General cognitive ability assessment in the German National Cohort (NAKO) – The block-adaptive number series task

Journal:	*The World Journal of Biological Psychiatry*
Manuscript ID	Draft
Manuscript Categories:	Original Investigation/Summaries of Original Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Schmiedek, Florian; DIPF Leibniz Institute for Research and Information in Education; Goethe University Frankfurt Institute of Psychology; Center for Mind, Brain and Behavior, University of Marburg and Justus Liebig University Giessen, Germany Kroehne, Ulf; DIPF Leibniz Institute for Research and Information in Education Goldhammer, Frank; DIPF Leibniz Institute for Research and Information in Education; Center for International Student Assessment (ZIB) Prindle, John; University of Southern California Suzanne Dworak-Peck School of Social Work Lindenberger, Ulman; Max-Planck-Institute for Human Development, Lifespan Psychology; Max Planck UCL Centre for Computational Psychiatry and Ageing Research Klinger-König, Johanna; University Medicine Greifswald, Department of Psychiatry and Psychotherapy Grabe, Hans; University Medicine Greifswald Department of Psychiatry and Psychotherapy; German Centre for Neurodegenerative Diseases (DZNE), Partner Site Rostock/Greifswald Riedel-Heller, Steffi; Institute of Social Medicine, Occupational Health and Public Health (ISAP), Medical Faculty, University of Leipzig Pabst, Alexander; Institute of Social Medicine, Occupational Health and Public Health (ISAP), Medical Faculty, University of Leipzig

	Streit, Fabian; Central Institute of Mental Health, Division of Genetic Epidemiology in Psychiatry Zillich, Lea; Central Institute of Mental Health, Division of Genetic Epidemiology in Psychiatry Kleineidam, Luca; University Hospital Bonn, Department of Neurodegenerative Diseases and Geriatric Psychiatry; German Center for Neurodegenerative Diseases (DZNE), Wagner, Michael; University Hospital Bonn, Department of Neurodegenerative Diseases and Geriatric Psychiatry; German Center for Neurodegenerative Diseases (DZNE) Rietschel, Marcella; Central Institute of Mental Health, Division of Genetic Epidemiology in Psychiatry Rujescu, Dan; Martin-Luther-University Halle-Wittenberg, University Clinic for Psychiatry, Psychotherapy and Psychosomatic Schmidt, Börge; University Hospital Essen, Institute for Medical Informatics, Biometry and Epidemiology Berger, Klaus; Westfälische Wilhelms-Universität Münster Fachbereich 05 Medizinische Fakultät, Institute of Epidemiology and Social Medicine Investigators, NAKO; Westfälische Wilhelms-Universität Münster
Keywords:	Reasoning, Rasch Model, Neuropsychological tests, Prospective study, Neuropsychiatry

General cognitive ability assessment in the German National Cohort (NAKO) – The block-adaptive number series task

Florian Schmiedek[1,2,3], Ulf Kroehne[1,2], Frank Goldhammer[1,2,4], John J. Prindle[5], Ulman Lindenberger[6,7], Johanna Klinger-König[8], Hans J. Grabe[8,9], Steffi G. Riedel-Heller[10], Alexander Pabst[10], Fabian Streit[11], Lea Zillich[11], Luca Kleineidam[12,13], Michael Wagner[12,13], Marcella Rietschel[11], Dan Rujescu[14], Börge Schmidt[15], NAKO Investigators[16], Klaus Berger[17]

[1]*Department of Education and Human Development, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany*
[2]*Institute of Psychology, Goethe University, Frankfurt am Main, Germany*
[3]*Center for Mind, Brain and Behavior, University of Marburg and Justus Liebig University Giessen, Germany*
[4]*Center for International Student Assessment (ZIB), Frankfurt am Main, Germany*
[5]*School of Social Work, University of Southern California Suzanne Dworak-Peck, USA*
[6]*Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany*
[7]*Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany, and London, UK*
[8]*Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany*
[9]*German Centre for Neurodegenerative Diseases (DZNE), Partner Site Rostock/Greifswald, Greifswald, Germany*
[10]*Institute of Social Medicine, Occupational Health and Public Health (ISAP), Medical Faculty, University of Leipzig, Germany*
[11]*Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Heidelberg, Medical Faculty Mannheim, Mannheim, Germany*
[12]*Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany*
[13]*German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*
[14]*University Clinic for Psychiatry, Psychotherapy and Psychosomatic, Martin-Luther-University Halle-Wittenberg, Germany*
[15]*Institute of Medical Informatics, Biometry, and Epidemiology (IMIBE), University of Duisburg-Essen, University Hospital Essen, Essen, Germany*
[16]*NAKO Investigators: listed at the end of the manuscript*
[17]*Institute of Epidemiology and Social Medicine, University of Münster, Germany*

Corresponding author:
Name: Florian Schmiedek
Address: DIPF | Leibniz Institute for Research and Information in Education,
Rostocker Str. 6, 60323 Frankfurt am Main, Germany
Email: schmiedek@dipf.de
Telephone: +49 69 24708 820
Fax: +49 69 24708 601

General cognitive ability assessment in the German National Cohort (NAKO) – The block-adaptive number series task

Objectives. Evaluate the block-adaptive number series task of reasoning, as a time-efficient proxy of general cognitive ability in the Level-2 sample of the German National Cohort (NAKO), a population-based mega cohort.

Methods. The number series task consisted of two blocks of three items each, administered as part of the touchscreen-based assessment. Based on performance on the first three items, a second block of appropriate difficulty was automatically administered. Scoring of performance was based on the Rasch model. Relations of performance scores to age, sex, education, study centre, language proficiency, and scores on other cognitive tasks were examined.

Results. Except for one very difficult item, the data of the remaining 14 items showed sufficient fit to the Rasch model (Infit: 0.89–1.04; Outfit: 0.80–1.08). The resulting performance scores ($N = 21{,}056$) had a distribution that was truncated at very high levels of ability. The reliability of the performance estimates was satisfactory. Relations to age, sex, education, and the executive function factor of the other cognitive tasks in the NAKO supported the validity.

Conclusions. The number series task provides a valid proxy of general cognitive ability for the Level-2 sample of the NAKO, based on a highly time-efficient assessment procedure.

Keywords: reasoning, Rasch model, cognitive aging, education, German National Cohort

Introduction

General cognitive ability, or intelligence, is a psychological construct of phenotypical differences that are of utmost importance for a wide range of outcomes, from educational attainment (Deary et al. 2007), to vocational achievement (Schmidt and Hunter 1998; Strenze 2007), and problem solving in everyday situations (Allaire and Marsiske 1999). Importantly, general cognitive ability has also been linked to health literacy (Fawns-Ritchie et al. 2018), health behaviour (Wraw et al. 2018; Davies et al. 2019), health outcomes (Wraw et al. 2015), and mortality risk (Lindenberger et al. 2002; Batty et al. 2007; Calvin et al. 2011, 2017; Christensen et al. 2016). Based on such findings, Deary and Batty (2007) have put forward the concept of "cognitive epidemiology" and proposed intelligence to be an important factor for disease outcomes. In a mega cohort study, like the German National Cohort (NAKO) (German National Cohort (NAKO) Consortium 2014), an important aim is to capture health and health-related behaviour across a broad range of age, education, and health status in the general population. An assessment of general cognitive ability is therefore highly desirable, besides more specific assessments of neuropsychological functions (like memory or cognitive control) that are of neuropsychological importance for, for example, depression, schizophrenia, or dementia (see Kleineidam et al. 2021, submitted).

Over one-hundred years of psychometric research on the structure of cognitive abilities have resulted in a hierarchical structure of intelligence (Carroll 1993) with narrowly defined specific abilities at the bottom, which are nested in broad abilities (like fluid and crystallized intelligence) in the middle, and a general factor (g) on top of the hierarchy. Of the broad abilities, reasoning (often termed fluid intelligence) shows a high latent relationship with g. In some theoretical conceptualizations, it is even considered identical to it (Gustafsson 1984) and also highly related to working memory capacity (Conway et al. 2003). When attempting to measure general cognitive functioning, reasoning therefore arguably is the most important broad psychometric ability to focus on. Reasoning itself is also hierarchically structured (Carroll 1993), so that a comprehensive assessment typically involves tasks of different paradigms (like the completion of series, matrices, and analogies) and content (i.e., verbal, numerical, or figural-spatial material) of inductive and deductive reasoning. The Woodcock-Johnson III (WJ-III) Test (Woodcock et al. 2001), for example, comprises five tasks of reasoning and the Berlin Intelligence Structure Test (Jäger et al. 1997) even 15 such tasks, five for each content domain (verbal, numerical, figural-spatial). Such comprehensive assessments require substantial testing time, which is prohibitive for study protocols like that of the NAKO (see Berger et al. 2021, submitted). In fact, even single tasks from such standard tests would often require too much time.

Computerized adaptive testing offers an elegant way to make testing more time-efficient (Gershon 2005; van der Linden 2018). In adaptive testing, not all participants are given the same set of items. Instead, items are chosen based on the performance on previous items such that the item difficulty matches the current estimate of the participant's ability level. A simple way to implement adaptive testing is to use a block-adaptive design, in which all participants work on the same initial set of items and then – based on how many items they answered correctly – are given a second set of items whose average difficulty depends on the previous performance (Yan et al. 2014). Such a procedure was chosen for the number series task that was used in the NAKO as a proxy of reasoning.

Number series tasks require inductive reasoning to identify the arithmetic rule behind a sequence of numbers and use this rule to calculate a missing element (e.g., identify the rule "alternate between +2 and +3" in the sequence "3 5 8 10 13 __" and complete it with the correct answer "15"). They are among the most widely used indicators of reasoning in the history of psychometric research on intelligence (Carroll 1993). Here, an adaptation of a block-adaptive number series task from the WJ-III is used, which has been applied in several studies and versions. Examples are a telephone-interview-based version in the Health and Retirement Survey (HRS; Fisher et al. 2013), an internet-based online version in the American Life Panel (ALP; Angrisani et al. 2016; Pollard and Baird 2017), a non-adaptive computer-based version in the Programme for the International Assessment of Adult Competencies-Longitudinal PIAAC-L; Engelhardt and Goldhammer 2018, 2019), and a block-adaptive computer-based version in the Berlin Aging Study II (BASE II; Bertram et al. 2014). This latter version has been implemented and used in a subsample of the NAKO (i.e., the L2 assessment, see Materials and Methods of Berger et al. 2021, submitted) as part of a touchscreen-based computerized assessment. It comprises two blocks of three items each. Based on whether 0, 1, 2, or 3 items are answered correctly in the first block, one of four blocks of three (very easy, easy, medium, or hard) additional items is chosen by the computer program. This way, each participant has to work on just six items in total, which, on average, cover the ability level of any given individual participant more accurately than if everybody would get the identical fixed set of six items.

When different participants work on different subsets of items, the scoring of overall performance becomes a challenge. A possible solution is to use psychometric models based on Item Response Theory (IRT; Embretson and Reise, 2000), which allow estimating individual ability based on different subsets of items – given known item parameters and provided that certain model assumptions are met. For the number series task, we use the Rasch model (Rasch 1960), which links observed performance (items answered correctly or incorrectly) to a latent continuous ability via a logistic link function. As model parameters, a difficulty parameter (estimated from the data) is used for each item. Based on these difficulty parameters and the information which items were answered (in)correctly, estimates of the latent ability of each participant can be generated that are comparable across participants working on different (combinations of) blocks of items. An important assumption of the model is that the probabilities of correct responses follow logistic functions of the difference between person ability and item difficulty (located on the same continuous latent dimension). Herein, a lower asymptote of zero (i.e., correct responses due to guessing are negligible), an upper asymptote of one (i.e., with increasing ability, the probability of answering items correctly converges to perfect accuracy), and the same slope of the function (i.e., items discriminate between differences in ability equally well) is assumed. Furthermore, it is assumed that it does not matter which (subsets of) items the participants work on (i.e., that there are no interaction effects between items, like, e.g., order effects). Model fit can be evaluated for each item using item fit measures (e.g., Infit and Outfit, Linacre 2012). Such fit to the Rasch model has been demonstrated for the number series task items in the development of the WJR tests (Woodcock and Johnson 1977; Woodcock et al. 2001) and in successful applications of its adaptive versions in the HRS and ALP samples (Prindle 2012; Fisher et al. 2013; McArdle 2013; Prescott et al. 2018)

In the NAKO, a timeout procedure was used to keep the task administration within feasible time limits and prevent participants from working on single items for a long time. A maximum of 40sec was available for the processing of each item, limiting the total time for the number series task (including instruction and an example item) to about 5min. This “speeded” procedure introduces a possible contribution of mental speed, in addition to reasoning, into the ability assessment (Wilhelm and Schulze 2002; Goldhammer 2015).

Objectives of the present work were to a) apply the Rasch model to generate performance scores for the block-adaptive number series task, b) validate these scores with other variables known to be related to reasoning, and c) to provide general descriptive information regarding task performance in the NAKO (i.e., relations to sex, language proficiency, study centre). Main variables for validation were age, education, and a set of other cognitive tasks that were assessed in the NAKO. Based on findings for reasoning from cross-sectional adult lifespan samples (e.g., Schaie 1993; McArdle et al. 2002; Li et al. 2004; Salthouse 2019), a monotonic negative age relation of average performance on the number series task was expected for the NAKO. Regarding educational level, positive associations with average task performance could be expected based on theoretical propositions of reciprocal relations between education and cognitive ability (Lövdén et al. 2020), as well as based on findings from metanalyses (Strenze, 2007; Ritchie and Tucker-Drob, 2018). Other cognitive variables assessed in the NAKO can be combined into performance scores for memory and executive functioning (Kleineidam et al. 2021, submitted). Strong relations to reasoning have been demonstrated for factors of executive functions (comprising different selections of tasks from the domains of working memory, attention, inhibition, task switching, and verbal fluency) in healthy populations, with these relations arguably being predominantly produced by working memory tasks (Friedman et al. 2006; Rey-Mermet et al. 2019). For the number series task, we did therefore expect substantial correlations with the other cognitive tasks (most strongly with a working memory task) as well as a substantial loading on the executive functioning factor.

Materials and Methods

Participants

The NAKO is a population-based cohort study, examining 205,000 randomly selected participants in 18 study centres across Germany. Baseline examination took place between 2014 and 2019. This analysis is based on data from the first 101,663 participants summarized in the NAKO data freeze 100,000 (DF100K; data set NAKO-399). At baseline, data was acquired in the study centres at two levels: Level-1 (L1; ~3-4 hours) assessment was undergone by all participants, and a subset of ~20% of the subjects underwent the more detailed Level-2 assessment (L2; ~5 hours). The number series task was included in the L2 extended examination program and data for it were available from 21,073 participants aged 20-72 (see Berger et al. 2021, submitted, for details on the procedures of sampling, recruitment, and examination in the 18 study centres). Of those, 17 participants were excluded because they did not answer any of the six items in time. The total sample used for analyses was therefore *N* = 21,056 (see Table 1 for descriptive information on the sample). An overview of the assessment of neuropsychiatric functions and conditions (Berger et al. 2021, submitted) and detailed analyses of specific measures can be found elsewhere (Erhardt et al. 2021, submitted; Kleineidam et al. 2021, submitted; Klinger-König et al. 2021, submitted; Streit et al. 2021, submitted).

Block-adaptive numbers series task

The number series task was administered as part of the touchscreen assessment of the L2 sample. The task started with a screen explaining the task by providing an example item together with its solution ("1 2 __ 4"; solution: "3") and instructing participants always to provide an answer (i.e., guessing one if no solution is found) and to try to focus on accuracy rather than on speed of responding. Next, another example item was provided ("2 4 6 __") and participants were instructed to enter the correct answer ("8") using a number block on the touchscreen. Correct answers were one- or two-digit numbers. Then the first block of three items started with number sequences always shown on the top of the screen and the number block below. Incorrectly entered numbers could be deleted with an "erase" button. After entering an answer within the time limit, a "continue" button had to be pressed to get to the next item. If not answering within 20sec, a question appeared at the top of the screen ("Please answer the item. Do you need more time?"), which participants could answer with "yes" or "no". In case of an affirmative answer, up to an additional 20sec were provided after which the item was terminated with a message that the next item will now be presented. Items not answered within this timeout window were scored as incorrect. The maximum total time for each item was 40sec. After the first block, the program automatically chose the second block of the remaining three items (i.e., very easy, easy, medium, or hard). The maximum total time for all six items was 265sec (6*40s + 5*5s breaks in between items), implying a maximum testing time of about 5min (including instruction and example items).

Sociodemographics and other cognitive tests

As covariates for descriptive analyses of performance on the number series task, we used age (retrieved from participants' identity cards and rounded to full years), sex, and education (see Table 1). Based on the International Standard Classification of Education 97 (ISCED97 2003; available for $N = 19{,}359$), education was classified as lower (ISCED97 Level 1&2), intermediate (ISCED97 Level 3&4), and higher (ISCED97 Level 5&6) in accordance with Dragano and colleagues (2020). Participants who did not finish their education or could not be classified were coded as an additional group but not considered when assessing effects of education. German language proficiency was categorized according to self-reported mother tongue (German; German and another language, i.e., bilingual; non-German native speaker). In addition, for non-native speakers, language proficiency was rated by the study nurse using five categories (very high*,* high*,* average, low*,* and very low). Cognitive variables assessed in the L1 sample included a word list recall test (immediate, repeated, and delayed) of verbal memory, a verbal fluency test (animal names), a digit span backwards test of working memory, and the Stroop interference task of selective attention (see Kleineidam et al. 2021, submitted, for details).

Analysis methods

Data processing and statistical analyses were conducted using R (R Core Team 2017), using the dedicated packages TAM (Robitzsch et al. 2018) for IRT analyses, mgcv (Wood 2017) for generalized additive models, and lme4 (Bates et al. 2015) for mixed models. Confirmatory factor analyses were conducted using Mplus version 7.3 (Muthén and Muthén 2003). The conventional α level was set to .05.

Results

Accuracies and timeouts for single items

Accuracies

In Block 1, the first item (working as a warm-up and motivator) was answered correctly by almost everybody, while the second and third items were only answered correctly by 70% and 58% of the participants (Table 2). Accordingly, only a relatively small group was given the very easy item set, and the largest subgroup of participants (49%) got the hardest item set in Block 2. The average percentages of correctly answered items were 67%, 63%, 33%, and 30% in the increasingly difficult versions of Block 2. The last item in the hardest version (Item O1) turned out to be too difficult (<1% correct). Other than that, the adaptive procedure successfully kept the performance levels for single items from floor and ceiling effects.

Number of timeouts

The number of timeouts was moderate (<20%) for most of the items, with some exceptions (Items F1, J1[very first item], and M1/N1/O1[three hardest items]; see Table 2). A timeout was reached at least once by 16,064 participants, but only 294 participants (1.4%) reached three or more timeouts. Only 17 participants did not answer any item within time limits, and were excluded from the analyses. Importantly, the average number of timeouts only very weakly correlated with age (r = .020; t[21054] = 2.937; p = .003) and was comparable across subsamples grouped by age decades (Range: 1.190–1.359), indicating that older age groups were not overly disadvantaged in their opportunity to respond in time.

Rasch model

Fitting the Rasch model resulted in acceptable item fit statistics (Infit and Outfit measures within the recommended range of 0.80–1.20; see Table 2) for all items except for the overly difficult Item O1. We therefore re-estimated the model without this item, leading to only slight changes in item difficulties and fit measures of the remaining items (Table 2). The following analyses are based on, and performance scores derived from, these remaining 14 items.

Scoring and distribution of ability scores

Reliability estimates for the ability scores derived from the Rasch model were satisfactory (weighted maximum likelihood reliability: .60; expected-a-posteriori reliability: .70; see Adams 2005). Ability scores were rescaled to “W Scores” based on the linear transformation W = 500 + 9.1024*Estimate (Woodcock and Dahl 1971). As there are 15 possibilities of how many items are solved correctly within the different versions of Block 2 (four possibilities for the blocks of very easy to medium difficulty plus three possibilities for the hardest block), 15 different possible W scores result.

Figure 1 shows the distribution of W scores for the full sample, indicating some skew to the left, and a ceiling effect at the upper end. Supplemental Table S1 provides quantiles of the performance score distributions, split up by sex and ten-year age groups. Performance scores based on the item difficulty parameter estimates reported in Fisher et al. (2013) correlate almost perfectly (r = .943, 95%-CI: .942, .945, t[21054] = 412.62, p < .001) with the scores based on item difficulties estimated in the NAKO L2 sample, indicating that sample-based variation in estimated item difficulty has little effect on performance scores.

Differences across study centres

Inspected visually, differences in the distribution of performance scores in the different study centres appeared small (see Figure 2), except for the study centre Neubrandenburg, which had somewhat lower average performance scores (d = 0.259 for the comparison of this to all other study centres). However, the intra-class correlation (ratio of systematic variance between study centres to total variance) was very small (ICC = .017), indicating little systematic variation across study centres overall.

Age-related differences

The age correlation of the number series scores was r = -.286 (95%-CI: -.298, -.274, t[21054] = 43.31, p < .001). Figure 3 shows average performance levels by year of age together with the fitted linear regression. As visual inspection indicated slight non-linearity of the age trend, we also fitted a generalized additive model (with the R package mgcv, using the maximum likelihood estimator and thin-plate regression splines). This resulted in a significantly better fit than the linear model (χ^2[effective df = 3.19] = 27881; p < .001; R^2 = .086 vs. .082) and indicated age-related differences becoming more pronounced with advancing age (see Figure 3). Controlling for sex and education did not substantially alter these findings (regression coefficient for age: -.421 vs. -.423 after controlling).

Education-related differences

Performance in the intermediate (M = 495.757; SD = 17.726) and high education (M = 503.711; SD = 16.668) groups was significantly higher (t[423.16] = 11.348; p < .001; d

= 0.618, and $t[411.54] = 19.706$; $p < .001$; $d = 1.131$, respectively) than in the lower education group ($M = 484.778$; $SD = 18.708$). The higher education group also performed above the mean of the intermediate education group ($t[16927] = 31.377$; $p <$ .001; $d = 0.464$).

Relations to other cognitive tasks

To investigate the relation of the number series task to the other cognitive tasks administered in the L1 sample, the confirmatory two-factor model reported in Kleineidam et al. (2021, submitted) was re-estimated for the L2 sample. Importantly, the number series task was included as an additional, continuous indicator variable with factor loadings on the executive functioning and the memory factors. Standardized factor loadings were large (.604) on the executive functioning factor and negligible (-.058) on the memory factor (see Figure 4). Accordingly, when correlating the number series task with factor scores based on the scoring weights from the analyses with the L1 sample (see Kleineidam et al. 2021, submitted), the correlation with the executive functions score ($r = .451$, 95%-CI: .440, .462, $t[20411] = 72.14$, $p < .001$) was larger than for the memory score ($r = .337$ (95%-CI: .324, .349, $t[20411] = 51.05$, $p < .001$). Table 3 shows correlations with individual neuropsychological tests, indicating that the correlation of the number series task was highest for the working memory task (digit span backwards).

Sex differences

There was a statistically significant sex difference favoring male participants ($M_{male} =$ 501.7; $M_{female} = 498.2$; $SD_{male} = 18.033$; $SD_{female} = 17.247$; $t[21006] = 14.736$; $p < .001$). With an effect size of $d = 0.203$ (95%-CI: .176, .230), this difference can be considered small. The variance of the male group was significantly larger than that of the female group (F[10545, 10509] = .915; $p < .001$).

Language proficiency

Performance scores were significantly associated with language proficiency (F[6, 20991] = 56.92, $p < .001$) with effect sizes indicating higher proficiency of native German speakers and ranging from very small (for bilingual speakers: $d = 0.148$, $N =$ 603; for very highly proficient non-native speakers: $d = 0.148$; $N = 519$), over small (for highly proficient non-native speakers: $d = 0.499$, $N = 535$), to large (for average non-native speakers: $d = 0.812$, $N = 236$; for low-proficiency non-native speakers: $d = 0.955$, $N = 58$).

Discussion

In summary, the presented results met expectations in that the Rasch model could be applied for scoring performance on the number series tasks and that psychometric analyses and relations with other variables provided support for its reliability and validity. With the number series task, the L2 sample of the NAKO therefore provides an appropriate indicator for reasoning, which, in turn, is known to be central to general cognitive ability (Marshalek et al. 1983). Given the strong relation of reasoning to many real-world outcomes (Schmidt and Hunter 1998; Allaire and Marsiske 1999; Deary et al. 2007; Strenze 2007), this is a precious measure to characterize cognitive functioning that is relevant to health behaviour, functional health, and predictive of diseases and mortality (Lindenberger et al. 2002; Batty et al. 2007; Calvin et al. 2011, 2017; Wraw et al. 2015, 2018; Christensen et al. 2016; Fawns-Ritchie et al. 2018; Davies et al. 2019). However, several aspects and limitations due to the very brief assessment of reasoning by just one task need to be considered when using the number series task in further analyses of data from the NAKO.

First, the distribution of performance scores indicates that the implemented version of the number series task does not cover well very high levels of ability (see truncation of ability distribution in Figure 1). Whenever the possibility of discriminating well among very high-ability participants should be relevant for health-related research questions, this needs to be considered. Given the relations of the performance scores to age and education, this is particularly relevant for subsamples of younger and more highly educated participants.

Second, due to the small number of items, the task only differentiates between 15 levels of ability, some of which are quite close to one another. In comparison to standard tests of cognitive ability (i.e., IQ tests), the measurement of continuous individual differences therefore is less fine grained.

Third, the reliability of the task is also lower than for full-length tests of reasoning. When using the number series task together with other cognitive measures from the NAKO as one of several indicators of a latent factor (e.g., executive functioning), this unreliability is accounted for by the measurement model. When using the number series task as a single observed proxy for reasoning, its reliability needs to be considered.

Fourth, direct convergent validity information (i.e., relations to other measures of reasoning) is not available within the NAKO, so that one has to rely on the number series paradigm having been well validated as an indicator task for reasoning in decades of psychometric research on intelligence. Importantly, the number series task meets the expectation of being substantially correlated with the other cognitive tasks in the NAKO. At the task level, this relation was highest for the digit span backwards task and at the factor score level, the correlation was much higher for the executive functioning than for the memory factor. These relations are in line with the psychometric literature showing that particularly working memory is strongly related to reasoning (Conway et al. 2003). Furthermore, the descriptive results reported here support the validity of the task. The strength of the age relation was in line with other cross-sectional adult lifespan studies (e.g., Schaie 1993; McArdle et al. 2002; Li et al. 2004; Salthouse 2019). Besides age-related changes, these differences may reflect cohort effects, which may favor (i.e., the "Flynn effect" of secular increases in reasoning; Flynn 1984) as well as disadvantage (i.e., decreasing arithmetic skills; Sundet et al. 2004) older cohorts. Similarly, the sex difference favoring men was within the range of heterogeneous findings reported for sex effects on mathematics performance in meta-analyses (Lindberg et al. 2010). Given this effect, however, sex should be considered as a covariate in analyses using the number series task. Education-related differences were also in line with results from meta-analyses (Strenze, 2007; Ritchie and Tucker-Drob, 2018).

Fifth, notwithstanding the present empirical support for the validity of the number series task, it has to be noted that assessing a broad ability construct, like reasoning, with only one task (and thereby only one paradigm and only for one content domain) is deficient in comparison to a comprehensive assessment (Little et al. 1999). A bias towards quantitative reasoning (and lack of coverage of verbal and figural-spatial reasoning) needs to be considered when interpreting the results. Furthermore, extant experience of participants with just this one kind of task (e.g., from quiz books), resulting in knowledge of typical rules used, may reduce the validity of the task for individual participants.

Sixth, the time pressure set by the timeout procedure potentially confounds reasoning ability to some degree with processing speed. Given the well-known strong age relation of processing speed (Verhaeghen and Salthouse 1997), older adults may be somewhat disadvantaged. The finding that the frequency of timeouts did vary only little across age groups, however, indicates that this potential source of confounding may not be large.

Seventh, performance was related to language proficiency. However, numbers of non-native German speaker of different proficiency levels were relatively small and the observed effect sizes for different levels of language proficiency among non-native speakers of German (when compared to native speakers) were smaller than those of for the executive function score (see Kleineidam et al. 2021, submitted). Nevertheless, results indicate that potential effects of this covariate need to be considered for the number series task as well.

Finally, the performance scores reported here depend on the estimated item difficulties. Once the full L2 sample (~40.000 participants) of the NAKO becomes available, these difficulties can be re-estimated, which could result in slightly different item difficulties, and accordingly slightly different performance scores. However, it can be expected that these differently estimated performance scores will correlate almost perfectly to the present ones, so that this should not lead to any substantial differences with regards to relations to outcome measures. In line with this, we could show that using item difficulty parameters from an external cohort yielded almost identical performance scores. In the re-assessment of the L2 sample, a parallel version of the number series task is included, which will allow for longitudinal analyses of changes in reasoning performance.

Conclusion

The block-adaptive number series task provides a highly time-efficient proxy of reasoning and thereby complements the L1 cognitive battery of the NAKO. This additional task allows in-depth analyses of the role of general cognitive ability for short- and long-term health outcomes in healthy, at-risk, and diseased participants across the adult lifespan and a broad range of ability levels.

NAKO Investigators (Study Centre, Name, Affiliation)

Augsburg	Annette Peters[1,2]
Regensburg	Michael Leitzmann[3], Beate Fischer[3]
Mannheim	Karin Halina Greiser[4], Kira Trares[5,6]
Freiburg	Karin B. Michels[7], Claus-Werner Franzke[7]
Essen	Karl-Heinz Jöckel[8], Sara Schramm[8]
Münster	André Karch[9], Heike Minnerup[9]
Berlin North	Tobias Pischon[10,11,12], Insa Feinkohl[10]
Berlin Center	Julia Fricke[13], Lilian Krist[13]
Berlin South	Matthias B. Schulze[14,15]
Hannover	Stefanie Castell[16], Max J. Hassenstein[16,17]
Hamburg	Nadia Obi[18], Heiko Becher[18]
Bremen	Stefan Rach[19], Kathrin Günther[19]
Kiel	Wolfgang Lieb[20]
Greifswald	Claudia Meinke-Franze[21], Wolfgang Hoffmann[22,23],

[1]Institute of Epidemiology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany (peters@helmholtz-muenchen.de)

[2]Department of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Munich, Germany (peters@helmholtz-muenchen.de)

[3]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany (michael.leitzmann@klinik.uni-regensburg.de, Beate.Fischer@klinik.uni-regensburg.de)

[4]Division of Cancer Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany (H.greiser@dkfz.de)

[5]Network Aging Research, Heidelberg University, Heidelberg, Germany (k.trares@dkfz-heidelberg.de)

[6]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany (k.trares@dkfz-heidelberg.de)

[7]Institute for Prevention and Cancer Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany (tumorepidemiologie@uniklinik-freiburg.de)

[8]Institute of Medical Informatics, Biometry and Epidemiology (IMIBE), University of Duisburg-Essen, University Hospital Essen, Essen, Germany (k-h.joeckel@uk-essen.de, sara.schramm@uk-essen.de)

[9]Institute for Epidemiology and Social Medicine, University of Münster, Münster, Germany (akarch@uni-muenster.de, h.minnerup@uni-muenster.de)

[10]Molecular Epidemiology Research Group, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany (tobias.pischon@mdc-berlin.de, insa.feinkohl@mdc-berlin.de)

[11]Charité – University Medical Center Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health (BIH), Berlin, Germany (tobias.pischon@mdc-berlin.de)

[12]MDC/BIH Biobank, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), and Berlin Institute of Health (BIH), Berlin, Germany (tobias.pischon@mdc-berlin.de)

[13]Institute for Social Medicine, Epidemiology, and Health Economics, Charité – University Medical Center Berlin, Berlin, Germany (julia.fricke@charite.de, lilian.krist@charite.de)

[14]Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany (mschulze@dife.de)

[15]Institute of Nutritional Science, University of Potsdam, Potsdam, Germany (mschulze@dife.de)

[16]Department for Epidemiology, Helmholtz Centre for Infection Research (HZI), Braunschweig, Germany (Stefanie.Castell@helmholtz-hzi.de, Max.Hassenstein@helmholtz-hzi.de)

[17]PhD Programme "Epidemiology", Braunschweig-Hannover, Germany (Max.Hassenstein@helmholtz-hzi.de)

[18]Institute for Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany (n.obi@vke.de, h.becher@uke.de)

[19]Leibniz Institute for Prevention Research and Epidemiology (BIPS), Bremen, Germany (rach@leibniz-bips.de, kguenth@leibniz-bips.de)

[20]Institute of Epidemiology, Kiel University, Kiel, Germany (wolfgang.lieb@epi.uni-kiel.de)

[21]Institute for Community Medicine, Section Clinical-Epidemiological Research, University Medicine Greifswald (UMG), Greifswald, Germany (claudia.meinke-franze@uni-greifswald.de)

[22]Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald (UMG), Greifswald, Germany (wolfgang.hoffmann@uni-greifswald.de)

[23]German Center for Neurodegenerative Diseases (DZNE) Site Rostock/Greifswald, Germany (wolfgang.hoffmann@uni-greifswald.de)

Acknowledgments

We thank all study participants and staff at the NAKO study centres, the data management and integration centre, and the NAKO head office who enabled the conduction of the study and made the collection of all data possible.

Funding

The German National Cohort (www.nako.de) is funded by the Federal Ministry of Education and Research (project numbers: 01ER1301A/B/C and 01ER1511D), the federal states, and the Helmholtz Association, with additional financial support from the participating universities and the participating institutes of the Leibniz Association and the Helmholtz Association.

This work was supported by the German Federal Ministry of Education and Research (BMBF) through ERA-NET NEURON, "SysMed - Linking synaptic dysfunction to disease mechanisms in schizophrenia - a multilevel investigation" [01EW1810 to M. Rietschel], through ERA-NET NEURON "Impact of Early life MetaBolic and psychosocial strEss on susceptibility to mental Disorders; from converging epigenetic signatures to novel targets for therapeutic intervention" [01EW1904 to M. Rietschel].

Conflict of Interest Statement

H. J. Grabe has received travel grants and speakers honoraria from Fresenius Medical Care, Neuraxpharm, Servier, and Janssen Cilag as well as research funding from Fresenius Medical Care.

Wolfgang Hoffmann received lecture fees and/or travel support from Roche, AMGEN, Pfizer, and a royalty from Janssen-Cilag.

References

Adams RJ. 2005. Reliability as a measurement design effect. Stud Educ Eval. 31:162–172.

Allaire JC, Marsiske M. 1999. Everyday cognition: age and intellectual ability correlates. Psychol Aging. 14(4):627–44.

Angrisani M, Kapteyn A, Lusardi A. 2016. The national financial capability study: Empirical findings from the American Life Panel survey. FINRA Report.

Bates D, Mächler M, Bolker B, Walker, S. 2015. Fitting linear mixed-effects models using lme4. J Stat Softw. 67(1):1–48.

Batty GD, Deary IJ, Gottfredson LS. 2007. Premorbid (early life) IQ and later mortality risk: systematic review. Ann Epidemiol. 17:278–288.

Berger K, Rietschel M, Rujescu D. 2021. The value of 'mega cohorts' for psychiatric research. Submitted to The World Journal of Biological Psychiatry.

Bertram L, Böckenhoff A, Demuth I, Düzel S, Eckardt R, Li SC, Lindenberger U, Müller L, Pawelec G, Siedler T, Wagner GG, Steinhagen-Thiessen E. 2014. Cohort profile: The Berlin Aging Study II (BASE II). Int J Epidemiol. 43:703–712.

Calvin CM, Batty GD, Der G, et al. 2017. Childhood intelligence in relation to major causes of death in 68 year follow-up: Prospective population study. BMJ 357.

Calvin CM, Deary IJ, Fenton C, Roberts BA, Der Geoff, Leckenby N, Batty GD. 2011. Intelligence in youth and all-cause mortality: systematic review and meta-analysis. Int J Epidemiol. 40:626–644.

Carroll JB. 1993. Human cognitive abilities. Cambridge: Cambridge University Press.

Christensen GT, Mortensen EL, Christensen K, Osler M. 2016. Intelligence in young adulthood and cause-specific mortality in the Danish Conscription Database – A cohort study of 728,160 men. Intelligence. 59:64-71.

Conway ARA, Kane MJ, Engle RW. 2003. Working memory capacity and its relation to general intelligence. Trends Cog Sci. 7:547–552.

Davies NM, Hill WD, Anderson EL, Sanderson E, Deary IJ, Smith GD. 2019. Multivariable two-sample mendelian randomization estimates of the effects of intelligence and education on health. Elife 8. doi:10.7554/eLife.43990

Deary IJ, Batty GD. 2007. Cognitive epidemiology. J Epidemiol Commun H. 61(5):378–384.

Deary IJ, Strand S, Smith P, Fernandes C. 2007. Intelligence and educational achievement. Intelligence. 35:13–21.

Dragano N, Reuter M, Greiser KH, et al. 2020. Socio-demographic and employment-related factors in the German National Cohort (NAKO Gesundheitsstudie). Bundesgesundheitsblatt - Gesundheitsforsch – Gesundheitsschutz. 63(3):267–278. doi:10.1007/s00103-020-03098-8

Erhardt A, Gelbrich G, Klinger-König J, Streit F, Kleineidam L, Riedel-Heller SG, NAKO Investigators, Schmidt B, Schmiedek F, Wagner M, et al. 2021. Generalized anxiety and panic symptoms in the German National Cohort (NAKO). Submitted to The World Journal of Biological Psychiatry.

Embretson SE, Reise SP. 2000. Item Response Theory (1st ed.). Psychology Press.

Engelhardt L, Goldhammer F. 2018. Number Series Study (DIPF): Technical Report. (GESIS Papers, 2018/01). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften.

Engelhardt L, Goldhammer F. 2019. Validating Test Score Interpretations Using Time Information. Front Psychol. 10:1131.

Fawns-Ritchie C, Starr JM, Deary IJ. 2018. Role of cognitive ability in the association between functional health literacy and mortality in the Lothian Birth Cohort 1936: A prospective cohort study. BMJ Open. 8(9):22502.

Fisher GG, McArdle JJ, McCammon RJ, Sonnega A, Weir DR. 2013. New Measures of Fluid Intelligence in the HRS, HRS Documentation Report DR-027. Survey Research Center, University of Michigan.

Flynn JR. 1984. The mean IQ of Americans: Massive gains from 1932 to 1978. Psychol Bull. 95:29–51.

Friedman NP, Miyake A, Corley RP, Young SE, DeFries JC, Hewitt JK. 2006. Not all executive functions are related to intelligence. Psychol Sci. 17(2):172–1789.

German National Cohort (NAKO) Consortium. 2014. The German National Cohort: aims, study design and organization. Eur J Epidemiol. 29:371–382.

Gershon RC. 2005. Computer adaptive testing. J Appl Meas. 6(1):109–127.

Goldhammer F. 2015. Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. Meas. 13:133–164.

Gustafsson J-E. 1984. A unifying model for the structure of intelligence. Intelligence. 8:179–203.

International Standard Classification of Education, ISCED 1997. 2003. In: Advances in Cross-National Comparison. doi:10.1007/978-1-4419-9186-7_10

Jäger AO, Süß H-M, Beauducel A. 1997. Der Berliner Intelligenzstruktur-Test (BIS-Test; Form 4). Test manual [The Berlin Intelligence Structure Test (BIS Test; Form 4): Test manual]. Göttingen: Hogrefe.

Kleineidam L, Stark M, Riedel-Heller S, Pabst A, Schmiedek F, Streit F, Rietschel M, Klinger-König J, Grabe HJ, Erhardt A, et al. 2021. The assessment of cognitive functions in the German National Cohort (NAKO) – Associations of demographics and

psychiatric symptoms on test performance. Submitted to The World Journal of Biological Psychiatry.
Klinger-König J, Streit F, Erhardt A, Kleineidam L, Schmiedek F, Schmidt B, Peters A, Fischer B, Leitzmann M, Kaaks R, et al. 2021. The assessment of childhood maltreatment and its associations with affective symptoms in adulthood: Results of the German National Cohort (NAKO). Submitted to The World Journal of Biological Psychiatry.
Li S-C, Lindenberger U, Hommel B, Aschersleben G, Prinz W, Baltes PB. 2004. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychol Sci. 15(3):155–163.
Linacre JM. 2012. What do infit and outfit, mean-square and standardized mean? Rasch Meas Transactions. 16(2):878.
Lindberg SM, Hyde JS, Petersen JL, Linn MC. 2010. New trends in sex and mathematics performance: A meta-analysis. Psychol Bull. 136(6):1123–1135.
Lindenberger U, Singer T, Baltes PB. 2002. Longitudinal selectivity in aging populations: Separating mortality-associated versus experimental components in the Berlin Aging Study (BASE). J Gerontol B-Psychol. 57B(6):474–482.
Little TD, Lindenberger U, Nesselroade JR. 1999. On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. Psychol Meth. 4(2):192–211.
Lövdén M, Fratiglioni L, Glymour MM, Lindenberger U, Tucker-Drob EM. 2020. Education and cognitive functioning across the life span. Psychol Sci Publ Int. 21(1):6–41.
Marshalek B, Lohman DF, Snow RE. 1983. The complexity continuum in the radex and hierarchical models of intelligence. Intelligence. 7:107–127.
McArdle JJ. 2013. Adaptive testing of the number series test using standard approaches and a new decision tree analysis approach. In: McArdle JJ, Ritschard G. Contemporary issues in exploratory data mining in the behavioral sciences. New York: Routledge; p. 312–342.
McArdle JJ, Ferrer-Caja E, Hamagami F, Woodcock R. 2002. Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. Dev Psychol. 38(1):115–142.
McArdle JJ, Woodcock RW. 2009. Adaptive testing with selected Woodcock cognitive *tests* [Powerpoint slides]. Charlottesville, VA: Longitudinal Research Institute.
Muthén L, Muthén B. 2007. Mplus User's Guide (Version 7).
Pollard MS, Baird MD. 2017. The RAND American Life Panel: Technical Description. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR1651.html.
Prescott C, Walters EE, McArdle JJ, Lapham SJ. 2018. The Project Talent Twin & Sibling Study: Understanding effects of family rearing environment on cognitive abilities. Innov Aging. 2(Suppl 1):859.
Prindle JJ. 2012. A functional use of response time data in cognitive assessment. Dissertation. University of Southern California. http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll3/id/19203
R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/
Rasch G. 1960. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rey-Mermet A, Gade M, Souza AS, von Bastian CC, Oberauer K. 2019. Is executive control related to working memory capacity and fluid intelligence? J Exp Psychol Gen. 148(8):1335–1372.
Ritchie SJ, Tucker-Drob EM. 2018. How much does education improve intelligence? A meta-analysis. Psychol Sci. 29(8):1358–1369.
Robitzsch R, Kiefer T, Wu M. 2018. TAM: Test analysis modules. https://CRAN.R-project.org/package=TAM
Salthouse T. 2019. Trajectories of normal cognitive aging. Psychol Aging. 34(1):17–24.
Schaie KW. 1993. The course of adult intellectual development. Am Psychol. 49(4):304–313.
Schmidt FL, Hunter JE. 1998. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol Bull. 124:262–274.
Streit F, Zillich L, Frank J, Kleineidam L, Wagner M, Baune BT, Klinger-König J, Grabe HJ, Pabst A, Riedel-Heller SG, et al. 2021. Lifetime and current depression in the German National Cohort (NAKO).
Strenze T. 2007. Intelligence and socioeconomic success: a meta-analytic review of longitudinal research. Intelligence. 35:401–426.
Sundet JM, Barlaug DG, Torjussen TM. 2004. The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. Intelligence. 32:349–362.
Van der Linden WJ. 2018. Handbook of item response theory. CRC Press.
Verhaeghen P, Salthouse T. 1997. Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. Psychol Bull. 122(3):231–249.
Wilhelm O, Schulze R. 2002. The relation of speeded and unspeeded reasoning with mental speed. Intelligence. 30(6):537–554.
Wood SN. 2017. Generalized additive models: An introduction with R. Chapman and Hall.
Woodcock RW, Dahl MN. 1971. A common scale for the measurement of person ability and test item difficulty (AGS Paper No. 10). Circle Pines, MN: American Guidance Service.
Woodcock RW, Johnson MB. 1977. Woodcock-Johnson Psycho-Educational Battery. Allen, TX: DLM.
Woodcock RW, McGrew KS, Mather N. 2001. Woodcock Johnson III. Rolling Meadows, IL: Riverside.
Wraw C, Deary IJ, Gale CR, Der G. 2015. Intelligence in youth and health at age 50. Intelligence. 53:23–32.
Wraw C, Der G, Gale CR, Deary IJ. 2018. Intelligence in youth and health behaviours in middle age. Intelligence. 69:71–86.
Yan D, von Davier AA, Lewis C. 2014. Computerized multistage testing: Theory and applications. CRC Press.

Table 1. Descriptive Statistics for the Interim NAKO L2 Sample with Assessment of the Number Series Task.

	Mean / N	SD / %	Sample size
Age (M/SD)	50.90	12.04	21,056
Female sex (N/%)	10,546	50.1%	21,056
Education groups (N/%)			21,056
low	390	1.9%	
middle	8,138	38.7%	
high	10,831	51.4%	
unclassified	1,697	8.1%	
Language proficiency (N/%)			21,056
Native speaker	19,035	90.4%	
Bilingual	603	2.9%	
Non-native speaker: very high	519	2.5%	
Non-native speaker: high	535	2.5%	
Non-native speaker: average	236	1.1%	
Non-native speaker: low	58	0.3%	
Non-native speaker: very low	12	0.1%	
Non-native speaker: unknown proficiency level	58	0.3%	

Note. Education groups were defined following the International Standard Classification of Education 97 (ISCED97) as described for the NAKO by Dragano et al. (2020), with low = ISCED97 level 1/2, middle = ISCED97 level 3/4, high = ISCED97 level 5/6. Unclassified participants resulted from the classification of job education not having been finalized at the time of data analyses.

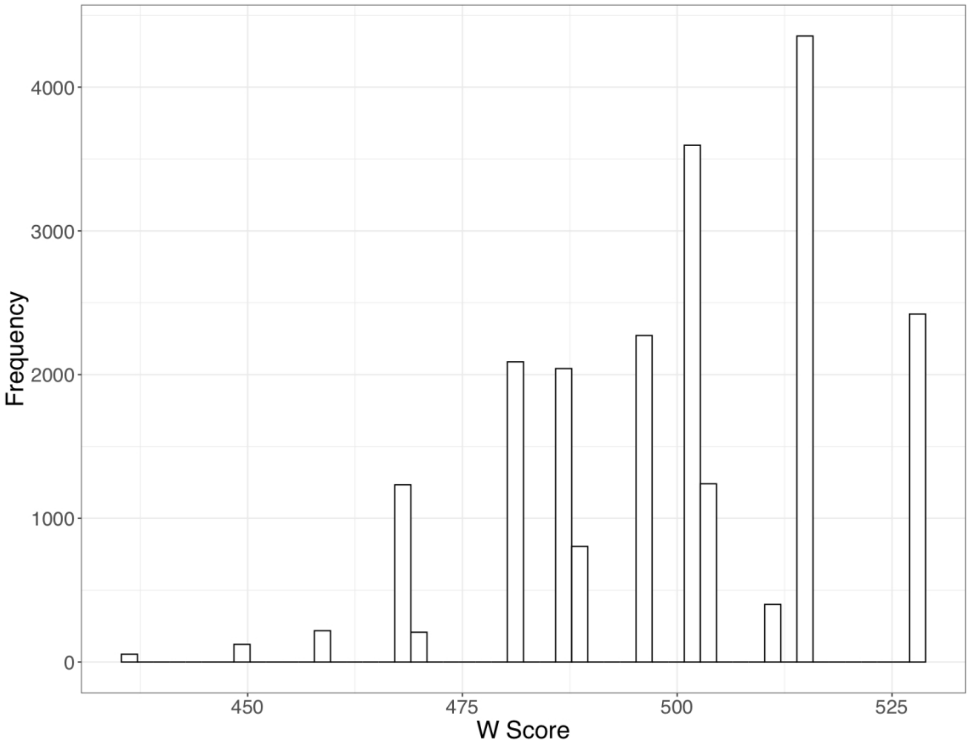
Table 2. Descriptive Information on Correctly Solved Items, Item Timeouts, Item Difficulties, and Item Fit.

			All 15 items			**Excluding last item**		
Item	N(%) correct	N(%) timeout	Difficulty	Infit	Outfit	Difficulty	Infit	Outfit
Block 1 items (worked on by all 21,056 participants)								
G1	20,284 (96%)	348 (2%)	-4.586	0.982	0.987	-4.593	0.983	0.990
H1	14,641 (70%)	2,505 (12%)	-1.264	0.982	0.983	-1.266	0.982	0.984
I1	12,266 (58%)	3,906 (19%)	-0.508	0.923	0.866	-0.509	0.923	0.866
Block 2: very easy items (worked on by 566 participants)								
A1	472 (83%)	35 (6%)	-5.274	0.924	0.846	-5.1287	0.924	0.846
B1	434 (77%)	31 (5%)	-4.752	0.891	0.803	-4.765	0.891	0.803
C1	238 (42%)	52 (9%)	-2.796	0.977	0.963	-2.808	0.977	0.963
Block 2: easy items (worked on by 4,162 participants)								
D1	4,097 (98%)	23 (1%)	-6.597	1.000	0.955	-6.606	1.000	0.956
E1	2,528 (61%)	304 (7%)	-2.359	1.028	1.046	-2.367	1.028	1.046
F1	1,198 (29%)	871 (21%)	-0.673	1.034	1.075	-0.679	1.035	1.075
Block 2: medium items (worked on by 5,955 participants)								
J1	2,554 (43%)	1,341 (23%)	-0.090	1.035	1.050	-0.092	1.035	1.050
K1	1,956 (33%)	584 (10%)	0.450	1.024	1.039	0.448	1.025	1.040
L1	1,445 (24%)	502 (8%)	0.975	1.041	1.081	0.973	1.041	1.082
Block 2: hard items (worked on by 10,373 participants)								
M1	5,217 (50%)	3,736 (36%)	1.120	1.036	1.056	1.124	1.035	1.053
N1	3,981 (38%)	3,342 (32%)	1.773	1.037	1.057	1.778	1.037	1.058
O1	55 (<1%)	8,325 (80%)	7.362	1.013	1.514	---	---	---

Table 3. Correlations of the Number Series Task with Other Cognitive Tasks in the L2 Sample.

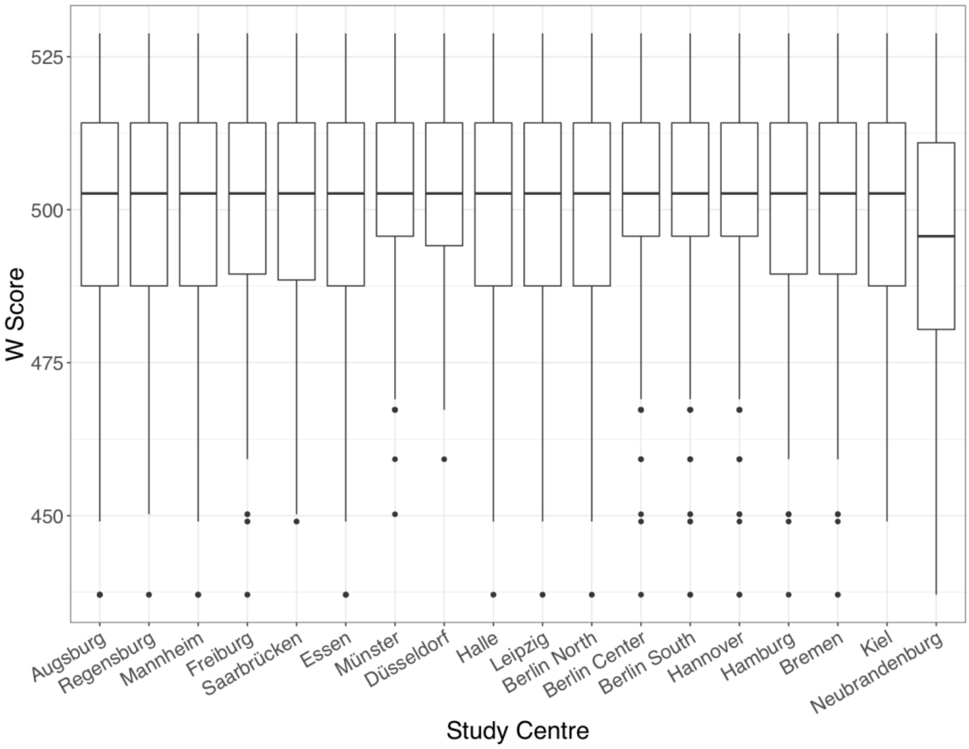
	r	**95%-CI**	***t*-Test**
Executive Function			
Verbal Fluency	.277	.265, .290	*t*[20371] = 41.21*
Stroop Task 2	-.251	-.264, -.238	*t*[20071] = 36.81*
Stroop Difference	-.311	-.323, -.298	*t*[20060] = 46.30*
Digit Span Backwards	.345	.333, .358	*t*[20315] = 52.47*
Memory			
Word List recall			
Immediate	.285	.272, .298	*t*[20198] = 42.26*
Repeated	.288	.275, .300	*t*[20179] = 42.70*
Delayed	.288	.275, .300	*t*[20114] = 42.62*

Note. * = *p* <.001



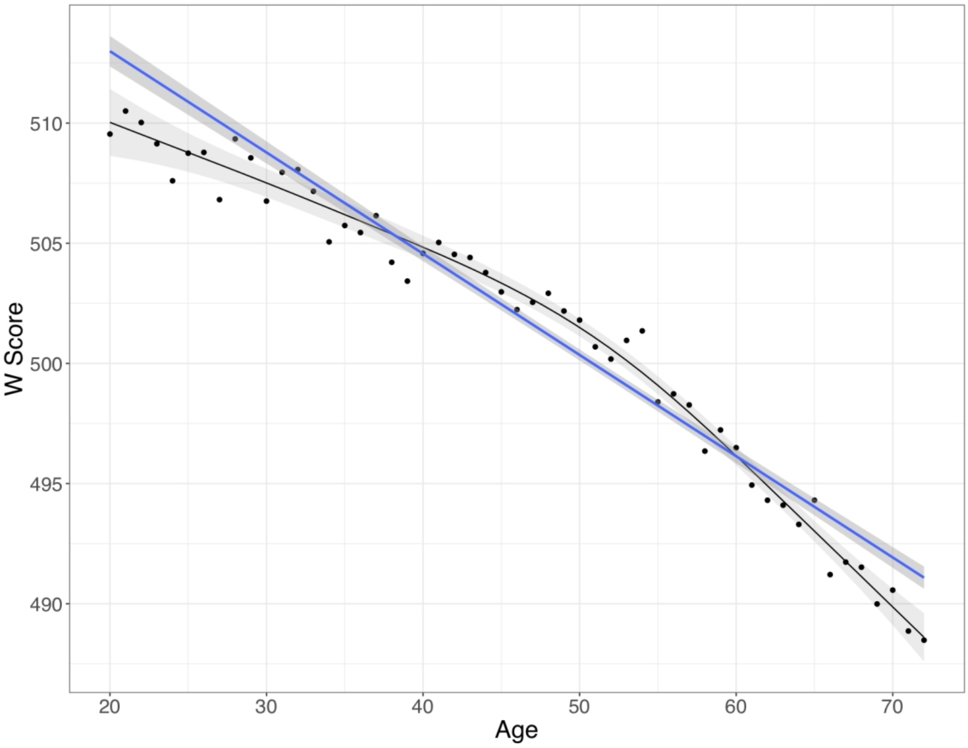
Distribution of Ability Estimates (W Scores).
Note. N = 21,056. Ability estimates result from Rasch scoring of the 15 possible response patterns. Scores of 449 and 450 are collapsed into one bar.

710x548mm (59 x 59 DPI)



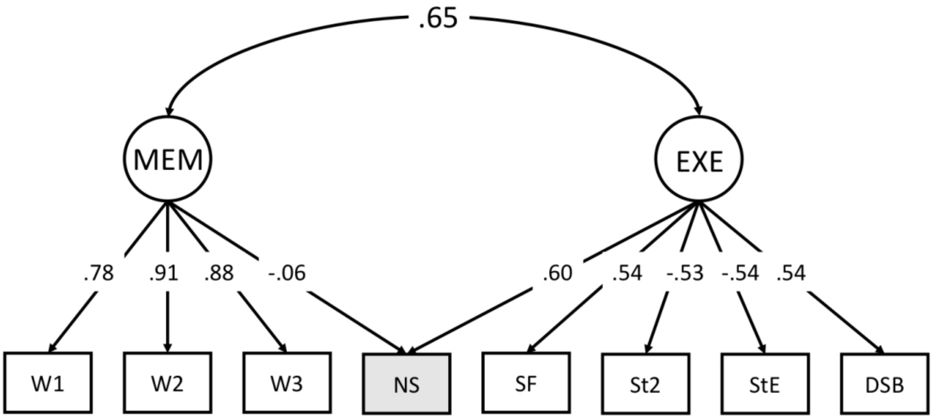
Distribution of Performance Scores Across Study Centres.
Note. Boxplot with vertical line showing the median, box ends indicating the first and third quartiles, whisker ends extending 1.5 times the interquartile range from the box ends, and points denoting outliers beyond this range.

710x548mm (59 x 59 DPI)



Age-Related Differences in Performance on the Number Series Task.
Note. Dots: average values per year. Blue line: linear regression. Black line: fitted function from generalized additive model. Shaded areas: 95% confidence bands.

710x548mm (59 x 59 DPI)



Confirmatory Factor Analysis Model of the Number Series Task and the Other Cognitive Tasks Included in the NAKO.
Note. Squares denote observed variables. Circles denote latent variables. Factor loadings and correlation are standardized estimates. W1 = Immediate word list recall trial 1; W2 = Immediate word list recall trial 2; W3 = Delayed word list recall trial 3; NS = Number series task performance score; SF = Semantic fluency (animals); St2 = Stroop task 2; StE = Stroop effect (task 2 – task 1); DSB = Digit span backwards; MEM = Memory factor; EXE = Executive function factor; estimated using maximum likelihood estimation in Mplus; N = 21,056.

645x484mm (59 x 59 DPI)

Table S1: Raw Percentiles of the Number Series Task Performance Score by Sex and Age Group

Sex		**Percentile**						
Age group	**N**	**2.5th**	**10th**	**25th**	**50th**	**75th**	**90th**	**97.5th**
Men								
20-29	693	480.44	487.55	502.66	514.19	528.83	528.83	528.83
30-39	1,084	467.31	487.55	495.67	510.94	514.19	528.83	528.83
40-49	2,487	467.31	480.44	495.67	502.94	514.19	528.83	528.83
50-59	3,135	467.31	480.44	487.55	502.94	514.19	528.83	528.83
60-72	3,088	459.23	467.31	480.44	495.67	510.94	514.19	528.83
Women								
20-29	783	480.44	487.55	502.66	514.19	528.83	528.83	528.83
30-39	1,071	467.31	487.55	495.67	502.94	514.19	528.83	528.83
40-49	2,725	467.31	480.44	489.49	502.94	514.19	528.83	528.83
50-59	3,215	467.31	480.44	487.55	502.94	514.19	514.19	528.83
60-72	2,729	459.23	467.31	480.44	489.49	502.94	514.19	528.83