Educational Methods & Psychometrics—Vol. 4, Article No. 22
January 2026
https://doi.org/10.65301/emp.2026.262

ARE INTELLIGENCE MEASURES STABLE ACROSS TIME AND GROUPS? A STUDY OF INVARIANCE IN A SYNONYMS TEST

Jeanette Melin*

Swedish Defence University, Department of Leadership and Command & Control, Karlstad, Sweden

Emelie Wahlkrantz

Swedish Defence University, Department of Leadership and Command & Control, Karlstad, Sweden

Maria Fors Brandebo

Swedish Defence University, Department of Leadership and Command & Control, Karlstad, Sweden

Gerry Larsson

Swedish Defence University, Department of Leadership and Command & Control, Karlstad, Sweden

The Inland University, Department of Health and Welfare, Elverum, Norway

Stefan Annell

Swedish Defence University, Department of Leadership and Command & Control, Stockholm, Sweden

Intelligence tests have a long tradition across different fields, yet little attention has been paid to how their measurement properties may change over time and impact selection outcomes and intelligence trends. This study illustrates these gaps and consequences, using a synonym test from the Swedish Enlistment Battery (SEB) as an example. The test aims to measure lexical ability, a narrow verbal ability. First, the measurement properties of the synonym test across years (2014, 2018, and 2022) and between subgroups were evaluated using Rasch analysis. Secondly, the implications for changes in the synonym test's measurement properties were evaluated for basic military training in Sweden. Misfitting items and invariance issues were identified. With the removal of misfitting items, improved measurement properties were achieved, but this did not significantly impact the possibilities of identifying group differences. However, the lexical ability of some test takers at the lower end of the scale was likely overestimated, whereas it was likely underestimated at the upper end of the scale if all original items were used. These findings underscore the importance of ensuring unbiased measurement properties and invariance over time, highlighting the need for regular evaluations and refinements of intelligence tests to maintain fairness and accuracy.

Keywords: cognitive ability, intelligence, crystallized intelligence (Gc), invariance, differential item functioning (DIF), Rasch

1. Introduction

If the word *pandemic* – a word rarely used in everyday language before the Covid-19 outbreak of 2020 – were to be included in a word test, it would likely be an easy word for test

Correspondence should be made to Jeanette Melin, Department of Department of Leadership and Command & Control, Karlstad, Sweden. Email: jeanette.melin@fhs.se

takers in tests administered from 2020 forward without implying an improved lexical ability. What we are exposed to over time and in different cultures and societies makes some knowledge or skills more common or easy to acquire (Muthukrishna & Henrich, 2016). Nevertheless, a key factor in making reliable comparisons between groups and over the years is invariance (Andrich et al., 2012; Leitgöb et al., 2022; Somaraju et al., 2022), meaning possible differences reflect true differences and not differences in the composition of test items being used (Borsboom, 2006; Fisher, 2009; Hagquist & Andrich, 2017).

In the context of military selection in Sweden, the composition of test takers in recent years has been changed from all voluntary to primarily mandatory, and are becoming more heterogeneous, including an increased proportion of women and individuals with foreign background. These changes in test taker composition can contribute to invariance issues.

The presence of some non-invariant items does not necessarily render the test unusable; instead, understanding these issues allows for methods to enable comparisons and/or suggest improvements to the test (Goldammer et al., 2024). Thus, regular evaluations of measurement properties – including potential item drift – in tests of general mental ability and intelligence are warranted, for instance, when these tests serve as the basis for key selection decisions, or when test results are compared over time to monitor trends or changes. This is rarely seen, but a few examples exist. For instance, more than 25 years ago, through a three-parameter item response theory (IRT) model, Chan et al. (1999) concluded significant item drift in the US Armed Services Vocational Aptitude Battery (ASVAB) over the duration of 16 years and that the semantically laden items were more susceptible to item-level changes than items focused on general skills and principles. A similar study by Shiu et al. (2013) applied a two-parameter IRT model to the Estonian version of the National Intelligence Tests (NIT) across cohorts from 1934 and 2006 reported that one-third of the synonyms and antonyms items showed non-invariance. Furthermore, recently, it was shown that items and response options in the Norwegian Armed Forces' word similarities tests use many outdated words (Nordmo et al., 2025). These findings, specifically addressing invariance issues in word tests, together with a general understanding of nonequivalence over time (Somaraju et al., 2022; Vandenberg & Lance, 2000) and between groups (Helland-Riise et al., 2024), suggest that more work is needed to understand the scope of the problem.

While IRT models generally are considered useful for evaluating measurement properties, one specific type — the Rasch model — offers unique features that make it particularly effective as an evaluation tool. At one point, the distinction between IRT models in general and the Rasch model is subtle. Mathematically, the Rasch model is a kind of one-parameter logistic (1PL) model within the IRT family. Nevertheless, there is a crucial distinction between IRT models and the Rasch model. IRT proponents are typically driven by finding a model that best explains the data, and when the data does not fit the model, they seek another model to explain the data better (Embretson, 2000; Reckase, 2009; Sijtsma & Junker, 2006). Conversely, with the Rasch model, observed data are examined against the model to determine the extent to which satisfying measurements can be obtained (Rasch, 1960; Cano & Hobart, 2011), and in turn, the Rasch model (1960) has been accentuated as *a specifically metrological approach to human-based measurement* (Pendrill, 2014). While the choice of IRT model varies (Andrich, 2004; McNamara & Knoch, 2012), the choice of the Rasch model – instead of either classical test theory-based models or other kinds of IRT models – can be justified for two main reasons. First, other models typically do not state or test for sufficiency together with measurement invariance, while the Rasch model does (Wilson, 2013). Secondly, we deliberately refrain from forcing the data to fit more complex models, models that offer better statistical fit but at the cost of reduced measurement traceability and comparability (Cano et al., 2019). Moreover, as a component of the Rasch analysis, *Differential Item Functioning* (DIF) analysis is a powerful

tool for identifying how different items may vary over time and/or across subgroups (c.f., Andrich & Hagquist, 2012). In case items fail to meet the requirements of invariance, specific objectivity cannot be obtained, and consequently, the validity when comparing measures of persons' abilities will be distorted (Babcock & Albano, 2012; Hagquist & Andrich, 2017). Nevertheless, identifying and understanding DIF provides grounds for compensating for it and still enabling valid comparisons.

Aims and research questions

To our knowledge, little is known about how the measurement properties of intelligence assessments may change over time and how such changes affect the interpretation of intelligence trends and their broader consequences. Therefore, using a synonym test in the Swedish Enlistment Battery (SEB) as an example, the overall aim of the present paper was to contribute to filling this knowledge gap. The synonym test intends to measure lexical ability, a narrow verbal ability, below the broad cognitive ability, crystalized intelligence (Gc), which in turn is an indicator of general mental ability (; e.g., Carroll, 1993; Flanagan et al., 2013; McGrew, 2009). Thus, in the SEB, the synonyms test is considered an important indicator of both Gc and GMA.

We examined data from the synonym test administered in 2014, 2018, and 2022 from test takers for admission tests for basic military training in Sweden. The study addressed two research questions. First, what are the measurement properties of a SEB synonym test, with a specific focus on investigating potential invariance issues over the years (2014, 2018, and 2022) and between subgroups (gender, age, and selection group [i.e., musters or volunteers]), using Rasch analysis? Secondly, what are the implications of changed measurement properties for the assessment of lexical ability?

2. Methods

Setting

Sweden introduced compulsory military conscription at the beginning of the 20th century (Jonsson et al., 2024; Nordlund, 2022). For most of the last century, all Swedish males had to muster (i.e., undergo compulsory testing for military service), typically at age 18, and nearly all had to complete basic military training as conscripts. Women have formally been allowed to join the Swedish military on a voluntary basis since 1980, with some restrictions remaining until 1989. The number of women applying for basic military training has been limited but slowly increasing with time; between 2000 and 2006, only about 2% of all musters were women, rising to about 6% from 2007 to 2009. In 2010, conscription was replaced by a voluntary system, in which all recruitment was completely voluntary for both men and women (Bäckström, 2022; Jonsson et al., 2024; Ludvigsson et al., 2022). This changed the composition of the group that underwent tests and examinations; the proportion of individuals aged 20 or older increased. Furthermore, the proportion of women continued to increase. However, since then, the security situation in Europe has worsened. Along with the Swedish Armed Forces facing difficulties recruiting enough personnel, the Swedish government decided to reinstate conscription in 2017, which is currently compulsory for both men and women. Notably, alongside conscription, the possibility to apply voluntarily remains. Since 2017, the numbers that muster have increased steadily, mirroring the growing needs of the Swedish Armed Forces.

The admission tests for basic military training in Sweden entail medical, physiological, and psychological examinations that aim to select suitable candidates for military service and to allocate them to appropriate positions (Jonsson et al., 2024; Ludvigsson et al., 2022). Sweden

began to use psychological tests to assess the intelligence among test takers during the 1940s (Husén, 1948). In 2000, a semi-adaptive technology was introduced to the SEB, in which the difficulty of the subtests depends on performance in previous subtests (Ullstadius et al., 2008). The 2000s version of the SEB, which has recently been replaced by the new SEB-2025, was developed based on Gustafsson's hierarchical model of cognitive abilities, the so-called NF model (c.f., Gustafsson & Balke, 1993), and comprising ten subtests with inductive reasoning, spatial, verbal, and technical comprehension tasks (Carlstedt & Gustafsson, 2005).

This study is a part of the Swedish Defence Conscription and Assessment Agency's (SDCAA) work to update the SEB-2000. The project was approved by the Swedish Ethical Review Authority, ref. 2019-03576.

Sample

Data on SEB-2000 from the admission procedure was made available by the Swedish Defense Conscription and Assessment Agency. The sample in the present study is test takers who took the SEB-2000 in 2014, 2018, and 2022, respectively. Table 1 shows the distributions of sex, age, and selection group across the years for the test takers included in this study. As expected, there are more men than women, but the proportion of women has increased from 2014 to 2022. The majority of the test takers were between 17 and 19 years old in 2018 and 2022, compared to 2014, when the proportions of 17-19-years old and those older were equal. This mirrors the reintroduction of compulsory military service by conscription to include more women and younger test takers.

In 2018 and 2022, nearly all musters (i.e., test takers subject to compulsory military conscription) were aged 17-19, accounting for 99.9%. Among the voluntary test takers, only 23% were 17-19 years old. Given the large overlap between age and whether test takers are taking admission tests for basic military training as musters or voluntarily, older test takers are most often voluntary and vice versa, the main analyses are based on age group rather than selection group. This also enables a more intuitive division in the cohort structure, even before the reintroduction of compulsory military service by conscription. Unfortunately, we do not have information about the test takers' foreign background.

Measure

The first subtest in SEB-2000 – SYNONYMS A – was used for the present study. Each item contains a target word and five response options, where the test taker is asked to select the correct synonym. For example, which word has the same meaning as hybrid (i.e., target word), a. counterfeit, b. crossbreed, c. duplicate, d. replica, or e. intersection (i.e., response options). A correct response to a test item is scored as one [1], and an incorrect response to a test item is scored as zero [0]. In total, SYNONYMS A comprises 20 items with varying content and difficulty; however, its construction lacks a documented or explicitly articulated theoretical basis for its composition. Nevertheless, confirmatory factor analyses have shown that the verbal subtests load on one specific factor in addition to the general factor (G, i.e., GMA), namely crystallized intelligence (Gc; Carlstedt & Gustafsson, 2005). Due to test confidentiality, item content cannot be presented in further detail.

Table 1.
Sex, age, and selection group distributions 2014, 2018, and 2022.

			2014	2018	2022	Total
			n=5 281	n = 13 396	n = 20 609	n = 39 299
Gender	Male	n	4 460	10 260	16 430	31 150
		% within year	84	77	80	79
	Female	n	821	3 140	4 188	8 149
		% within year	16	23	20	21
Age		Mean (SD)	20.9 (3.6)	18.9 (2.5)	18.5 (2.0)	19.0 (2.6)
		Median (IQR)	20 (19-22)	18 (18-18)	18 (18-18)	18 (18-19)
Age groups	17-19 years	n	2 490	11 054	18 999	32 543
		% within year	47	82	92	83
		Mean (SD) within year	18.6 (0.5)	18.1 (0.3)	18 (0.3)	18.1 (0.3)
		Median (IQR) within year	19 (18-19)	18 (18-18)	18 (18-18)	18 (18-18)
	>20 years	n	2 791	2 346	1 619	6 756
		% within year	53	18	8	17
		Mean (SD) within year	22.8 (4.0)	23.0 (3.8)	23.9 (4.3)	23.2 (4.0)
		Median (IQR) within year	22 (20-24)	22 (20-24)	23 (21-26)	22 (20-24)
Selection group	Musters	n	-	9 388	17 649	27 037
		% within year	-	70	86	69
	Volunteers 2018 or 2022	n	-	4 012	2 965	6 977
		% within year	-	30	14	18
	Volunteers 2014	n	5 276	-	-	5 276
		% within year	100	-	-	100
	Unkown	n	5	0	4	9

SD=Standard deviation
IQR = Interquartile range

Data Analysis

The Rasch model provides separate estimates of person ability and item difficulty on a conjoint logit scale. A Rasch analysis was conducted to examine if the observed data fit the model well enough to be considered satisfying measurements (Rasch, 1960; Cano & Hobart, 2011). For a more comprehensive presentation of the Rasch model and the Rasch analysis methods, we refer the reader to other sources (c.f., Hobart & Cano, 2009; Andrich & Marais, 2019).

For identifying items underfitting or overfitting, we evaluated bootstrap-based item-restscore, which iterate sampling with replacement from the test takers and reports the percentage of iterations where misfit is indicated by statistically significant misfit after Benjamini-Hochberg correction (Johansson, 2025). To account for our large data set, we used

200 iterations with a sample size of 800 to minimize the risk of inflated statistical significance while retaining power to detect item misfit. Because there is no established detection rate for the proportion of bootstrap replications indicating misfit, we inspect the conditional item characteristic curves (CICC; Buchardt et al., 2023) iteratively for each item. Based on these two pieces of evidence, we decided whether items were sufficiently misfitting to exclude or could be retained. Analyses of patterns in Yen's Q^3 residuals (i.e., correlations between item residuals after accounting for the latent trait) were conducted for the evaluation of local dependency (LD; Christensen et al., 2017). Since the sampling distribution for Yen's Q^3 residuals is unknown, we used a simulation-based cutoff (Johansson, 2024b): the 99.5% percentile value from 1000 simulations.

Invariance was evaluated using DIF analyses based on Rasch trees, which apply score-based tests of parameter instability and recursively partition the sample by covariates (Strobl et al., 2015). Splits were retained when Bonferroni-adjusted p-values were ≤ .05, and each flagged item was then evaluated for DIF magnitude. DIF size is the absolute difference in item difficulty between subgroups; values ≥ 0.50 logits generally are considered noteworthy (Bond et al., 2020). DIF was primarily assessed for three subgroups: cohort year (2014, 2018, and 2022), sex (male and female), and age (17-19 or 20 years and older, due to a limited number of older test takers), and their interaction effects. Corresponding analyses with selection groups can be found in the Supplementary.

While reliability is contingent upon fulfillment of the other measurement properties, it provides complementary information about how the test works. A person separation index (PSI) was used, which indicates how reliably persons are separated by the test (Andrich, 1982).

To examine whether conclusions about subgroup and temporal differences in lexical ability are robust to changes in measurement properties, we estimated linear models using both the full item set and a refined item composition. The models included year, sex, age group, and selection group as fixed effects. After evaluating candidate interaction terms, the final model retained year × age group and sex × age group to allow subgroup trends to differ over time and by sex. Further details on model selection and diagnostics are provided in the Supplementary.

Secondly, to examine whether changes in item composition affected individual estimates of lexical ability, we compared estimates of person abilities derived from the full item set and a refined item composition using the Bland–Altman approach (Bland & Altman, 1999). Prior to the comparison, person ability estimates were anchored on a common latent scale using a Rasch calibration performed in the *mirt* package (Chalmers, 2012). Subsequently, the proportion of test takers with significantly higher or lower estimates was tabulated by year, sex, age group, and selection group to evaluate whether discrepancies between the full and refined item sets were systematically associated with subgroup membership or time period. Differences in distributions were tested using Pearson's chi-square tests and summarized with Cramér's V to quantify association strength.

All analyses were conducted in R version 4.4.1 and R Studio version 2024.09.0+375. The main package used for the Rasch analysis was easyRasch version 0.3.2 (Johansson, 2024a) along with its conditional packages. To evaluate possible differences in lexical ability between subgroups and across years, linear mixed models were fitted using the lme4 (Bates et al., 2015) and marginaleffects (Arel-Bundock et al., 2025) packages. Fully documented analysis reports, including codes and specific packages, were made using the scientific publication system Quarto (Allaire et al., 2022) and are available in the Supplementary at https://osf.io/dapwr/files/b8gwq.

3. Results

The results are presented in two sections. The first section responds to the study's first aim, providing the result of the measurement properties and invariance tests over time and between subgroups and the second section addresses the second aim by reporting the implications of changed measurement properties when assessing lexical ability among test takers for admission tests for basic military training in Sweden in 2014, 2018, and 2022.

Measurement Properties

In the first Rasch analysis, comprising the full sample of 39 299 test takers, there were major issues with misfitting items. Four items showed fit statistics that misfitted in almost all iterations of item-rest scores (>99%); SYNONYMS 11, 12, and 20 showed underfit, and SYNONYM 2 overfit (Table 2). This was also corroborated by the CICC plots. Furthermore, several DIF issues were identified (Table 3). In particular, SYNONYMS 16 and 19 showed the largest DIF sizes (1.177 logits for gender and 1.136 logits for years; Supplementary), and further issues with DIF were evident as an interaction effect by year and sex (Table 3). In this first analysis, most issues with LD were identified with SYNONYMS 1-5 and SYNONYM 7 (Supplementary).

Table 2.
Summary of bootstrapped item-rest score fit analyses based on 200 iterations with a sample size of 800.

Item	20 items (original)		14 items		8 items		6 items (refined model)	
	Item-rest score results	% of iterations	Item-rest score results	% of iterations	Item-rest score results	% of iterations	Item-rest score results	% of iterations
SYNONYM1	Overfit	82.0	Overfit	47.5	Overfit	12.0		*
SYNONYM2	Overfit	100.0						
SYNONYM3	Overfit	65.5	Overfit	24.5				
SYNONYM4	Overfit	27.5		*		*		
SYNONYM5	Overfit	36.0	Overfit	11.0				
SYNONYM6		*		*	Underfit	9.5		
SYNONYM7	Overfit	68.5	Overfit	33.0				
SYNONYM8		*	Underfit	7.0				
SYNONYM9	Overfit	82.0	Overfit	14.5	Overfit	18.0	Overfit	13.0
SYNONYM10	Underfit	18.5	Underfit	66.5				
SYNONYM11	Underfit	100.0						
SYNONYM12	Underfit	99.5						
SYNONYM13	Overfit	41.5		*		*		*
SYNONYM14	Overfit	89.0	Overfit	24.5	Overfit	22.0	Overfit	9.0
SYNONYM15	Overfit	16.0		*		*		*
SYNONYM16	Underfit	16.5						
SYNONYM17	Underfit	13.5	Underfit	65.5				
SYNONYM18	Overfit	19.0		*		*		*
SYNONYM19		*						
SYNONYM20	Underfit	99.5						

* Less than 5% iterations showed overfit or under

Table 3.
DIF analyses with mean location and locations separated for 2014, 2018 and 2022.

	20 items (original)									6 items (refined model)							
Item	Mean location	StDev	Male 2014 location	Female 2014 location	Male 2018 location	Female 2018 location	Male 2022 location	Female 2022 location	MaxDiff	Mean location	StDev	Male 2014 location	Female 2014 location	Male 2018 location	Male 2022 location	Female 2018 & 2022 location	MaxDiff
SYNONYM1	-2.112	0.276	-2.267	-2.583	-1.821	-2.014	-1.916	-2.073	0.762 *	-2.063	0.365	-2.234	-2.589	-1.719	-1.733	-2.038	0.871*
SYNONYM2	-2.517	0.338	-3.074	-2.449	-2.690	-2.125	-2.521	-2.247	0.949 *								
SYNONYM3	-1.355	0.526	-1.896	-1.825	-1.593	-1.027	-1.267	-0.520	1.376 *								
SYNONYM4	-0.496	0.455	-0.594	-1.222	-0.196	-0.665	0.114	-0.410	1.336 *								
SYNONYM5	-2.250	0.322	-1.836	-1.853	-2.384	-2.374	-2.481	-2.569	0.733 *								
SYNONYM6	0.468	0.335	0.142	0.069	0.445	0.475	0.730	0.943	0.874 *								
SYNONYM7	-1.334	0.467	-1.875	-0.813	-1.811	-0.937	-1.543	-1.024	1.062 *								
SYNONYM8	-1.034	0.519	-1.136	-1.867	-0.691	-1.220	-0.341	-0.946	1.526 *								
SYNONYM9	-0.817	0.353	-0.402	-0.379	-0.911	-0.954	-1.265	-0.993	0.886 *	-0.719	0.289	-0.333	-0.522	-0.788	-1.038	-0.923	0.705*
SYNONYM10	0.564	0.115	0.743	0.462	0.664	0.516	0.533	0.462	0.282								
SYNONYM11	1.115	0.221	1.267	0.715	1.298	1.014	1.236	1.159	0.583								
SYNONYM12	1.348	0.317	1.118	1.000	1.720	1.331	1.747	1.174	0.748 *								
SYNONYM13	-0.122	0.251	-0.151	0.380	-0.285	-0.200	-0.205	-0.274	0.665 *	0.021	0.179	-0.064	0.283	-0.100	0.127	-0.143	0.426
SYNONYM14	0.087	0.335	-0.032	0.462	-0.217	0.363	-0.353	0.297	0.815 *	0.170	0.238	0.064	0.374	-0.025	-0.037	0.475	0.513*
SYNONYM15	0.952	0.146	0.793	1.109	0.800	1.084	0.871	1.053	0.315	1.164	0.154	0.968	1.104	1.107	1.337	1.302	0.369
SYNONYM16	0.058	0.752	1.147	-0.236	0.607	-0.694	0.267	-0.746	1.893 *								
SYNONYM17	2.977	0.206	2.623	2.865	2.985	3.077	3.128	3.182	0.560 *								
SYNONYM18	1.153	0.185	1.350	1.324	1.159	1.208	0.876	1.003	0.474	1.427	0.123	1.599	1.349	1.514	1.344	1.327	0.272
SYNONYM19	1.257	0.645	1.680	2.297	1.003	1.273	0.608	0.679	1.690 *								
SYNONYM20	2.060	0.325	2.400	2.544	1.915	1.869	1.781	1.851	0.763 *								

* DIF size > 0.5 logits

Thus, the implications of removing items were iteratively evaluated and can be summarized into three steps of action. Supplementary shows all analyses in each of the four steps: first, for all 20 items; second, a slight improvement with 14 items; third, further refinements with eight items; and last, a six-item version with almost satisfactory measurement properties. The six-item version is hereafter referred to as the refined model. Between the first and second analytical steps, six items were deleted: those with the largest underfit (SYNONYM 11, 12, and 20), overfit (SYNONYM 2) and DIF (SYNONYMS 16 and 19). Between the second and third analytical step, six additional items were deleted; two items due to large underfit (SYNONYMS 10 and 17) and four items due to interaction effects between year and sexes in the DIF analyses (SYNONYMS 4, 7, 8, and 9). Between the third and fourth analytical steps, two further items were removed: SYNONYM 6 due to DIF interaction effects between year and sexes, and SYNONYM 5 due to LD, resulting in the refined model.

Furthermore, the sample-to-item was slightly off-target in all steps (person abilities for 20 items: *M* 0.38, *SD* 1.28; with 14 items: *M* 0.66, *SD* 1.53; with 8 items: *M* 0.80, *SD* 1.57; and for 6 items: *M* 0.62, *SD* 1.59). Finally, the reliability PSI decreased when items were removed; from 0.79 with 20 items, to 0.73 with 14 items to 0.51 with 8 items, and further to 0.33 in the refined model.

Lexical Ability

Figures 1a-b show trendlines of estimated lexical ability from the refined model for sexes and age groups over time. The patterns are similar when the original 20 items are used to estimate lexical ability (Supplementary). Irrespective of whether the refined 6-item version with improved measurement properties or the original version with 20 items were used to estimate lexical ability, the same general developmental and group-level trends were observed when comparing differences in lexical ability in the linear regression models with interaction effects between year, and the three other subgroup variables (sex, age, and selection group [i.e., musters or volunteers]) respectively (Table 5, Table 6). However, the differences are systematically smaller when the original 20-item version is used then when to the refined 6-item version is used.

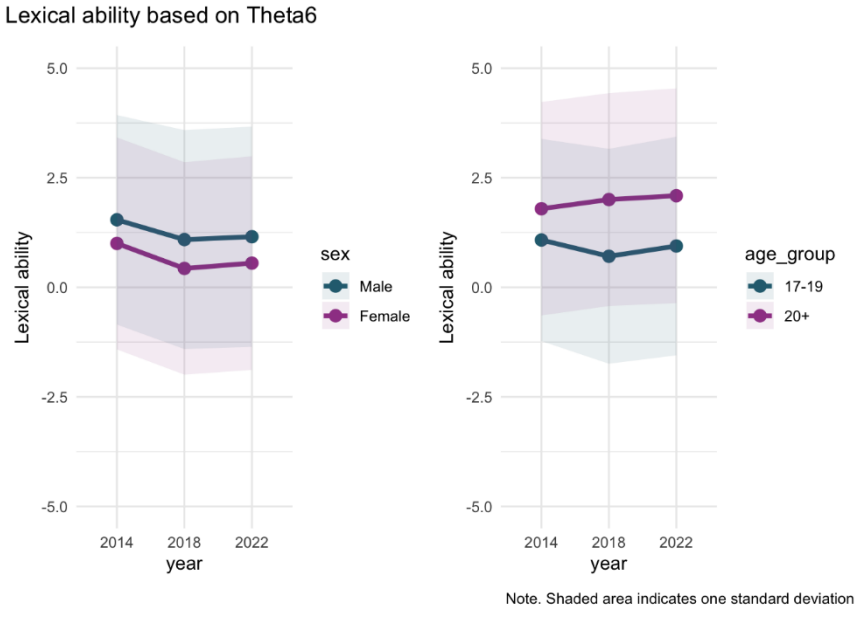

Figure 1a-b.
Trendlines of estimated lexical ability from the refined model for sexes (right figure) and age groups (left figure) over time.

Table 5.
Predicted means and marginal effects for lexical ability from the linear regression models with interaction effects based on the refined model (6 items), by subgroup and year.

Subgroup	Year	Predicted Mean (95% CI)	Marginal Effect (95% CI)	*p*-value
Sex	2014	Male = 1.52 [1.46–1.59]; Female = 1.09 [0.99–1.19]	Female – Male = -0.43 [-0.52, -0.34]	<.001
	2018	Male = 1.09 [1.05–1.16]; Female = 0.42 [0.36–0.48]	Female – Male = -0.53 [-0.60, -0.48]	<.001
	2022	Male = 1.16 [1.12–1.19]; Female = 0.54 [0.48–0.60]	Female – Male = -0.57 [-0.63, -0.51]	<.001
Age group	2014	17–19 = 1.08 [0.98–1.18]; 20 years or older = 1.79 [1.70–1.88]	20+ – 17–19 = 0.71 [0.58, 0.84]	<.001
	2018	17–19 = 0.71 [0.66–0.75]; 20 years or older = 2.00 [1.90–2.10]	20+ – 17–19 = 0.55 [0.40, 0.70]	<.001
	2022	17–19 = 0.94 [0.91–0.98]; 20 years or older = 2.09 [1.97–2.21]	20+ – 17–19 = 0.77 [0.59, 0.94]	<.001

Note: Estimates come from lm(Theta6 ~ year + sex + age_group + sel_group + year:age_group + sex:age_group). Predicted means represent the model-based expected values of lexical ability for each subgroup and year, holding other variables constant. Marginal effects indicate the estimated change in lexical ability associated with a one-unit difference in the variable of interest, averaged across the sample.

Table 6.
Predicted means and marginal effects for lexical ability from the linear regression models with interaction effects based on the original 20 items, by subgroup and year.

Subgroup	Year	Predicted Mean (95% CI)	Marginal Effect (95% CI)	*p*-value
Sex	2014	Male = 0.83 [0.80–0.87]; Female = 0.74 [0.69–0.80]	Female – Male = −0.09 [−0.14, −0.04]	<.001
	2018	Male = 0.41 [0.39–0.43]; Female = 0.17 [0.13–0.20]	Female – Male = −0.16 [−0.20, −0.13]	<.001
	2022	Male = 0.35 [0.33–0.37]; Female = 0.14 [0.11–0.17]	Female – Male = −0.18 [−0.22, −0.15]	<.001
Age group	2014	17–19 = 0.58 [0.53–0.63]; 20 years or older = 1.03 [0.98–1.08]	20+ – 17–19 = 0.45 [0.38, 0.53]	<.001
	2018	17–19 = 0.21 [0.19–0.24]; 20 years or older = 1.01 [0.95–1.06]	20+ – 17–19 = 0.39 [0.31, 0.47]	<.001
	2022	17–19 = 0.26 [0.24–0.27]; 20 years or older = 0.95 [0.88–1.01]	20+ – 17–19 = 0.45 [0.36, 0.55]	<.001

Note: Estimates come from lm(Theta20 ~ year + sex + age_group + sel_group + year:age_group + sex:age_group). Predicted means represent the model-based expected values of lexical ability for each subgroup and year, holding other variables constant. Marginal effects indicate the estimated change in lexical ability associated with a one-unit difference in the variable of interest, averaged across the sample.

As shown in the Bland-Altman plot (Figure 2), some test takers have their lexical abilities overestimated or underestimated when the original 20 items are used compared to the six-item version with improved measurement properties. While most test takers' lexical ability was not significantly over- or underestimated, some differences were observed across gender and years. Chi-square test confirmed a statistically significant but negligible association between sex $\chi^2(2) = 27.03$, $p < .001$, Cramér's V = 0.03 and years $\chi^2(4) = 57.79$, $p < .001$, Cramér's V = 0.03. Males were slightly more likely to have underestimated lexical ability with the original 20 items, whereas females were more likely to have overestimated lexical ability with the original

20 items. Pairwise tests revealed that the proportion of test takers underestimated with the original 20 items was slightly higher in 2022 compared to 2014 and 2018, while differences in overestimation were small but significant in 2022 compared to earlier years.

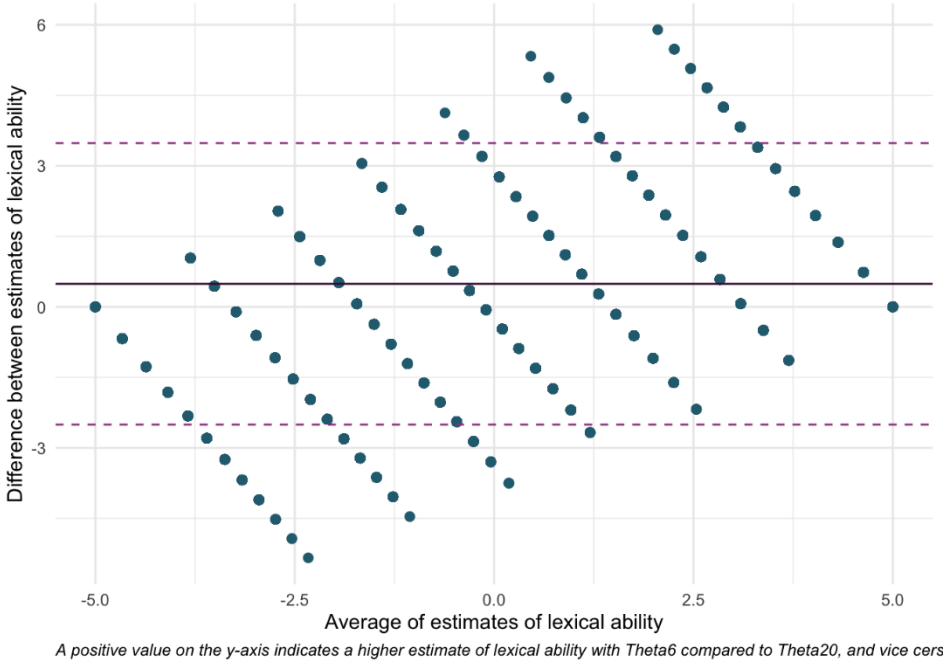

Figure 2.
Bland-Altman plot comparing estimated lexical ability based on the refined model with six items and the original 20 items.

4. Discussion

This study aimed to enhance current knowledge of changes in measurement properties may influence results from intelligence tests. By using data from a synonyms test from the Swedish SEB-2000 test, used for selection to basic military training, as an example, the main findings in this paper display fit properties and invariance issues in the synonym test. Yet it shows that the measurement properties could be improved by removing malfunctioning items. While the refined item composition improved measurement properties, except for reliability, it had only a minor impact at the group level when comparing lexical ability over time and across subgroups of test takers for admission tests to basic military training in Sweden.

The major issues of misfit and invariance in the synonym test studied here could be addressed by refining item composition, though this approach yielded lower reliability and a risk of construct underrepresentation. The implications of improved measurement properties for group-level comparisons were small, but it should be noted that individuals at the upper and lower ends are more affected than those in the middle. Specifically, the lexical ability of some test takers at the lower end of the scale was likely overestimated, whereas it was likely underestimated at the upper end of the scale if all items were used, including those not showing satisfactory measurement properties. This pattern is expected, as reliability and precision are lower at the extremes of the scale, which explains why the observed differences occur at the ends of the scale. Currently, when test takers are selected for military training and various positions, the information from SEB (i.e., not only a single synonym test as used here) is primarily based on cut-off scores that fall in the middle or slightly above or below. Thus, differences between upper and lower-end individuals may not have any substantial influence on selections today. Nevertheless, such effects may be important in the future, for instance, if

intelligence tests are used to match test takers to different positions or if selection thresholds must be lowered to further reinforce the Swedish Armed Forces. While it is also likely that similar problems with misfitting items and invariance issues may arise in other testing contexts, the magnitude and practical implications of such effects can naturally vary across settings and types of decision-making. Furthermore, regarding the effects of improved measurement properties when adding multiple subtests together to provide a measure of GMA, it is recommended to conduct further studies, for instance, by so-called differential test functioning (Stark et al., 2004), focusing on measurement properties at the test level (e.g., when results from subtests are combined to a global estimate).

Our results corroborate previous work, such as on the ASVAB and NIT, showing item drift across multiple verbal items (Chan et al., 1999; Shiu et al., 2013). Importantly, although the other studies covered longer periods, 16 and 72 years, respectively, this study only spanned eight years, while still indicating risks of item drift. Moreover, as discussed in the study of the Norwegian Armed Forces' word similarities test, certain vocabulary and concepts may have been more familiar and relevant at the time of the test's development, making them more challenging for newer generations (Nordmo et al., 2025). However, why certain words are easier or more difficult over time and in different groups remains to be explained, for instance, by constructing specification equations inspired by related word-based tests (Stenner et al., 2006; Melin et al., 2022; Bulut et al., 2023; Svetina et al., 2011) or through qualitative approaches that explore contextual and cultural factors influencing word familiarity. This can enable better designs of verbal items and guide how often verbal items need to be re-examined or replaced in test batteries.

The composition of subgroup characteristics differed across the annual cohorts in the analyses. The composition and representativeness of 17-19-year-old men in the admission testing were rather stable until 2006, but since then, with periods of compulsory or non-compulsory admission tests for basic military training and stressed gender equality in the Swedish Armed Forces, the estimated coverage has changed (Ludvigsson et al., 2022). Thus, the present study did not aim to evaluate trends in lexical ability but rather to examine the consequences of measurement properties when evaluating such trends. However, it is worth noticing a major difference in lexical ability between 2014 and 2018, when compulsory admission tests were reintroduced in 2017. Whether this is evidence for a general change in lexical ability or reflects a changed composition of subgroups remains to be further evaluated, preferably with a similar study design and including additional subtests of other intelligence components. Moreover, it should be noted that variations in subgroup composition, as demonstrated in this study, are likely to occur in other contexts as well, highlighting the necessity of carefully considering both the demographic structure under investigation and the broader societal or contextual factors that may exert influence.

Females generally perform better than males in the Swedish language subject in school (Swedish Association of Local Authorities and Regions, 2019; Swedish National Agency for Education, 2023) and verbal cognitive test (see summary in Roebuck-Spencer et al., 2008), yet males perform better on the synonym test here. Moreover, as reported by McGeown et al. (2016), females read more frequently, have more positive attitudes toward reading, higher reading motivation, greater confidence in their reading skills, and superior reading abilities than males. The fact that males perform better can be a sampling issue in the present study, possibly due to a first selection based on the mustering questionnaire (an online web survey including biodata questions). Nevertheless, it may also be related to, and questioning, the synonym test's validity. Could the synonyms used in the SEB-2000 test be "male-attributed" words, thereby introducing bias and disadvantaging female test takers? Lexical ability, as measured by the synonym test, is related to reading ability, which is strongly shaped by

motivation and interest in texts. For instance, in elementary school, females generally read more stories about relationships and romance, whereas males generally consume more action-packed content of science fiction or sports featuring male protagonists (c.f. Clark & Foster, 2005; Lepper et al., 2022; McGeown et al., 2016). However, at this stage, we have not investigated whether the synonyms used in the SEB-2000 test could be considered as typically "male-female-attributed" words. Thus, with possible new synonym tests and with the proportion of females increasing in the admission testing, deeper investigations are required to avoid biased disqualification of certain groups in military selection.

The Flynn Effect, that is, that more recent cohorts show higher test scores in the intelligence domains (Flynn, 2012; Williams, 2013), is well known. On the other hand, more recent studies have also presented a *negative Flynn Effect* (Dutton & Lynn, 2013; Meisenberg & Lynn, 2023), and it must also be noted that intelligence trends are too rich and diverse to be captured by any single construct (Flynn, 2012). Previously, Swedish musters between 1970 and 1993 showed greater improvement over the years in test takers' visual-spatial abilities, but little change and a slight decline in crystalized intelligence (Rönnlund et al., 2013). Others have proposed differences in trends across intelligence domains, which could be explained by societal cognitive development (Alwin & Pacheco, 2012) and a shift toward greater emphasis on rational thinking and analytic abilities (Flynn, 2012). Suppose the decrease in lexical ability presented here reflects a negative Flynn Effect. In that case, differences in one of many intelligence components or changes in group composition remain to be further studied. Importantly, a recent work on this topic on the Norwegian Armed Forces concludes that: *previous studies using the Norwegian GMA data must be interpreted with more caution, but that the test should measure males and females equally fairly* (Helland-Riise et al., 2024). This is of critical importance, likely also applicable in Sweden and other countries, and further work is warranted to ensure efficient and fair testing.

Methodological Considerations

First, it is worth noticing that many of the 20 synonym items studied here, until recently used for admission tests for basic military training in Sweden, do not fit the Rasch model. To our knowledge, no Rasch analysis has been conducted to evaluate measurement properties when implementing SEB to date. Thus, we cannot know when such misfitting issues have occurred or ascertain the reasons why so many items show poor fit to the measurement model. Furthermore, it must be noted that not even the refined model with six items fits the Rasch model perfectly, although a too-hardline statistical approach is not always desirable (Morel & Cano, 2017). Moreover, at the earlier analytical stages, alternative approaches could have included splitting items with DIF and forming testlets for locally dependent items, thereby retaining items and improving reliability. In the present case, where we sought to shed light on the importance of continuous evaluation of measurement properties – in particular examinations of possible item drift – and its implications, the refined model was considered satisfactory for our illustrative purposes.

As stated above, the removal of more than two-thirds of the items in the refined item set raises concerns about potential construct underrepresentation, that is, the extent to which a test fails to comprehensively capture the intended construct (Joint Committee on the Standards for Educational and Psychological Testing, 2014). It is important to emphasize that this study does not propose simply omitting the 14 excluded items in today's testing. Rather, it forms part of the SDCAA's ongoing efforts to modernize and improve the SEB. Drawing on historical data, our findings highlight the need for continuous test evaluation to ensure fairness and validity in selection procedures. However, decisions regarding how SDCAA should replace malfunctioning items or subtests lie beyond the scope of this study. Moreover, while item

removal may introduce risks of construct underrepresentation, it may simultaneously enhance construct-irrelevant variance (Messick, 1995), particularly when outdated items are removed. For example, a word such as pandemic—which was rarely used in everyday language before 2020—may now be easily recognized by most test takers, not because of improved lexical ability, but because of increased societal exposure. Over time, certain words may become more or less relevant to the intended construct and may therefore warrant removal or replacement due to construct relevance. Verbal ability tests must account for such linguistic and cultural shifts in order to preserve construct validity and minimize the introduction of irrelevant variance arising from changing patterns of language familiarity.

A close inspection of the CICC plots becomes important with a high misfit. For instance, SYNONYMS 11 and 17 showed a particularly high misfit and a significant deviation from the curve, with dots flatter than expected. This implies that the item does not discriminate enough and can be a sign of guessing (Andrich et al., 2012). In contrast to other IRT models, such as the 3PL model, the Rasch model does not include a separate guessing parameter. Notably, in the 3PL model, guessing is attributed to the item. On the other hand, it is persons and not items that guess (Andrich & Marais, 2019), and therefore, guessing is more likely to be considered an interaction between the propensity of an item to provoke guessing and the proclivity of a person to guess (Smith, 1993). While potential guessing could be explored further, we decided to remove those items from the refined model in our study.

In contrast to the work on the entire ASVAB (Chan et al., 1999) and the Estonian version of NIT (Shiu et al., 2013), this paper examined item drift only in one subtest in the SEB-2000. However, this paper also introduced the role of using Rasch analysis for including longitudinal studies of measurement properties of lexical ability intelligence tests to provide an enhanced understanding of how measurement properties may have changed over time and their implications for understanding trends in lexical ability. Nevertheless, we welcome similar studies focusing on the other subtests in the SEB-2000 and the newly launched SEB-2025, as well as on additional test batteries, in future research. We also invite further research incorporating additional data to evaluate the test's predictive validity and practical impact in selection contexts. Such efforts could contribute to the continued refinement of its measurement properties over time, as well as to longitudinal studies examining other domains of intelligence and the overall selection process.

5. Conclusion

This study aimed to enhance current knowledge of how changes in measurement properties may influence results from intelligence tests. Using data from the Swedish SEB-2000, the study identified fit and invariance issues in the synonym test, yet it also shows that the measurement properties could be improved by removing malfunctioning items. After identifying misfit and invariance issues, removing malfunctioning items improved the measurement properties of a synonym test used in the SEB-2000, but also lowered its reliability. While the implications of improved measurement properties for group-level comparisons were small, it must be noted that individuals at the upper and lower ends are more affected than those in the middle, which may threaten fairness in the selection process. Moreover, it is likely that similar problems with misfitting items and invariance issues may arise in other testing contexts, although the magnitude and practical implications of such effects can naturally vary across settings and types of decision-making.

Citing Chan et al. (1999, p. 617): "*Test development and revision is an ongoing process. Test developers and users should constantly evaluate the psychometrics of their items and tests to guard against the possibility of DIF over time*", which is corroborated in this study.

Importantly, the SDCAA has initiated such work to refine the existing SEB by replacing malfunctioning items and subtests to ensure high quality and fairness in their decision-making for admission to basic military training in Sweden. It is also likely to be needed in other countries and authorities that use admission tests over time. Continuous revision of admissions tests would not only benefit the individual organization by making test administration fairer and more efficient, but also enable the sharing of results, creating opportunities to design and trustfully manage tests better.

Given the evolving composition of test takers for basic military training in Sweden, including changes in sex, age, and selection groups, ongoing evaluations like this kind of study are essential to maintain fair selections. Furthermore, this study underscores the need for regular reassessments of measurement properties – and actions to ensure invariance –to better understand how intelligence may change over time. The results in this study primarily pertain to lexical ability among test takers for admission to basic military training in Sweden. However, the broader implications for test validity and fairness are relevant across many fields where standardized testing plays a critical role.

Funding and Conflict of Interest

This study is a part of the Swedish Defence Conscription and Assessment Agency's work to update the existing Swedish Enlistment Battery. The work has been partially funded by the Swedish Defence Conscription and Assessment Agency. Beyond that, the authors have no conflicts of interest to declare.

Data Availability Statement

The data that support the findings of this study have been retrieved from the Swedish Defence Conscription and Assessment Agency. Restrictions apply to the availability of these data, which were used under license for this study.

How to Cite

Melin, J., Wahlkrantz, E., Brandebo, M. F., Larsson, G., & Annell, S. (2026). Are intelligence measures stable across time and groups? A study of invariance in a synonyms test. *Educational Methods & Psychometrics, 4*: 22. https://doi.org/10.65301/emp.2026.262

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2022). *Quarto* (Version 1.2) [Computer software]. https://doi.org/10.5281/zenodo.5960048

Alwin, D. F., & Pacheco, J. (2012). Population trends in verbal intelligence in the United States. In *13. Population Trends in Verbal Intelligence in the United States* (pp. 338–368). Princeton University Press. https://doi.org/10.1515/9781400845569-015

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95–104. https://doi.org/10.3316/aeipt.13636

Andrich, D. (2004). Controversy and the Rasch Model: A characteristic of incompatible paradigms? *Medical Care*, *42*(Supplement), I–7. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics*, *37*(3), 387–416. https://doi.org/10.3102/1076998611411913

Andrich, D., & Marais, I. (2019). *A Course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice Items. *Journal of Educational and Behavioral Statistics*, *37*(3), 417–442.

Arel-Bundock [aut, V., cre, cph, Greifer, N., Bacher, E., McDermott, G., & Heiss, A. (2025). *marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests* (Version 0.30.0) [R package]. https://cran.r-project.org/web/packages/marginaleffects/index.html

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565–580. https://doi.org/10.1177/0146621612455090

Bäckström, P. (2022). Self-selection and recruit quality in Sweden's all volunteer force: Do civilian opportunities matter? *Defence and Peace Economics*, *33*(4), 438–453. https://doi.org/10.1080/10242694.2021.1903284

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

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Borsboom, D. (2006). When does measurement invariance matter?: *Medical Care*, *44*(Suppl 3), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Buchardt, A.-S., Christensen, K. B., & Jensen, N. (2023). Visualizing Rasch item fit using conditional item characteristic curves in R. *Psychological Test and Assessment Modeling*, *65*(2), 206–219.

Bulut, H. C., Bulut, O., & Arikan, S. (2023). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing*, *23*(1), 10–33. https://doi.org/10.1080/15305058.2022.2044821

Cano, S., & Hobart, J. (2011). The problem with health measurement. *Patient Preference and Adherence*, 279. https://doi.org/10.2147/PPA.S14399

Cano, S. J., Pendrill, L. R., Melin, J., & Fisher, W. P. (2019). Towards consensus measurement standards for patient-centered outcomes. *Measurement*, *141*, 62–69. https://doi.org/10.1016/j.measurement.2019.03.056

Carlstedt, B., & Gustafsson, J.-E. (2005). Construct validation of the Swedish Scholastic Aptitude Test by means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology*, 13.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

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

Chan, K.-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, *84*(4), 610–619. https://doi.org/10.1037/0021-9010.84.4.610

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's *Q3*: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Clark, C., & Foster, A. (2005). Children's and young people's reading habits and preferences: The who, what, why, where and when. In *National Literacy Trust*. National Literacy Trust. https://eric.ed.gov/?id=ED541603

Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland, 1997–2009. *Intelligence*, *41*(6), 817–820. https://doi.org/10.1016/j.intell.2013.05.008

Embretson, S. E. (2000). Polytomous IRT. In S. E. Embretson & S. P. Reise (Eds.), *Item Response Theory for psychologists* (2nd ed.).

Fisher, W. P. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287. https://doi.org/10.1016/j.measurement.2009.03.014

Flanagan, D. P., Alfonso, V. C., Ortiz, S. O., & Dynda, A. M. (2013). Cognitive assessment: Progress in psychometric theories of intelligence, the structure of cognitive ability tests, and interpretive approaches to cognitive test performance. In *The Oxford handbook of child psychological assessment* (pp. 239–285). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199796304.001.0001

Flynn, J. R. (2012). *Are we getting smarter?: Rising IQ in the twenty-first century*. Cambridge University Press.

Goldammer, P., Annen, H., Lienhard, C., & Jonas, K. (2024). An examination of model fit and measurement invariance of general mental ability and personality measures used in the multilingual context of the Swiss armed forces: A Bayesian structural equation modeling approach. *Military Psychology*, *36*(1), 96–113. https://doi.org/10.1080/08995605.2021.1963632

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407–434. https://doi.org/10.1207/s15327906mbr2804_2

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes* 15, 181. https://doi.org/10.1186/s12955-017-0755-0

Helland-Riise, F., Norrøne, T. N., & Andersson, B. (2024). Large-scale item-level analysis of the figural matrices test in the Norwegian armed forces: Examining measurement precision and sex bias. *Journal of Intelligence*, *12*(9). https://doi.org/10.3390/jintelligence12090082

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessessments*. 13(12):iii, ix-x, 1-177. doi: 10.3310/hta13120.

Husén, T. (1948). *Konstruktion och standardisering av svenska krigsmaktens inskrivningsprov: 1948 års version: redogörelse utarbetad vid CVB*. Centrala värnpliktsbyrån.

Johansson, M. (2024a). *easyRasch* (Version 0.3.1) [R package]. https://github.com/pgmj/easyRasch.

Johansson, M. (2024b). *Simulation based cutoff values for Rasch item fit and residual correlations*. R, Rasch, Etc. https://pgmj.github.io/simcutoffs.html

Johansson, M. (2025). Detecting item misfit in Rasch models. *Educational Methods & Psychometrics*, *3*. https://dx.doi.org/10.61186/emp.2025.5

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Jonsson, E., Salo, M., Lillemäe, E., Steder, F. B., Ferst, T., Kasearu, K., Novagrockiene, J., Österberg, J., Sederholm, T., Svensén, S., Tresch, T. S., & Truusa, T.-T. (2024). *Multifaceted conscription: A comparative study of six European countries* (1). *7*(1), Article 1. https://doi.org/10.31374/sjms.166

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Lepper, C., Stang-Rabrig, J., & McElvany, N. (2022). Gender differences in reading: Examining text-based interest in relation to text characteristics and reading comprehension. *Learning and Instruction*, *82*, 101680. https://doi.org/10.1016/j.learninstruc.2022.101680

Ludvigsson, J. F., Berglind, D., Sundquist, K., Sundström, J., Tynelius, P., & Neovius, M. (2022). The Swedish military conscription register: Opportunities for its use in medical research. *European Journal of Epidemiology*, *37*(7), 767–777. https://doi.org/10.1007/s10654-022-00887-0

McGeown, S. P., Osborne, C., Warhurst, A., Norgate, R., & Duncan, L. G. (2016). Understanding children's reading activities: Reading motivation, skill and child characteristics as predictors. *Journal of Research in Reading*, *39*(1), 109–125. https://doi.org/10.1111/1467-9817.12060

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Meisenberg, G., & Lynn, R. (2023). Ongoing trends of human intelligence. *Intelligence*, *96*, 101708. https://doi.org/10.1016/j.intell.2022.101708

Melin, J., Cano, S., Flöel, A., Göschel, L., & Pendrill, L. (2022). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests: Extension to word lists. *Entropy*, *24*(7). https://doi.org/10.3390/e24070934

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morel, T., & Cano, S. J. (2017). Measuring what matters to rare disease patients – reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. *Orphanet Journal of Rare Diseases*, *12*, 171. https://doi.org/10.1186/s13023-017-0718-x

Muthukrishna, M., & Henrich, J. (2016). Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1690), 20150192. https://doi.org/10.1098/rstb.2015.0192

Nordlund, P. (2022). Sweden and Swedish defence – Introduction to the special issue. *Defence and Peace Economics*, *33*(4), 387–398. https://doi.org/10.1080/10242694.2021.2003529

Nordmo, M., Norrøne, T. N., & Lang-Ree, O. C. (2025). Reevaluating the Flynn effect, and the reversal: Temporal trends and measurement invariance in Norwegian armed forces intelligence scores. *Intelligence*, *110*, 101909. https://doi.org/10.1016/j.intell.2025.101909

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35. https://doi.org/10.1080/19315775.2014.11721702

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reckase, M. D. (2009). Unidimensional item response theory models. In M. D. Reckase (Ed.), *Multidimensional item response theory* (pp. 11–55). Springer. https://doi.org/10.1007/978-0-387-89976-3_2

Roebuck-Spencer, T. M., Reeves, D. L., Bleiberg, J., Cernich, A. N., Schwab, K., Ivins, B., Salazar, A., Harvey, S., Brown, F., & Warden, D. (2008). Influence of demographics on computerized cognitive testing in a military sample. *Military Psychology*, *20*(3), 187–203. https://doi.org/10.1080/08995600802118825

Rönnlund, M., Carlstedt, B., Blomstedt, Y., Nilsson, L.-G., & Weinehall, L. (2013). Secular trends in cognitive test performance: Swedish conscript data 1970–1993. *Intelligence*, *41*(1), 19–24. https://doi.org/10.1016/j.intell.2012.10.001

Shiu, W., Beaujean, A. A., Must, O., te Nijenhuis, J., & Must, A. (2013). An item-level examination of the Flynn effect on the national intelligence test in Estonia. *Intelligence*, *41*(6), 770–779. https://doi.org/10.1016/j.intell.2013.05.007

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, *33*, 75–102. https://doi.org/10.2333/bhmk.33.75

Skolverket. (2023). *PIRLS 2021 Läsförmågan hos svenska elever i årskurs 4 i ett internationellt perspektiv* (SKOLV-R-2023:4-SE). Skolverket. https://www.skolverket.se/publikationer?id=11490

Smith, R. (1993). Guessing and the Rasch model. *Rasch Measurement Transactions*, *6*(4), 262–263.

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: what's old, what's new, what's next? *Organizational Research Methods*, *25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting Differential Item Functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Sveriges kommuner och landsting. (2019). *Könsskillnader i skolresultat*. Sveriges kommuner och landsting. https://skr.se/skr/tjanster/rapporterochskrifter/publikationer/konsskillnaderiskolresultat.65164.html

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*(1), 1–23. https://doi.org/10.1080/15305058.2010.518261

Ullstadius, E., Carlstedt, B., & Gustafsson, J.-E. (2008). The multidimensionality of verbal analogy items. *International Journal of Testing*, *8*(2), 166–179. https://doi.org/10.1080/15305050802001243

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, *41*(6), 753–764. https://doi.org/10.1016/j.intell.2013.04.010

Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236. https://doi.org/10.1007/s11336-013-9327-3

Manuscript Received: 09 SEPT 2025
Final Version Received: 11 DEC 2025
Published Online Date: 15 JAN 2026