GOETHE TEST PRO: GERMAN FOR PROFESSIONALS

INVESTIGATION OF PSYCHOMETRIC QUALITY

18/01/2024 – GOETHE INSTITUTE

ARON FINK, ANDREAS FREY, & LARA WEIß

GOETHE INSTITUT

Sprache. Kultur. Deutschland.

Content

ABSTRACT

The online German test – the *Goethe Test PRO: German for Professionals* – has been administered since April 2017 at the test centers of the Goethe Institute and at the sites of corporate customers around the world. Devised as an adaptive online test, it assesses language skills in reading and listening at the workplace at the A1–C2 levels of the Common European Framework of Reference for Languages. In this study, based on the test data of N = 5,636 test takers from 17 countries, the psychometric quality of the Goethe Test PRO was examined with regard to various criteria. The Goethe Test PRO delivered reliable test results, with a reliability of > .9. The psychometric quality of the test was comparable between the different countries in which the Goethe Test PRO is administered.

1. INTRODUCTION

The *Goethe Test PRO: German for Professionals* (GTP) is a computer-based test offered by the Goethe Institute, which assesses reading and listening competence at the workplace, based on the Common European Framework of Reference for Languages (CEFR) (see, e.g., https://www.goethe.de/Z/50/commeuro/101.htm). The test can be administered flexibly at one of Goethe Institutes or directly at the company, and it can be used as a basis for decision-making, for example, for further advanced training.

The GTP applies the method of computerized adaptive testing (CAT; e.g., Frey, 2020). In CAT, unlike in traditional testing procedures, not all test takers have to respond to the same set of items that has been compiled in advance. Instead, the item selection is based on the test takers' response pattern over the course of the test. In simplified terms, when CAT is used, the difficulty of the items provided is adjusted to the test takers' ability level. Therefore, only those items are administered to the test taker that provide as much diagnostic information as possible about the individual characteristics to be measured, which, in the case of the GTP, means the individual language ability level. Compared to traditional, nonadaptive tests, this method leads to higher measurement precision and/or shorter test length (e.g., Segall, 2005). In particular, CAT makes it possible to measure the individual ability levels with a comparable amount of precision across the entire ability distribution (Frey & Ehmke, 2007). This means that, as opposed to conventional test methods, the accuracy of the test scores and, specifically in the case of the GTP, the reliability of the classification of a test taker into a CEFR level is largely independent of the individual ability level. Because not all test takers respond to the same set of items, individual ability levels cannot be calculated based on the number of correct responses. Item characteristics such as item difficulties have to be considered in the calculation of the ability level. For this purpose, models from the Item Response Theory (IRT; e.g., van der Linden, 2016) are used in CAT. These models describe the probability of an individual with a certain ability giving a correct response to an item with certain characteristics. The GTP uses the Rasch model (Rasch, 1960) as an IRT model.

It takes 60 to 90 minutes to complete the GTP. All test takers respond first to 15 tasks from the subdimension of reading and then to 15 tasks from the subdimension of listening. Here, tasks refer to

one or more items that share a common stimulus. The GTP has various innovative item formats. Directly after the completion of the test, the test takers receive a report of the separate results for reading and listening, as well as the overall result, including the CEFR level reached. In addition, test takers get a digital certificate with a detailed description of their language ability level.

2. RESEARCH QUESTIONS

This study examined the psychometric quality of the GTP. In addition, the results obtained in the countries in which the GTP is offered were compared. The following research questions were examined and answered in this study:

1. What psychometric quality does the Goethe Test PRO provide?
    - 1.1 Can the items of the GTP be regarded as one-dimensional?
    - 1.2 How reliable are the test results of the GTP?
    - 1.3 To what extent is the GTP able to differentiate across the ability distribution?
    - 1.4 What degree of adaptivity does the GTP provide?
2. Does the psychometric quality differ between countries?

3. PROCEDURE AND METHOD

Sample

The analyses were based on the test results of the N = 5,636 test takers who had completed the test since April 2017. The test results were extracted from the platform Moodle, which is used for the administration of the test. The distribution of the test takers across the examined countries is presented in Table 1. The majority of test takers (78.66%) completed the test in Germany, France, or the Netherlands. Table 2 shows the number of test takers per CEFR level.

Research question 1

Research question 1.1: Can the items of the GTP be regarded as one-dimensional?

The overall result of the GTP is calculated by averaging the individual test results for reading and listening. This procedure only makes sense if the two subject areas represent one common scale, which, in turn, means that the items of the GTP need to be one-dimensional. Therefore, with regard to Research question 1.1, it was necessary to check whether the items of the GTP can be regarded as being one-dimensional. For this purpose, a one-dimensional and a two-dimensional model with the two correlating subdimensions of reading and listening comprehension were estimated on the basis of the test data. The two models were compared to each other by means of a likelihood ratio test. The likelihood ratio test is an inferential statistical test that allows the comparison of the fit of two competing statistical models. In addition, the models were compared to each other based on the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). The lower the value, the better the model fit for both coefficients.

Table 1

Number and percentage of test takers per location

	N	%
Argentina	20	0.35
Brazil	6	0.11
China	6	0.11
Germany	1,273	22.59
Finland	20	0.36
France	1,572	27.89
Great Britain	56	0.99
Greece	21	0.37
Italy	11	0.20
Netherlands	1,588	28.18
Poland	247	4.38
Russia	23	0.41
Switzerland	463	8.21
Spain	113	2.00
Taiwan	34	0.60
Turkey	99	1.76
Uzbekistan	84	1.49
Total	5,636	100.00

Table 2

Number of test takers per CEFR level for the subdimensions of reading and listening comprehension and for the overall test

		$N_{Reading}$	$N_{Listening}$	$N_{Overall}$
CEFR level	Pre-A1	5	21	7
	A1	131	78	50
	A2	2,786	819	1,600
	B1	1,750	2,646	2,665
	B2	605	1,345	875
	C1	218	538	343
	C2	141	189	96

Note: CEFR = Common European Framework of Reference for Languages; Pre-A1 = level below A1.

Research question 1.2: How reliable are the test results of the GTP?

The psychometric quality criterion of reliability provides information on the precision of the test results. To be regarded as reliable, a repeated test conducted under the same conditions should lead to the same results. There are different reliability coefficients. The reliability coefficient used in this study is derived from the definition of reliability as the squared correlation between the true ability level θ and the

estimated ability level $\hat{\theta}$ (squared correlation reliability; $p_{\theta\hat{\theta}}{}^2$; Kim, 2012), which is frequently used to estimate the reliability in simulation studies. An estimate based on empirical data can be derived from the quotient of the variance of the theta estimates $\sigma^2_{\hat{\theta}_j}$ and the sum of the variance of the theta estimates and the mean squared standard error $SE^2_{\hat{\theta}_j}$ of the individual ability estimates $\hat{\theta}_j$:

$$p^2_{\theta\hat{\theta}} = \frac{\sigma^2_{\hat{\theta}_j}}{\sigma^2_{\hat{\theta}_j} + \frac{1}{N}\sum_{j=1}^{N} SE^2_{\hat{\theta}_j}}. \quad (1)$$

An alternative reliability coefficient is based on the approach of determining reliability with the correlation $p_{\hat{\theta}\hat{\theta}}$ of the ability estimation $\hat{\theta}$, assessed with two parallel test forms. The analyses in this study revealed that comparable results were achieved with both reliability coefficients. Therefore, we limit the presentation of the results on reliability to the reliability coefficient shown in Equation 1.

Research question 1.3: To what extent is the GTP able to differentiate across the ability distribution*?*

According to the Standards for Educational and Psychological Tests (AERA, APA, & NCME, 2014) and specifically referring to Research question 1.3, the conditional standard error of the ability estimates needs to be calculated to investigate the degree to which the GTP can differentiate between different ability levels. For the calculations, we used the range of the latent ability scale, which (a) covers all levels of the CEFR and (b) includes at least 99% of the test takers

Research question 1.4: What degree of adaptivity does the GTP provide?

Referring to Research question 1.4, the recently proposed Engineering Optimal Information Index (EOI; Kingsbury & Wise, 2020) for adaptive tests was calculated on the basis of the Rasch model to examine the level of adaptivity. The EOI indicates the proportion of information actually received from the test compared to the maximum achievable information at the final ability level. This index assumes a hypothetically perfect item base under the Rasch model, which means that items that exactly match the estimated ability level are administered to the test taker. It represents a theoretical value that quantifies the hypothetically most informative test that could be conducted and it sets it in relation to the test information actually observed. The EOI for the Rasch model is calculated as follows:

$$EOI = 100\frac{\sum_{j=1}^{N}(IA_j/0.25K)}{N}, \quad (2)$$

where IA_j is the test information of person *j* for their current ability estimate*, K* is the number of items answered, and 0.25 is the maximum achievable item information of an item in the Rasch model. The EOI has a maximum value of 100. This value implies that the most informative number of items, given the final ability estimate, has been presented to each test taker in the tested group. The EOI is also suitable for comparing the degree of a test's adaptivity between groups of test takers (e.g., CEFR levels, countries).

In this way, the potential strengths or weaknesses of a test instrument can be examined in a more differentiated way.

Research question 2

As the GTP is applied in different countries, it is important to ensure that the test works in the same way across countries and delivers reliable results. Therefore, to answer Research question 2, the aforementioned reliability coefficient as well as the EOI were calculated for the different countries and compared to each other. For this purpose, only countries in which at least 20 test takers had completed the GTP were included in the analysis.

All calculations were conducted using the statistics software R (R Core Team, 2020). To answer Research question 1.1, the R-package "mirt" (Chalmers, 2012) was used.

4. RESULTS AND ANALYSES

Descriptive statistics

The descriptive statistics for the two item pools are shown in Table 3. Figures 1 and 2 show the absolute frequencies of the items per difficulty level (left *y*-axis) and the relative frequency of test takers per ability level (right *y*-axis) in logits. Items with difficulty parameters in close proximity to each other have been merged. For an adaptive test such as the GTP, it is preferable that a sufficient number of items is available for all ranges in which test takers are placed on the logit scale. An adaptive test can adapt ideally when, for each person, at least as many items are available to match the length of the test. This optimum was reached to different extents for reading and listening. While the item pool for reading lacks very easy items, the item pool for listening lacks difficult and very difficult items.

Table 3
Descriptive statistics of the item pools for the reading and listening subdimensions

	Reading	Listening
Number of tasks	420	168
Number of items	449	289
M Item difficulty	0.035	0.043
SD Item difficulty	1.339	1.245

Note: M = mean; *SD* = standard deviation.

Regarding research question 1 – Psychometric quality of the GTP

The results for the different research questions are presented below. Based on the standardized test administration, the evaluation, and the reporting of the test results, the objectivity of the GTP can be regarded as given.

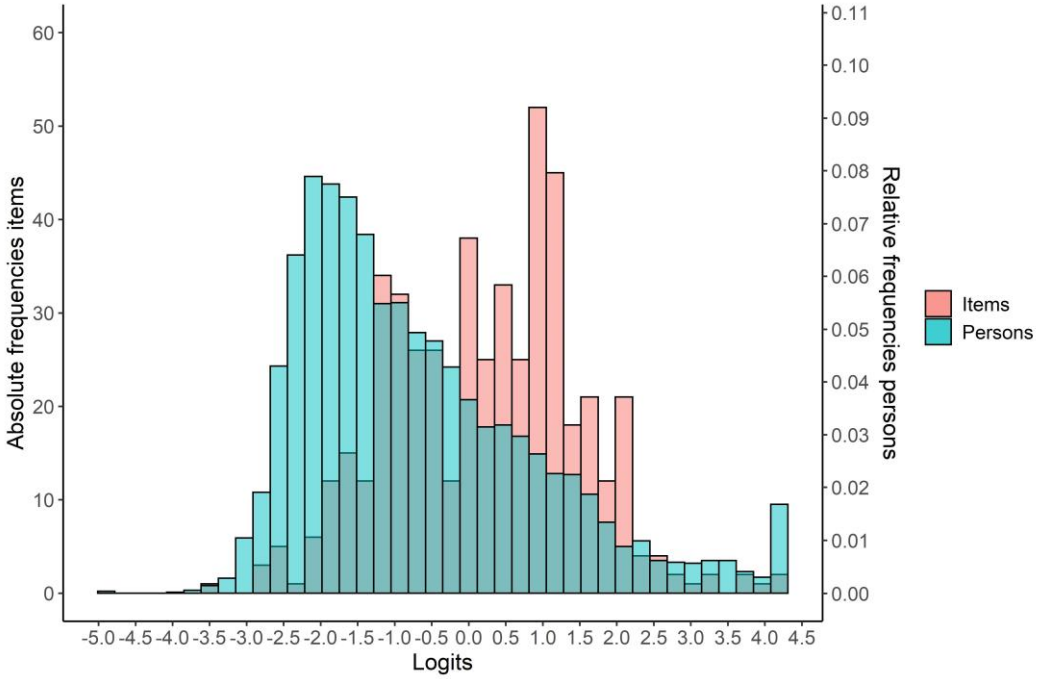

Figure 1. Absolute frequencies of items (left) and relative frequencies of the estimated abilities (right) per difficulty level in logits for the reading subdimension.

Regarding research question 1.1 – Dimensionality of the items

The results for the model comparison between the one-dimensional and the two-dimensional model are presented in Table 4. The likelihood ratio test identified the two-dimensional model as significantly better fitting compared to the one-dimensional model. Given the large sample size and the associated very high test power, this is not surprising. The AIC and BIC information criteria also identified the two-dimensional model as the slightly better-fitting model. The two subdimensions for reading and listening, however, correlated almost perfectly with each other, with a latent correlation of .936. Based on these findings, it is possible not only to identify the two subdimensions separately but also to merge the results for the two subdimensions in order to report the GTP results on one common scale.

Table 4

Results of the model comparison between the one-dimensional and the two-dimensional model

Model	Log likelihood	χ^2	*df*	*p*	AIC	BIC
one-dimensional	-132905	-	-	-	267268	272107
two-dimensional	-132802	206.479	2	< .001	267066	271918

Note: AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

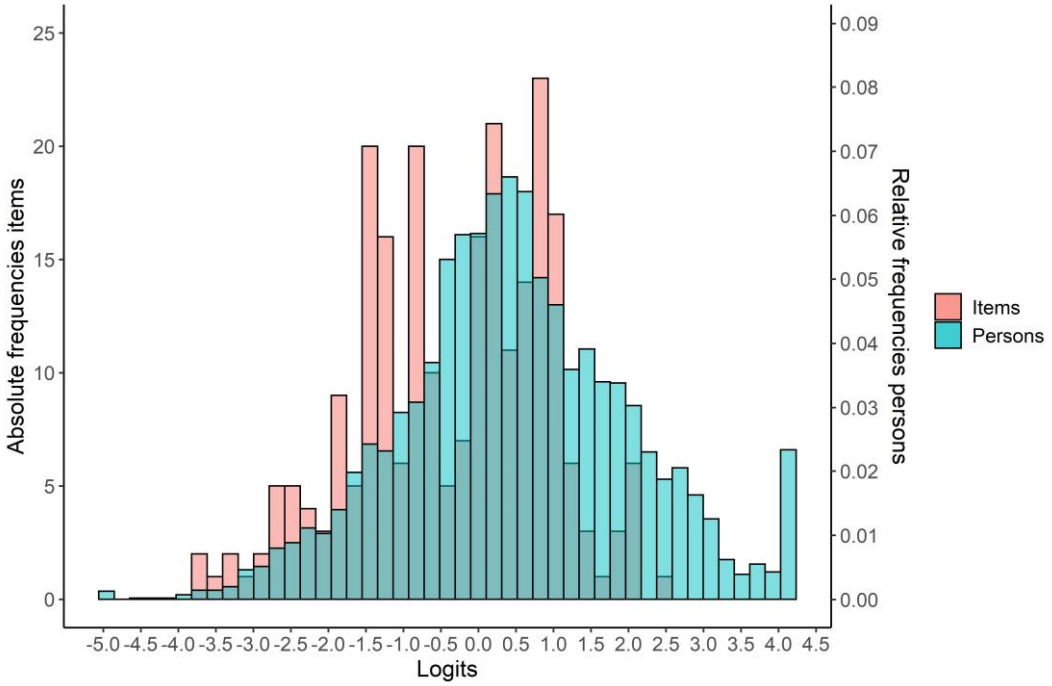

Figure 2. Absolute frequencies of items (left) and relative frequencies of the estimated abilities (right) per difficulty level in logits for the listening subdimension.

Regarding research question 1.2 – Reliability

Table 5 shows the reliability coefficients for the reading and listening subdimensions and for the overall test. With over .9, the reliability of the overall test can be considered to be high. Both subdimensions showed good reliability. The reliability coefficient was slightly higher for the reading subdimension than for the listening subdimension.

Table 5
Descriptive statistics and reliability coefficients for the Goethe Test PRO

	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$
Reading	-0.664	1.609	0.892
Listening	0.478	1.530	0.868
Total	-0.093	1.445	0.923

Note: $M(\hat{\theta})$ = mean of skills estimates; $SD(\hat{\theta})$ = standard deviation of skills estimates; $p^2_{\theta\hat{\theta}}$ = reliability.

Regarding research question 1.3 – Differentiability

Figure 3 shows the conditional standard errors of the ability estimates for the overall test and the reading and listening subdimensions. The standard error for the overall test is at a comparably low level, namely, within the range of -3.0 to 2.0 logits. This demonstrates that the test measures a broad range of ability levels with comparable precision. As expected, higher standard errors resulted for the reading and

listening subdimensions due to the shorter test length, which translates to a measurement precision slightly lower than that of the overall test. The standard errors of the ability estimates were largely consistent on the partial scales, that is, ranging from -3.0 to 1.5. Thus, even at the level of the partial scales, the measurement precision was comparable across a broad range of ability levels. However, Figure 3 illustrates that the shortage of difficult and very difficult items in the subdimension of listening entailed a stronger rise in standard errors for high-ability test takers.

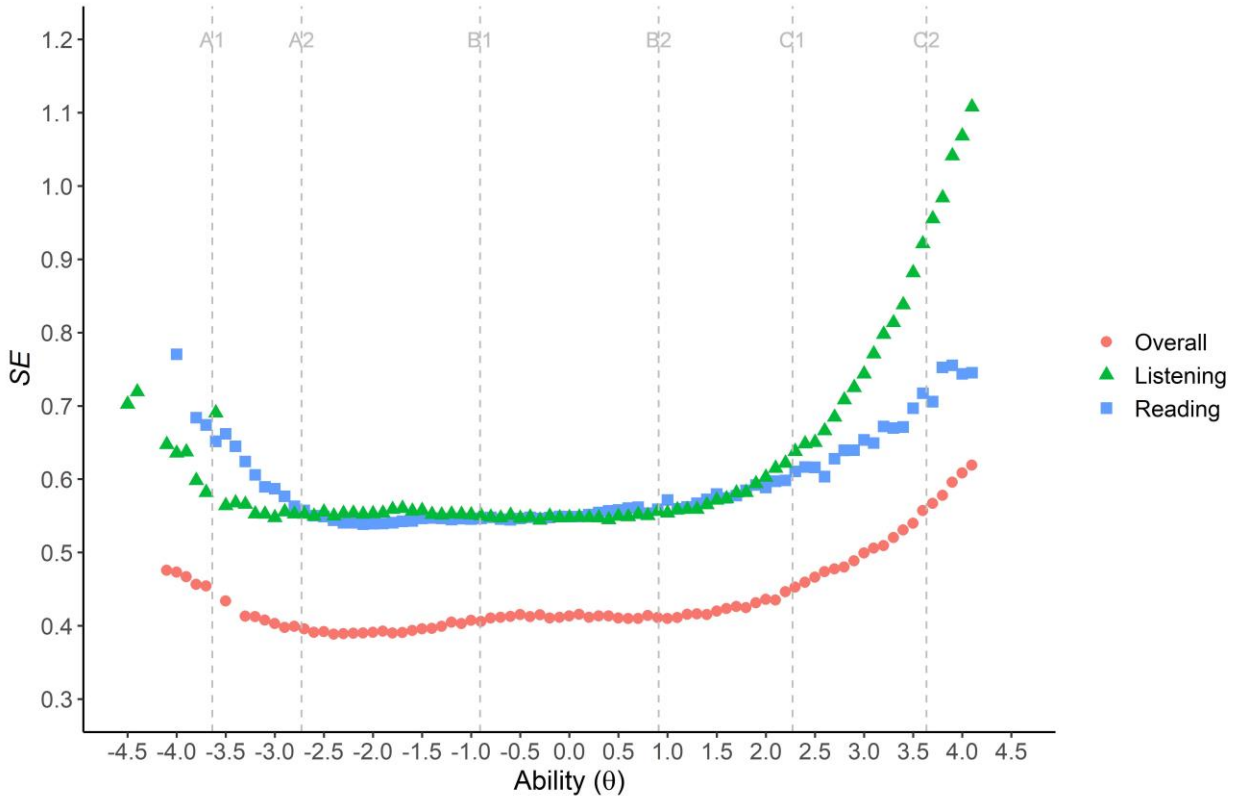

Figure 3. Conditional standard error (*SE*) of the ability estimates for the overall test and for the reading and listening subdimensions. The vertical dashed lines represent the limits of the CEFR levels.

Regarding research question 1.4 – Adaptivity

The EOI approximated its theoretical maximum of 100 for both the overall test (78.44) and the reading (86.55) and listening (82.64) subdimensions. Consequently, the adaptivity of the GTP can be considered to be good. As before, the results for the reading subdimension were slightly better than those for the listening subdimension This can be explained in part by the significantly smaller item pool for listening, which is lacking very difficult items. This becomes even clearer in an examination of the EOIs when separated by CEFR levels. The results for the EOIs separated by CEFR levels are shown in Table 6. Test takers classified below the A1-level were not included in the analyses due to the very small sample size (N = 7).

Table 6
EOI separated by CEFR levels for the Goethe Test PRO

CEFR levels	*N*	$EOI_{Reading}$	$EOI_{Listening}$	$EOI_{Overall}$
A1	50	82.32	81.55	82.19
A2	1,600	89.34	88.08	85.00
B1	2,665	89.51	87.12	79.39
B2	875	83.88	71.79	76.22
C1	343	68.35	46.96	57.28
C2	96	52.92	24.04	37.77

Note: EOI = Engineering Optimal Information Index.

The results show that the EOIs from A1 to B2 were comparably high for the reading subdimension. The EOI started to decrease at the C1 level. The EOIs for the listening subdimension were below the EOIs for reading. For Level C2, with 24.04, the EOI can be considered to be very low. At this level, the test seems to administer comparably few items that correspond to the final ability level (as these are not included in the item pool). This is also reflected in the EOI for the overall test, which was by far the lowest for the CEFR Level C2.

Regarding research question 2 – Comparison of the psychometric quality between countries

Table 7 shows the results of the reliability analyses and the EOIs separated by countries for the overall test and for the reading and listening subdimensions. The reliability of the overall test can be considered as being good or very good. Thus, despite small differences, a high to very high measurement precision is assured in all countries. The same applies to the reading subdimension. For this subdimension, the reliability was between .778 and .923 and was thus within a good to very good range for all countries. The reliability for listening was slightly lower compared to this, as could be expected based on the results across all countries. However, the reliability can be considered to be good to very good in nearly all countries. In the Netherlands, Turkey, Russia, and Argentina, the reliability can be regarded to be acceptable.

With reference to the EOI, a comparable degree of adaptivity was achieved in nearly all countries. Again, the EOI for the reading subdimension was above that for the listening subdimension. The lowest EOI resulted for Russia, irrespective of the subdimension. However, the average performance in Russia for reading (1.691) and for listening (2.352) was clearly higher than the total mean. As can be seen in Table 6, the EOI for both subdimensions was considerably lower at the upper ability range. This could probably explain the lower EOIs in Russia. Overall, however, the EOI for the complete test and the two subdimensions was at a comparably high level for all countries.

Table 7

Country-specific descriptive statistics, reliability coefficients, and EOIs for the Goethe Test PRO

		Reading				Listening				Overall			
Country	N	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI
Netherlands	1.588	-1.234	1.026	.778	89.30	0.374	1.061	.780	86.25	-0.430	0.923	.833	79.40
France	1.572	-0.490	1.604	.892	86.60	0.622	1.596	.874	79.85	0.066	1.478	.926	78.50
Germany	1.273	-0.796	1.664	.897	86.01	0.004	1.560	.877	83.39	-0.396	1.474	.927	80.02
Switzerland	463	-0.398	1.829	.912	84.96	0.510	1.681	.882	79.27	0.056	1.640	.938	77.73
Poland	247	0.092	2.024	.923	81.29	1.152	1.934	.893	71.07	0.622	1.869	.947	72.59
Spain	113	0.544	1.889	.914	81.24	1.417	1.810	.876	69.76	0.980	1.747	.941	72.71
Turkey	99	1.245	1.218	.815	81.33	2.305	0.972	.649	61.55	1.775	0.945	.818	70.06
Uzbekistan	84	-0.778	1.467	.875	88.09	-0.432	1.200	.824	87.24	-0.605	1.211	.902	84.19
Great Britain	56	-0.555	1.574	.886	85.42	1.109	1.444	.842	77.21	0.277	1.387	.912	74.47
Taiwan	34	-0.256	1.324	.850	86.79	0.289	1.322	.844	83.66	0.017	1.222	.901	82.05
Russia	23	1.691	1.767	.891	74.04	2.352	1.345	.747	57.72	2.022	1.436	.899	63.98
Greece	21	0.938	1.551	.875	80.65	1.431	1.274	.808	74.48	1.185	1.332	.906	75.96
Argentina	20	-0.128	1.646	.896	86.57	1.210	1.043	.755	79.15	0.541	1.272	.901	76.15
Finland	20	0.007	1.752	.906	85.60	1.217	1.499	.847	74.22	0.612	1.524	.927	75.93

Note: $M(\hat{\theta})$ = Mean of ability estimates; $SD(\hat{\theta})$ = Standard deviation of ability estimates; $p^2_{\theta\hat{\theta}}$ = reliability; EOI = Engineering Optimal Information Index.

5. CONCLUSION

The objective of Research question 1 was to analyze the psychometric quality of the GTP. For this purpose, the dimensionality of the GTP was investigated. In addition to the descriptive statistics for item pools and test takers, reliabilities, conditional standard errors, and the EOIs were calculated. Based on the results of the analysis of dimensionality, the merging of the two subdimensions into one common scale, which is used for the reporting of GTP results, proved to be appropriate. With a total of 738 items, the item pool of the GTP is very large and permits a very fine adjustment of the items administered to the response pattern of the test takers. Accordingly, the EOI values were high. The adaptivity, with a value of 86.55, was close to the hypothetical perfect adaptivity of 100 for the reading subdimension. Correspondingly, in total, the reliability was also very good and varied only slightly between countries at the level of the overall test and at the level of the subdimensions. The observed differences can most likely be traced back to differences in the ability distributions between countries. However, in the upper extremes of the ability distribution, there is still potential for optimization in the subdimension of listening. This is also reflected in the higher standard errors of the ability estimates obtained for this

subdimension. This could be remedied by expanding the item pool to include more difficult and very difficult items. In future studies, the EOI could provide useful information on how well the GTP maximizes the test information for the different CEFR levels. The term "engineering" reflects the potential of the EOI to indicate the effects of changes in the item pool, test design, or the adaptive algorithm on the adaptivity of the test.

Based on the data available in this study, no statements can be made regarding the validity of the test score interpretations derived (e.g., Hartig, Frey, & Jude, 2020). Analyses of validity could therefore be the subject of future studies.

In conclusion, the GTP represents a computer-based adaptive test instrument that provides precise measures of reading and listening at the workplace across countries and across a broad range of abilities within a short period of time and in an efficient manner.

REFERENCES

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing*. Washington: AERA Publication Sales.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. https://doi.org/10.1109/TAC.1974.1100705

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1-29. https://doi.org/10.18637/jss.v048.i06

Frey A. (2020). Computerisiertes adaptives Testen [Computerized adaptive testing]. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 501-524). Springer. https://doi.org/10.1007/978-3-662-61532-4_20

Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 8*, 169-184. https://doi.org/10.1007/978-3-531-90865-6_10

Hartig, J., Frey, A., & Jude, N. (2020). Validität von Testwertinterpretationen [Validity of test score interpretations]. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 529-545). Springer. https://doi.org/10.1007/978-3-662-61532-4_21

Kingsbury, G. & Wise, S. L. (2020) Three measures of test adaptation based on optimal test information. *Journal of Computerized Adaptive Testing*. 8(1), 1–19. https://doi.org/10.7333/2002-0801001

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77,* 153–162. https://doi.org/10.1007/S11336-011-9238-0

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Studies in mathematical psychology. Copenhagen:* Danmarks Paedagogiske Institut.

R Core Team (2020). R: A language and environment for statistical computing [Software]. R Foundation for Statistical Computing. Available from www.r-project.org

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464, https://doi.org/10.1214/aos/1176344136

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Boston: Elsevier Academic. https://doi.org/10.1016/b0-12-369398-5/00444-8

van der Linden, W. J. (Hrsg.). (2016). *Handbook of item response theory. Volume one: Models.* Boca Raton: Chapman & Hall/CRC.