of the labour force] (pp. 12–39). Münster: Waxmann.

Grotlüschen, A. (2018). Global competence: Does the new OECD competence domain ignore the Global South? *Studies in the Education of Adult, 50*(2), 185–202.

Grotlüschen, A., Dessinger, Y., Heinemann, A.M.B., & Schepers, C. (2010). Alpha-Levels Schreiben [Alpha levels writing]. *lea.-Verlinkungsstudie Blog* [blogpost July]. Hamburg: Hamburg University. Retrieved 2 August 2016 from https://blogs.epb.uni-hamburg.de/lea/files/2009/09/Kompetenzmodell-Schreiben.pdf.

Grotlüschen, A., Riekmann, W., & Buddeberg, K. (2012). Leo. – Level-One Studie: Methodische Herausforderungen [Leo. – Level-one study: Methodological challenges]. In A. Grotlüschen, & W. Riekmann (Eds), *Funktionaler Analphabetismus in Deutschland. Ergebnisse der ersten leo. – Level-One Studie* [Functional illiteracy in Germany: Results of the first Leo. – Level-one study] (pp. 54–76). Münster: Waxmann.

Guadalupe, C., & Cardoso, M. (2011). Measuring the continuum of literacy skills among adults: Educational testing and the LAMP experience. *International Review of Education, 57*(1–2), 199–217.

Hamilton, M., Maddox, B., & Addey, C. (Eds.). (2015). *Literacy as numbers: Researching the politics and practices of international literacy assessment. The Cambridge Research series*. Cambridge: Cambridge University Press.

Kelava, A., & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen [Descriptive statistical evaluation of items (item analysis) and test value distributions]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Testing theory and questionnaire construction]* (pp. 75–102). Berlin/Heidelberg: Springer.

Kirsch, I., & Thorn, W. (2013). Foreword. The Programme for International Assessment of Adult Competencies: An overview. In: OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (pp. 1–20). Paris: OECD Publishing. Pre-publication copy retrieved 7 February 2020 from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf.

Kretschmann, R. (2005). *Prozessdiagnose der Schriftsprachkompetenz in den Schuljahren 1 und 2* [Process diagnosis of written language competence in primary Grades 1 and 2]. Bergedorfer Förderprogramme. Horneburg: Persen.

Kretschmann, R. (2011). Kompetenzverfahren „Leseverständnis“ (KLV) [Competence procedure “reading comprehension” (KLV)]. In A. Grotlüschen, R. Kretschmann, E. Quante-Brandt, & K. D. Wolf (Eds.), *Literalitätsentwicklung von Arbeitskräften [Literacy development of the labour force]* (pp. 41–57). Münster: Waxmann.

Lunz, M. (2010). Using the very useful Wright map [dedicated web page]. Chicago, IL: Measurement Research Associates, Inc. Retrieved 27 February 2020 from https://www.rasch.org/mra/mra-01-10.htm.

Moosbrugger, H. (2012). Item-response-theory (IRT). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Testing theory and questionnaire construction]* (pp. 227–274). Berlin/Heidelberg: Springer.

Murray, S. (2001). *Understanding the skills of low-literate adults: A proposal*. Ottawa, ON: Statistics Canada.

Murray, T. S., Jones, S., Willms, D., Shillington, R., McCracken, M., & Glickman, V. (2008). *Reading the future: Planning to meet Canada's future literacy needs*. Ottawa, ON: Canadian Council on Learning.

OECD (Organisation for Economic Co-Operation and Development). (2013). OECD Skills outlook 2013: First results from the survey of adult skills. *OECD Publishing*. https://doi.org/10.1787/9789264204256-en.

OECD & Statistics Canada. (2000). *Literacy in the information age. Final report of the International Adult Literacy Survey.* Paris: OECD.

OECD & Statistics Canada. (2005). *Learning a living: First results of the Adult Literacy and Life Skills Survey*. Ottawa/Paris: Statistics Canada/OECD. Retrieved 6 February 2020 from https://www150.statcan.gc.ca/n1/en/pub/89-603-x/2005001/pdf/4200878-eng.pdf?st=0qyXmqEp.

Ordinate Corporation. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate.

QCA (Qualifications and Curriculum Authority) (2005). *National standards for adult literacy, numeracy and ICT.* London: QCA. Retrieved 7 February 2020 from https://set.et-foundation.co.uk/media/131246/2005_national_standards_for_adult_literacy_numeracy_ict.pdf.

Rammstedt, B., Zabal, A., Martin, S., Perry, A., Helmschrott, S., Massing, N., Ackermann, D., & Maehler, D. (2015). *ZA5845: Programme for the International Assessment of Adult Competencies (PIAAC), Germany – Reduced version.* ZA5845 Data file Version 2.0.0. Cologne: GESIS Data Archive, Cologne. https://doi.org/10.4232/1.12385 [restricted access].

Reuter-Liehr, C. (2008). Eine Einführung in das Training der phonemischen Strategie auf der Basis des rhythmischen Syllabierens mit einer Darstellung des Übergangs zur morphemischen Strategie [An introduction to phonemic strategy training on the basis of rhythmic syllabi with a presentation of the transition to morphic strategy]. Bochum: Winkler.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion [Textbook of testing theory and test construction]*. Berne: Huber.

Sabatini, J.P. (2015). *Understanding the basic reading skills of U.S. adults:. Reading components in the PIAAC literacy survey.* Princeton, NJ: Educational Testing Service (ETS). Retrieved 7 February 2020 from https://www.ets.org/s/research/report/reading-skills/ets-adult-reading-skills-2015.pdf.

Sabatini, J.P., & Bruce, K.M. (2009). *PIAAC reading components: A conceptual framework.* OECD Education Working Papers, no. 33. Paris: OECD Publishing. https://doi.org/10.1787/220367414132

Schermelleh-Engel, K., & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung [Methods of reliability determination]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion [Testing theory and questionnaire construction]* (pp. 119–141). Berlin/Heidelberg: Springer.

Spitta, G. (1997). *Kinder schreiben eigene Texte: Klasse 1 und 2. Lesen und Schreiben im Zusammenhang; spontanes Schreiben; Schreibprojekte* [Children write their own texts: Grades 1 and 2. Reading and writing in context; spontaneous writing; writing projects]. Frankfurt am Main: Cornelsen.

Strucker, J., & Davidson, R. (2003). *Adult Reading Components Study (ARCS).* Harvard, MA: National Center for the Study of Adult Learning and Literacy (NCSALL). Retrieved 2 August 2016 from https://www.ncsall.net/fileadmin/resources/teach/prac_res_guide_read2.pdf.

Strucker, J., Yamamoto, K., & Kirsch, I. (2007). *The relationship of the component skills of reading to performance on the International Adult Literacy Survey (IALS).* Harvard, MA: National Center for the Study of Adult Learning and Literacy (NCSALL). Retrieved 2 August 2016 from https://www.ncsall.net/fileadmin/resources/research/reading_ials_rb.pdf.

Tamassia, C., Lennon, M.L., & Yamamoto, K. (2013). Scoring reliability studies. In: OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (chapter 12). Pre-Publication Copy retrieved 7 February 2020 from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). *TOWRE: Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.

UIL (UNESCO Institute for Lifelong Learning) (2019). *Progress on indicator 4.6.1.* GAML 6 Meeting 27–28 August, Yerevan, Armenia [presentation slides]. Hamburg: UIL. Retrieved 11 February 2020 from https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/TF4.6.GAML6_UILPresentation_clean.pdf.

UIS (United Nations Institute for Statistics) (2017). *Global Alliance to Monitor Learning* [dedicated webpage]. Retrieved 12 February 2020 from https://gaml.uis.unesco.org/.

UN (United Nations) (2015). *Transforming our world: The 2030 Agenda for sustainable development*. A/RES/70/1. New York: UN. Retrieved 11 February 2020 from https://sustainabledevelopment.un.org/post2015/transformingourworld/publication.

UN (2016). *Sustainable Development Goal 4: Targets and indicators* [dedicated webpage]. New York: UN. Retrieved 11 February 2020 from https://sustainabledevelopment.un.org/sdg4#targets.

Wechsler, D. (1997). *WMS-III: Wechsler Memory Scale administration and scoring manual*. New York: Psychological Corporation.

Wilson, M. (2005). *Constructing measures. An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0. Generalised item response modelling software*. Camberwell, VIC: ACER Press.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (chapter 17). Paris: OECD Publishing. Pre-publication copy retrieved 7 February 2020 from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., et al., & Rammstedt, B. (Eds.). (2014). *PIAAC Germany 2012: Technical report*. Münster: Waxmann.



**Anke Grotlüschen**, PhD, is professor for lifelong learning at Hamburg University, Germany. She works in the field of adult education and lifelong learning. She is in charge of the German nationwide "Level-One Surveys" (LEO; conducted in 2010 and 2018) and she is the spokeswoman for the "Hamburg Numeracy Project" research association, a cooperation with partners at the Helmut-Schmidt-University (Hamburg), the Hamburg University of Applied Sciences and the UNESCO Institute for Lifelong Learning (also in Hamburg).

**Barbara Nienkemper**, PhD, publishes on literacy, adult basic education and diagnostics, especially from the perspective of test-takers. She uses the theoretical approach of the New Literacy Studies and applies it to data from large-scale assessments on skills use. She currently works with the Hamburg Adult Education Centre (Hamburger Volkshochschule).

**Caroline Duncker-Euringer**, PhD, focuses on the interlink between literacy assessment and language assessment in her quantitative analyses. Her qualitative work asks about the definitions of adult basic education in Germany from the point of view of educational administration. She currently works with the Hamburg Police Academy (Akademie der Polizei Hamburg).