HHS Public Access

Author manuscript

Psychol Bull. Author manuscript; available in PMC 2024 December 09.

Published in final edited form as:

Psychol Bull. 2024 April ; 150(4): 399–439. doi:10.1037/bul0000425.

The Stability of Cognitive Abilities: A Meta-Analytic Review of Longitudinal Studies

Moritz Breit[a], **Vsevolod Scherrer**[a], **Elliot M. Tucker-Drob**[b], **Franzis Preckel**[a]

[a]University of Trier

[b]Department of Psychology, University of Texas at Austin

Abstract

Cognitive abilities, including general intelligence and domain-specific abilities such as fluid reasoning, comprehension knowledge, working memory capacity, and processing speed, are regarded as some of the most stable psychological traits, yet there exist no large-scale systematic efforts to document the specific patterns by which their rank order stability changes over age and time interval, or how their stability differs across abilities, tests, and populations. Determining the conditions under which cognitive abilities exhibit high or low degrees of stability is critical not just to theory development, but to applied contexts in which cognitive assessments guide decisions regarding treatment and intervention decisions with lasting consequences for individuals. In order to supplement this important area of research, we present a meta-analysis of longitudinal studies investigating the stability of cognitive abilities. The meta-analysis relied on data from 205 longitudinal studies that involved a total of 87,408 participants, resulting in 1,288 test-retest correlation coefficients among manifest variables. For an age of 20 years and a test-retest interval of 5 years, we found a mean rank-order stability of ρ =.76. The effect of mean sample age on stability was best described by a negative exponential function, with low stability in preschool children, rapid increases in stability in childhood, and consistently high stability from late adolescence to late adulthood. This same functional form continued to best describe age trends in stability after adjusting for test reliability. Stability declined with increasing test-retest interval. This decrease flattened out from an interval of approximately 5 years onward. According to the age and interval moderation models, minimum stability sufficient for individual level diagnostic decisions (r_{tt} = .80) can only be expected over the age of seven and for short time intervals in children. In adults, stability levels meeting this criterion are obtained for over five years.

Keywords

cognitive ability; intelligence; stability; rank-order; life span

Individual differences in cognitive ability are considered to be highly stable over time (e.g., Hunt, 2010; Mackintosh, 1998; Neisser et al., 1996). Researchers describe cognitive ability as “the most stable psychological trait” (Plomin & Stumm, 2018, p. 149) and numerous studies document a high stability of cognitive ability (e.g., Deary et al., 2000;

Correspondence concerning this article should be addressed to Moritz Breit, Department of Psychology, University of Trier, Giftedness Research and Education, Universitaetsring 15, 54296 Trier, Germany. breitm@uni-trier.de.

Larsen et al., 2008; Schalke et al., 2013). This stability is crucial because cognitive ability tests are frequently used to inform treatment and intervention decisions that have long-term consequences for individuals. For example, cognitive assessments are used for long-term decisions in educational settings, for instance with respect to school tracking, provision of special education services for struggling learners, and provision of gifted and talented enrichment opportunities (Gottfredson & Saklofske, 2009; Nettelbeck & Wilson, 2005). Similarly, cognitive assessments frequently inform admissions and personnel selection decisions within work settings (Ones et al., 2017; Salgado et al., 2002), and commonly inform decisions with respect to the most suitable among alternative therapeutic interventions within clinical settings (Taylor et al., 2008). All of these applications build on the assumption that cognitive ability will remain stable over the period that the decision is effective. Otherwise, the fit between the individual and the test-based decision (e.g., selected environment or treatment) may deteriorate (Cronbach & Snow, 1977).

High stability of cognitive ability may not apply equally to all populations, circumstances, and cognitive abilities. It is widely recognized that the stability of cognitive ability varies with age (e.g., McArdle et al., 2002; McCall et al., 1977) and test-retest interval (e.g., Watkins & Smith, 2013), but the exact influence of these moderators over the life span is less clear. Further moderators of stability such as the mean ability level of the sample (Breit, Scherrer, & Preckel, 2021) and the utilized test instrument (Villado et al., 2016) have rarely been examined systematically. Quantifying the stability of cognitive ability and the influence of moderating variables is best achieved by meta-analysis (Deeks et al., 2008). With two exceptions, one confined to 15 longitudinal twin studies and to formal comparisons of general intelligence, broad fluid abilities, and broad crystallized abilities (Tucker-Drob & Briley, 2014), and one confined to general intelligence scores in Wechsler and Standford-Binet tests before 1990 (Schuerger & Witt, 1989), there has been no comprehensive meta-analysis investigating the stability of cognitive ability. The present study aims to close this gap. Specifically, we integrate the results of longitudinal studies investigating the rank-order stability of cognitive ability and examine the moderating effects of age, test-retest interval, measured cognitive ability, ability level, test instrument, and geographic location. Knowledge of the stability of cognitive ability and the factors that influence it is of great benefit for both application and basic research. Understanding variations in stability, that is to what extent and when there are individual differences in cognitive change, advances test-based decision making and our understanding of cognitive development, healthy cognitive aging (Deary, 2014), and the nature of cognitive ability in general (Rinaldi & Karmiloff-Smith, 2017).

Cognitive Abilities in the Psychometric Framework

Cognitive abilities can be defined as any ability that substantially involves mental functions needed for the correct or appropriate processing of mental information (Carroll, 2009). Their measures are frequently based on psychometric models of cognitive abilities (Mackintosh, 2011). The first psychometric model was devised by Spearman (1904), who ascribed the generally positive correlations between different measures of cognitive ability to one common factor that is associated with every measure, albeit to varying degrees. This factor is called “general intelligence” or “general cognitive ability” (*g*) and comprises

"the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredson, 1997, p. 13). Over the course of the 20th century, a general factor ultimately proved insufficient to explain all covariance between different measures (e.g., Thurstone, 1938). Therefore, current psychometric models of cognitive ability are hierarchical models that include both a set of lower-order factors representing more specific abilities and *g* at the apex, reflecting the correlations between those factors (Jensen, 1998; Mackintosh, 2011). The most recent structural model, which integrates prominent precursor models, is the Cattell-Horn-Carroll model of cognitive ability (CHC-model; McGrew, 1997; Schneider & McGrew, 2018). The CHC-model differentiates narrow, intermediate, and broad abilities and puts *g* at the apex of the hierarchy. Of the 17 broad abilities, six are classified as "tentative" because they require further research to be included (i.e., emotional intelligence, kinesthetic abilities, olfactory abilities, tactile abilities, psychomotor abilities, psychomotor speed). The remaining, more established broad abilities are described in Table 1; the resulting top two levels of the CHC model are depicted in Figure 1.

Many modern tests of cognitive ability are constructed based on the CHC model or locate their scales in the model. Usually, the tests assess *g* and broad abilities from the CHC, but not all broad abilities are equally represented. A recent analysis showed that virtually all subtests of major cognitive ability tests (i.e., Differential Abilities Scale, Second Edition; Kaufman Assessment Battery for Children, Second Edition; Wechsler Intelligence Scale for Children, Third, Fourth, and Fifth Editions; Woodcock–Johnson III Tests of Cognitive Abilities) can be factor analytically assigned to *g* and one of six broad abilities. These are: Comprehension Knowledge, Fluid Reasoning, Learning Efficiency, Processing Speed, Visual Processing, and Working Memory Capacity (Caemmerer et al., 2020). Some of the ability tests were not explicitly developed based on the CHC model, which demonstrates the usefulness of the model to classify the subtests of different instruments.

The Stability of Cognitive Abilities

When investigating the stability of cognitive abilities, four different types of stability can be distinguished (Breit, Scherrer, & Preckel, 2021; Fryer & Elliot, 2007): mean-level change, ipsative continuity, individual-level change, and rank-order stability. First, *mean-level change* describes the average change in a score over time. Cognitive ability generally increases during childhood and adolescence (e.g., McArdle et al., 2002; Schroeders et al., 2015), with abilities that require effortful processing (e.g., Fluid Reasoning, Learning Efficiency, Processing Speed, Visual Processing, Working Memory Capacity) increasing faster than knowledge-based abilities (e.g., Comprehension Knowledge) (Baltes et al., 1999). After young adulthood, effortful-processing-based abilities begin to decrease whereas Comprehension Knowledge increases further into late adulthood and only declines toward the end of life (Horn & Cattell, 1967; Tucker-Drob et al., 2022; Wang & Kaufman, 1993). These changes may not be independent, but instead linked through developmental couplings between different cognitive abilities (e.g., Li et al., 2004). Interindividual variance in early cognitive development or cognitive aging (Lövdén et al., 2005) is not captured by this type of stability. Second, *ipsative continuity* represents the stability of an individual's configuration of different scores – their ability profile – over time. Research on age

differentiation effects has shown that the strength of cognitive profiles remains relatively uniform across childhood and adolescence and decreases towards the end of life (Breit, Brunner, et al., 2022). There is some tentative evidence that cognitive ability profiles have little temporal stability in low-ability samples (McDermott et al., 1992; Watkins & Smith, 2013) but moderate stability in high-ability samples (Breit, Scherrer, & Preckel, 2021). Third, *individual-level change* represents the change in a test score of an individual. Individual-level change is directly relevant to many practical applications of cognitive ability testing that involve the placement of an individual in an appropriate educational, vocational, or social environment. If an individual exhibits substantial change in their abilities in any direction, the selected environment may become inappropriate (Cronbach & Snow, 1977). Nevertheless, individual-level change is rarely investigated directly in cognitive ability research. Instead, research mostly focuses on the last and most relevant stability to the present work, *rank-order stability*, also known as differential continuity. It represents the stability of individual differences in cognitive ability and is usually assessed with test-retest Pearson product–moment correlations (e.g., Breit, Scherrer, & Preckel, 2022). With the investigation of rank-order correlations and potential influencing factors, one investigates individual differences in cognitive development over the life span (Deary, 2014). Rank-order stability analyses can also be regarded as an indirect analysis of the frequency of substantial individual-level change. If a sizable percentage of individuals in a sample exhibit large changes in cognitive ability over time, this will affect the overall rank-order. Analogously, a high rank-order stability indicates that the proportion of individuals with major changes relative to other individuals is low.

Genetic influences have been shown to strongly drive the rank-order stability of cognitive ability, with some additional contribution of shared environmental influences that are responsible for both stability and change in cognitive abilities (Bartels et al., 2002). Deary et al. (2012) reported the genetic correlation between cognitive ability in childhood and old age to be as high as .62. A meta-analysis of 15 longitudinal twin and adoption studies supported the notion that both genetic and, to a lesser extent, environmental factors contribute to the rank-order stability of cognitive ability (Tucker-Drob & Briley, 2014). According to their results, genetic influences on phenotypic stability increase during early cognitive development and account for up to 75% of the stability in adulthood, whereas shared environmental influences decrease and non-shared environmental influences slightly increase over the life span. These trajectories mostly conform to theories proposing gene×environment interactions in the development of cognitive ability. Such theories propose that a) based on genetic differences, individuals respond differently to the same environment, and in turn, there may be experience-activated epigenetic processes that are robust over time, and b) individuals systematically experience different environments as a consequence of their genotypes, caused by passive, evocative, and active selection processes (for an in-depth discussion, see Tucker-Drob & Briley, 2014). The findings therefore imply that moderators of rank-order stability can operate both genetically and environmentally.

Early work by Bayley (1949) showed that rank-order stability of cognitive ability increases sharply during early and middle childhood. In line with this finding, Tucker-Drob and Briley (2014) found an increasing stability (test-retest correlation) over the course of childhood in their meta-analysis of twin and adoption studies, from just above .2 in infancy to above .7

by the age of 12 for a test-retest interval of six years. Beyond early childhood, many studies have reported substantial stabilities for very long test-retest intervals, including impressive long-term longitudinal studies such as the Scottish Mental Surveys, the Seattle Longitudinal Study, and the Vietnam Era Twin Study of Aging. Findings from the Scottish Mental Surveys, in which children's IQ was first measured at age 11, showed test-retest correlations of .67 for a time span of 59 years and .54 for a time span of 79 years (Deary, 2014; Deary et al., 2013). Based on a different long-term study, Schwartzman et al. (1987) reported a test-retest correlation of .78 between the ages of 25 and 65. Rönnlund et al. (2015) found that cognitive ability level at age 18 accounted for 90% of ability variance at age 50 (r = .95) and 74% at age 65 (r = .86). Similarly high test-retest correlations for considerable test-retest intervals in adults were found, for example, by Hertzog and Schaie (1988) and Larsen et al. (2008). Schurger and Witt (1989) conducted a meta-analysis on the rank-order stability of Wechsler and Stanford-Binet IQ test scores based on 34 studies and reported a mean test-retest correlation of .82. Even in very old adults, studies have reported evidence for substantial stability. For example, based on data from the Virginia Cognitive Aging Project, Salthouse (2012a) reported three-year stability estimates in adults aged 80 to 97 ranging from .63 to .80 for different cognitive abilities.

The available empirical findings therefore point to a substantial long-term rank-order stability of cognitive abilities, but the precise estimates vary. A number of potential moderating variables like age, time span between measurements, or assessed ability may explain this heterogeneity. Therefore, a comprehensive meta-analysis is needed to reliably quantify the stability of cognitive abilities and the influence of potential moderator variables.

Potential Moderators of the Rank-Order Stability of Cognitive Abilities

Age

Systematic changes in the rank-order stability of cognitive ability across the life span have been demonstrated many times. The predominant view in the literature is that cognitive tests taken in infancy have little predictive value for later cognitive ability, but that the stability rapidly increases during early cognitive development until almost no reordering occurs anymore by early adulthood (Tucker-Drob & Briley, 2014). The rapid increase in stability during childhood was already described by Bayley (1949), who found hardly any stability in infancy but stabilities over .90 at age 13. Schuerger and Witt (1989) investigated age as a moderator in their meta-analysis of rank-order stability of *full scale IQ test* scores, finding a positive and statistically significant logarithmic trend (solid line in Figure 2). This finding implies a steep increase in stability between ages 3 to 15 that slowly diminishes thereafter but is maintained throughout adulthood. Tucker-Drob and Briley (2014) tested the fit of different functional forms for the age trend in rank-order stability of a variety of cognitive abilities, finding the best fit for an exponential function that again implied a rapid increase of stability during childhood closely approaching an asymptote after early adulthood with no further increase (dashed line Figure 2). Notably, both analyses did not include samples of very old participants, with maximum sample ages of 65 and 73 years, respectively. Yet, it is possible that some reordering occurs in old age, as different factors may contribute to cognitive decline than those that contribute to cognitive changes in earlier

development (e.g., Li et al., 2004). Moreover, in their two-component theories of intellectual development, Lindenberger and Baltes (Baltes et al., 1999; Lindenberger & Baltes, 1994) predicted that cognitive aging should be associated with a partial reordering of individual differences in cognitive abilities. They hypothesized that increased reordering will occur whenever there is greater mean-level change in cognitive development (i.e., early childhood and late adulthood) because greater mean-level change is expected to be associated with a greater increases in novel variation per unit of time. This reordering would imply a decrease in stability in old age, meaning that the stability trajectory may in fact be best described by an inverse U shaped function (dotted line in Figure 2) with increases in childhood and decreases in old age. In line with this idea, some studies reported only modest stabilities in very old samples (e.g., Ghisletta & Lindenberger, 2003; Gregory et al., 2009). Conversely, there is also some evidence for high stability of cognitive ability in very old adults (e.g., Hopp et al., 1997), underscoring the need for meta-analytic investigation across the life span.

Test-Retest Interval and Its Interaction with Age

A general effect of the duration of the test-retest interval on the rank-order stability estimate is well established. Interestingly, the effects of the interval duration do not appear to be linear. Instead, at first there is a steep decrease in stability with increasing interval, after which even large increases in the test-retest interval have little additional effect on stability (Schuerger & Witt, 1989; Tucker-Drob & Briley, 2014). Whereas Schuerger and Witt (1989) investigated age and interval effects separately, Tucker-Drob and Briley (2014) stressed the importance of considering the effect of the interval in relation to the age of the sample. This dependence was already observed by Bayley (1949), who found that stability deteriorates quickly in infants and young children with increasing test-retest interval but is much more persistent in adolescents. Tucker-Drob and Briley (2014) focused on childhood and adolescence in their interaction analyses because of their sparse data in adulthood. They found a more pronounced time decay effect for younger ages, albeit not statistically significant. Comparable analyses for older samples or larger age ranges are lacking.

Cognitive Ability Captured

The stability of cognitive ability may depend not only on the age of the participants and the test-retest interval but also on the captured cognitive ability. First, *g* might be more stable than more specific abilities, i.e., the broad abilities of the CHC model. Higher level abilities such as *g* that are measured by a multitude of different tasks are less likely to be affected by specific learning experiences or activities, whereas lower level abilities may be more dependent on acquired knowledge and skills and therefore more easily changed by experience, leading to lower stability (Reeve & Bonaccio, 2011; Tucker-Drob & Briley, 2014). The lower stability of broad ability scores as compared to *g* has been shown in individual studies (McDermott et al., 1992; Watkins & Smith, 2013) and meta-analytically for children (Tucker-Drob & Briley, 2014). Moreover, Tucker-Drob and Briley (2014) proposed that various broad abilities are dependent on genetic and environmental factors to different degrees, potentially also leading to differences in stability between them. When grouping abilities as either “fluid”, characterized by effortful processing, or “crystallized”, characterized by knowledge dependency, they found no significant differences in their stability. However, this analysis was based on relatively sparse data (i.e., 33 and 21

effect sizes for fluid and crystallized abilities, respectively). Following the CHC model, it is necessary to investigate all broad abilities individually, as they are assumed to rely on somewhat distinct neurological functions and serve different purposes in human survival and reproduction (Schneider & McGrew, 2018). Broad abilities may differ in their developmental trajectories and stabilities beyond an effortful-processing vs. knowledge divide. Ideally, not only the mean stabilities but also the stability trajectories across the lifespan are investigated for each broad ability. This kind of meta-analytic investigation naturally requires a very broad database.

General Cognitive Ability (g) Level

The ability differentiation hypotheses states that the structure of cognitive ability changes across the continuum of the general cognitive ability *g*. Whereas the overall factor structure appears to be constant, the relative importance of *g* for individual test performances decreases with increasing *g*-level (Breit, Brunner, et al., 2022; Tucker-Drob, 2009). This ability differentiation effect implies a decrease in systematic variance in measures of general intelligence with increasing *g*-level, leading to a gradual reduction of test-score reliability (Breit, Brunner, et al., 2022). Because test-retest correlations cannot exceed the reliability of the utilized test, this may cause lower estimates of rank-order stabilities of *g* in higher ability samples. Up to now, there are no systematic investigations of differences in the stability of *g* depending on the ability level.

Test Instrument

Cognitive ability tests vary considerably in content. One important distinction is that between unidimensional and multidimensional tests. Unidimensional tests aim to measure *g*, often by using a single type of task like figural reasoning tasks (e.g., Raven, 1938). Multidimensional test batteries aim to measure a variety of cognitive abilities with specific subtests; usually they also provide a *g*-score as average score (e.g., Wechsler, 2003). Multidimensional test batteries show heterogeneity in terms of test size (i.e., number of subtests and tasks), task presentation (i.e., oral by the administrator or in writing), test setting (i.e., individual or group test), test reliability, as well as which specific abilities are measured, in which order, and how they are weighted. All these factors may affect the test-retest correlation of test scores, but systematic investigations are lacking. In one study, greater stability was found for the full test scores of the multidimensional Wonderlic Personnel Test than for the unidimensional Raven's Advanced Progressive Matrices (Villado et al., 2016).

Sometimes, test administrators use a different test at a second administration either to avoid retest effects (i.e., training or memory effects) or because the test used at first testing cannot be used for the age range of the tested sample at retest. In the latter case, it is often possible to use a test from the same test family that is constructed similarly but was adjusted to the target age (e.g., different Wechsler tests). Stability estimates may differ systematically between the same test, a test from the same test family, or a different test being used at retest because the test-retest correlation is limited by the size of the concurrent correlation of both tests.

Finally, it is important whether a whole test battery is used or only parts of it. In many scientific investigations, subtests from one or more tests are selected to measure the constructs of interest economically. Test batteries may not only be shortened in testing practice but also augmented, although this happens less often. The CHC cross-battery assessment approach entails "augment or supplement any major ability test to ensure measurement of a wider range of broad and narrow cognitive abilities in a manner that is consistent with contemporary theory and research and that is predicated upon sound psychometric principles" (Flanagan et al., 2012, p. 459). The stability estimates may be affected by such a shortening or extension of existing test instruments because test reliability varies with test length.

Geographic Location

Geographic location is often accompanied by differences in culture, language, educational systems, economic opportunities, nutrition, health care, social mobility, or physical living environment. As environmental factors contribute to stability, these differences and their respective combinations can affect the stability of individual differences in cognitive ability (Tucker-Drob & Briley, 2014). To our knowledge, there are no systematic comparisons of differences in rank-order stabilities between different countries, regions, or cultures.

Test Reliability

Test-retest correlations (i.e., rank-order stabilities) rely on the reliability of the tests to measure the construct at both times of measurement. They cannot exceed the square root of the product of both reliability coefficients. Alternatively put, stability coefficients are reduced by measurement error (i.e., unreliability of the tests). This fact has important implications for the present meta-analysis, as, without adjustment for test reliability, stability will be underestimated. For example, Hopkins and Bibelheimer (1971) reported substantially larger disattenuated test-retest correlations (.55 to .94) than uncorrected correlations (.33 to .81). Moreover, moderator analyses may be biased by test reliability if there is a systematic relationship between the moderator variable and test reliability. For example, cognitive ability tests and the testing situation must be modified in various ways to be appropriate for young children. If these modifications result in systematic differences in test reliability between different age groups, and reliability is not adjusted for, these age trends in reliability may be confused for age trends in stability. Similarly, reliability may also differ between tests of different abilities. Systematic differences in captured cognitive abilities, length, speededness, and response format in tests of different abilities may affect test reliability (Hong & Cheng, 2019; Symonds, 1928). For example, Processing Speed tasks are constructed very differently from tests of Comprehension Knowledge. Like individual subtests for different abilities, entire test batteries vary in content, length, speededness, and response format, leading to differences in reliability. That is, the scores from different test instruments or batteries (e.g., the Wechsler tests vs. the Woodcock Johnson tests) may appear to differ in stability because of differences in their reliabilities. Approaches for disattenuating stability estimates from measurement error are therefore important both for estimating stability of true scores and for obtaining unbiased estimates in moderator analyses. We are not aware of any meta-analytic investigation of the stability of cognitive abilities that adjusted for test reliability.

The Current Meta-Analysis

Previous meta-analyses of the rank-order stability of cognitive ability provided valuable insights into the stability of *g* in Wechsler and Stanford-Binet tests (Schuerger & Witt, 1989) and the contributions of genetic and environmental influences herein (Tucker-Drob & Briley, 2014). However, due to their specific research questions and associated narrow selection of studies, they were limited in the age range (maximum baseline age of 65 and 73 years, respectively), captured cognitive ability (*g*, Gf, Gc), and test instruments (Wechsler and Stanford-Binet tests). Relatedly, some moderators were investigated with relatively sparse data (e.g., cognitive ability captured) or not at all (e.g., mean *g*-level of the sample, geographic location). Further, in both meta-analyses there was no adjustment for test reliability.

The aim of the present study therefore is to provide a comprehensive analysis of the available evidence by meta-analyzing the findings of longitudinal studies that investigate the rank-order correlation (i.e., rank-order stability) of cognitive ability and that cover the entire life span from 1 to 90 years. All hypotheses and analyses were preregistered except H10 and the associated analyses (https://osf.io/2pn3x). We investigated the following hypotheses:

H1: (a) Rank-order stability decreases with increasing *test-retest interval*. (b) The decrease per additional year diminishes with increasing intervals. In addition, we exploratory tested whether this effect can be better described with a linear, a quadratic, or an exponential function.

H2: Rank-order stability varies with *age*. (a) Over all life stages, *age* is nonlinearly related to stability. We expected that rank-order stability increases with age in preschool children and in school-aged children and adolescents, remains constant in adults, and decreases in elderly individuals. We tested whether this effect can be better described with a linear function, a quadratic function, a connected-linear-spline, or an exponential function. (b) As an open research question, we investigated interaction effects between *test-retest interval* and *age*.

H3: The average rank-order stability of cognitive ability is higher than that of other (noncognitive) personality traits across the lifespan, controlling for the test-retest interval.

H4: The rank-order stability varies with *captured cognitive ability*. The stability is higher for measures of *g* than for measures of CHC broad abilities.

H5: The rank-order stability of *g* decreases with increasing *g-level*. The relation of rank-order stabilities of CHC broad abilities and *g*-level was investigated as an open research question.

H6: As an open research question, we investigated whether rank-order stability of cognitive ability varies with the *test instrument* used.

H7: (a) The rank-order stability of cognitive ability is higher when the same test was used for both measurements compared to varying measurement instruments from the same test

families or varying tests. (b) The rank-order stability of cognitive ability is higher when measurement instruments from the same test families were used compared to varying tests.

H8: The rank-order stability of cognitive ability is higher when the complete test battery is used compared to a selection of subtests.

H9: As an open research question, we investigated whether rank-order stability of cognitive ability varies between countries.

H10: As an open research question, we investigated whether the results of the analyses for H2, H3, H4, and H6 are replicated in a sample of effect sizes corrected for test reliability.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Because no human or animal participants were involved in the study, ethics committee approval was not sought. We adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines for systematic reviews (Page et al., 2021). All data and research materials are available at https://osf.io/ajufs/. This meta-analysis was preregistered (https://osf.io/2pn3x). Any deviations from the preregistration are reported in Supplemental Materials.

Identification and Screening Process

The literature search for longitudinal studies reporting rank-order stability of cognitive abilities over time was conducted following the guidelines by Johnson and Hennessy (2019) and in consultation with a research librarian at the University of Trier. Figure 3 depicts a PRISMA chart (Page et al., 2021) of the preliminary identification and screening process.

Overall, we identified 4,077 different records based on five search strategies: 1. search for peer-reviewed articles in APA PsycInfo; 2. search for books in APA PsycInfo; 3. search for dissertations in APA PsycInfo; 4. search for peer-reviewed articles in ERIC; 5. reference lists of relevant reviews and meta-analyses identified in steps 1 – 4. In Search Strategies 1 – 4, we used the following key words as the search string: *(Intelligence* or cognitive abilit* or mental abilit* or IQ or g factor or mental test* or gma or fluid reasoning or comprehension knowledge or gf or gc).ti,ab. AND (longitudinal or stabilit* or retest* or repeated measur* or cross-lagged or autocorrelation* or long term or "change over time").ti,ab. AND (not autism not dement* not artificial not emotional intelligence* not social intelligence* not schizophrenia not infant not disorder not lesion* not animal* not disabilit* not sexual abuse not depress* not defect* not injur* not ADHD not experiment* not medication* not drug* not stimulant* not mental illness* not senile* not disease* not patient* not therap* not concussion* not stroke* not malnourish* not syndrome* not robot* not manipulat* not aphasia).ti,ab.*

In Search Strategy 5, we used the references of four relevant reviews or meta-analyses that were identified by Search Strategies 1 – 4.

The following inclusion and exclusion criteria were applied to the identified records:

1. Cognitive ability was assessed and ability scores were: (a) full scale test scores from tests developed and standardized to measure cognitive ability; (b) composite scores comprised of subtests measuring one specific cognitive ability, taken from a test developed and standardized to measure cognitive ability; or (c) composite scores derived from test batteries comprised of individual scales all developed and standardized to measure cognitive ability. Screenings reported to be shorter than 10 minutes were excluded.
2. Cognitive ability was assessed longitudinally, and the same ability was assessed at all times of measurement (e.g., Gf). The minimum acceptable test-retest interval was one day.
3. The study reports primary analyses as opposed to integrative secondary analyses in meta-analyses or reviews.
4. The study did not investigate a clinical sample and did not apply an intervention to all participants between measurement.
5. The study examined a sample that covered a sufficiently narrow age range. An age range was considered sufficiently narrow if it did not include more than two of seven predefined, partially overlapping life stages: infants [0-3], preschool [3-7], elementary school [5-14], secondary school [10-20], young adults [16-25], adults [25-70], old age [70-100].
6. The test-retest intervals in the study were sufficiently homogeneous. The range of test-retest intervals was considered too large if the standard deviation of the interval was more than half the size of the mean interval length.
7. The study examined human participants.

All 4,077 articles, dissertations or book chapters were first screened for eligibility based on titles and abstracts by two of the project researchers and four trained student assistants. One project researcher validated the ratings of the coauthors in a randomly assigned subsample of 250 records. Interrater agreement was 91.20%, and all discrepancies were resolved through discussion. Next, two of the project researchers and four trained student assistants assessed 965 studies for eligibility based on the full text. Note that we contacted the corresponding authors of studies published 2010 or later if rank-order stabilities were not reported to obtain these missing values. We contacted the authors of 100 studies, 33 of which responded, leading to inclusion in the meta-analysis. A table presenting the studies included and excluded during the full-text screening process can be found in the OSF directory (https://osf.io/ajufs/). One researcher validated the ratings of the coauthors in a randomly assigned subsample of 50 records. Interrater agreement was 94% and all discrepancies were resolved through discussion. A total of 190 different records remained at the conclusion of the screening process; 185 of them included 205 samples and provided 1,288 rank-order stability effect sizes based on manifest values. Six records including six samples provided 50 rank-order stability effect sizes based on latent factor scores. We only included studies reporting latent correlations if standard errors (*SE*s) for the relevant effects were available. Note that rank-order stability effect sizes based on manifest values and latent

factor scores were investigated separately because these effect sizes are not comparable (i.e., in contrast to manifest values, latent factor scores are corrected for measurement error).

Coding Procedure, Study Variables, and Subdatasets

All records were coded on the variables described below by four trained student assistants, and one project researcher controlled each coding. As many studies contained multiple scales that were assigned to different broad CHC abilities and/or several measurement points, we often estimated multiple effect sizes based on one sample. All variables were coded on the effect-size level. The complete dataset including all moderator variables is available as an Excel sheet at https://osf.io/ajufs/. Detailed information on the 205 samples providing rank-order stability effect sizes based on manifest values is presented in the Supplemental Materials (Table S1). The frequencies of the study variables are reported in Table S2.

Study Variables

Effect Size.—Autocorrelations r_{tt} of participants' cognitive ability scores over time were coded as effect sizes of rank-order stability (e.g., r_{tt} of Gf at T1 with Gf at T2). If a study reported several measurement points, all possible combinations of r_{tt} were coded. For example, if three measurement points were reported, three r_{tt} effect sizes corresponding to the correlations of T1 with T2, T1 with T3, and T2 with T3 were coded. In manifest correlations, effect size variance was estimated according to the following formula (Borenstein et al., 2009, p. 41; Hedges & Olkin, 2014).

$$V_{r_{tt}} = \frac{\left(1 - r_{tt}^2\right)^2}{n - 1} \quad (1)$$

In latent correlations, effect size variance was calculated by squaring the *SE*. In the following, we labeled the averaged rank-order stability ρ because ρ is the population parameter of r (Borenstein et al., 2009; Hedges & Olkin, 2014).

Publication Type.—Publication type was coded as a dichotomous variable (i.e., 0 = peer-reviewed journal article; 1 = no peer-reviewed journal article).

Sample Size *n*.—For each effect size, we coded the respective sample size n as a continuous variable. If available, we coded the overlapping n of T1 and T2. If only n of T1 and n of T2 were available, we coded the smaller n as an estimate of the overlapping n. If only n of T1 or n of T2 was available, we coded the available n as an estimate of the overlapping n.

Test-Retest Interval.—For each effect size, we coded the test-retest interval between the measurement points as a continuous variable in years. In cases where the duration was reported precisely to the day, we coded it accordingly (i.e., a one-month interval was defined as .083 years). We subtracted 5 from the test-retest interval and used this new

variable instead of test-retest interval in all subsequent analyses. Thus, the parameters in our meta-regressions represented a 5-year interval instead of a 0-year interval thereby allowing a more meaningful interpretation.

Age.—The mean age of the study participants at the first measurement point was coded as a continuous variable in years. If the exact information regarding participants' age was missing, we used other available information from the sample description (e.g., grade level) to estimate a plausible age if possible. We subtracted 20 from age and used this new variable instead of age. By this operation, the parameters in subsequent meta-regressions corresponded to an age of 20 years instead of an age of 0 years. This rescaling allowed for a more meaningful interpretation of the parameters because the stability at age 20 is more relevant to practical testing contexts than the stability in newborns where cognitive ability testing is not feasible. In addition, we calculated the quadratic form of age by squaring this variable.

Life Stage.—Life stages were coded to allow modeling the age moderation effect in the terms of a connected linear spline. The effect sizes referred either to preschool children, school-aged children and adolescents, adults, older adults, or to elderly individuals at T1. We assorted the effect sizes to one of these categories based on the reported age. Samples with age ≤ 6 years were assorted as preschool children. Samples with ages > 6 and ≤ 18 years were assorted as school-aged children and adolescents. Samples with ages >18 and ≤ 65 years were assorted as adults. Samples with ages > 65 and ≤ 80 years were assorted as older adults. Samples aged > 80 years were assorted as elderly (Lindenberger & Staudinger, 2018). For the categories preschool children, school-aged children, adolescents, adults, older adults, and elderly, we calculated five connected linear-spline variables by following the approach of Tucker-Drob and Briley (2014). These variables represent the linear age differences within these life stages, meaning that a linear age moderation trend was modeled separately for each spline section (i.e., age group). They were calculated as follows:

Preschool children spline (age_1).—If age is ≤ 6, then age_1 = age; if age is > 6, then age_1 = 6.

School-aged children and adolescents spline (age_2).—If age is ≤ 6, then age_2 = 0; if age is between > 6 and ≤ 18, then age_2 = age – 6; if age is > 18, then age_2 = 12.

Adult spline (age_3).—*If* age is ≤ 18, then age_3 = 0; if age is between > 18 and ≤ 65, then age_3 = age – 18; if age is > 65, then age_2 = 47.

Old adult spline (age_4).—If age is ≤ 65, then age_4 = 0; if age is between > 65 and ≤ 80, then age_4 = age - 65; if age is > 80, then age_4 = 15.

Elderly spline (age_5).—If age is ≤ 80, then age_5 = 0; if age is > 80, then age_5 = age – 80.

This procedure transformed the age values of the effect sizes in each spline section to values ranging from zero to the age range value of the spline (e.g., 18 to 65 years results in an age

range of 0 to 47 years). Age values less than the minimum value of the respective spline were set to zero, and age values greater than the maximum value of the spline were set to the value of the age range of the spline.

Cognitive Ability Captured.—The effect sizes were classified into subcategories of specific CHC broad abilities or general intelligence. We computed one dichotomous dummy variable for each subcategory of captured cognitive ability (e.g., Gf: 1 = captured cognitive ability is Gf; 0 = captured cognitive ability is not Gf). In subsequent moderator analyses, we used *g* as the reference category.

General Cognitive Ability Level at T1.—If available, we coded the mean IQ score (M = 100; SD = 15) at the T1 score as a continuous variable. Sometimes the mean general cognitive ability level was reported as z or t scores. We converted these scores into IQ scores.

Test Instrument.—Different test instruments that were used in at least four samples were coded as separate dichotomous dummy variables (e.g., CFT: 1 = instrument is CFT; 0 = instrument is not CFT). All instruments that were used in less than four samples were subsumed into one further dichotomous dummy variable representing all other instruments. In the subsequent moderator analyses, the most frequently used test instrument was used as the reference category.

Varying Measurement Instruments.—Test and retest measurements of cognitive abilities were either carried out with the same instrument (e.g., CFT at T1 and T2), with instruments from the same test family (e.g., WISC-R at T1 and WAIS-III at T2), or with different instruments (e.g., Stanford-Binet at T1 and WISC-III at T2). We coded one dichotomous dummy variable for each of these categories (e.g., same test family: 1 = T1 and T2 were measured with tests from same test family; 0 = T1 and T2 were not measured with tests from same test family). In the subsequent moderator analyses, the category referring to the use of the same instrument was used as the reference category.

Complete Test Battery.—Measurements of cognitive abilities were either carried out with a complete test battery or based on a selection of subtests from one or more tests. Note that we excluded all studies that used just one single subtest as an estimate of cognitive ability because these did not satisfy our criteria for a valid cognitive ability measure (e.g., only the mosaic cube test from WISC as an estimate of Gv). Complete test was coded as a dichotomous dummy variable (1 = cognitive ability was estimated based on a selection of subtests; 0 = cognitive ability was estimated based on the complete test).

Geographic Location.—Studies were conducted in different countries. We subsumed these countries into the following four separate dichotomous variables according to their geographic location: North America (United States and Canada), Europe (all European countries; e.g., France), and Asia (all Asian countries; e.g., Japan). Countries from South and Central America, Oceania, and Africa were carried out in less than four samples each and therefore were coded together as one dichotomous dummy variable representing other locations (i.e., Australia, Congo, Ecuador, Guatemala, and New Zealand). In the subsequent

moderator analyses, the category North America was used as the reference category because the most effect sizes referred to this geographic location.

Reliability.—Test scales have varying degrees of reliability. To control an effect size for reliability, reliability estimates must be available for both the test and the retest. We first searched all records for estimates of reliability (internal consistency Cronbach's α or split-half reliability) based on the sample studied. As these were rarely available, we also used the information provided in the test manuals. We determined that the reliability estimates given in the manuals were only appropriate if the respective test scales were used and scored in accordance with the manuals.

Subdatasets

Before conducting analyses, we transformed the complete dataset into ten subdatasets based on *g* and the CHC broad abilities that referred to the effect sizes (i.e., *g*, Ga, Gc, Gf, Gl, Gq, Grw, Gs, Gv, and Gwm). This transformation was essential because we aimed to report all analyses not only for the entire dataset but also for each broad ability separately. The frequencies of the study variables in the full dataset and the subdatasets are reported in Table S2. Detailed descriptive statistics of the dataset are provided at the outset of the results section.

Analyses

The goal of this meta-analysis was to summarize findings from longitudinal studies reporting rank-order stability in cognitive abilities throughout the life span. In many cases, we estimated more than one effect size based on one sample, leading to a clustered data structure with a large number of partly dependent effect sizes within a smaller number of samples. To address this nested data structure, we applied robust variance estimation (RVE) with the robumeta package in R to control for the dependency of effect sizes within studies (Hedges et al., 2010; Tipton, 2013). Of note, some of our analyses were not possible in robumeta. Therefore, when necessary, we also used other R packages or Mplus and clearly denote these occurrences in the following paragraphs.

Moderator analyses were only calculated in the subdatasets that included at least four study samples for moderators because a minimum of four *df* is recommended for RVE meta-regressions (Tipton, 2015). Similarly, we only reported the results of moderators if the *df* of an effect was ≥ 4.

Pre-Analyses

Outlier analyses.—Because methods for outlier analyses with RVE are not yet available, we performed outlier analyses using the influence function of the metafor package in R (Viechtbauer, 2010) based on a simple random effects model in the complete dataset including all effect sizes. Outliers were identified using the studentized residuals, which are the ratio of a raw residual and the sampling variance of the raw residual (Viechtbauer & Cheung, 2010). All effect sizes with absolute studentized residuals larger than 1.96 were checked for coding errors and plausibility. The influence of the identified outliers was tested by calculating Cook's distance, which tests whether the average effect changes after

excluding the considered outlier (Viechtbauer & Cheung, 2010). Viechtbauer and Cheung (2010) recommend using the identified influential outliers for sensitivity analyses rather than routinely deleting effect sizes based on influence statistics (i.e., based on a significant Cook's distance). Therefore, we did not routinely exclude effect sizes with significant Cook's distance but used these values for robustness tests of the magnitude of rank-order stabilities (for more information, see H3 analyses). We only excluded outlier effect sizes that were implausible (i.e., negative autocorrelations).

Publication Bias.—Publication bias occurs in meta-analyses when an unrepresentative proportion of significant studies showing a positive direction are included in the analyses (Duval & Tweedie, 2000; Egger et al., 1997). We applied multiple methods to test for publication bias. We calculated funnel plot analyses and trim-and-fill analyses of effect sizes in the complete dataset and the subdatasets. Because funnel plot analyses using RVE are not yet available, we aggregated multiple effect sizes from the same studies and conducted funnel plot analyses with the funnel function of the metafor package. Before calculating funnel plots and performing trim-and-fill analyses, we residualized the effect sizes for age and test-retest interval effects to ensure that any identified asymmetry in the plots was not caused by these moderating variables. We used the best fitting age and test-retest interval functions (see H1 and H2). In addition, r effect sizes were transformed into Fishers' z effect sizes because r variance (V_r) depends on the magnitude of r and therefore would lead to biased funnel plots (see Formula 1). Note that Fishers' z variance (V_z) does not depend on the magnitude of z (see Formula 2).

$$V_z = \frac{1}{n-3} \tag{2}$$

Lastly, we examined whether the type of publication (i.e., peer-reviewed journal article vs. non-peer-reviewed journal article) significantly moderated stability. We calculated a random-effect RVE meta-regression analysis based on the complete dataset, as well as based on the subdatasets of g and Gf including type of publication as a dichotomous moderator. Note that the other subdatasets contained fewer than four samples from non-peer-reviewed journal articles.

Main Analyses

H1: Test-Retest Interval.—We used Mplus and followed the approach by Tucker-Drob and Briley (2014) to investigate linear (H1a) and nonlinear effects (i.e., quadratic, exponential; H1b) of the test-retest interval on stability. Equations for these models are presented in Table S5 in the Supplemental Materials. We had to use Mplus instead of RVE in these analyses because exponential functions in RVE are not yet available in robumeta. Note that the approach of Tucker-Drob and Briley (2014) also applies cluster robust standard errors to account for nested data by the type is complex command in Mplus. Furthermore, this approach provides model fit indices that allow for a comparison of the different effect forms (i.e., to compare linear, quadratic, and exponential effects). To make these results

comparable to the RVE analyses, we used the weighting formula from robumeta (Hedges et al., 2010; Tipton, 2013) in Mplus (Formula 3).

$$w = \frac{1}{k\left(Var + \tau^2\right)} \tag{3}$$

In the original analyses by Tucker-Drob and Briley (2014), the weighting formula did not include the between sample variance τ^2 (Formula 4).

$$w = \frac{1}{k^* Var} \tag{4}$$

To enter τ^2 into the Mplus analyses, we estimated the same models with robumeta and used the estimated τ^2 from these analyses. To keep model fit indices comparable between the linear, quadratic, and exponential models, we entered the same τ^2 as an approximation into all Mplus models (i.e., the τ^2 from the quadratic model). To find the best fitting test-retest interval form, models were evaluated based on the Bayesian Information Criteria (BIC) fit index (smaller BIC values indicate a better fit to the data).

H2: Age.—We used Mplus and followed the approach of Tucker-Drob and Briley (2014) to estimate the effects of age on stability. In the first step, linear, quadratic, connected-linear-spline, and exponential effects of age were investigated without including further predictor variables. To find the best fitting age form, models were evaluated based on the BIC fit index. In the second step, the models with the best fitting age form were controlled by the best fitting test-retest interval form. That is, stability was simultaneously predicted by age and test-retest interval. In the third step, the continuous interaction effects between test-retest interval and age were also entered into the model. Interaction effects were evaluated by comparing the BIC fit indices of models from step 2 and step 3. Finally, the same three steps were repeated based on subsamples of preschool children, school-aged children, and adolescents to explore whether interaction effects of age and test-retest interval can be discovered in these more restricted age groups.

Residualizing out Test-Retest and Age Effects before Conducting H3 to H9.—Before testing H3 to H9, we took into account differences between the effect sizes on test-retest interval and age so that the variation in rank-order stability between datasets, samples, and effect sizes could not be explained by varying test-retest interval and age. This modification was done by residualizing each effect size for test-retest interval and age effects based on the functional forms that were found to be best fitting in H1 and H2 (for the exact formula, see the result sections).

After this process, each effect size referred to an expected stability for a test-retest interval of five years and a sample age of 20 years. That is, the parameters of the RVE meta-regressions

based on the residualized effect sizes represent the average stability ρ for a test-retest interval of five years and for a sample age of 20 years. Note that the residualization process was only carried out in the complete dataset, whereupon the complete dataset was divided into subdatasets based on the CHC broad abilities that referred to the effect sizes. Residualizing out test-retest interval and age separately in each subdataset would sometimes lead to implausible results because in some small subdatasets that comprised only a few samples, implausible functions for test-retest interval or age were estimated (e.g., in Grw, test-retest interval positively moderated the stability. For more detail, see H1 results section).

H3: Magnitudes of rank-order stabilities.—We estimated the magnitude of rank-order stabilities by calculating intercept-only random effect RVE meta-regressions based on the complete dataset and the subdatasets. We conducted three robustness checks of the magnitude of rank-order stabilities. First, we repeated the analyses without controlling for test-retest interval and age. Second, we repeated the analyses after excluding all influential outlier effect sizes. Third, we repeated the analyses after excluding all effect sizes causing asymmetry (i.e., publication bias observed by Egger's regression).

H4: Cognitive Ability Captured.—To test whether captured cognitive ability was associated with the magnitude of rank-order stability, we calculated one random effect RVE meta-regression in the complete dataset and included all dichotomous dummy variables for each cognitive ability except for the *g*-factor category of general intelligence. Thus, the *g*-factor category was used as the reference category.

H5: General cognitive ability level.—To test whether the stability of *g* decreases with increasing *g*-level, we conducted one random effect RVE meta-regression in the subdataset that referred to *g* and included the mean *g*-level at T1 as a continuous predictor of stability. To test whether the stability of CHC broad abilities increased with increasing *g*-level, we conducted several random effect RVE meta-regressions in the subdatasets that referred to the particular abilities. In these analyses, stability was predicted by the continuous moderator mean general intelligence level at T1.

H6: Test Instrument.—To test whether stability varied by test instrument, we conducted several random effect RVE meta-regressions and included all dichotomous dummy variables for each test instrument as predictors of stability. In the analyses based on the complete dataset, *g*, Gc, Gv, and Gwm, the WISC test was used as the reference category because it was the category with the most effect sizes. In the analyses based on Gf, the Ravens matrices test was used as the reference category because it was the category with the most effect sizes. In the analyses based on Gq and Gs, the Woodcock Johnson test was used as the reference category because it was the category with the most effect sizes.

H7: Varying Measurement Instruments.—To test whether stability differed based on whether varying measurement instruments were used at T1 and T2, we conducted several random effect RVE meta-regressions and included the dichotomous dummy variables of the use of the same test, same test family, and different tests as predictors of stability. In these analyses, the use of the same test was used as a reference category.

H8: Complete Test.—To test whether stability differed depending on whether the complete test battery was used as an estimate of cognitive ability, we calculated several random effect RVE meta-regressions and included the dichotomous dummy variable complete test as a predictor in these analyses. The complete test variable was used as a reference category.

H9: Geographic Location.—To test whether stability varied between geographic locations, we calculated random effect RVE meta-regressions including the dichotomous dummy variable that referred to different geographic location categories. In these analyses, North America was used as a reference category, as it was the category with the most effect sizes.

H10: Using Reliability Estimates to Disattenuate for Measurement Error.—Like all correlations, stability coefficients are attenuated by measurement error. If hypothesized moderators of stability (e.g., age) are themselves associated with variation in measurement error, this can lead to spurious results. When estimates of test reliability (e.g., internal consistency) are available, such estimates can be used to correct correlations for measurement error. To test whether the findings from H2, H3, H4, and H6 are robust to corrections for measurement error, we replicated these analyses within a subsample of effect sizes where reliability information was available. Each of these analyses was performed twice. First, we carried out the analyses without adjusting for reliability to demonstrate the original effects within this subsample of effect size. Second, we conducted the analyses after correcting both the effect sizes and the effect variance for measurement error by the following formulas (Borenstein et al., 2009, p. 341).

$$r_{adjusted} = \frac{r_{original}}{\alpha} \tag{5}$$

$$V_{adjusted} = \frac{V_{original}}{\alpha^2} \tag{6}$$

Where α represents the averaged *reliability* ($r_{tt1} + r_{tt2}/2$) of the first and the second measurement.

Exploratory Analyses

1. **Simultaneous Inclusion of all Categorical Moderators.—**To assess the robustness of our initial findings, we conducted a meta-regression analysis that incorporated all categorical moderators simultaneously. This comprehensive approach was driven by Chi-square tests that indicated an uneven distribution among these moderators (see Data Description section).

2. Age Moderation Analysis Based on Age Homogenous Samples.—Given that some of the assessed effect sizes were derived from very age heterogeneous samples, we conducted an age analysis replication with a subset of effect sizes that were exclusively drawn from age homogeneous samples. In this procedure, we excluded samples with an age *SD* greater than 5 years, or, in cases where the age *SD* was not reported, samples spanning an age range exceeding 20 years. Samples lacking any information about age variation were also excluded. The final analysis incorporated 1038 effect sizes drawn from 153 samples.

Results

Description of Studies

Table S2 reports the descriptive statistics and frequencies of the study variables for the complete dataset and for subsets referring to different abilities or latent estimates. Effect sizes were obtained from a wide range of sources including many well-known longitudinal studies of cognitive development and cognitive aging such as the Lothian Birth Cohorts of 1921 and 1936, the Berlin Aging Study, the Seattle Longitudinal Study, the Virginia Cognitive Aging Project, the BETULA Study, the Victoria Longitudinal Study, the Colorado Adoption Project, the Twins Early Development Study, the LOGIC Study, Project Head Start, and the Fullerton Longitudinal Study. We included a substantial number of effect sizes corresponding to test-retest correlations among manifest (i.e., not latent) variables (e = 1,288) corresponding to a large number of samples (h = 205) with a total sample size of n = 87,508 across all abilities. In addition, we coded 50 latent effect sizes based on six samples with a total sample size of n = 18,107. The main meta-analysis only included the manifest effect sizes. Thus, the following sections present the results based on the datasets of manifest correlations, while meta-analytic results derived from latent correlations are reported in Supplemental Materials (Tables S8 and S9). Unfortunately, the dataset based on latent correlations was too small (i.e., $df < 4$) for most moderator analyses. Therefore, these results should be interpreted with great caution.

The mean number of participants per effect size (M = 440.53) and the associated standard deviation (SD = 1391.10) indicate that our meta-analysis includes, on average, rather large samples and that there is considerable variation in sample size across studies, ranging from n = 9 to n = 15,496. Most effect sizes referred to g (650 effect sizes, 151 samples) followed by Gc (195 effect sizes, 76 samples) and Gf (172 effect sizes, 60 samples). Effect sizes based on Ga, Gl, and Grw each referred to fewer than 10 samples, which hinders complex moderator analyses based on these abilities. The data integrated effect sizes over a wide range of test-retest intervals (1day to 79 years), with an average of 6.52 years (SD = 10.81). The age of the participants at first testing in the studies also ranged widely (1.00 to 88.50 years), with a mean age of 18.07 years (SD = 21.37). The percentage of females in the sample ranged from 0.00 to 100%, with an average of 50.05% (SD = 23.19%). Average general cognitive ability level was available for 456 effect sizes based on 79 samples. These effect sizes indicated an average cognitive ability level of M = 104.19 IQ (SD = 10.37).

Most of the records included into our meta-analysis were peer-reviewed studies (1,202 effect sizes, 197 samples). The test instruments varied, with the Wechsler tests being the most common (338 effect sizes, 85 samples), followed by the Stanford Binet test (74 effect

sizes, 24 samples). Measurements were predominantly derived from identical tests at both measurement points (751 effect sizes, 172 samples) and from complete tests (928 effect sizes, 161 samples). Furthermore, the majority of studies originated from North America (808 effect sizes, 134 samples) or Europe (420 effect sizes, 62 samples) and only a small minority of studies came from Asia, Africa, or South America (52 effect sizes and 8 samples combined). This indicates that a large majority of studies originated in Western, educated, industrialized, rich, and democratic (WEIRD) societies. The studies span a wide publication period, ranging from 1921 to 2022 (M = 1998.70, SD = 18.64), reflecting the long-standing interest in this aera of research. A closer look at the data reveals that 773 out of 1288 effect sizes, or approximately 60%, were published after the year 2000, underscoring the growing attention and continued development of this research area over the past two decades. Finally, a reliability estimate for the measurements was available for only 250 effect sizes, based on 45 samples. The average reliability was found to be quite high (α = .89, SD = 0.07), with the low SD indicating minimal variation between studies. Reliability was available for at least four samples for each ability, allowing us to conduct robustness checks for the main analyses by controlling for reliability. Nonetheless, this subset of data is considerably less comprehensive and not necessarily representative of the complete dataset. Therefore, we subject analyses based on the reliability-adjusted dataset to additional scrutiny, comparing the results to those in the same reduced dataset without adjusting for reliability (i.e., all effect sizes for which reliability estimates were available).

Table S3 in the Supplemental Materials presents a correlation matrix of all continuous variables based on the complete dataset. Effect sizes (r) showed a positive correlation with participant age (r = .42, p < .001) and a negative correlation with both the interval duration (r = −.22, p < .001) and the year of publication (r = −.19, p < .001). The number of participants per effect size (n) did not show a significant correlation with r effect sizes (r = .05, p = .103) nor with the duration of test-retest intervals (r = .01, p = .779). Conversely, it showed a positive correlation with the year of publication (r = .11, p < .001) and negative correlations with sample age (r = −.08, p = .003), general cognitive ability level (r = −.15, p = .002), and reliability of measurements (r = −.14, p = .030). Test-retest interval length was positively correlated with general cognitive ability level (r = .09, p = .048) and the year of publication (r = .18, p < .001). Sample age showed a negative relationship with reliability (r = −.18, p < .001), and a positive relationship with the year of publication (r = .13, p < .001). General cognitive ability level, the year of publication, and reliability did not show significant relationships with each other (p > .05). Taken together, the correlations indicate that age, test-retest-interval and samples size are associated with various other moderators including sample size, and that age and test-retest-interval are associated with stability. These results underscore the importance of controlling for the duration of the test-retest interval and sample age, while also considering sample size when conducting moderator analyses for other variables.

Table S4 in the Supplemental Materials presents a matrix of Chi-square tests conducted across all categorical variables based on the complete dataset. This additional analysis served to uncover possible confounds of the categorical moderator variables due to uneven distribution of one variable in the categories of another. The Chi-square tests revealed that none of the categorical variables demonstrated an even distribution across the other

categorical variables. To rule out or examine any confounding influences on the moderator analyses, we conducted an additional review of the significant categorical moderators while controlling for the other significant categorical moderators.

Pre-Analyses

Outlier analyses.—In the complete dataset, 80 effect sizes showed an absolute studentized residual larger than 1.96 and were therefore identified as significant outliers. Cook's distance analyses indicated that 13 of these outlier effect sizes were influential. These effect sizes were noted for robustness check analyses. Four outlier effect sizes were excluded from all further analyses because they constituted negative and therefore implausible rank-order stability coefficients (i.e., Bauer & Smith, 1988, reported a r_{tt} of −.25; Jankowska et al., 2014, reported a r_{tt} of −.51; McArdle & Wang, 2008, reported r_{tt} of −.05 and −.02).

Publication Bias.—Funnel plots and trim-and-fill plots based on Fishers' z effect sizes for the complete dataset and all subdatasets are presented in Figures S1 and S2 in the Supplemental Materials, respectively (for funnel plots and trim-and-fill plots based on r, see Figures S3 and S4 in the Supplemental Materials). Visually, no extraordinary asymmetry was noted. The trim-and-fill R_0 estimator indicated no significant publication bias in either dataset ($p > .05$). Egger's regression analyses indicated significant asymmetry in the complete dataset ($z = 2.30$, $p = .021$) and g ($z = 2.01$, $p = .044$) and no significant asymmetry in all other subdatasets ($p > .05$). Two samples causing significant asymmetry in the complete dataset and g were noted for robustness check analyses. Finally, publication type was not significantly related with stability in either dataset ($p > .05$) and thus indicated no publication bias.

Main Analyses

H1: Test-Retest Interval.—Model fit indices for models testing linear, quadratic, and exponential test-retest interval effects on stability are reported in Table 2. In the complete dataset, and for g, Gc, and Gv, the exponential test-retest interval function showed the best fit (i.e., lowest BIC). In Ga, Gc, Gl, Grw, Gs, and Gwm, the linear test-retest interval function indicated the best fit. In Gq, the quadratic function indicated the best fit. However, these linear and quadratic trends were based on small datasets ($df < 4$) and therefore cannot be interpreted substantively. Figure 4 depicts the exponential test-retest interval function based on the complete dataset. The model parameters of the best fitting models in all datasets are reported in Table 3 (parameters of the remaining models are reported Table S6 in the Supplemental Materials) and depicted in Figure S5 in the Supplemental Materials. The exponential test-retest interval functions in the complete dataset, g, Gc, and Gv, indicate that stability initially steeply decreased with each additional year of the test-retest interval, whereas after approximately five years, the stability did not further decrease with increasing duration of the test-retest interval and approximated a fixed asymptote of .69 (complete dataset), .67 (g), .73 (Gc), and .66 (Gv).

H2: Age.—Model fit indices for models testing different functional forms of the age effects on stability and their interaction with test-retest interval in steps 1, 2, and 3 are reported in

Table 2. In step 1, the exponential age function showed the best model fit in the complete dataset, *g*, Gc, Gf, Gs, Gv, and Gwm (i.e., lowest BIC). The linear age function indicated the best model fit in Ga, Gl, and Gq. The linear spline function showed the best model fit in Grw.

Figure 5 depicts the exponential age function as well as linear spline age function based on the complete dataset. The model parameters of the best fitting models are reported in Table 3 (parameters of the remaining models are reported Table S5 in the Supplemental Materials) and depicted in Figure S6 in the Supplemental Materials. The exponential age functions in the complete dataset, *g*, Gc, Gf, Gs, Gv, and Gwm, indicate that stability initially increased with each additional year of age, whereas after approximately 20 years, the stability did not further increase with increasing age and approximated a fixed asymptote of .79 (complete dataset), .84 (*g*), .85 (Gc), .78 (Gf), .82 (Gs), .79 (Gv), and .76 (Gwm). In Gl and Gq, linear models indicated increasing stability with increasing age over the entire lifespan. In Ga, the linear distribution could not be interpreted because of too few samples. In Grw, the linear spline function could not be interpreted because $df < 4$.

In the complete dataset, the comparison of the step 2 and step 3 analyses indicated that model fit improved after the additional inclusion of the interaction with test-retest interval duration. This interaction is illustrated in Figure 6, demonstrating that the effects of the test-retest interval are more pronounced in young children compared to adolescents or adults. In all other datasets, the interaction of age with test-retest interval duration did not enhance the model fit. The model parameters of the best fitting models are reported in Table 3. The parameters of the best-fitting models are presented in Table 3. It is important to note that, with only a few exceptions, the direction and significance of age effects remained consistent, even after controlling for the test-retest interval.

Additional exploratory analyses in a single subsample comprised of preschool children, school-aged children and adolescents did not indicate interaction effects of age and test-retest interval (see Supplemental Materials, Table S7 for details).

Residualization of Test-Retest and Age Effects Before Conducting H3 to H10.—The best fitting function in analyses of H2 for the complete dataset included exponential effects of test-retest interval and age and their interaction. We used estimates from this model to residualize all effect sizes as follows:

$$r_{residualized} = r_{observed} - \left(.746 - .005e^{-.223 * age - .002 * age * interval} + .025e^{-.265 * interval}\right) + \left(.746 - .005e^{-.223 * 0 - .002 * 0 * 0} + .025e^{-.265 * 0}\right) \quad (7)$$

That is, we first subtracted the expected *r* value (i.e., the first bracket) from the observed *r* value (i.e., $r_{observed}$) to get the deviation from the expected value at a given interval and age. We then added the expected *r* value at interval = 0 and age = 0 (i.e., the second bracket) to this deviation. Thus, $r_{residualized}$ reflects the *r* value which we predicted according to our statistical model for $r_{observed}$ at interval = 0 and age = 0. Note that interval = 0 represents

the interval of five years and age = 0 represents the age of 20 years because 5 and 20 were subtracted from the interval and age variables respectively before we conducted these analyses. By applying this formula, each effect size was transformed into a corresponding model implied effect size for age 20 with a test-retest interval of five years.

H3: Magnitudes of rank-order stabilities.—Table 4 reports the magnitudes of rank-order stabilities at age 20 years and a test-retest interval of 5 years, as implied by the best fitting model for each dataset. Estimates ranged from $\rho = .65$ in Ga to $\rho = .80$ in *g*. The mean effect across all abilities was $\rho = .76$. The robustness checks indicated no noticeable differences from the main analyses (see Table S10 in Supplemental Materials).

H4: Cognitive Ability Captured.—The captured cognitive ability significantly moderated stability. We found a significantly lower stability in Ga ($\rho = -.15$, $p = .043$), Gf ($\rho = -.11$, $p < .001$), Gs ($\rho = -.05$, $p = .018$), Gv ($\rho = -.05$, $p = .007$), and Gwm ($\rho = -.13$, $p = .002$) than in the reference category *g*, whereas Gc and Gq did not significantly differ from *g* ($p > .05$; for more details, see Table 3).

H5: General cognitive ability level.—The general cognitive ability level of the sample was not significantly related to stability in any dataset ($p > .05$; for more details, see Table 3). For some abilities, the analyses were not conducted or could not be interpreted because of too few samples.

H6: Test Instrument.—In the complete dataset and the subdatasets of *g*, Gf, and Gv, the test instrument significantly moderated stability, whereas it was not related to stability in the subdatasets of Gc, Gq, Gs, and Gwm ($p > .05$; for more details, see Table 3). In Ga and Gl, these analyses were not conducted because of too few samples. In the complete dataset, CFT ($\rho = -.09$, $p = .019$), Raven's matrices ($\rho = -.17$, $p < .001$), Woodcock Johnson ($\rho = -.08$, $p = .004$), and mixed instruments ($\rho = -.04$, $p = .044$) demonstrated significantly lower stability than the reference category WISC. In *g*, Woodcock Johnson ($\rho = -.08$, $p = .004$) had a significantly lower stability than the reference category WISC. In Gf, CFT ($\rho = .09$, $p = .034$), other instruments ($\rho = .13$, $p = .003$), and mixed instruments ($\rho = .16$, $p < .001$) demonstrated a higher stability than the reference category Raven's matrices. In Gv, Woodcock Johnson ($\rho = -.10$, $p = .036$) and the other instruments category ($\rho = -.10$, $p = .031$) showed significantly lower stabilities than the reference category WISC.

H7: Varying Measurement Instruments.—In the complete dataset, *g*, Gc, Gf, and Gwm varying measurement instruments significantly moderated stability. The use of different tests at the two times of measurement was associated with a lower stability compared to the reference category same test. In Ga, Gl, Gq, Grw, Gs, and Gv these analyses were not conducted or interpreted because of too few samples.

H8: Complete Test.—In the complete dataset and *g*, incomplete testing was associated with lower stability than the reference category complete testing (complete dataset: $\rho = -.04$, $p = .009$; *g*: $\rho = -.04$, $p = .034$; for more details, see Table 3). In Gc, Gf, Gq, Gs, Gv, and Gwm, incomplete testing was not significantly related to stability. In Ga, Gl, and Grw, these analyses were not conducted because of too few samples.

H9: Geographic Location.—In the complete dataset and *g*, we found a slightly lower stability in Europe than in the reference category North America (complete dataset: $\rho = -.05$, $p = .002$; *g*: $\rho = -.05$, $p = .024$; for more details, see Table 3). In Gc, Gf, Gq, Gs, and Gv, geographic location was not significantly related to stability. In Ga, Gl, and Grw, these analyses were not conducted because of too few samples.

H10: Using Reliability Estimates to Disattenuate for Measurement Error.—Exponential age effects were confirmed in the subset of effect sizes with available reliability information (see Table 3 and Figure 7). Both, the model unadjusted for reliability and the model adjusted for reliability demonstrated that stability initially increased with each additional year of age, whereas after approximately 20 years, the stability did not further increase with additional aging. The two models only substantially differed in the fixed asymptote that was reached after approximately 20 years (unadjusted for reliability: asymptote = .80 $p < .001$; adjusted for reliability: asymptote = .90, $p < .001$). The age scaling factor and growth rate were almost identical in both models (for more details, see Table 3).

In the model not adjusted for reliability, the stability estimates for the different cognitive abilities were nearly identical to the analysis based on all effect sizes. The most significant difference was found in Gl (all effect sizes: $\rho = .69$, $p < .001$; only effect sizes with available reliability information: $\rho = .64$, $p < .001$). Thus, the subset of effect sizes with available reliability information appears representative of the complete dataset in terms of stability. The estimates of stability in the model adjusted for reliability were substantially higher than those unadjusted for reliability, with ρ differences ranging from $\rho = .06$ in Gl to $\rho = .11$ in Gv (for more details, see Table 3). Therefore, as expected, not accounting for reliability leads to an underestimation of the stability of cognitive abilities.

In the moderator analysis of cognitive ability captured, a slightly different pattern emerged after adjusting for reliability compared to the complete, unadjusted dataset, particularly regarding Gf, Gs, and Gwm (see Table 3). Without adjusting for reliability, significant negative effects were observed for Gf ($\rho = -.12$, $p = .001$), Gs ($\rho = -.09$, $p = .023$), and Gwm ($\rho = -.18$, $p = .033$) indicating lower stability compared to general intelligence. Yet, after adjusting for reliability, these effects were no longer statistically significant: Gf ($\rho = -.07$, $p = .055$), Gs ($\rho = -.06$, $p = .200$), and Gwm ($\rho = -.17$, $p = .073$). It is important to note that despite these changes in statistical significance, the descriptive values remained negative, though to a lesser extent, even after adjusting for reliability.

The moderator analysis for test instrument also showed substantial shifts when adjusting for reliability (see Table 3). In the model without adjusting for reliability, the Woodcock Johnson test showed significantly negative effects compared to the Wechsler test, with $\rho = -.07$ ($p = .045$). Yet, after adjusting for reliability, this effect, although they remained negative, did not reach statistical significance ($\rho = -.06$, $p = .109$). This finding supports the assumption that the observed different stabilities of the tests can be explained by differences in their reliability.

Exploratory Sensitivity Analyses

1. Simultaneous Inclusion of all Categorical Moderators.—In the comprehensive model that simultaneously included all categorical moderators, the effects of captured cognitive ability showed minimal variation from the initial analyses, except for Ga, which did not retain its statistical significance. Certain effects of test instruments demonstrated shifts in statistical significance within this comprehensive model. Notably, the CFT and Woodcock-Johnson Tests did not retain their initial significant difference from the Wechsler test, while the effect sizes for Raven's Progressive Matrices remained relatively consistent. This finding can be partly attributed to the fact that CFT tests (37 out of 37 effect sizes, 100%) are used for the assessment of Gf, which showed significantly lower stability than *g* in the initial analyses. Furthermore, in all instances (57 out of 57 effect sizes, 100%), the Woodcock-Johnson test scales were utilized incompletely, which generally demonstrates lower stability than complete tests. The comprehensive model confirmed that using different tests leads to lower stability compared to using the same test, while the effect of using tests from the same test family remained statistically non-significant. The effects of incomplete tests versus complete tests lost significance in the comprehensive model, which can be attributed to the inclusion of all Woodcock-Johnson assessments (57 out of 57 effect sizes, 100%). Lastly, the effect of European samples showing less stability compared to North American samples was not maintained in the comprehensive model. This shift can be partly attributed to the fact that a majority of Raven assessments (25 out of 41 effect sizes, 61%) were conducted with European samples, and that European effect sizes more often referred to incomplete assessments (147 out of 418 effect sizes, 35% in Europe versus 186 out of 808 effect sizes, 23% in North America).

2. Age Analysis Based on Age Homogenous Samples.—The exponential age model based on age homogenous samples ($SD < 5$ years) revealed stability parameters and age effects that were nearly identical to those in the complete dataset (for more details, see Table 3). This finding implies that the inclusion of age heterogeneous samples did not significantly distort the results of the age analyses performed in H2.

Discussion

The current study provides the first comprehensive meta-analysis on the stability of cognitive ability over the life span. We investigated the rank-order correlations of cognitive abilities and their moderators in 205 samples. The analyses were conducted on the overall dataset as well as on subdatasets pertaining to general intelligence and broad cognitive abilities, thereby covering the top two strata of the CHC model, the most recent psychometric model of cognitive ability.

Overall, cognitive abilities were exceedingly stable over considerable time spans, with stabilities ranging from .65 to .80 for a 5-year interval and an age of 20 depending on the specific ability. The highest stability was observed for general intelligence. Interestingly, the more knowledge-based abilities of Comprehension Knowledge, Quantitative Knowledge, and Reading and Writing were similarly stable to general intelligence. In contrast, abilities that are based on effortful processing, such as Fluid Reasoning, Learning Efficiency,

or Working Memory capacity, tended to exhibit lower stabilities. This finding may seem counterintuitive, as effortful processing-based abilities are usually thought to be less dependent on environmental influences (e.g., Baltes et al., 1999) and therefore less susceptible to environmental changes. On the other hand, in a somewhat different argument, Tucker-Drob and Briley (2014) hypothesized that environmental experiences relevant for knowledge-based abilities may produce lasting stores of declarative knowledge. A similar argument can be made from the perspective of investment theory (Cattell, 1986). This theory proposes that during cognitive development, fluid (or effortful-processing-based) abilities are invested in the acquisition of crystallized (or knowledge-based) abilities. As the result of years of cumulative investment, these crystallized abilities are acquired and automated, such that they are better maintained even as currently available processing power wanes with aging (see also Baltes et al., 1999) or varies from day to day. Indeed, effortful processing abilities are known to begin to decline at much earlier periods in adulthood than are knowledge-based abilities (Baltes et al., 1999; Tucker-Drob, 2019). Heterogeneity in trajectories of the aging of processing abilities at earlier period of adulthood may contribute to their lower overall stability as compared to knowledge-based abilities (although see Tucker-Drob et al., 2022).

We observed that the differences in stability across abilities are diminished when test reliability is adjusted. This finding implies that the differences are at least partially due to tests of knowledge-based abilities being more reliable than tests of effortful-processing-based abilities. This difference in reliability may be due to a potentially lower emphasis on speededness in knowledge tests compared to processing-based tests, which often have strict time limits that can attenuate reliability (Hong & Cheng, 2019). Of course, other test properties such as the average test length or item difficulty distribution may also contribute to systematic differences in the reliability of knowledge-based and effortful-processing-based tests.

Moderators

Test-Retest Interval—As expected, stability declined with increasing test-retest interval, and this decay leveled off with increasing intervals. This trajectory was best described by an exponential function in those datasets that included enough long-interval effect sizes, namely, the overall dataset, general intelligence, Comprehension Knowledge, and Visual Processing. For all other abilities, there were no or very few effect sizes with very long test-retest intervals, which makes the resulting curve trajectories difficult to interpret and less trustworthy. The exponential trend is consistent with the results of previous meta-analytic investigations (Schuerger & Witt, 1989; Tucker-Drob & Briley, 2014).

There was evidence for a small but significant interaction effect between age and test-retest interval. This interaction effect implies that the impact of test-retest interval is larger in young children than in adolescents or adults. In two-year-olds, an increase of test-retest interval from one to five years results in a decrease in stability by .08. In five-year-olds, the same increase in interval results in a decrease in stability by .06. From the age of eight onward there is little additional change, with an increase of test-retest interval from one to five years resulting in a decrease in stability by .05. This interaction is in line with the

early findings by Bayley (1949) who observed steeper declines in stability with increasing test-retest intervals in younger children than in adolescents. As can be seen in Figure 6, the interaction further implies that until the age of 8, the interval effect is not asymptotic like in adolescents and adults but instead continuously decreasing, especially in very young children.

Age—As hypothesized, stability increased with age and most markedly in early childhood, with the increase leveling off over the course of adolescence until no further increase was observable in adulthood. This pattern is consistent with the results by Tucker-Drob and Briley (2014), whereas the models by Schuerger and Witt (1989) implied a further increase in stability until late adulthood (Figure 2). An exponential curve best described the age trajectories of stability for the complete dataset and for all CHC abilities except Auditory Processing, Learning Efficiency, and Reading and Writing, where the effect sizes were lacking for older samples and exponential models often did not converge.

Stability was very high in old age (asymptoting at .77 in the complete dataset), and there was little evidence for a decrease in stability in late adulthood, implying that cognitive change does not appear to be more heterogeneous in late adulthood than in early and middle adulthood. A high degree of stability in cognitive ability into late adulthood has theoretical implications as it suggests that the same factors or developmental mechanisms may play a role in cognitive decline that previously influenced individual differences in cognitive abilities. For example, the prefrontal cortex, which is associated with executive functions and working memory, may be a central determinant of both cognitive ability in adolescence and adulthood (Kane & Engle, 2002) and cognitive decline in late adulthood (Nyberg et al., 2022). Environmental selection processes may also play a role, leading individuals to remain in similarly cognitively challenging living environments throughout their lives (Harden et al., 2007; van der Sluis et al., 2008), which can also affect cognitive decline (e.g., Frick & Benoit, 2010). Importantly, the studies meta-analyzed were composed of individuals from the general population, and we excluded studies focusing on clinical samples. Studies of older adults in the general population also tend to exclude individuals with mild cognitive impairment and dementia. Moreover, individuals with mild cognitive impairment and dementia are more likely to drop out of longitudinal studies compared to those experiencing milder trajectories of cognitive decline. Thus, it is likely that the full range of heterogeneity in aging-related trajectories among older adults was restricted in many of the studies composing the meta-analytic dataset. It remains an open question whether, in a fully representative study of cognitive aging, stability would begin to decrease in old age, as sizable proportions of individuals undergo precipitous declines toward impaired levels of cognitive functioning (Lövdén et al., 2005; Tucker-Drob, 2019).

Interestingly, age trends in stability were very similar for both Comprehension Knowledge (which shows little decline or even mean-level gains over the course of adulthood) and processing-based abilities such as Fluid Reasoning and Visual Processing (which start to decline in early adulthood) (Baltes et al., 1999; Tucker-Drob, 2019). Given the different mean-level trajectories over the lifespan, one might also expect different mechanisms of cognitive change and thus different patterns of reordering between these abilities. Yet, a recent study showed that changes in knowledge-based and processing-based abilities are

correlated over the lifespan: individuals who experience a greater decline in processing-based abilities also experience smaller gains or even a decline in knowledge-based abilities (Tucker-Drob et al., 2022). This finding implies that the same underlying mechanisms drive change in all cognitive abilities (see also Li et al., 2004) or that the underlying mechanisms are closely related, which is consistent with the present observation of similar stability patterns over the lifespan in all abilities. Of course, even if all long-term cognitive changes in adulthood are driven by the same mechanisms, there may be differences between specific abilities in their susceptibility to change by shorter-term environmental influences.

General Cognitive Ability Level—The effect of the cognitive ability level of the sample on stability could only be investigated for some abilities because of insufficient data in the others. In both the overall dataset and the individual abilities where the analysis was possible, including general intelligence, there was no significant effect of mean cognitive ability level on stability. This finding is somewhat inconsistent with predictions based on the ability differentiation hypothesis. As the ability differentiation hypothesis states that general intelligence accounts for less systematic variance of cognitive performances in high-ability individuals, it would predict that the stability of composite measures of general intelligence decreases with increasing ability level (e.g., Breit, Brunner, et al., 2022). The absence of this effect may be due to the ability differentiation effect being not very consistent in children (Breit, Brunner, & Preckel, 2021) and given the large number of studies based on samples of children in the present meta-analysis, this may significantly reduce the overall impact of the ability differentiation effect. Even more importantly, the impact of the ability differentiation effect may also not be particularly evident in meta-analytic investigations because the variance of cognitive ability within samples is much larger than that between samples. The ability differentiation effect implies rather large differences in g-factor variance between the average and the very high or very low regions of the ability distribution, but small differences within few IQ points around the population average (Breit, Brunner, et al., 2022). Nevertheless, in the present analysis, there is no evidence for the hypothesized effect of ability level – and therefore of ability differentiation – on the stability of general intelligence.

Test Instruments—The effect of the test instrument was investigated across three moderator analyses. The results were consistent with our hypotheses. When the same test or tests from the same test family were used at both times of assessment, this led to greater stability estimates than when the test instrument changed between test and retest. This difference remained in the analysis that included all categorical moderators. Moreover, using complete test batteries led to greater stability than using only a selection of subtests from one or several test batteries. We also found that unidimensional tests such as the Raven tests displayed less stability than the Wechsler tests, whereas other multidimensional tests such as the Stanford-Binet tests were comparable to the Wechsler tests in stability. The only exception to this rule was the Woodcock Johnson test, which displayed lower stability than the Wechsler tests, but was never used completely. Taken together, the highest stability would be expected if the same multidimensional intelligence test was used in its entirety at both times of measurement. Multidimensional tests are usually more reliable than unidimensional tests due to their greater length, and longer test forms (e.g., complete

tests) are more reliable than short forms (e.g., a selection of subtests). We found that when adjusting for test reliability, the differences in stability between tests were diminished indicating that stability differences between tests can largely be explained by test reliability.

When different tests are used at the different times of measurement, the stability is further limited by the magnitude of the concurrent correlation between the different tests (which cannot be exceeded by the test-retest correlation). In addition, memory effects could contribute to the higher stability when the same test is used at both times of measurement. When interpreting these findings, however, it must be kept in mind that unidimensional tests generally measure Fluid Reasoning, which shows less stability than knowledge-based abilities. Thus, it cannot be completely ruled out that the lower stability of unidimensional procedures is partly or completely due to the lower stability of Fluid Reasoning. Looking at the comparison only within Fluid Reasoning, Raven's matrices are descriptively less stable than multidimensional tests, but not significantly so.

Geographic Location—The last investigated moderator was the geographic origin of the sample. The number of samples for each continent only allowed the comparison between North America and Europe. The stability was generally higher in North American samples, which was statistically significant in the full dataset and general intelligence. One explanation is that studies based on North American samples may on average use more stable test instruments than studies based on European samples. Indeed, in the analysis including all categorical moderators, the difference between Europe and North America no longer reached statistical significance.

Adjusting for Reliability—A subset of effect sizes was reanalyzed while adjusting for test reliability. Three main findings emerged from these analyses. First, the mean stability estimate for age 20 and a test-retest-interval of five years increased from $\rho = .76$ in the full, unadjusted, dataset to $\rho = .86$ for the disattenuated correlations. This value suggests that even when adjusting for reliability in adults, cognitive abilities are not perfectly stable, although the stability is very high. Second, the age moderation curve was mostly unaffected by reliability. The low stability in young children does not appear to be primarily due to lower test reliability in this age group. This may seem surprising, as one might expect lower test-reliability in young children, given that cognitive testing is more challenging in this age group, but this problem is usually compensated for by individual testing with extensive, age-appropriate test batteries. Conversely, in adults, researchers often rely on shorter scales, sometimes administered in group settings. This practice might explain the overall low correlation between test reliability and mean sample age ($r = .05$). Third, the differences in stability between knowledge-based and effortful-processing-based abilities diminished when adjusting for reliability. While the difference in stability between general intelligence and Comprehension Knowledge and Reading and Writing was largely unaffected by adjusting for reliability, the difference between general intelligence and Fluid Reasoning decreased from −.12 to −.07. Similar reductions were observed for Processing Speed and Visual Processing. Thus, reliability differences between tests of knowledge-based and effortful processing-based abilities in part account for the tests' differences in stability.

The Stability of Cognitive Abilities in Comparison to Other Constructs

Similar meta-analyses to the present one have been conducted for other personality constructs. Six major meta-analyses on various constructs and the present meta-analysis are summarized in Table 5. The mean time intervals in these meta-analyses (median 4.88 years, range 1.65-7.06 years) were comparable to the chosen reference interval in the present meta-analysis (5 years). The stability of cognitive abilities was generally higher than that of Big Five personality traits, self-esteem, vocational interests, work values, and motivational constructs. One notable exception was the stability of work values in 25- to 30-year-olds (ρ = .83). Yet, this specific estimate was based on only two effects and can therefore not be regarded as reliable.

The age moderation of the stability of cognitive abilities implied a strong increase in stability over the course of childhood and adolescence, leveling off around age 18, with stability remaining constant in adulthood. The stability of personality traits increases with age not only during childhood and adolescence but also into adulthood until the age of 50 (Roberts & DelVecchio, 2000), although there appears to be some heterogeneity between traits (Bleidorn et al., 2022). Both vocational interests and self-esteem also increased in stability during childhood and adolescence, but the stability decreased again after early adulthood (Low et al., 2005; Trzesniewski et al., 2003). Last, no systematic age trend was observable in the stability of work values (Jin & Rounds, 2012), but this meta-analysis did not include studies with samples of young children. Taken together, stability generally increased during childhood and adolescence across all constructs. Differences between constructs were observable in adulthood, with constancy of stability in cognitive abilities and work values, decrease of stability in vocational interests and self-esteem after early adulthood, and a further increase in personality traits until middle adulthood.

The overall higher stability of cognitive ability compared to the other constructs suggests that cognitive ability may be a more trait-like construct, particularly in adolescence and adulthood, whereas many other psychological constructs may be somewhat more state-like (Geiser et al., 2017). State-like constructs have been conceptualized as being more immediately responsive to—potentially fluctuating—contexts and experiences (e.g., Conley, 1984), whereas trait-like constructs have been conceptualized as relatively stable, and more slowly changing (Nesselroade & Liben, 1991). Genetic contributions to both cognitive abilities and personality are especially stable from middle to late adulthood (Briley & Tucker-Drob, 2017; Tucker-Drob & Briley, 2014). The particularly high heritability of cognitive abilities in adulthood may help to account for its higher overall stability during this period (Tucker-Drob & Briley, 2014). Of course genetic influences on cognitive abilities are likely to be contingent on environmental contexts (Tucker-Drob et al., 2013) and there is strong evidence that a multitude of environmental factors, such as educational attainment, also contribute to lifelong cognitive function (Lövdén et al., 2020; Ritchie & Tucker-Drob, 2018).

Differential developmental trends in stability of different psychological constructs may be explained by influences specific to those constructs. Personality traits are hypothesized to reach peak stability with the highest levels of identity certainty in life, which is around middle age (Bleidorn et al., 2022; Roberts & DelVecchio, 2000). Conversely,

self-esteem was suggested to undergo changes later in life when individuals review their accomplishments and experiences (Trzesniewski et al., 2003), whereas the stability of vocational interest may change depending on the educational or professional stage the individual is in (Low et al., 2005). The relatively constant high levels of stability of cognitive abilities throughout adulthood suggest that cognitive abilities are either relatively unaffected by such life events, or that the events and contexts relevant for adult cognitive function are highly stable over time or correlated with one another over time.

Implications for Applied Assessment and Longitudinal Research

The findings on the age and interval dependence of stability have important implications for cognitive ability testing practice. Testing is often done to inform treatment and intervention decisions or to provide guidance regarding educational or vocational decisions. These applications often presume that the relevant cognitive functions are either stable over the intervention window or period of time relevant to the educational or vocational decision, or that they would have otherwise been stable absent the intervention (Cronbach & Snow, 1977). There are no clear conventions regarding the minimum level of stability needed, but it has been suggested that a stability of .70 may be sufficient for group decisions, whereas a stability of at least .80 should be required for individual diagnostic decisions (Watkins & Smith, 2013). The best-fitting curves for describing the age and interval moderation of the stability of cognitive abilities determined in the present analyses make it possible to calculate for each age the maximum time interval for which these specific stabilities can still be expected. Using these curves, we may estimate the ages at which a certain rank-order stability of test scores of cognitive abilities may be achievable, along with the corresponding time interval beyond which a retest is warranted because sufficient stability can no longer be assumed. It should be noted, of course, that rank-order stability is not an individual-level metric. Even high rank-order stability does not preclude the possibility of large changes in intelligence over short intervals for specific individuals. Especially in circumstances of serious illness, neurological trauma, psychosocial stress, or dramatic changes in educational or social experiences, shorter retest intervals may be warranted than would be recommended based on the results of the present meta-analysis. Importantly, however, the decision to reassess individuals frequently, or over short retest intervals, must pay particular attention to validity threats associated with retest effects (Salthouse & Tucker-Drob, 2008).

We provide estimates for general intelligence stability thresholds based on the age and interval moderator analyses (see Supplemental Material for computational details). Figure 8 presents the maximum test-retest interval for which the stability criteria of .70 and .80 can still be satisfied, depending on the age of the tested person. When applying a strict criterion of at least r_{tt} = .80 as suitable for individual level decisions, this stability is not obtained in children younger than six years old. In eight-year-olds, this stability can be assumed for almost two years; in twelve-year-olds, it can already be assumed for approximately four years. At age 18 and beyond, a stability of .80 can generally be assumed for approximately six years, after which retesting would be recommended. When applying a more liberal criterion of r_{tt} =.70 that may be suitable for group-level decisions, this stability can already be obtained starting at age four. The maximum interval for this stability increases rapidly

with age, reaching five years at age six, twelve years at age nine, and 18 years at age eleven. After the age of 14, $r_{tt} > .70$ stability can be assumed for the full life span.

The resulting curves divide the age-interval space into three zones. In the leftmost zone, beyond the .70 line, adequate stability cannot be expected. This zone mainly concerns children under four years of age as well as children and adolescents between 4 and 18 years of age in the case of excessively long intervals. In the middle zone between the two lines, only moderate stability can be assumed. This zone concerns children under seven years of age, children and adolescents with medium length to long intervals (increasing strongly with age), and adults with time intervals longer than six years. Finally, in the bottom right zone, high stability can be assumed. It should be noted that the .70 and .80 values delimiting these zones are chosen somewhat arbitrarily, they present minimum values, and other values could be selected based on the specific question or application context.

The interval estimations may also be used in the planning of longitudinal studies. Cognitive abilities are often used as control variables or predictors of performance, achievement, or motivation. The present results can be used to determine when or how often cognitive abilities are to be measured to obtain results that are still relevant at the time of testing other variables. For example, school achievement at age 16 may be predicted by general intelligence. The results suggest that general intelligence measures as early as age eleven can still be regarded as very relevant due to a stability above .80. Measurements between the ages of eight and ten are still somewhat relevant (i.e., stability >.70), whereas measurements from age seven or below are not stable above .70 until age 16.

A second and related implication concerns the differences in stability between general intelligence and specific ability scores. Previous studies have sometimes found a large discrepancy between a very stable general intelligence score and much less stable specific ability scores, calling into question the diagnostic utility of the latter (e.g., Ryan et al., 2010; Watkins & Smith, 2013). Our results provide only partial support for this notion. Specifically, we found that knowledge-based abilities were comparably stable to general intelligence, suggesting that their relative diagnostic utility is not particularly limited by their stability. Conversely, scores relating to effortful-processing abilities may sometimes be insufficiently stable to support diagnostic decisions with long-term consequences, especially in young children, where stability is generally lower. We observed particularly low stabilities in Learning Efficiency and Working Memory, the latter even when adjusting for test reliability. Thus, special caution seems warranted if these scores are to be used as the basis of interventions or counseling.

A third implication of the present results is based on the finding that multidimensional cognitive ability tests such as the Wechsler and Stanford-Binet tests have a higher stability than unidimensional tests such as CFT or Raven tests. Thus, when making decisions with long-term consequences, multidimensional test batteries may be preferred. Of course, unidimensional tests have advantages, such as shorter test times, specificity of information about individual domains of functioning, and the capability to assess many abilities nonverbally. In practice, these advantages must be weighed against the disadvantage of lower stability. It should be noted that the lower stability of effortful processing-based

abilities compared to knowledge-based abilities may be at least partly responsible for the lower stability of unidimensional tests as the unidimensional tests in this meta-analysis were classified as Fluid Reasoning tests. More frequent retesting may generally be necessary for effortful processing-based abilities. Of course, retest effects need to be taken into account when planning repeated testing, especially when using the same test instrument (Hausknecht et al., 2007; Salthouse & Tucker-Drob, 2008; Scharfen et al., 2018). There is some evidence that when using the same test instrument for test and retest, interval durations of at least two years are required to reduce the retest effects to the level of retest effects with different test instruments (Hausknecht et al., 2007), but retest effects have also been detected in longitudinal studies after periods of 7 years or more (Horn & Donaldson, 1976; Salthouse et al., 2004; Thorvaldsson et al., 2006).

Remaining Gaps in our Knowledge about the Stability of Cognitive Abilities

In the present meta-analysis, we were able to draw on a very broad evidence base, covering a wide age range, very short to extremely long test-retest intervals, and many different cognitive abilities and tests of cognitive ability. The large number of studies and their heterogeneity along these dimensions allow for a deeper understanding of the stability of cognitive ability. Still, the evidence base is also limited in several ways.

First, effect sizes are not evenly distributed across age groups. In general, there are fewer studies with adult samples than with children and adolescents, especially between the ages of 30 and 50 and over 70. More data from older adults in particular would be helpful to improve our understanding of cognitive aging. The current results suggest that stability remains very high even in the oldest adults, but more research is needed to confirm this finding and to investigate possible moderators. Importantly, we did not include studies from clinical populations, and longitudinal studies of nonclinical populations (of the sort included in the current meta-analysis) typically exclude individuals with conditions that may interfere with cognitive function (e.g., those with dementia). It may be the case that stability of cognitive function decreases as the prevalence of such conditions increases with advancing old age. Longitudinal research taking an inclusive approach to following individuals over time, regardless of incidence of disorders of aging, would be needed to empirically test this hypothesis.

Second, studies have predominantly been conducted with WEIRD samples. WEIRD samples are in many ways unrepresentative of humanity as a whole (Henrich et al., 2010; Nielsen et al., 2017). The search was limited to English reports, was conducted in English, and conducted in predominantly English databases. This reduced the likelihood of finding and including appropriate studies from non-WEIRD samples, contributing to the somewhat biased study selection. More data from other cultural backgrounds are needed to understand the impact of cultural, economic, and educational factors on the stability of cognitive abilities. There is some tentative evidence for the generalizability of factor analytic models of intelligence across cultures (Wilson et al., 2023), but it is unclear to what extent other properties observed in WEIRD samples, such as high stability, are universal. In the present meta-analysis, there were insufficient data to conduct moderator analyses comparing African or South American samples with European or North American samples and there were

insufficient Asian samples to compare with European or North American samples on specific cognitive abilities. Although some non-WEIRD samples (k = 8; 4%) are included in the meta-analysis, the generalizability of the results is therefore somewhat limited.

Third, the reliability of the tests was not adequately documented in the included studies. In general, it is good scientific practice to report the psychometric properties of the instruments used. Yet, our study reveals that this is not routinely done for cognitive ability tests. When published, standardized tests are used true to manual reliability estimates can be derived from these manuals, but the manuals are usually not freely available. Moreover, the reliability of the samples studied may differ from that of the norming sample due to characteristics of the sample itself (e.g., sample homogeneity). Even more problematic are studies that use only a subset of subtests and combine them into short scales that are not described in the manual, or even combine self-developed tests and subtests of established tests. In these cases, there is little evidence on the reliability of the final test score. This unavailability of reliability estimates limits the interpretability of low stability estimates, as they may be due to actual rank-order changes in cognitive ability, low test reliability, or a combination of both.

Strengths and Limitations

The present meta-analysis has several strengths. It investigated a very broad set of studies published in the 100 years from 1921 to 2020, with mean baseline ages ranging from one to 88 years, test-retest intervals ranging from one day to 79 years, and data collected in 29 countries and five continents. These data were the result of an extensive literature search based on two bibliographic databases and preexisting reviews on the subject matter. The statistical analyses represented the current state of the art and included wide-ranging moderator analyses involving theoretically relevant moderator variables. The analyses also included a detailed examination of the stability of the broad abilities in the CHC model and adjusting for test reliability in a subset of effect sizes.

The meta-analysis also has several limitations that need to be considered when interpreting the results. Where possible, specific measures were taken to minimize the consequences of these limitations.

The first limitation is that our age moderator analyses were based on the convergence assumption (Bell, 1953), assuming that differences in age between samples are informative about changes within the population. No one study investigated the stability of cognitive abilities from early childhood to late adulthood. Instead, we relied on the combination of information derived from across-sample age differences and from within-sample across-occasion stability to investigate age trends. This approach potentially confounds age differences, cohort differences, and methodological differences between studies, but this is only a major concern if these confounds are systematically associated with age (Tucker-Drob & Briley, 2014). Additionally, pertaining to the age moderator analyses, the utilized functional forms only represent a subset of possible mathematical functions that could be applied to approximate the empirical result patterns. Other functions, such as cubic or logarithmic functions, could have been applied additionally. Nevertheless, the utilized functions already covered all theoretically plausible trajectories, and the connected linear

spline was included as a flexible function that could approximate result patterns that strongly deviated from the other functions.

The second limitation is that any minimum requirement used to accept or reject measures of cognitive abilities is necessarily somewhat arbitrary. Convincing arguments could be made for allowing the use of individual subtests or, conversely, for accepting only the use of complete tests. Our criteria were selected to allow the inclusion of a large and heterogeneous study base for different cognitive abilities while still maintaining a reasonable quality of the cognitive ability measurements. These criteria were preregistered. Moreover, we included a moderator analysis in which we compared the use of complete test batteries with the use of some form of subtest selection, thereby also applying a more conservative criterion.

The third limitation is that we did not investigate socioeconomic status (SES) as a moderator. SES could not be included because it was rarely reported, and the few available reports varied drastically in the operationalization of the construct (e.g., income, education level, ISEI). This heterogeneity made it impossible to integrate the results. Previous findings suggest that the influence of genetic and environmental factors on cognitive abilities varies strongly with SES (Tucker-Drob & Bates, 2016; Turkheimer et al., 2003) and that both genetic and environmental factors contribute to the stability of cognitive abilities (Tucker-Drob & Briley, 2014). It is therefore quite plausible to assume that SES has some influence on stability. Future studies may address this research question systematically.

The fouth limitation is that we did not include the studies that reported latent correlations in the primary meta-analysis. Latent correlations are not directly comparable to manifest correlations because they are controlled for measurement error and therefore generally larger. To avoid losing the information contributed by these studies, we conducted a smaller, separate meta-analysis of latent stability estimates. The meta-analysis based on latent stability estimates generally indicated very similar stability to that indicated in our primary meta-analysis.

The fifth limitation is that most studies that investigate cognitive development in older adults exclude participants close to and beyond the clinical range of functioning. As a consequence, older adults are likely positively selected and also somewhat restricted in the range of included cognitive trajectories. Depending on sample age this may constitute a substantial exclusion rate, with a prevalence of dementia of 17% in those aged 75-84 years and 32% in those aged 85 years or older in the United States (Hebert et al., 2013). This selection bias may limit the reordering in the older samples, as the individuals that contribute most to reordering are excluded. In turn, this may result in an overestimation of rank-order stability in this age range.

The sixth limitation is that our results and their implications may, in many respects, be restricted to the nonclinical population. We do not, for example, recommend the use of cognitive testing for long-term predictions in young children. However, many neurodevelopmental disorders are characterized, among other things, by specific cognitive deficits (Thapar et al., 2015) that can be detected early on and that remain throughout cognitive development. In these cases, cognitive testing may serve as a robust and valid

aspect of diagnostic screening in early childhood, despite the low stability of cognitive abilities in the general population in this age range. These and other special cases of deviating stability in clinical subpopulations likely did not affect the results of this meta-analysis because of their rarity and because clinical samples were excluded from the analyses.

Conclusion

In summary, this meta-analysis of longitudinal studies showed that cognitive abilities exhibit high rank order stability, reaching its peak around age 20 and remaining at this high level throughout adulthood and old age. Mean stability estimate for age 20 and a test-retest-interval of five years were $\rho = .77$ for the observed correlations and $\rho = .86$ for the disattenuated correlations. Stability is much lower in young children. Before age 4, stability never exceeds .70, whereas in late adolescence, the stability no longer drops below this value for any test-retest interval. The low stability in young children cannot be explained by lower test reliability in this age group; overall, the correlation between test reliability and mean sample age is low. General intelligence and knowledge-based abilities were found to be somewhat more stable than abilities based on effortful processing ($\rho = .77$-$.80$ vs. $\rho = .65$-$.75$ for an age of 20 and an interval of 5 years). Reliability differences between tests of knowledge-based and effortful processing-based abilities in part account for the tests' differences in stability. Multidimensional intelligence tests may generally be preferred over unidimensional tests when the goal is a high stability of the test result, especially when there is no specific need for assessing individual ability domains. Our findings indicate that the use of cognitive testing in diagnostic decision making in younger children in particular may require repetition over relatively short intervals of time, whereas retest intervals can be longer in adults. We have provided information on age-dependent maximum intervals (in years) for which a stability of .70 and .80 can be expected for cognitive ability measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Moritz Breit and Vsevolod Scherrer share first-authorship of this article. The research materials can be accessed via https://osf.io/ajufs/. E.M.T.D. was supported by research grants RF1AG073593, R01MH120219, and R01AG054628 from the National Institutes of Health (NIH). Additionally, E.M.T.D. is a member of the University of Texas Center on Aging and Population Sciences and the University of Texas Population Research Center, which are supported by NIH grants P30AG066614 and P2CHD042849, respectively.

References

Note: References marked with an asterisk were a part of the meta-analysis.

*Adkins DC (1937). The efficiency of certain intelligence tests in predicting scholarship scores. Journal of Educational Psychology, 28(2), 129–134. 10.1037/h0058951

*Allan ME, & Young FM (1943). The constancy of the intelligence quotient as indicated by retests of 130 children. Journal of Applied Psychology, 27(1), 41–60. 10.1037/h0054264

*Allen MM (1945). Relationship between the indices of intelligence derived from the Kuhlmann-Anderson intelligence tests for grade I and the same tests for grade IV. Journal of Educational Psychology, 36(4), 252–256. 10.1037/h0058184

*Almas AN, Degnan KA, Nelson CA, Zeanah CH, & Fox NA (2016). Iq at age 12 following a history of institutional care: Findings from the Bucharest Early Intervention Project. Developmental Psychology, 52(11), 1858–1866. 10.1037/dev0000167 [PubMed: 27709994]

*Arthur W Jr, Glaze RM, Villado AJ, & Taylor JE (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. International Journal of Selection and Assessment, 18(1), 1–16. 10.1111/j.1468-2389.2010.00476.x

*Aschwanden D, Martin M, & Allemand M (2017). Cognitive abilities and personality traits in old age across four years: More stability than change. Journal of Research in Personality, 70, 202–213. 10.1016/j.jrp.2017.08.002

Baltes PB, Staudinger UM, & Lindenberger U (1999). Lifespan psychology: Theory and application to intellectual functioning. Annual Review of Psychology, 50, 471–507. 10.1146/annurev.psych.50.1.471

*Bartels M, Rietveld MJH, van Baal GC, & Boomsma DI (2002). Genetic and environmental influences on the development of intelligence. Behavior Genetics, 32(4), 237–249. 10.1023/A:1019772628912 [PubMed: 12211623]

*Basso MR, Carona FD, Lowery N, & Axelrod BN (2002). Practice effects on the WAIS-III across 3-and 6-month intervals. The Clinical Neuropsychologist, 16(1), 57–63. 10.1076/clin.16.1.57.8329 [PubMed: 11992227]

*Baudson TG, & Preckel F (2013). Development and validation of the German Test for (Highly) Intelligent Kids-T(H)INK. European Journal of Psychological Assessment, 29, 171–181. 10.1027/1015-5759/a000142

*Bauer JJ, & Smith DK (1988). Stability of the K-ABC and SB: 4 with Preschool Children.

*Beaver KM, Schwartz JA, Connolly EJ, Nedelec JL, Al-Ghamdi MS, & Kobeisy AN (2013). The genetic and environmental architecture to the stability of IQ: Results from two independent samples of kinship pairs. Intelligence, 41(5), 428–438. 10.1016/j.intell.2013.06.012

Bayley N (1949). Consistency and variability in the growth of intelligence from birth to 18 years. The Journal of Genetic Psychology, 75(2), 165–196. 10.1080/08856559.1949.10533516 [PubMed: 15403674]

Bell RQ (1953). Convergence: An Accelerated Longitudinal Approach. Child Development, 24(2), 145. 10.2307/1126345 [PubMed: 13141335]

*Bergold S, & Steinmayr R (2016). The relation over time between achievement motivation and intelligence in young elementary school children: A latent cross-lagged analysis. Contemporary Educational Psychology, 46, 228–240. 10.1016/j.cedpsych.2016.06.005

*Bishop E, Cherny SS, Corley R, Plomin R, DeFries JC, & Hewitt JK (2003). Development genetic analysis of general cognitive ability from 1 to 12 years in a sample of adoptees, biological siblings, and twins. Intelligence, 31(1), 31–49. 10.1016/S0160-2896(02)00112-5

Bleidorn W, Schwaba T, Zheng A, Hopwood CJ, Sosa S, Roberts B, & Briley DA (2022). Personality Stability and Change: A Meta-Analysis of Longitudinal Studies. 10.31234/osf.io/eq5d6

*Bonney ME (1943). The Relative Stability of Social, Intellectual, and Academic Status in Grades II to IV, and the Inter-relationships between these Various Forms of Growth. Journal of Educational Psychology, 34(2), 88–102. 10.1037/h0056937

Borenstein M, Hedges LV, Higgins JP, & Rothstein HR (2009). Introduction to meta-analysis. Wiley. http://www.Meta-Analysis.com

*Bradshaw DH (1964). Stability of California test of mental maturity IQ's from the second to the fourth grade. Educational and Psychological Measurement, 24(4), 935–939. 10.1177/001316446402400421

*Bradway KP, & Thompson CW (1962). Intelligence at adulthood: A twenty-five year follow-up. Journal of Educational Psychology, 53(1), 1–14. 10.1037/h0045764

*Breeman LD, Jaekel J, Baumann N, Bartmann P, & Wolke D (2015). Preterm Cognitive Function Into Adulthood. Pediatrics, 136(3), 415–423. 10.1542/peds.2015-0608 [PubMed: 26260714]

Breit M, Brunner M, Molenaar D, & Preckel F (2022). Differentiation hypotheses of intelligence: A systematic review of the empirical evidence and an agenda for future research. Psychological Bulletin. Advance online publication, 10.1037/bul0000379

Breit M, Brunner M, & Preckel F (2021). Age and ability differentiation in children: A review and empirical investigation. Developmental Psychology, 57(3), 325–346. 10.1037/dev0001147 [PubMed: 33539120]

Breit M, Scherrer V, & Preckel F (2021). Temporal stability of specific ability scores and intelligence profiles in high ability students. Intelligence, 86, 101538. 10.1016/j.intell.2021.101538

Breit M, Scherrer V, & Preckel F (2022). Temporal stability and change in manifest intelligence scores: Four complementary analytic approaches. MethodsX, 9, 101613. 10.1016/j.mex.2021.101613 [PubMed: 35004234]

Breit M, Scherrer V, Tucker-Drob EM, & Preckel F (2023, November 17). The Stability of Intelligence: A Meta-Analysis. Retrieved from osf.io/ajufs

Briley DA, & Tucker-Drob EM (2017). Comparing the Developmental Genetics of Cognition and Personality over the Life Span. Journal of Personality, 85(1), 51–64. 10.1111/jopy.12186 [PubMed: 26045299]

*Brouwer RM, van Soelen ILC, Swagerman SC, Schnack HG, Ehli EA, Kahn RS, Hulshoff Pol HE, & Boomsma DI (2014). Genetic associations between intelligence and cortical thickness emerge at the start of puberty. Human Brain Mapping, 35(8), 3760–3773. 10.1002/hbm.22435 [PubMed: 24382822]

*Bryant CK, & Roffe MW (1978). A reliability study of the McCarthy Scales of Children's Abilities. Journal of Clinical Psychology, 34(2), 401–406. 10.1002/1097-4679(197804)34:2<401::AID-JCLP2270340230>3.0.CO;2-V

*Bryant P, Maclean M, & Bradley L (1990). Rhyme, language, and children's reading. Applied Psycholinguistics, 11(3), 237–252. 10.1017/S0142716400008870

*Bub KL, Buckhalt JA, & El-Sheikh M (2011). Children's sleep and cognitive performance: a cross-domain analysis of change over time. Developmental Psychology, 47(6), 1504–1514. 10.1037/a0025535 [PubMed: 21942668]

Caemmerer JM, Keith TZ, & Reynolds MR (2020). Beyond individual intelligence tests: Application of Cattell-Horn-Carroll Theory. Intelligence, 79, 101433. 10.1016/j.intell.2020.101433

*Capwell DF (1945). Personality patterns of adolescent girls: I Girls who show improvement in IQ. Journal of Applied Psychology, 29(3), 212–228. 10.1037/h0062853

Carroll JB (2009). Human Cognitive Abilities. Cambridge University Press. 10.1017/cbo9780511571312

*Cardon LR, & Fulker DW (1994). A model of developmental change in hierarchical phenotypes with application to specific cognitive abilities. Behavior Genetics, 24(1), 1–16. 10.1007/BF01067924 [PubMed: 8192616]

*Carioti D, Danelli L, Guasti MT, Gallucci M, Perugini M, Steca P, Stucchi NA, Maffezzoli A, Majno M, Berlingeri M, & Paulesu E (2019). Music Education at School: Too Little and Too Late? Evidence From a Longitudinal Study on Music Training in Preadolescents. Frontiers in Psychology, 10, 2704. 10.3389/fpsyg.2019.02704 [PubMed: 31920782]

*Catron DW, & Thompson CC (1979). Test-retest gains in WAIS scores after four retest intervals. Journal of Clinical Psychology, 35(2), 352–357. 10.1002/1097-4679(197904)35:2<352::AID-JCLP2270350226>3.0.CO;2-2 [PubMed: 457898]

Cattell RB (1986). Intelligence: Its structure, growth and action (Vol. 35). North-Holland.

*Cherny SS, Fulker DW, Emde RN, Robinson J, Corley RP, Reznick JS, Plomin R, & DeFries JC (1994). A developmental-genetic analysis of continuity and change in the Bayley Mental Development Index from 14 to 24 months: The MacArthur Longitudinal Twin Study. Psychological Science, 5(6), 354–360. 10.1111/j.1467-9280.1994.tb00285.x

*Colom R, Quiroga MÁ, Solana AB, Burgaleta M, Román FJ, Privado J, Escorial S, Martínez K, Álvarez-Linera J, Alfayate E, García F, Lepage C, Hernández-Tamames JA, & Karama S (2012). Structural changes after videogame practice related to a brain network associated with intelligence. Intelligence, 40(5), 479–489. 10.1016/j.intell.2012.05.004

*Colom R, Román FJ, Abad FJ, Shih PC, Privado J, Froufe M, Escorial S, Martínez K, Burgaleta M, Quiroga MA, Karama S, Haier RJ, Thompson PM, & Jaeggi SM (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. Intelligence, 41(5), 712–727. 10.1016/j.intell.2013.09.002

Conley JJ (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. Personality and Individual Differences, 5(1), 11–25. 10.1016/0191-8869(84)90133-8

*Cowan R, Hurry J, & Midouhas E (2018). The relationship between learning mathematics and general cognitive ability in primary school. The British Journal of Developmental Psychology, 36(2), 277–284. 10.1111/bjdp.12200 [PubMed: 28801949]

*Crano WD, Kenny DA, & Campbell DT (1972). Does intelligence cause achievement? A cross-lagged panel analysis. Journal of Educational Psychology, 63(3), 258–275. 10.1037/h0032639

*Croake JW, Keller JF, & Catlin N (1973). WPPSI, Rutgers, Goddenough, Goodenough-Harris IQ's for lower socioeconomic, black, preschool children. Psychology, 10(2), 58–69.

*Crockett BK, Rardin MW, & Pasewark RA (1975). Relationship between WPPSI and Stanford-Binet IQs and subsequent WISC IQs in Headstart children. Journal of Consulting and Clinical Psychology, 43(6), 922. 10.1037/0022-006x.43.6.922

Cronbach LJ, & Snow RE (1977). Aptitudes and instructional methods: A handbook for research on interactions. Irvington.

*Dauphinais SM, & Bradley RW (1979). IQ change and occupational level: A longitudinal study with Third Harvard Growth Study participants. Journal of Vocational Behavior, 75(3), 367–375. 10.1016/0001-8791(79)90030-7

*Deary IJ (1995). Auditory inspection time and intelligence: What is the direction of causation? Developmental Psychology, 31(2), 237–250. 10.1037/0012-1649.31.2.237

Deary IJ (2014). The Stability of Intelligence From Childhood to Old Age. Current Directions in Psychological Science, 23(4), 239–245. 10.1177/0963721414536905

*Deary IJ, Allerhand M, & Der G (2009). Smarter in middle age, faster in old age: A cross-lagged panel analysis of reaction time and cognitive ability over 13 years in the West of Scotland Twenty-07 Study. Psychology and Aging, 24(1), 40–47. 10.1037/a0014442 [PubMed: 19290736]

*Deary IJ, Batty GD, Pattie A, & Gale CR (2008). More intelligent, more dependable children live longer: A 55-year longitudinal study of a representative sample of the Scottish nation. Psychological Science, 19(9), 874–880. 10.1111/j.1467-9280.2008.02171.x [PubMed: 18947352]

*Deary IJ, & Brett CE (2015). Predicting and retrodicting intelligence between childhood and old age in the 6-Day Sample of the Scottish Mental Survey 1947. Intelligence, 50, 1–9. 10.1016/j.intell.2015.02.002 [PubMed: 26207078]

*Deary IJ, Pattie A, & Starr JM (2013). The stability of intelligence from age 11 to age 90 years: the Lothian birth cohort of 1921. Psychological Science, 24(12), 2361–2368. 10.1177/0956797613486487 [PubMed: 24084038]

*Deary IJ, Whalley LJ, & Crawford JR (2004). An 'instantaneous' estimate of a lifetime's cognitive change. Intelligence, 32(2), 113–119. 10.1016/j.intell.2003.06.001

*Deary IJ, Whalley LJ, Lemmon H, Crawford J, & Starr JM (2000). The Stability of Individual Differences in Mental Ability from Childhood to Old Age: Follow-up of the 1932 Scottish Mental Survey. Intelligence, 25(1), 49–55. 10.1016/SO160-2896(99)00031-8

*Deary IJ, Whalley LJ, & Starr JM (2009). Recruiting the Aberdeen Birth Cohorts 1921 and 1936 and the Lothian Birth Cohort 1921 and assessing the stability of intelligence across the life span. In Deary IJ, Whalley LJ, & Starr JM (Eds.), A lifetime of intelligence: Follow-up studies of the Scottish mental surveys of 1932 and 1947 (pp. 101–114). American Psychological Association. 10.1037/11857-006

*Deary IJ, Whiteman MC, Starr JM, Whalley LJ, & Fox HC (2004). The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. Journal of Personality and Social Psychology, 56(1), 130–147. https://psycnet.apa.org/doi/10.1037/0022-3514.86.1.130

*Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, Luciano M, Lopez LM, Gow AJ, & Corley J (2012). Genetic contributions to stability and change in intelligence from childhood to old age. Nature, 482(7384), 212–215. 10.1038/nature10781 [PubMed: 22258510]

Deeks JJ, Higgins JPT, & Altman DG (2008). Analysing data and undertaking meta-analyses. In Higgins J & Green S (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.0 (pp. 243–296). John Wiley & Sons.

*Demetriou A, Spanoudis G, Shayer M, Mouyi A, Kazi S, & Platsidou M (2013). Cycles in speed-working memory-G relations: Towards a developmental–differential theory of the mind. Intelligence, 41(1), 34–50. 10.1016/j.intell.2012.10.010

*Denno D, Meijs B, Nachshon I, & Aurand S (1982). Early cognitive functioning: Sex and race differences. International Journal of Neuroscience, 16(3-4), 159–172. 10.3109/00207458209147143 [PubMed: 7169281]

*D'Souza S, Backhouse-Smith A, Thompson JMD, Slykerman R, Marlow G, Wall C, Murphy R, Ferguson LR, Mitchell EA, & Waldie KE (2016). Associations Between the KIAA0319 Dyslexia Susceptibility Gene Variants, Antenatal Maternal Stress, and Reading Ability in a Longitudinal Birth Cohort. Dyslexia, 22(4), 379–393. 10.1002/dys.1534 [PubMed: 27465261]

*Dudek SZ, Lester EP, & Goldberg JS (1969). Relationship of Piaget measures to standard intelligence and motor scales. Perceptual and Motor Skills, 28(2), 351–362. 10.2466/pms.1969.28.2.351 [PubMed: 5803452]

*Dunkel CS, & Woodley Of Menie MA (2019). Maternal sensitivity and performance and verbal intelligence in late childhood and adolescence. Journal of Biosocial Science, 57(1), 48–58. 10.1017/S0021932017000669

Duval S, & Tweedie R (2000). A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis. Journal of the American Statistical Association, 95(449), 89–98. 10.1080/01621459.2000.10473905

Egger M, Davey Smith G, Schneider M, & Minder C (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ (Clinical Research Ed.), 315(7109), 629–634. 10.1136/bmj.315.7109.629

*Ellzey JT, & Karnes FA (1990). Test-retest stability of WISC-R IQs among young gifted students. Psychological Reports, 66(3), 1023–1026. 10.2466/pr0.1990.66.3.1023 [PubMed: 2377683]

*Estrada E, Ferrer E, Abad FJ, Román FJ, & Colom R (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. Intelligence, 50, 93–99. 10.1016/j.intell.2015.02.004

Flanagan DP, Alfonso VC, & Ortiz SO (2012). The cross-battery assessment approach An overview, historical perspective, and current directions. In Flanagan DP & Harrison PL (Eds.), Contemporary intellectual assessment: Theories, tests, and issues. (pp. 459–483). The Guilford Press.

*Franz CE, O'Brien RC, Hauger RL, Mendoza SP, Panizzon MS, Prom-Wormley E, Eaves LJ, Jacobson K, Lyons MJ, & Lupien S (2011). Cross-sectional and 35-year longitudinal assessment of salivary cortisol and cognitive functioning: the Vietnam Era twin study of aging. Psychoneuroendocrinology, 36(7), 1040–1052. 10.1016/j.psyneuen.2011.01.002 [PubMed: 21295410]

Frick KM, & Benoit JD (2010). Use it or lose it: Environmental enrichment as a means to promote successful cognitive aging. TheScientificWorldJournal, 10, 1129–1141. 10.1100/tsw.2010.111

*Frischkorn GT, Greiff S, & Wüstenberg S (2014). The development of complex problem solving in adolescence: A latent growth curve analysis. Journal of Educational Psychology, 106(4), 1007–1020. 10.1037/a0037114

Fryer JW, & Elliot AJ (2007). Stability and change in achievement goals. Journal of Educational Psychology, 99(4), 700–714. 10.1037/0022-0663.99.4.700

*Gathercole SE, Willis CS, Emslie H, & Baddeley AD (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. Developmental Psychology, 28(5), 887–898.

Geiser C, Götz T, Preckel F, & Freund PA (2017). States and Traits. European Journal of Psychological Assessment, 33(4), 219–223. 10.1027/1015-5759/a000413

*Ghisletta P, & Lindenberger U (2003). Age-based structural dynamics between perceptual speed and knowledge in the Berlin Aging Study: Direct evidence for ability dedifferentiation in old age. Psychology and Aging, 18(4), 696–713. 10.1037/0882-7974.18.4.696 [PubMed: 14692858]

*Ghisletta P, Mason F, Dahle CL, & Raz N (2019). Metabolic risk affects fluid intelligence changes in healthy adults. Psychology and Aging, 34(7), 912–920. 10.1037/pag0000402 [PubMed: 31589057]

*Ghisletta P, Rabbitt P, Lunn M, & Lindenberger U (2012). Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared. Intelligence, 40(3), 260–268. 10.1016/j.intell.2012.02.008

*Giangrande EJ, Beam CR, Carroll S, Matthews LJ, Davis DW, Finkel D, & Turkheimer E (2019). Multivariate analysis of the scarr-rowe interaction across middle childhood and early adolescence. Intelligence, 77, 101400. 10.1016/j.intell.2019.101400

*Gjerde PF, Block J, & Block JH (1985). Longitudinal consistency of Matching Familiar Figures Test performance from early childhood to preadolescence. Developmental Psychology, 21(2), 262–271. 10.1037/0012-1649.21.2.262

*Gold DP, Andres D, Etezadi J, Arbuckle T, Schwartzman A, & Chaikelson J (1995). Structural equation model of intellectual change and continuity and predictors of intelligence in older men. Psychology and Aging, 10(2), 294–303. 10.1037/0882-7974.10.2.294 [PubMed: 7662188]

*Gorbach T, Pudas S, Lundquist A, Orädd G, Josefsson M, Salami A, Luna X. de, & Nyberg L (2017). Longitudinal association between hippocampus atrophy and episodic-memory decline. Neurobiology of Aging, 51, 167–176. 10.1016/j.neurobiolaging.2016.12.002 [PubMed: 28089351]

Gottfredson L (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence, 24(1), 13–23. 10.1016/S0160-2896(97)90011-8

Gottfredson L, & Saklofske DH (2009). Intelligence: Foundations and issues in assessment. Canadian Psychology/Psychologie Canadienne, 50(3), 183–195. 10.1037/a0016641

*Gottfried AE, Gottfried AW, Reichard RJ, Guerin DW, Oliver PH, & Riggio RE (2011). Motivational roots of leadership: A longitudinal study from childhood through adulthood. The Leadership Quarterly, 22(3), 510–519. 10.1016/j.leaqua.2011.04.008

*Gow AJ, Johnson W [Wendy], Pattie A, Brett CE, Roberts B, Starr JM, & Deary IJ. (2011). Stability and change in intelligence from age 11 to ages 70, 79, and 87: the Lothian Birth Cohorts of 1921 and 1936. Psychology and Aging, 26(1), 232–240. 10.1037/a0021072 [PubMed: 20973608]

*Gow AJ, Mortensen EL, & Avlund K (2012). Activity Participation and Cognitive Aging from Age 50 to 80 in the Glostrup 1914 Cohort. Journal of the American Geriatrics Society, (60(10), 1831–1838. 10.1111/j.1532-5415.2012.04168.x [PubMed: 23035883]

*Gow AJ, Whiteman MC, Pattie A, & Deary IJ (2005). The personality–intelligence interface: Insights from an ageing cohort. Personality and Individual Differences, 39(4), 751–761. 10.1016/j.paid.2005.01.028

*Green CT, Bunge SA, Briones Chiongbian V, Barrow M, & Ferrer E (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. Journal of Experimental Child Psychology, 157, 125–143. 10.1016/j.jecp.2016.12.005 [PubMed: 28152390]

*Green Bartoi M, Issner JB, Hetterscheidt L, January AM, Kuentzel JG, & Barnett D (2015). Attention problems and stability of WISC-IV scores among clinically referred children. Applied Neuropsychology: Child, 4(3), 133–140. 10.1080/21622965.2013.811075 [PubMed: 25074427]

*Gregory T, Nettelbeck T, Howard S, & Wilson C (2009). A test of the cascade model in the elderly. Personality and Individual Differences, 46(1), 71–73. 10.1016/j.paid.2008.08.017

*Grønkjær M, Osier M, Flensborg-Madsen T, Sørensen HJ, & Mortensen EL (2019). Associations between education and age-related cognitive changes from early adulthood to late midlife. Psychology and Aging, 34(2), 177–186. 10.1037/pag0000332 [PubMed: 30829528]

*Grover DR, & Hertzog C (1991). Relationships between intellectual control beliefs and psychometric intelligence in adulthood. Journal of Gerontology, 46(3), 109–15. 10.1093/geronj/46.3.p109

Harden KR, Turkheimer E, & Loehlin JC (2007). Genotype by environment interaction in adolescents' cognitive aptitude. Behavior Genetics, 37(2), 273–283. 10.1007/s10519-006-9113-4 [PubMed: 16977503]

*Hart SA, Petrill SA, Deckard KD, & Thompson LA (2007). Ses and CHAOS as environmental mediators of cognitive ability: A longitudinal genetic analysis. Intelligence, 35(3), 233–242. 10.1016/j.intell.2006.08.004 [PubMed: 19319205]

Hausknecht JP, Halpert JA, Di Paolo NT, & Moriarty Gerrard MO (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. The Journal of Applied Psychology, 92(2), 373–385. 10.1037/0021-9010.92.2.373 [PubMed: 17371085]

*Haworth CMA, Harlaar N, Kovas Y, Davis OSP, Oliver BR, Hayiou-Thomas ME, Frances J, Busfield P, McMillan A, & Dale PS (2007). Internet cognitive testing of large samples needed in genetic research. Twin Research and Human Genetics, 10(4), 554–563. 10.1375/twin.10.4.554 [PubMed: 17708696]

Hebert LE, Weuve J, Scherr PA, & Evans DA. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology, 80(19), 1778–1783. 10.1212/WNL.0b013e31828726f5 [PubMed: 23390181]

Hedges LV, & Olkin I (2014). Statistical Methods for Meta-Analysis. Academic Press.

Hedges LV, Tipton E, & Johnson MC (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1(1), 39–65. 10.1002/jrsm.5 [PubMed: 26056092]

*Hegelund ER, Grønkjær M, Osler M, Dammeyer J, Flensborg-Madsen T, & Mortensen EL (2020). The influence of educational attainment on intelligence. Intelligence, 78, 101419. 10.1016/j.intell.2019.101419

*Heim AW, & Wallace JG (1949). The effects of repeatedly retesting the same group on the same intelligence test part 1: normal adults. The Quarterly Journal of Experimental Psychology, 1(4), 151–159. 10.1080/17470214908416760 [PubMed: 15392595]

*Helder EJ, Mulder E, & Gunnoe ML (2016). A longitudinal investigation of children internationally adopted at school age. Child Neuropsychology, 22(1), 39–64. 10.1080/09297049.2014.967669

*Henmon VAC, & Burns HM (1923). The Constancy of Intelligence Quotients with Borderline and Problem Cases. Journal of Educational Psychology, 14(4), 247–250. 10.1037/h0070797

Henrich J, Heine SJ, & Norenzayan A (2010). Most people are not WEIRD. Nature, 466(7302), 29. 10.1038/466029a [PubMed: 20595995]

*Hertzig ME, & Birch HG (1971). Longitudinal course of measured intelligence in preschool children of different social and ethnic backgrounds. American Journal of Orthopsychiatry, 41(3), 416–426. 10.1111/j.1939-0025.1971.tb01128.x [PubMed: 5549913]

*Hertzog C, Dixon RA, Hultsch DF, & MacDonald SWS (2003). Latent Change Models of Adult Cognition: Are Changes in Processing Speed and Working Memory Associated With Changes in Episodic Memory? Psychology and Aging, 18(4), 755–769. 10.1037/0882-7974.18.4.755 [PubMed: 14692862]

Hertzog C, & Schaie KW (1988). Stability and change in adult intelligence: 2. Simultaneous analysis of longitudinal means and covariance structures. Psychology and Aging, 3(2), 122–130. 10.1037/0882-7974.3.2.122 [PubMed: 3268250]

*Heymans P, & van Lieshout CFM (2013). Developing Talent Across the Lifespan, 67–102. Psychology Press, 10.4324/9780203775998

*Hindley CB, & Owen CF (1978). The extent of individual changes in I.Q. For ages between 6 months and 17 years, in a British longitudinal sample. Journal of Child Psychology and Psychiatry, 19(4), 329–350. 10.1111/j.1469-7610.1978.tb00480.x [PubMed: 711823]

*Hindley CB, & Owen CF (1979). An analysis of individual patterns of DQ and IQ curves from 6 months to 17 years. British Journal of Psychology, 70(2), 273–293. 10.1111/j.2044-8295.1979.tb01685.x

*Hoekstra RA, Bartels M, & Boomsma DI (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. Learning and Individual Differences, 17(2), 97–114. 10.1016/j.lindif.2007.05.005

Hong MR, & Cheng Y (2019). Clarifying the Effect of Test Speededness. Applied Psychological Measurement, 43(8), 611–623. 10.1177/0146621618817783 [PubMed: 31551639]

*Hopkins KD, & Bibelheimer M (1971). Five-Year Stability of Intelligence Quotients from Language and Nonlanguage Group Tests. Child Development, 2, 645–649. 10.2307/1127499

*Hopp GA, Dixon RA, Grut M, & Bäckman L (1997). Longitudinal and psychometric profiles of two cognitive status tests in very old adults. Journal of Clinical Psychology, 53(7), 673–686. [PubMed: 9356897]

Horn JL, & Cattell RB (1967). Age differences in fluid and crystallized intelligence. Acta Psychologica, 26, 107–129. 10.1016/0001-6918(67)90011-X [PubMed: 6037305]

Horn JL, & Donaldson G (1976). On the myth of intellectual decline in adulthood. American Psychologist, 37(10), 701–719. 10.1037/0003-066X.31.10.701

*Hülür G, Gasimova F, Robitzsch A, & Wilhelm O (2018). Change in Fluid and Crystallized Intelligence and Student Achievement: The Role of Intellectual Engagement. Child Development, 89(4), 1074–1087. 10.1111/cdev.12791 [PubMed: 28369877]

*Hülür G, Siebert JS, & Wahl H-W (2020). The role of perceived work environment and work activities in midlife cognitive change. Developmental Psychology, 56(12), 2345–2357. 10.1037/dev0001112 [PubMed: 33001669]

Humphreys LG, Davey TC, & Park RK (1985). Longitudinal correlation analysis of standing height and intelligence. Child Development, 56(6), 1465–1478. 10.2307/1130466 [PubMed: 4075869]

Hunt E (2010). Human Intelligence. Cambridge University Press.

*Irwin DO (1966). Reliability of the wechsler intelligence scale for children. Journal of Educational Measurement, 3(4), 287–292.

*Ivnik RJ, Smith GE, Malec JF, Petersen RC, & Tangalos EG (1995). Long-term stability and intercorrelations of cognitive abilities in older persons. Psychological Assessment, 7(2), 155–161. 10.1037/1040-3590.7.2.155

*Jankowska AM, Bogdanowicz M, & Takagi A (2014). Stability of WISC-R scores in students with borderline intellectual functioning. Health Psychology Report, 2(1), 49–59. 10.5114/hpr.2014.42789

*Jenni OG, Chaouch A, Locatelli I, Thoeni I, Diezi M, Werner H, Caflisch J, & Rousson V (2011). Intra-individual stability of neuromotor tasks from 6 to 18 years: A longitudinal study. Human Movement Science, 30(6), 1272–1282. 10.1016/j.humov.2010.12.002 [PubMed: 21813200]

Jensen AR (1998). The G Factor: The Science of Mental Ability. Prager.

Jin J, & Rounds J (2012). Stability and change in work values: A meta-analysis of longitudinal studies. Journal of Vocational Behavior, 80(2), 326–339. 10.1016/j.jvb.2011.10.007

Johnson BT, & Hennessy EA (2019). Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. Social Science & Medicine (1982), 233, 237–251. 10.1016/j.socscimed.2019.05.035 [PubMed: 31233957]

*Kanazawa S (2013). Childhood intelligence and adult obesity. Obesity, 21(3), 434–440. 10.1002/oby.20018 [PubMed: 23404798]

Kane MJ, & Engle RW (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. Psychonomic Bulletin & Review, 9(4), 637–671. 10.3758/bf03196323 [PubMed: 12613671]

*Kangas J, & Bradway KP (1971). Intelligence at Middle Age: A Thirty-Eight-Year Follow-Up. Developmental Psychology, 5(2), 333–337. 10.1037/h0031471

*Keage HAD, Muniz G, Kurylowicz L, van Hooff M, Clark L, Searle AK, Sawyer MG, Baghurst P, & McFarlane A (2016). Age 7 intelligence and paternal education appear best predictors of educational attainment: The Port Pirie Cohort Study. Australian Journal of Psychology, 68(1), 61–69. 10.1111/ajpy.12083

*Kieng S, Rossier J, Favez N, & Lecerf T (2017). Long-term stability of the French WISC-IV: Standard and CHC index scores. European Review of Applied Psychology, 67(1), 51–60. 10.1016/j.erap.2016.10.001

*Klonoff H (1972). IQ constancy and age. Perceptual and Motor Skills, 35(2), 527–534. 10.2466/pms.1972.35.2.527 [PubMed: 5081282]

*Koenis MMG, Brouwer RM, Swagerman SC, van Soelen ILC, Boomsma DI, & Hulshoff Pol HE (2018). Association between structural brain network efficiency and intelligence increases during adolescence. Human Brain Mapping, 39(2), 822–836. 10.1002/hbm.23885 [PubMed: 29139172]

*Koenis MMG, Brouwer RM, van den Heuvel MP, Mandl RCW, van Soelen ILC, Kahn RS, Boomsma DI, & Hulshoff Pol HE (2015). Development of the brain's structural network efficiency in

early adolescence: a longitudinal DTI twin study. Human Brain Mapping, 36(12), 4938–4953. 10.1002/hbm.22988 [PubMed: 26368846]

*Kogan N, & Pankove E (1972). Creative Ability over a Five-Year Span. Child Development, 43(2), 427–442. 10.2307/1127546 [PubMed: 5034728]

*Kremen WS, Beck A, Elman JA, Gustavson DE, Reynolds CA, Tu XM, Sanderson-Cimino ME, Panizzon MS, Vuoksimaa E, Toomey R, Fennema-Notestine C, Hagler DJ, Fang B, Dale AM, Lyons MJ, & Franz CE (2019). Influence of young adult cognitive ability and additional education on later-life cognition. Proceedings of the National Academy of Sciences of the United States of America, 116(6), 2021–2026. 10.1073/pnas.1811537116 [PubMed: 30670647]

*Krohn EJ, & Lamp RE (1999). Stability of the SB:FE and K-ABC for Young Children From Low-Income Families: A 5-Year Longitudinal Study. Journal of School Psychology, 37(3), 315–332. 10.1016/S0022-4405(99)00013-8

*Lamp RE, & Krohn EJ (1990). Stability of the Stanford-Binet Fourth Edition and K-ABC for Young Black and White Children from low Income Families. Journal of Psychoeducational Assessment, 8(2), 139–149. 10.1177/073428299000800204

*Larsen L, Hartmann P, & Nyborg H (2008). The stability of general intelligence from early adulthood to middle-age. Intelligence, 36(1), 29–34. 10.1016/j.intell.2007.01.001

*Lassiter KS, & Matthews TD (1999). Test-retest reliability of the General Ability Measure for Adults. Perceptual and Motor Skills, 88(2), 531–534. 10.2466/pms.1999.88.2.531

*Leong RLF, Lo JC, Sim SKY, Zheng H, Tandi J, Zhou J, & Chee MWL (2017). Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. NeuroImage, 147, 852–860. 10.1016/j.neuroimage.2016.10.016 [PubMed: 27742600]

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. Psychological Science, 15(3), 155–163. 10.1111/j.0956-7976.2004.01503003.x [PubMed: 15016286]

Lindenberger U, & Baltes PB (1994). Sensory functioning and intelligence in old age: A strong connection. Psychology and Aging, 9(3), 339–355. 10.1037/0882-7974.9.3.339 [PubMed: 7999320]

Lindenberger U, & Staudinger UM (2018). Hoheres Erwachsennalter. In Schneider W & Lindenberger U (Eds.), Entwicklungspsychologie: Mit Online-Material (Originalausgabe, 8., vollständig überarbeitete Auflage) (pp. 291–318). Beltz.

*Livingston R, Jennings E, Reynolds C, & Gray R (2003). Multivariate analyses of the profile stability of intelligence tests: high for IQs, low to very low for subtest analyses. Archives of Clinical Neuropsychology, 18(5), 487–507. 10.1016/S0887-6177(02)00147-6 [PubMed: 14591445]

*Lonner WJ, Thorndike RM, Forbes NE, & Ashworth C (1985). The Influence of Television on Measured Cognitive Abilities. Journal of Cross-Cultural Psychology, 16(3), 355–380. 10.1177/0022002185016003006

Lövdén M, Bergman L, Adolfsson R, Lindenberger U, & Nilsson L-G (2005). Studying individual aging in an interindividual context: Typical paths of age-related, dementia-related, and mortality-related cognitive development in old age. Psychology and Aging, 20(2), 303–316. 10.1037/0882-7974.20.2.303 [PubMed: 16029094]

Lövdén M, Fratiglioni L, Glymour MM, Lindenberger U, & Tucker-Drob EM (2020). Education and Cognitive Functioning Across the Life Span. Psychological Science in the Public Interest, 21(1), 6–41. 10.1177/1529100620920576 [PubMed: 32772803]

Low KSD, Yoon M, Roberts BW, & Rounds J (2005). The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. Psychological Bulletin, 131(5), 713–737. 10.1037/0033-2909.131.5.713 [PubMed: 16187855]

*Lowe JD, Anderson HN, Williams A, & Currie BB (1987). Long-term predictive validity of the WPPSI and the WISC-R with black school children. Personality and Individual Differences, 8(4), 551–559. 10.1016/0191-8869(87)90218-2

*Lyons MJ, Panizzon MS, Liu W, McKenzie R, Bluestone NJ, Grant MD, Franz CE, Vuoksimaa EP, Toomey R, Jacobson KC, Reynolds CA, Kremen WS, & Xian H (2017). A longitudinal

twin study of general cognitive ability over four decades. Developmental Psychology, 53(6), 1170–1177. 10.1037/dev0000303 [PubMed: 28358535]

*Lyons MJ, York TP, Franz CE, Grant MD, Eaves LJ, Jacobson KC, Schaie KW, Panizzon MS, Boake C, Xian H, Toomey R, Eisen SA, & Kremen WS (2009). Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. Psychological Science, 20(9), 1146–1152. 10.1111/j.1467-9280.2009.02425.X [PubMed: 19686293]

Mackintosh NJ (1998). IQ and Human Intelligence. Oxford University Press.

Mackintosh NJ (2011). History of Theories and Measurement of Intelligence. In Sternberg RJ & Kaufman SB (Eds.), The Cambridge Handbook of Intelligence (pp. 3–19). Cambridge University Press, 10.1017/cbo9780511977244.002

*Magnusson D[D], & Backteman G. (1978). Longitudinal Stability of Person Characteristics: Intelligence and Creativity. Applied Psychological Measurement, 2(4), 481–490. 10.1177/014662167800200402

*Mansukoski L, Bogin B, Galvez-Sobral JA, Furlán L, & Johnson W [William] (2020). Differences and secular trends in childhood IQ trajectories in Guatemala City. Intelligence, 80, 101438. 10.1016/j.intell.2020.101438 [PubMed: 32508371]

*Martin JD, Blair GE, Stokes EH, & Lester EH. (1977). A Validity and Reliability Study of the Slosson Intelligence Test and the Shipley Institute of Living Scale. Educational and Psychological Measurement, 37(4), 1107–1110. 10.1177/001316447703700441

*Matarazzo RG, Wiens AN, Matarazzo JD, & Manaugh TS (1973). Test-retest reliability of the WAIS in a normal population. Journal of Clinical Psychology, 29(2), 194–197. 10.1002/1097-4679(197304)29:2<194::aid-jclp2270290212>3.0.co;2-w

*McArdle JJ, & Epstein D (1987). Latent Growth Curves within Developmental Structural Equation Models. Child Development, 58(1), 110–133. 10.2307/1130295 [PubMed: 3816341]

*McArdle JJ, Ferrer-Caja E, Hamagami E, & Woodcock RW (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. Developmental Psychology, 38(1), 115–142. 10.1037//0012-1649.38.1.115 [PubMed: 11806695]

*McArdle JJ, & Wang L (2008). Modeling age-based turning points in longitudinal life-span growth curves of cognition. In Cohen P(Ed.), Applied data analytic techniques for turning points research (pp. 105–127). Routledge/Taylor & Francis Group.

McCall RB, Eichorn DH, Hogarty PS, Uzgiris IC, & Schaefer ES (1977). Transitions in Early Mental Development. Monographs of the Society for Research in Child Development, 42(3), 1. 10.2307/1165992

*McCoach DB, Yu H, Gottfried AW, & Gottfried AE (2017). Developing talents: A longitudinal examination of intellectual ability and academic achievement. High Ability Studies, 28(1), 7–28. 10.1080/13598139.2017.1298996

McDermott PA, Fantuzzo JW, Glutting JJ, Watkins MW, & Baggaley AR (1992). Illusions of Meaning in the Ipsative Assessment of Children's Ability. The Journal of Special Education, 25(4), 504–526. 10.1177/002246699202500407

*McGhee RL, & Lieberman LR (1990). Test-retest reliability of the test of nonverbal intelligence (TONI). Journal of School Psychology, 28(4), 351–353. 10.1016/0022-4405(90)90024-2

McGrew KS (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In Flanagan DP, Genshaft JL, & Harrison PL (Eds.), Contemporary intellectual assessment: Theories, tests, and issues, (pp. 151—179). Guilford.

*Milgram, Norman A. (1971). IQ Constancy in Disadvantaged Negro Children. Psychological Reports, 29(1), 319–326. 10.2466/pr0.1971.29.1.319

*Moore T (1967). Language and Intelligence: A Longitudinal Study of the First Eight Years Part I. Patterns of Development in Boys and Girls. Human Development, 10(2), 88–106. https://www.jstor.org/stable/26761864 [PubMed: 6036486]

*Mortensen EL, & Kleven M (1993). A WAIS longitudinal study of cognitive development during the life span from ages 50 to 70. Developmental Neuropsychology, 9(2), 115–130. 10.1080/87565649109540548

*Narvaez D, Gleason T, Wang L, Brooks J, Lefever JB, & Cheng Y (2013). The evolved development niche: Longitudinal effects of caregiving practices on early childhood psychosocial development. Early Childhood Research Quarterly, 28(4), 759–773. 10.1016/j.ecresq.2013.07.003

Neisser U, Boodoo G, Bouchard TJ, Boykin AW, Brody N, Ceci SJ, Halpern DF, Loehlin JC, Perloff R, Sternberg RJ, & Urbina S (1996). Intelligence: Knowns and unknowns. American Psychologist, 51(2), 77–101. 10.1037/0003-066X.51.2.77

Nesselroade JR, & Liben LS (1991). The warp and the woof of the developmental fabric. In Downs RM, Liben LS, & Palermo DS (Eds.), Visions of aesthetics, the environment, & development: The legacy of Joachim F. Wohlwill. Lawrence Erlbaum Associates.

Nettelbeck T, & Wilson C (2005). Intelligence and IQ: What teachers should know. Educational Psychology, 25(6), 609–630. 10.1080/01443410500344696

*Nieding G, Ohler P, Diergarten AK, Möckel T, Rey GD, & Schneider W (2017). The Development of Media Sign Literacy—A Longitudinal Study With 4-Year-Old Children. Media Psychology, 20(3), 401–427. 10.1080/15213269.2016.1202773

Nielsen M, Haun D, Kärtner J, & Legare CH (2017). The persistent sampling bias in developmental psychology: A call to action. Journal of Experimental Child Psychology, 162, 31–38. 10.1016/j.jecp.2017.04.017 [PubMed: 28575664]

*Nisbet JD (1957). Symposium: Contributions to intelligence testing and the theory of intelligence: IV. Intelligence and age: Retesting with twenty-four years' interval. British Journal of Educational Psychology, 27, 190–198. 10.1111/j.2044-8279.1957.tb01410.x

*Nkaya HN, Huteau M, & Bonnet J-P (1994). Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. Perceptual and Motor Skills, 78(2), 503–510. 10.2466/pms.1994.78.2.503

Nyberg L, Karalija N, Papenberg G, Salami A, Andersson M, Pedersen R, Vikner T, Garrett DD, Riklund K, Wåhlin A, Lövdén M, Lindenberger U, & Bäckman L (2022). Longitudinal stability in working memory and frontal activity in relation to general brain maintenance. Scientific Reports, 12(1), 20957. 10.1038/s41598-022-25503-9 [PubMed: 36470934]

O'Connor PA, Morsanyi K, & McCormack T (2019). The Stability of Individual Differences in Basic Mathematics-Related Skills in Young Children at the Start of Formal Education. Mind, Brain, and Education, 13(3), 234–244. 10.1111/mbe.12190

*Oelhafen S, Nikolaidis A, Padovani T, Blaser D, Koenig T, & Perrig WJ (2013). Increased parietal activity after training of interference control. Neuropsychologia, 57(13), 2781–2790. 10.1016/j.neuropsychologia.2013.08.012

*Okely JA, Akeroyd MA, Allerhand M, Starr JM, & Deary IJ (2019). Longitudinal associations between hearing loss and general cognitive ability: The Lothian Birth Cohort 1936. Psychology and Aging, 34(6), 766–779. 10.1037/pag0000385 [PubMed: 31393145]

Ones D. S [Deniz S.], Viswesvaran C [Chockalingam], & Dilchert S. (2017). Cognitive Ability in Personnel Selection Decisions. In Evers A, Anderson N, & Voskuijl O (Eds.), The Blackwell Handbook of Personnel Selection (pp. 143–173). Blackwell Publishing Ltd. 10.1002/9781405164221.ch7

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, … Moher D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ (Clinical Research Ed.), 372, n71. 10.1136/bmj.n71

*Paolo AM, & Ryan JJ (1993). Test-retest stability of the Satz-Mogel WAIS-R short form in a sample of normal persons 75 to 87 years of age. Archives of Clinical Neuropsychology, 8(5), 397–404. 10.1016/0887-6177(93)90003-J [PubMed: 14589709]

*Persson N, Ghisletta P, Dahle CL, Bender AR, Yang Y, Yuan P, Daugherty AM, & Raz N (2016). Regional brain shrinkage and change in cognitive performance over two years: The bidirectional influences of the brain and cognitive reserve factors. NeuroImage, 126, 15–26. 10.1016/j.neuroimage.2015.11.028 [PubMed: 26584866]

*Petrill SA, Lipton PA, Hewitt JK, Plomin R, Cherny SS, Corley R, & DeFries JC (2004). Genetic and environmental contributions to general cognitive ability through the first 16 years of life. Developmental Psychology, 40(5), 805–812. 10.1037/0012-1649.40.5.805 [PubMed: 15355167]

Plomin R, & Stumm S. von (2018). The new genetics of intelligence. Nature Reviews. Genetics, 19(3), 148–159. 10.1038/nrg.2017.104

*Pluck G (2019). Preliminary Validation of a Free-to-Use, Brief Assessment of Adult Intelligence for Research Purposes: The Matrix Matching Test. Psychological Reports, 122(2), 709–730. 10.1177/0033294118762589 [PubMed: 29540106]

*Polderman TJC, Gosso MF, Posthuma D, van Beijsterveldt TC, Heutink P, Verhulst FC, & Di Boomsma (2006). A longitudinal twin study on IQ, executive functioning, and attention problems during childhood and early adolescence. Acta Neurologica Belgica, 106(4), 191–207. [PubMed: 17323837]

*Quereshi MY (1968). Practice effects on the WISC subtest scores and IQ estimates. Journal of Clinical Psychology, 24(1), 79–85. 10.1002/1097-4679(196801)24:1<79::AID-JCLP2270240125>3.0.CO;2-C [PubMed: 5639467]

*Raguet ML, Campbell DA, Berry DTR, Schmitt FA, & Smith GT (1996). Stability of intelligence and intellectual predictors in older persons. Psychological Assessment, 8(2), 154–160. https://psycnet.apa.org/doi/10.1037/1040-3590.8.2.154

*Randall JG, Villado AJ, & Zimmer CU (2016). Is Retest Bias Biased? Examining Race and Sex Differences in Retest Performance. Journal of Personnel Psychology, 15(2), 45–54. 10.1027/1866-5888/a000149

*Rantalainen V, Lahti J, Henriksson M, Kajantie E, Tienari P, Eriksson JG, & Raikkonen K (2016). Apoe and aging-related cognitive change in a longitudinal cohort of men. Neurobiology of Aging, 44, 151–158. 10.1016/j.neurobiolaging.2016.04.024 [PubMed: 27318143]

Raven JC (1938). Progressive matrices: A perceptual test of intelligence. H. K. Lewis.

*Raykov T (2000). A Method for Examining Stability in Reliability. Multivariate Behavioral Research, 35(3), 289–305. 10.1207/S15327906MBR3503_01 [PubMed: 26745333]

*Raz N, Lindenberger U, Ghisletta P, Rodrigue KM, Kennedy KM, & Acker JD (2008). Neuroanatomical correlates of fluid intelligence in healthy adults and persons with vascular risk factors. Cerebral Cortex, 18(3), 718–726. 10.1093/cercor/bhm108 [PubMed: 17615248]

*Razavieh A, & Shahim S (1990). Retest reliability of the Wechsler Preschool and Primary Scale of Intelligence restandardized in Iran. Psychological Reports, 66(3), 865–866. 10.2466/pr0.1990.66.3.865 [PubMed: 2377704]

*Rees AH, & Palmer FH (1970). Factors Related to Change in Mental Test Performance. Developmental Psychology, 3(2), 1–57. https://oce.ovid.corn/article/00063061-197009001-00001/HTML

Reeve CL, & Bonaccio S (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. Intelligence, 39(5), 255–272. 10.1016/j.intell.2011.06.009

*Reeve CL, & Lam H (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. Intelligence, 33(5), 535–549. 10.1016/j.intell.2005.05.003

*Reuben A, Arseneault L, Belsky DW, Caspi A, Fisher HL, Houts RM, Moffitt TE, & Odgers C (2019). Residential neighborhood greenery and children's cognitive development. Social Science & Medicine, 230, 271–279. 10.1016/j.socscimed.2019.04.029 [PubMed: 31035206]

*Richerson LP, Watkins MW, & Beaujean AA (2014). Longitudinal Invariance of the Wechsler Intelligence Scale for Children–Fourth Edition in a Referral Sample. Journal of Psychoeducational Assessment, 32(7), 597–609. 10.1177/0734282914538802

Rinaldi L, & Karmiloff-Smith A (2017). Intelligence as a Developing Function: A Neuroconstructivist Approach. Journal of Intelligence, 5(2). 10.3390/jintelligence5020018

*Ritchie SJ, Bates TC, & Deary IJ (2015). Is education associated with improvements in general cognitive ability, or in specific skills? Developmental Psychology, 51(5), 573–582. 10.1037/a0038981 [PubMed: 25775112]

Ritchie SJ, & Tucker-Drob EM (2018). How Much Does Education Improve Intelligence? A Meta-Analysis. Psychological Science, 29(8), 1358–1369. 10.1177/0956797618774253 [PubMed: 29911926]

Roberts BW[BW], & DelVecchio WF (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. Psychological Bulletin, 126(1), 3–25. 10.1037/0033-2909.126.1.3 [PubMed: 10668348]

Rönnlund M, Sundström A, & Nilsson L-G (2015). Interindividual differences in general cognitive ability from age 18 to age 65years are extremely stable and strongly associated with working memory capacity. Intelligence, 53, 59–64. 10.1016/j.intell.2015.08.011

*Rose SA, Feldman JF, Wallace IF, & McCarton C (1991). Information processing at 1 year: Relation to birth status and developmental outcome during the first 5 years. Developmental Psychology, 27(5), 723–737. 10.1037/0012-1649.27.5.723

*Rudinger G, Andres J, & Rietz C (2010). Structural equation models for studying intellectual development. In Magnusson D, Bergman LR, Rudinger G, & Torestad B (Eds.), Problems and Methods in Longitudinal Research (pp. 274–307). Cambridge University Press. 10.1017/CBO9780511663260.015

*Rugg H, & Colloton C (1921). Constancy of the Stanford-Binet IQ as shown by retests. Journal of Educational Psychology, 12(6), 315–322. 10.1037/h0072291

*Ryan JJ, Glass LA, & Bartels JM (2010). Stability of the WISC-IV in a sample of elementary and middle school children. Applied Neuropsychology, 17(1), 68–72. 10.1080/09084280903297933 [PubMed: 20146124]

Salgado JF, Viswesvaran C[C], & Ones DS[DS] (2002). Predictors used for personnel selection: An overview of constructs, methods and techniques. In Anderson N, Ones DS, Sinangil HK, & Viswesvaran C (Eds.), Handbook of industrial, work and organizational psychology, Vol. 1. Personnel psychology (pp. 165–199). Sage Publications Ltd.

*Salthouse TA (2012). Does the direction and magnitude of cognitive change depend on initial level of ability? Intelligence, 40(4), 352–361. 10.1016/j.intell.2012.02.006 [PubMed: 22711949]

*Salthouse TA (2012). Psychometric properties of within-person across-session variability in accuracy of cognitive performance. Assessment, 19(4), 494–501. 10.1177/1073191112438744 [PubMed: 22389243]

*Salthouse TA (2014). Correlates of cognitive change. Journal of Experimental Psychology: General, 143(3), 1026–1048. 10.1037/a0034847 [PubMed: 24219021]

Salthouse TA, Schroeder DH, & Ferrer E (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. Developmental Psychology, 40(5), 813–822. 10.1037/0012-1649.40.5.813 [PubMed: 15355168]

Salthouse TA, & Tucker-Drob EM (2008). Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology, 22(6), 800–811. 10.1037/a0013091 [PubMed: 18999354]

*Sameroff AJ, Seifer R, Baldwin A, & Baldwin C (1993). Stability of intelligence from preschool to adolescence: The influence of social and family risk factors. Child Development, 64(1), 80–97. 10.1111/j.1467-8624.1993.tb02896.x [PubMed: 8436039]

*Sandu A-L, Staff RT, McNeil CJ, Mustafa N, Ahearn T, Whalley LJ, & Murray AD (2014). Structural brain complexity and cognitive decline in late life--a longitudinal study in the Aberdeen 1936 Birth Cohort. NeuroImage, 100, 558–563. 10.1016/j.neuroimage.2014.06.054 [PubMed: 24993896]

*Schaie KW (2012). Developmental Influences on Adult Intelligence. Oxford University Press, 10.1093/acprof:osobl/9780195386134.001.0001

*Schalke D, Brunner M, Geiser C, Preckel F, Keller U, Spengler M, & Martin R (2013). Stability and change in intelligence from age 12 to age 52: Results from the Luxembourg MAGRIP study. Developmental Psychology, 49(8), 1529–1543. 10.1037/a0030623 [PubMed: 23148935]

Scharfen J, Peters JM, & Holling H (2018). Retest effects in cognitive ability tests: A meta-analysis. Intelligence, 67, 44–66. 10.1016/j.intell.2018.01.003

Scherrer V, & Preckel F (2019). Development of Motivational Variables and Self-Esteem During the School Career: A Meta-Analysis of Longitudinal Studies. Review of Educational Research, 89(2), 211–258. 10.3102/0034654318819127

*Schmidt FL, & Crano WD (1974). A test of the theory of fluid and crystallized intelligence in middle-and low-socioeconomic-status children: A cross-lagged panel analysis. Journal of Educational Psychology, 66(2), 255–261. 10.1037/h0036093

*Schneider W, & Bullock M (2010). Human Development from Early Childhood to Early Adulthood, 7–34. Psychology Press, 10.4324/9780203888544

Schneider WJ, & McGrew KS (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In Flanagan DP & McDonough EM (Eds.), Contemporary intellectual assessment: Theories, tests, and issues. (4th ed., pp. 73–162). Guilford.

*Schneider W, Niklas F, & Schmiedeler S (2014). Intellectual development from early childhood to early adulthood: The impact of early IQ differences on stability and change over time. Learning and Individual Differences, 32, 156–162. 10.1016/j.lindif.2014.02.001

Schroeders U, Schipolowski S, & Wilhelm O (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. Intelligence, 48, 15–29. 10.1016/j.intell.2014.10.006

Schuerger JM, & Witt AC (1989). The temporal stability of individually tested intelligence. Journal of Clinical Psychology, 45(2), 294–302. 10.1002/1097-4679(198903)45:2<294::AID-JCLP2270450218>3.0.CO;2-N

*Schwartzman AE, Gold D, Andres D, Arbuckle TY, & Chaikelson J (1987). Stability of intelligence: A 40-year follow-up. Canadian Journal of Psychology, 41(2), 244–256. 10.1037/h0084155 [PubMed: 3502899]

*Segal NL, McGuire SA, Havlena J, Gill P, & Hershberger SL (2007). Intellectual similarity of virtual twin pairs: Developmental trends. Personality and Individual Differences, 42(7), 1209–1219. 10.1016/j.paid.2006.09.028 [PubMed: 17476320]

*Seidler AL, & Ritchie SJ (2018). The association between socioeconomic status and cognitive development in children is partly mediated by a chaotic home atmosphere. Journal of Cognition and Development, 19(5), 486–508. 10.1080/15248372.2018.1515077

*Sherman LE, Rudie JD, Pfeifer JH, Masten CL, McNealy K, & Dapretto M (2014). Development of the default mode and central executive networks across early adolescence: a longitudinal study. Developmental Cognitive Neuroscience, 10, 148—159. 10.1016/j.dcn.2014.08.002 [PubMed: 25282602]

*Smith DK, Buckley S, & Shine AE (1996). WISC-III/WISC-R Relationships in Native Alaskan Students.

*Smith GE, Bohac DL, Ivnik RJ, & Malec JF (1997). Using word recognition tests to estimate premorbid IQ in early dementia: Longitudinal data. Journal of the International Neuropsychological Society, 3(6), 528–533. 10.1017/S1355617797005286 [PubMed: 9448366]

*Snow WG, Tierney MC, Zorzitto ML, Fisher RH, & Reid DW (1989). Wais-R test-retest reliability in a normal elderly sample. Journal of Clinical and Experimental Neuropsychology, 11(4), 423–428. 10.1080/01688638908400903 [PubMed: 2760178]

*Sparks RL, Patton J, & Murdoch A (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of ‘10 years later’. Reading and Writing, 27(1), 189–211. https://psycnet.apa.org/doi/10.1007/s11145-013-9439-2

Spearman C (1904). ‘General Intelligence,’ Objectively Determined and Measured. The American Journal of Psychology, 15(2), 201. 10.2307/1412107

*Starkweather J (2009). Bidirectional effects between engaged lifestyle and cognition in later life: Exploring the moderation hypothesis for personality variables. Doctoral dissertation. University of North Texas.

*Starnawska A, Tan Q, Lenart A, McGue M, Mors O, Børglum AD, Christensen K, Nyegaard M, & Christiansen L (2017). Blood DNA methylation age is not associated with cognitive functioning in middle-aged monozygotic twins. Neurobiology of Aging, 50, 60–63. 10.1016/j.neurobiolaging.2016.10.025 [PubMed: 27889677]

*Strand S (2004). Consistency in reasoning test scores over time. British Journal of Educational Psychology, 74(4), 617–631. 10.1348/0007099042376445 [PubMed: 15530205]

*Stumm S. von, & Deary IJ (2013). Intellect and cognitive performance in the Lothian Birth Cohort 1936. Psychology and Aging, 28(3), 680–684. 10.1037/a0033924 [PubMed: 24041000]

Symonds PM (1928). Factors influencing test reliability. Journal of Educational Psychology, 19(2), 73–87. 10.1037/h0071867

*Taji W, Mandell B, & Liu J (2019). China's urban-rural childhood cognitive divide:evidence from a longitudinal cohort study after a 6-year follow up. Intelligence, 73, 1–7. 10.1016/j.intell.2019.01.002

Taylor JL, Lindsay WR, & Willner P (2008). CBT for People with Intellectual Disabilities: Emerging Evidence, Cognitive Ability and IQ Effects. Behavioural and Cognitive Psychotherapy, 36(6), 723–733. 10.1017/S1352465808004906

Thapar A, Pine DS, Leckman JF, Scott S, Snowling MJ, & Taylor E (Eds.). (2015). Rutter's Child and Adolescent Psychiatry. John Wiley & Sons, Ltd.

*Thompson AP, & Molly K (1993). The stability of WAIS-R IQ for 16-year-old students retested after 3 and 8 months. Journal of Clinical Psychology, 49(6), 891–898. 10.1002/1097-4679(199311)49:6<891::AID-JCLP2270490617>3.0.CO;2-8 [PubMed: 8300878]

*Thompson AP, & Sota DD (1998). Comparison of WAIS—R and WISC—III Scores with a Sample of 16-Year-Old Youth. Psychological Reports, 82(3_suppl), 1339—1346. 10.2466/pr0.1998.82.3c.1339 [PubMed: 9709537]

*Thorndike R (1977). Causation of Binet IQ Decrements. Journal of Educational Measurement, 14(3), 197–202. 10.1111/J.1745-3984.1977.TB00036.X

Thorvaldsson V, Hofer SM, Berg S, & Johansson B (2006). Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 61(6), P348–54. 10.1093/geronb/61.6.P348 [PubMed: 17114304]

Thurstone LL (1938). Primary mental abilities.(1), ix + 121.

*Tikhomirova T, Kuzmina Y, Lysenkova I, & Malykh S (2019). Development of approximate number sense across the elementary school years: A cross-cultural longitudinal study. Developmental Science, 22(4), e12823. 10.1111/desc.12823 [PubMed: 30811762]

Tipton E (2013). Robust variance estimation in meta-regression with binary dependent effects. Research Synthesis Methods, 4(2), 169–187. 10.1002/jrsm.1070 [PubMed: 26053656]

Tipton E (2015). Small sample adjustments for robust variance estimation with meta-regression. Psychological Methods, 20(3), 375–393. 10.1037/met0000011 [PubMed: 24773356]

Trzesniewski KH, Donnellan MB, & Robins RW (2003). Stability of self-esteem across the life span. Journal of Personality and Social Psychology, 84(1), 205–220. 10.1037/0022-3514.84.1.205 [PubMed: 12518980]

Tucker-Drob EM (2009). Differentiation of cognitive abilities across the life span. Developmental Psychology, 45(4), 1097–1118. 10.1037/a0015864 [PubMed: 19586182]

Tucker-Drob EM (2019). Cognitive Aging and Dementia: A Life Span Perspective. Annual Review of Developmental Psychology, 1, 177–196. 10.1146/annurev-devpsych-121318-085204

Tucker-Drob EM, & Bates TC (2016). Large Cross-National Differences in Gene χ Socioeconomic Status Interaction on Intelligence. Psychological Science, 27(2), 138—149. 10.1177/0956797615612727 [PubMed: 26671911]

Tucker-Drob EM, & Briley DA (2014). Continuity of genetic and environmental influences on cognition across the life span: A meta-analysis of longitudinal twin and adoption studies. Psychological Bulletin, 140(4), 949–979. 10.1037/a0035893 [PubMed: 24611582]

Tucker-Drob EM, Briley DA, & Harden KP (2013). Genetic and Environmental Influences on Cognition Across Development and Context. Current Directions in Psychological Science, 22(5), 349–355. 10.1177/0963721413485087 [PubMed: 24799770]

Tucker-Drob EM, La Fuente J. de, Köhncke Y, Brandmaier AM, Nyberg L, & Lindenberger U (2022). A strong dependency between changes in fluid and crystallized abilities in human cognitive aging. Science Advances, 8(5), eabj2422. 10.1126/sciadv.abj2422 [PubMed: 35108051]

*Tuma JM, & Appelbaum AS (1980). Reliability and Practice Effects of Wisc-R Iq Estimates in a Normal Population. Educational and Psychological Measurement, 40(3), 671–678. 10.1177/001316448004000310

Turkheimer E, Haley A, Waldron M, D'Onofrio B, & Gottesman II (2003). Socioeconomic status modifies heritability of IQ in young children. Psychological Science, 14(6), 623–628. 10.1046/j.0956-7976.2003.psci_1475.x [PubMed: 14629696]

*Uka F, Gunzenhauser C, Larsen RA, & von Suchodoletz A (2019). Exploring a bidirectional model of executive functions and fluid intelligence across early development. Intelligence, 75, 111–121. 10.1016/j.intell.2019.05.002

van der Sluis S, Willemsen G, Geus E. J. C. de, Boomsma DI, & Posthuma D (2008). Gene-environment interaction in adults' IQ scores: Measures of past and present environment. Behavior Genetics, 38(4), 348–360. 10.1007/s10519-008-9212-5 [PubMed: 18535898]

*Van der Stel M, & Veenman MVJ (2014). Metacognitive skills and intellectual ability of young adolescents: a longitudinal study from a developmental perspective. European Journal of Psychology of Education, 29(1), 117–137. 10.1007/s10212-013-0190-5

*Van Soelen ILC, Brouwer RM, van Leeuwen M, Kahn RS, Pol HEH, & Boomsma DI (2011). Heritability of verbal and performance intelligence in a pediatric longitudinal sample. Twin Research and Human Genetics, 14(2), 119–128. 10.1375/twin.14.2.119 [PubMed: 21425893]

*Vance H, Maddux CD, Fuller GB, & Awadh AM (1996). A longitudinal comparison of WISC-III and WISC-R scores of special education students. Psychology in the Schools, 33(2), 113–118. 10.1002/(SICI)1520-6807(199604)33:2<113.AID-PITS3>3.0.CO;2-S

*Verswijveren SJ, Wiebe SA, Rahman AA, Kuzik N, & Carson V (2020). Longitudinal associations of sedentary time and physical activity duration and patterns with cognitive development in early childhood. Mental Health and Physical Activity, 19, 100340. 10.1016/j.mhpa.2020.100340

Viechtbauer W (2010). Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software, 36(3). 10.18637/jss.v036.i03

Viechtbauer W, & Cheung MW-L (2010). Outlier and influence diagnostics for meta-analysis. Research Synthesis Methods, 1(2), 112–125. 10.1002/jrsm.11 [PubMed: 26061377]

*Villado AJ, Randall JG, & Zimmer CU (2016). The Effect of Method Characteristics on Retest Score Gains and Criterion-Related Validity. Journal of Business and Psychology, 31(2), 233–248. 10.1007/s10869-015-9408-7

Wang J-J, & Kaufman AS (1993). Changes in Fluid and Crystallized Intelligence Across the 20-to 90-Year Age Range on the K-Bit. Journal of Psychoeducational Assessment, 11(1), 29–37. 10.1177/073428299301100104

*Watkins MW, & Smith LG (2013). Long-term stability of the Wechsler Intelligence Scale for Children—Fourth Edition. Psychological Assessment, 25(2), 477–483. 10.1037/a0031653 [PubMed: 23397927]

Wechsler D (2003). Wechsler Intelligence Scale for Children, 4th ed. Harcourt Assessment.

*Welter MM, Jaarsveld S, & Lachmann T (2018). Problem Space Matters: Evaluation of a German Enrichment Program for Gifted Children. Frontiers in Psychology, 9, 569. 10.3389/fpsyg.2018.00569 [PubMed: 29740367]

*Whalley LJ, Fox HC, Starr JM, & Deary IJ (2004). Age at natural menopause and cognition. Maturitas, 49(2), 148–156. 10.1016/j.maturitas.2003.12.014 [PubMed: 15474759]

Wilson CJ, Bowden SC, Byrne LK, Joshua NR, Marx W, & Weiss LG (2023). The cross-cultural generalizability of cognitive ability measures: A systematic literature review. Intelligence, 98, 101751. 10.1016/j.intell.2023.101751

*Yuan P, Voelkle MC, & Raz N (2018). Fluid intelligence and gross structural properties of the cerebral cortex in middle-aged and older adults: A multi-occasion longitudinal study. NeuroImage, 172, 21–30. 10.1016/j.neuroimage.2018.01.032 [PubMed: 29360573]

*Zax M, Cowen EL, Beach DR, & Rappaport J (1972). Longitudinal relationships among aptitude, achievement, and adjustment measures of school children. The Journal of Genetic Psychology, 121(1), 145–154. 10.1080/00221325.1972.10533137

*Ziegler M, Danay E, Heene M, Asendorpf J, & Bühner M (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. Journal of Research in Personality, 46(2), 173–183. 10.1016/j.jrp.2012.01.002

Public Significance Statement

This meta-analytic review finds that cognitive abilities are highly stable from adolescence to late adulthood, but only moderately stable in young children. Stability decreases with increasing time intervals and varies across different cognitive abilities. General intelligence was found to be the most stable cognitive ability, but many specific cognitive abilities are similarly stable. These results provide important standards with respect to the "shelf-life" of cognitive test scores across development and time.

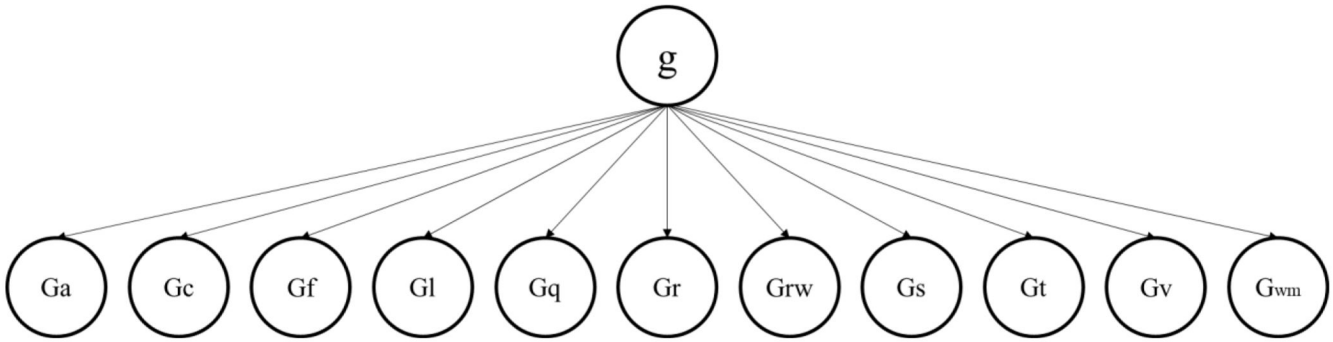

Figure 1. The Top Two Levels of the CHC-Model

Notes. The acronyms are defined in Table 1.

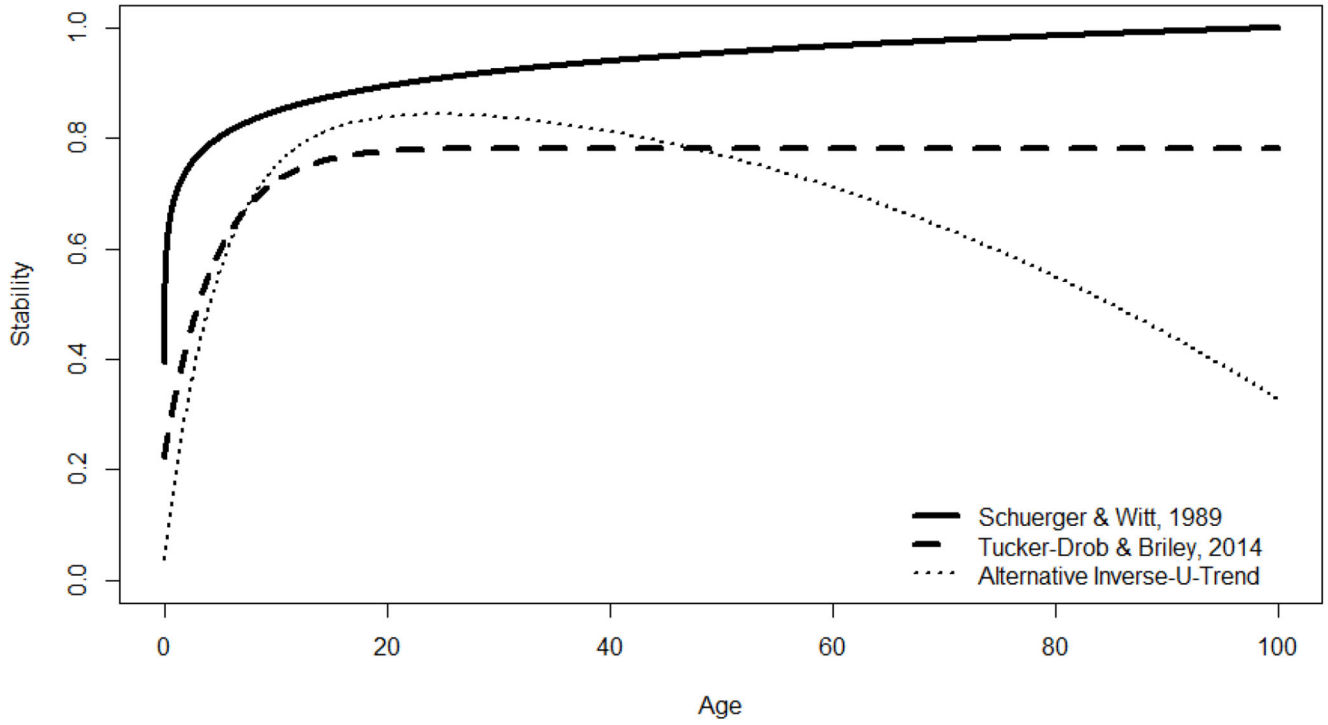

Figure 2. Some Possible Age Moderation Trends.

Notes. Age Moderation of Rank-Order Stability of Cognitive Abilities in Schuerger and Witt (1989), Tucker-Drob and Briley (2014), and an Example of an Alternative Inverse U Shaped Moderation Trend Expected Based on A Model in Which Stability is Inversely Related to the Absolute Magnitude of Mean Change.

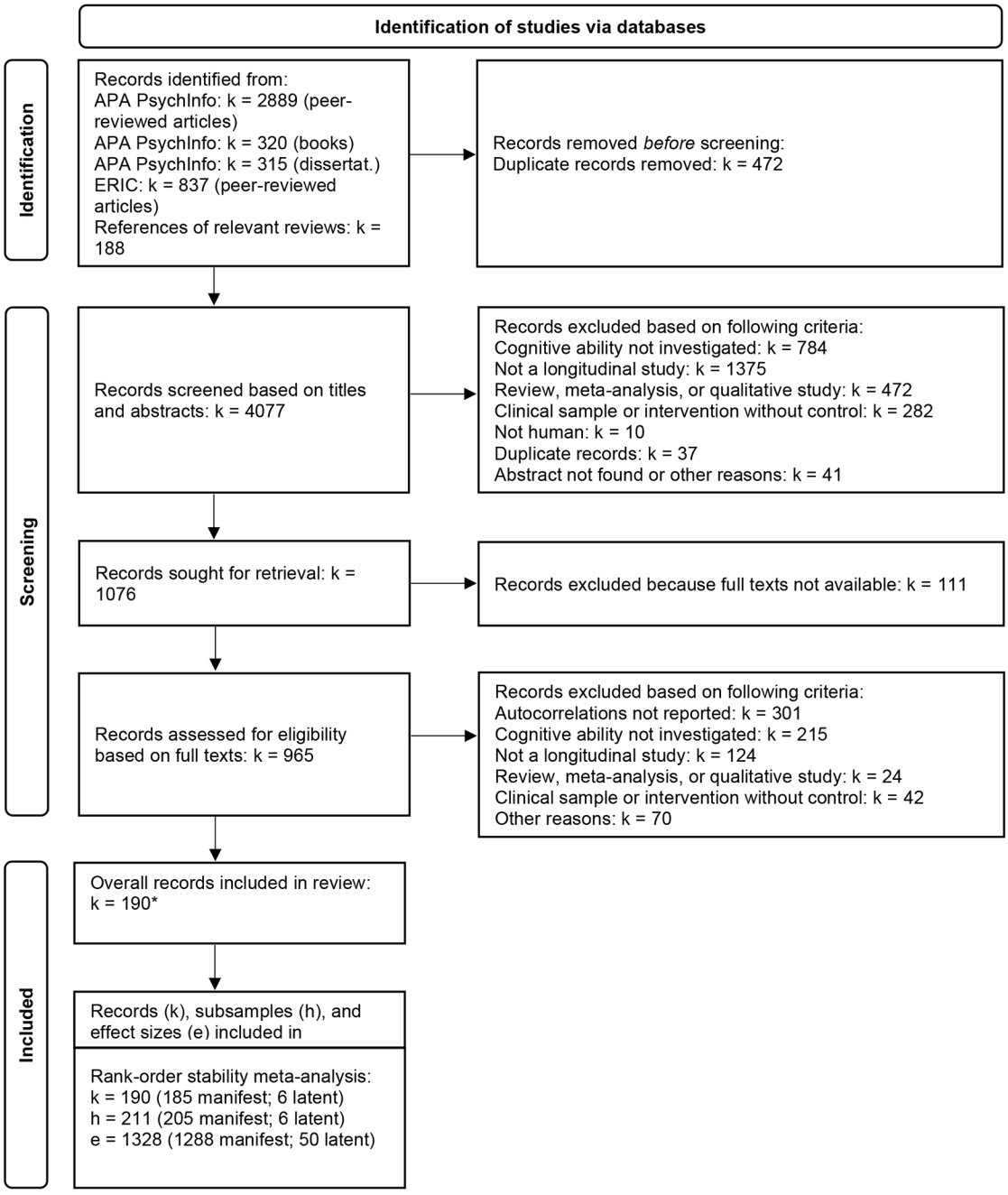

Figure 3. Study Identification and Screening Process as a PRISMA Chart

Notes. This figure presents the literature search and gives information on the number of coded studies and effect sizes. k = number of records, h = number of included samples, e = number of effect sizes. *One additional record (one subsample, six effect sizes) was added after suggestion by a reviewer.

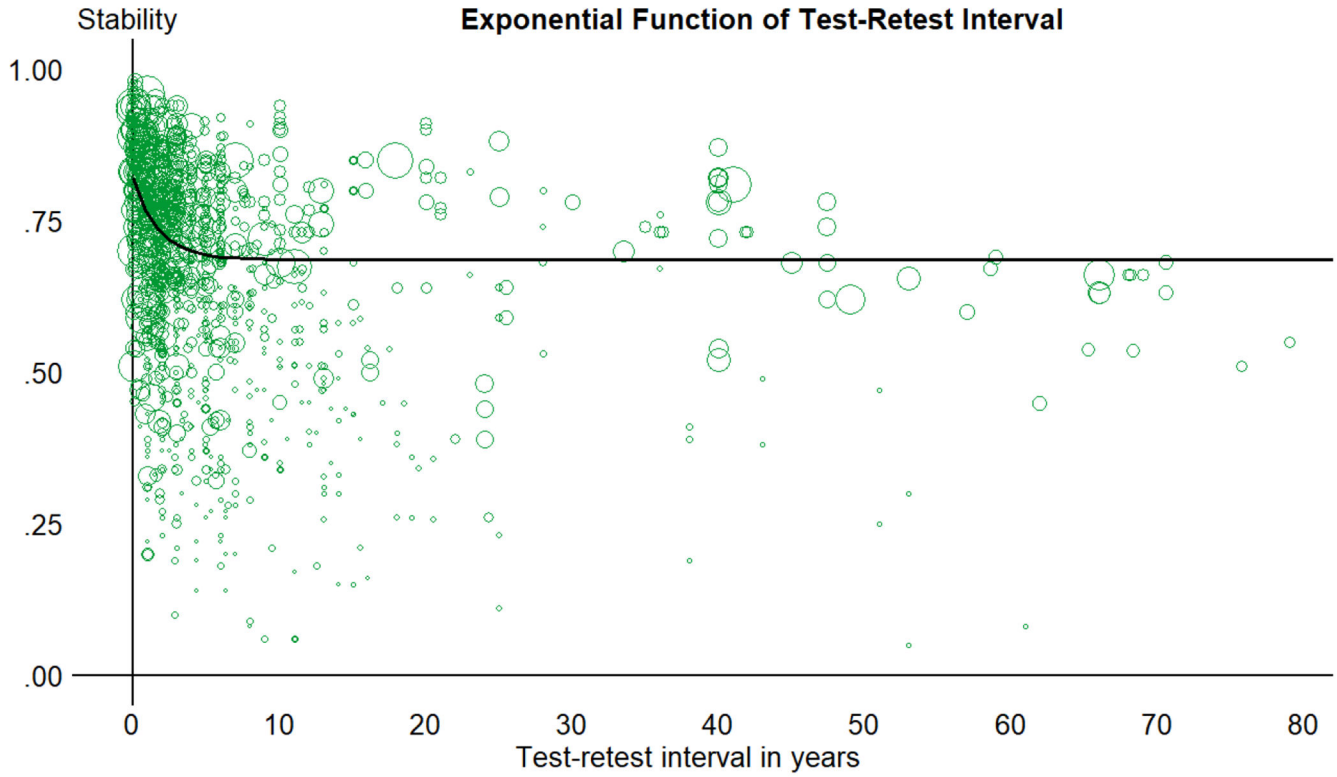

Figure 4. Rank-Order Stability as an Exponential Function of Test-Retest Interval Based on the Complete Dataset

Notes. Larger points represent larger weight of the effect sizes.

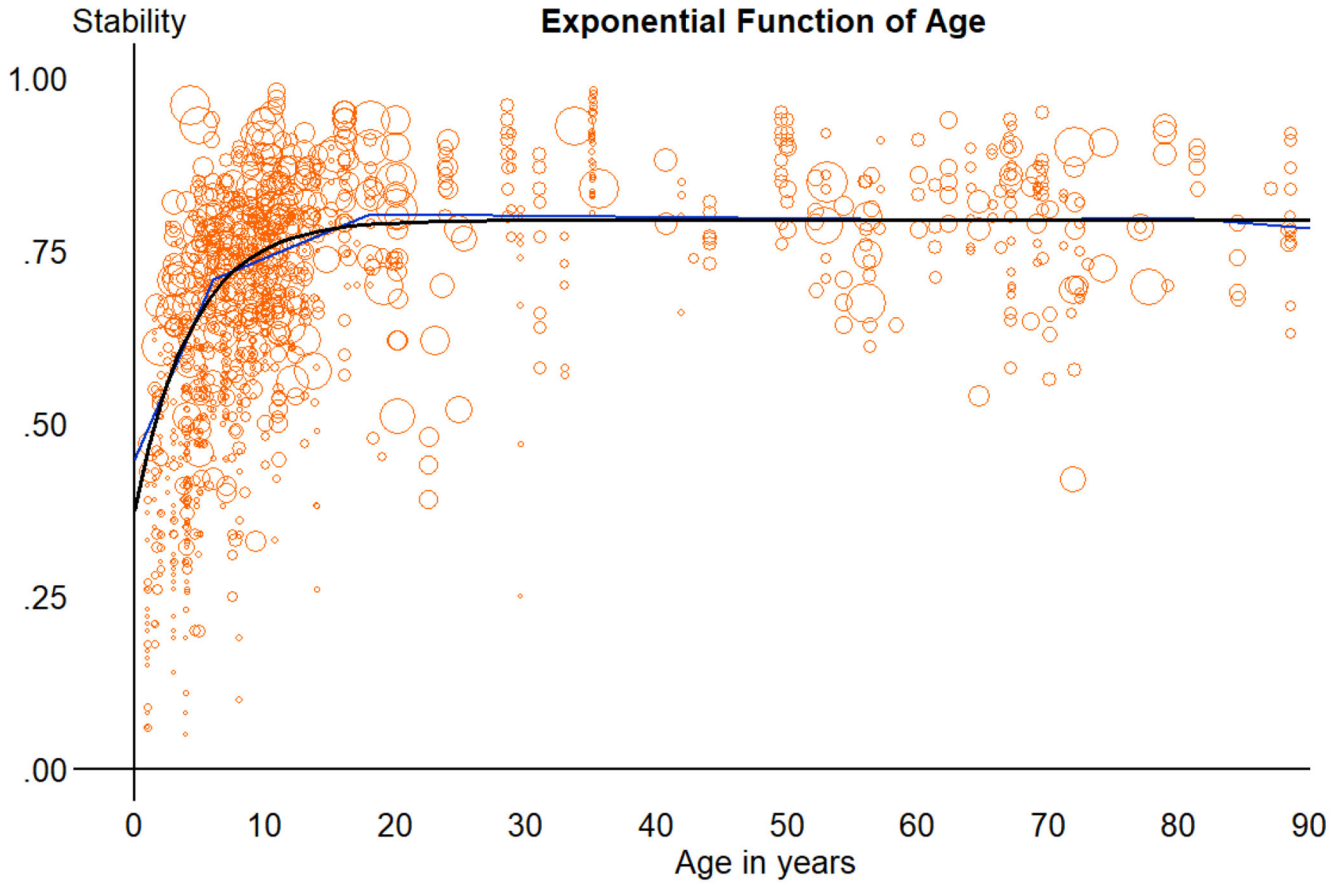

Figure 5. Rank-Order Stability as Exponential (black) and Connected Linear Spline (blue) Functions of Age Based on the Complete Dataset

Notes. Larger points represent larger weight of the effect sizes. As reported in Table S2, the mean test-retest interval was 6.52 years.

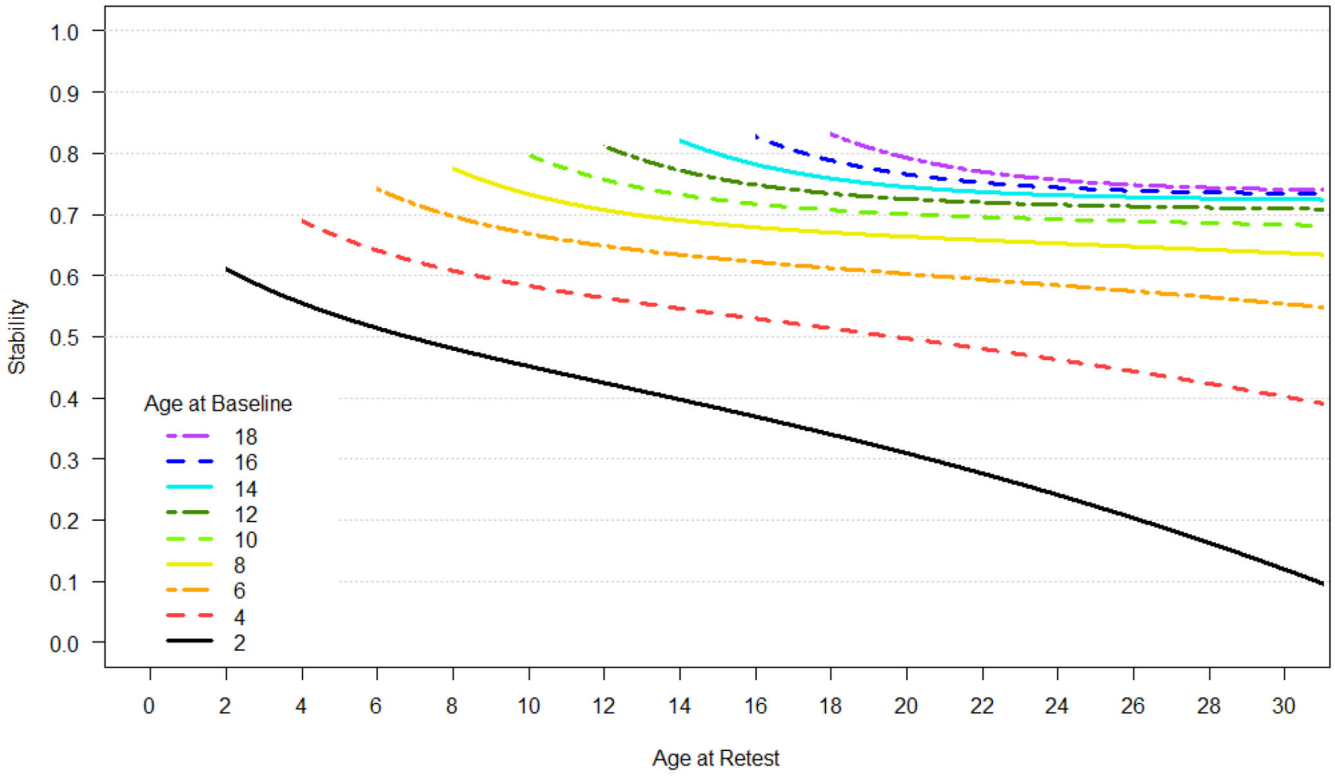

Figure 6.

Temporal Decay of Stability of Cognitive Abilities in Childhood and Adolescence Based on the Exponential Age and Interval Moderator Functions, where Each Plotted Line Represents a Different Baseline Age Followed With Increasing Time Lags

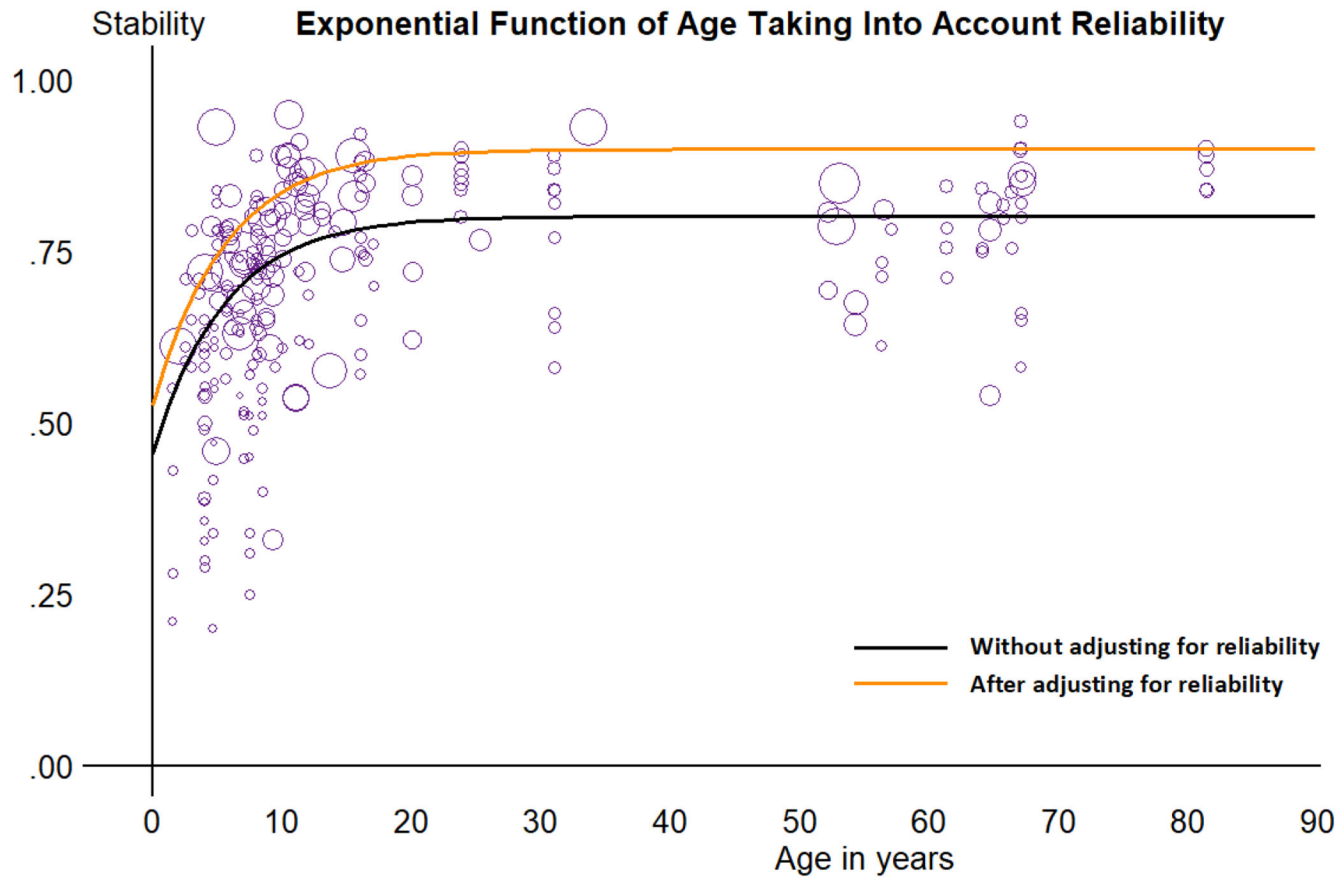

Figure 7. Rank-Order Stability as an Exponential Function of Test-Retest Interval Based on Effect Sizes for Which Reliability Information was Available

Notes. Larger points represent larger weight of the effect sizes. Depicted points represent unadjusted effect sizes.

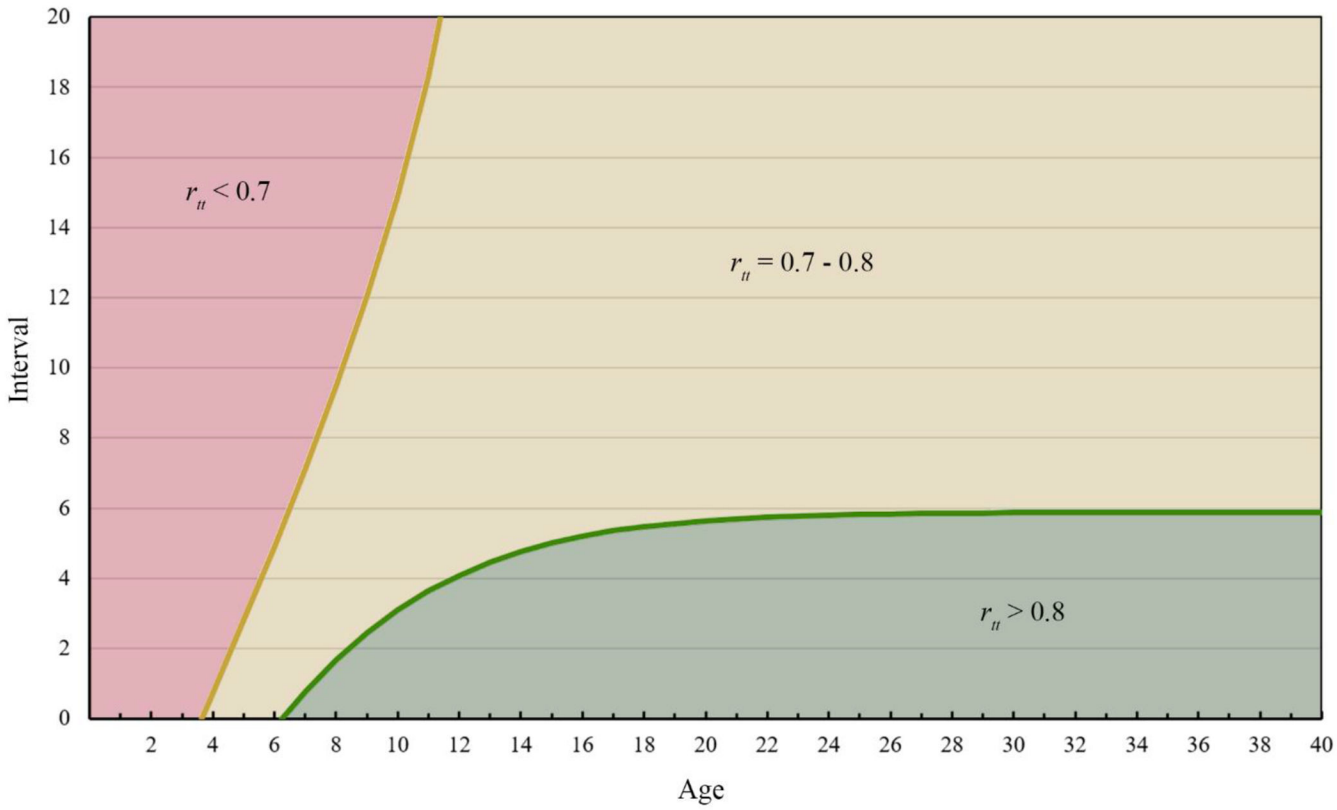

Figure 8.

Maximum Intervals (in Years) for Which a Stability of .70 and .80 is Obtained for Measures of General Intelligence, as Implied by the Age and Interval Duration Moderator Analyses of General Cognitive Ability (Exponential Models)

Table 1

CHC-Abilities and Their Definitions by W. J. Schneider and McGrew (2018)

Ability		Definition
Ga	Auditory Processing	The ability to discriminate, remember, reason, and work creatively (on) auditory stimuli, which may consist of tones, environmental sounds, and speech units.
Gc	Comprehension Knowledge	The ability to comprehend and communicate culturally valued knowledge. Gc includes the depth and breadth of knowledge and skills such as language, words, and general knowledge developed through experience, learning and acculturation.
Gf	Fluid Reasoning	The use of deliberate and controlled procedures (often requiring focused attention) to solve novel “on the spot” problems that cannot be solved by using previously learned habits, schemas, and scripts.
Gl	Learning Efficiency	The ability the ability to learn, store, and consolidate new information over periods of time measured in minutes, hours, days, and years.
Gq	Quantitative Knowledge	The depth and breadth of declarative and procedural knowledge and skills related to mathematics.
Gr	Retrieval Fluency	The rate and fluency at which individuals can produce and selectively and strategically retrieve verbal and nonverbal information or ideas stored in long-term memory.
Grw	Reading and Writing	The depth and breadth of declarative and procedural knowledge and skills related to written language.
Gs	Processing Speed	The ability to control attention to automatically, quickly and fluently perform relatively simple repetitive cognitive tasks. Attentional fluency or attentional speediness.
Gt	Reaction and Decision Speed	The speed of making very simple decisions or judgements when items are presented one at a time.
Gv	Visual Processing	The ability to make use of simulated mental imagery to solve problems – perceiving, discriminating, manipulating, and recalling images in the “mind’s eye.”
Gwm	Working Memory Capacity	The ability to maintain and manipulate information in active attention.

Table 2

Model Fit Indices of Models Testing Linear and Non-Linear Test-Retest Interval and Age Effects

Dataset	Model	LL	SCF	AIC	BIC
H1: Test-Retest Interval					
Complete	Linear interval model	−8800.756	12.804	17615.512	17651.617
Complete	Quadratic interval model	−8799.456	11.325	17614.912	17656.174
Complete	**Exponential interval model**	**−8791.710**	**11.398**	**17599.421**	**17640.683**
g	Linear interval model	−4795.353	12.230	9604.707	9636.045
g	Quadratic interval model	−4792.271	10.902	9600.543	9636.359
g	**Exponential interval model**	**−4781.632**	**10.837**	**9579.264**	**9615.080**
Ga	**Linear interval model**	**−9.727**	**0.735**	**33.454**	**30.720**
Ga	Quadratic interval model	−9.597	0.649	35.194	32.070
Ga	Exponential interval model	Model did not converge			
Gc	Linear interval model	−1571.223	5.228	3156.447	3179.322
Gc	Quadratic interval model	−1565.579	4.875	3147.158	3173.301
Gc	**Exponential interval model**	**−1561.793**	**4.760**	**3139.585**	**3165.728**
Gf	**Linear interval model**	**−1224.067**	**6.589**	**2462.134**	**2484.085**
Gf	Quadratic interval model	−1223.532	5.874	2463.064	2488.151
Gf	Exponential interval model	−1223.979	5.790	2463.957	2489.044
Gl	**Linear interval model**	**−71.896**	**0.944**	**157.792**	**161.746**
Gl	Quadratic interval model	−71.825	0.888	159.649	164.169
Gl	Exponential interval model	−71.877	0.858	159.755	164.274
Gq	Linear interval model	−121.045	1.951	256.090	261.046
Gq	**Quadratic interval model**	**−119.578**	**1.845**	**255.157**	**260.821**
Gq	Exponential interval model	−120.551	1.795	257.101	262.766
Grw	**Linear interval model**	**−61.573**	**1.623**	**137.147**	**136.768**
Grw	Quadratic interval model	−61.550	1.565	139.101	138.668
Grw	Exponential interval model	Model did not converge			
Gs	**Linear interval model**	**−347.390**	**3.024**	**708.780**	**722.439**
Gs	Quadratic interval model	−347.277	2.725	710.553	726.163
Gs	Exponential interval model	Model did not converge			
Gv	Linear interval model	−940.658	3.217	1895.317	1915.601
Gv	Quadratic interval model	−939.284	3.030	1894.569	1917.752
Gv	**Exponential interval model**	**−936.140**	**2.901**	**1888.280**	**1911.463**
Gwm	Linear interval model	−232.239	1.272	478.478	490.967
Gwm	**Quadratic interval model**	**−231.996**	**1.139**	**479.993**	**494.266**
Gwm	Exponential interval model	Model did not converge			
H2: Age					
Step 1. Age Without Additional Predictors					
Complete	Linear age model	−18514.095	23.271	37086.191	37235.765
Complete	Quadratic age model	−18501.550	22.529	37063.101	37217.833
Complete	Linear age spline model	−18489.451	20.529	37044.902	37215.107

Dataset	Model	LL	SCF	AIC	BIC
Complete	**Exponential age model**	**−18488.853**	**22.530**	**37037.707**	**37192.439**
g	Linear age model slope	−8322.772	20.700	16703.543	16833.375
g	Quadratic age model	−8309.768	20.0544	16679.535	16813.845
g	Linear age spline model	−8296.206	18.2360	16658.413	16806.153
g	**Exponential age model**	**−8295.635**	**20.0338**	**16651.270**	**16785.579**
Ga	**Linear age model slope**				
Ga	Quadratic age model	Model did not converge			
Ga	Linear age spline model	Model did not converge			
Ga	Exponential age model	Model did not converge			
Gc	Linear age model	−3004.839	4.906	6067.678	6162.446
Gc	Quadratic age model	−2992.633	4.795	6045.266	6143.301
Gc	Linear age spline model	−2975.651	4.436	6017.302	6125.141
Gc	**Exponential age model**	**−2970.189**	**4.716**	**6000.378**	**6098.414**
Gf	Linear age model	−2469.548	5.139	4997.097	5088.035
Gf	Quadratic age model	−2467.304	5.000	4994.609	5088.683
Gf	Linear age spline model	−2465.802	4.585	4997.603	5101.084
Gf	**Exponential age model**	**−2466.477**	**5.007**	**4992.955**	**5087.029**
Gl	**Linear age model**	**−162.592**	**1.657**	**369.185**	**381.614**
Gl	Quadratic age model	−162.034	1.584	370.069	383.062
Gl	Linear age spline model	−160.296	1.408	370.591	384.715
Gl	Exponential age model	Model did not converge			
Gq	**Linear age model**	**−183.205**	**1.255**	**410.411**	**425.988**
Gq	Quadratic age model	−183.182	1.269	412.364	428.650
Gq	Linear age spline model	−181.330	1.165	412.660	430.361
Gq	Exponential age model	−183.176	1.310	412.352	428.637
Grw	Linear age model	−58.659	0.974	161.318	160.128
Grw	Quadratic age model	−58.002	0.968	162.005	160.761
Grw	**Linear age spline model**	**−55.429**	**0.798**	**160.857**	**159.505**
Grw	Exponential age model	Model did not converge			
Gs	Linear age model	−815.906	2.279	1689.813	1746.399
Gs	Quadratic age model	−807.087	2.207	1674.174	1732.711
Gs	Linear age spline model	−793.624	1.920	1653.248	1717.639
Gs	**Exponential age model**	**−793.805**	**2.096**	**1647.609**	**1706.147**
Gv	Linear age model	−1964.122	5.149	3986.244	4070.281
Gv	Quadratic age model	−1961.850	5.009	3983.699	4070.634
Gv	Linear age spline model	−1955.352	4.581	3976.704	4072.333
Gv	**Exponential age model**	**−1955.853**	**5.001**	**3971.706**	**4058.641**
Gwm	Linear age model	−655.241	2.148	1368.481	1420.223
Gwm	Quadratic age model	−653.077	2.096	1366.154	1419.680
Gwm	Linear age spline model	−651.215	1.952	1368.430	1427.308
Gwm	**Exponential age model**	**−651.724**	**2.171**	**1363.449**	**1416.974**

Step 2 and 3. Test-Retest Interval and Age and Their Interaction

Dataset	Model	LL	SCF	AIC	BIC
Complete	**Exponential interval + exponential age model**	**−14780.884**	**9.580**	**29589.768**	**29661.976**
Complete	Exponential interval × exponential age model	−14777.659	8.950	29585.317	29615.036
g	**Exponential interval + exponential age model**	**−7625.202**	**9.783**	**15278.403**	**15341.081**
g	Exponential interval × exponential age model	−7622.524	9.138	15275.049	15342.203
Ga	**Linear interval + linear age model**	**−30.292**	**0.744**	**84.584**	**81.272**
Ga	Linear interval × linear age model	−29.963	0.677	85.925	79.898
Gc	**Exponential interval + exponential age model**	**−2456.068**	**3.608**	**4940.136**	**4985.886**
Gc	Exponential interval × exponential age model	−2455.322	3.377	4940.644	4989.662
Gf	**Linear interval + exponential age model**	**−1996.580**	**4.600**	**4019.160**	**4059.925**
Gf	Linear interval × exponential age model	−1995.610	4.257	4019.219	4063.121
Gl	**Linear interval + linear age model**	**−131.858**	**1.078**	**287.717**	**294.496**
Gl	Linear interval × linear age model	−131.059	0.956	288.118	295.462
Gq	**Quadratic interval + linear age model**	**−188.865**	**1.380**	**403.730**	**412.935**
Gq	Quadratic interval × linear age model	−187.593	1.172	403.186	413.099
Grw	**Linear interval + linear age spline model**	**−116.639**	**0.744**	**299.278**	**297.493**
Grw	Linear interval × linear age spline model	Model did not converge			
Gs	**Linear interval + exponential age model**	**−632.885**	**2.195**	**1291.770**	**1317.136**
Gs	Linear interval × exponential age model	−632.711	2.062	1293.423	1320.740
Gv	**Exponential interval + exponential age model**	**−1535.332**	**2.592**	**3098.663**	**3139.233**
Gv	Exponential interval × exponential age model	Model did not converge			
Gwm	**Linear interval + exponential age model**	**−1989.646**	**4.581**	**4005.292**	**4046.057**
Gwm	Linear interval × exponential age model	−1988.698	4.239	4005.397	4049.298

Note. All model analyses were conducted in Mplus. Complete = Complete dataset including *g* and CHC broad abilities. Best fitting models are in **bold.**

Table 3

Moderator Analyses of Rank-Order Stability in Cognitive Ability

Dataset	Predictor	Reference category	ρ	*df*	*p*	95% Confidence Interval	I^2	τ^2
H1: Test-Retest Interval								
Exponential interval model								
Complete	Horizontal asymptote (b_0)		.686	114.746	< .001	[.661, .711]	96.014	.011
	Test-retest interval scaling factor (b_1)		−.008	104.956	< .001	[−.010, −.005]		
	Test-retest interval growth rate (b_2)		−.576		.096			
Exponential interval model								
g	Horizontal asymptote (b_0)		.673	66.856	< .001	[.636, .710]	96.950	.009
	Test-retest interval scaling factor (b_1)		−.035	83.142	< .001	[−.045, −.024]		
	Test-retest interval growth rate (b_2)		−.340		.001			
Linear interval model								
Ga	Intercept (b_0)			< 4				
	Test-retest interval linear slope (b_1)			< 4				
Exponential interval model								
Gc	Horizontal asymptote (b_0)		.734	41.313	< .001	[.685, .782]	95.604	.005
	Test-retest interval scaling factor (b_1)		−.010	41.449	< .001	[−.015, −.005]		
	Test-retest interval growth rate (b_2)		−.527		.120			
Linear interval model								
Gf	Intercept (b_0)		.707	4.371	< .001	[.567, .847]	92.354	.008
	Test-retest interval linear slope (b_1)			< 4				
Linear interval model								
Gl	Intercept (b_0)			< 4				
	Test-retest interval linear slope (b_1)			< 4				
Quadratic interval model								
Gq	Intercept (b_0)		.734	5.37	< .001	[.594, .875]	98.203	.009
	Test-retest interval linear slope (b_1)			< 4				
	Test-retest interval quadratic slope (b_2)			< 4				
Linear interval model								
Grw	Intercept (b_0)		.800	11.88	< .001	[.624, .977]	96.389	.005
	Test-retest interval linear slope (b_1)			< 4				
Linear interval model								
Gs	Intercept (b_0)		.736	21.41	< .001	[.680, .792]	93.573	.010
	Test-retest interval linear slope (b_1)			< 4				
Exponential interval model								
Gv	Horizontal asymptote (b_0)		.661	30.127	< .001	[.602, .720]	91.615	.011
	Test-retest interval scaling factor (b_1)		−.018	38.032	< .001	[−.026, −.010]		
	Test-retest interval growth rate (b_2)		−.453		.122			
Linear interval model								

Dataset	Predictor	Reference category	ρ	*df*	*p*	95% Confidence Interval	I^2	τ^2
Gwm	Intercept (b_0)		.638	5.630	< .001	[.483, .792]	91.962	.014
	Test-retest interval linear slope (b_1)		−.016	7.657	0.427	[−.061, .028]		
H2: Age								
Step 1. Age without additional predictors								
Exponential age model								
Complete	Horizontal asymptote (b_0)		.794	133.114	< .001	[.776, .813]	94.246	.007
	Age scaling factor (b_1)		.004	42.479	< .001	[.003, .005]		
	Age growth rate (b_2)		−.230		< .001			
Exponential age model								
g	Horizontal asymptote (b_0)		.835	99.139	< .001	[.814, .856]	95.675	.006
	Age scaling factor (b_1)		.003	29.272	< .001	[.002, .003]		
	Age growth rate (b_2)		−.268		< .001			
Linear age model								
Ga	Intercept (b_0)			< 4				
	Age linear slope (b_1)			< 4				
Exponential age model								
Gc	Horizontal asymptote (b_0)		.847	52.020	< .001	[.821, .873]	94.259	.004
	Age scaling factor (b_1)		.003	11.860	< .001	[.002, .003]		
	Age growth rate (b_2)		−.312		< .001			
Exponential age model								
Gf	Horizontal asymptote (b_0)		.780	28.883	< .001	[.748, .812]	90.731	.007
	Age scaling factor (b_1)		.041	33.055	< .001	[.024, .057]		
	Age growth rate (b_2)		−.095		.069			
Linear age model								
Gl	Intercept (b_0)		.655	4.14	< .001	[.553, .757]	90.836	.015
	Age linear slope (b_1)		.002	4.22	.402	[−.004, .007]		
Linear age model								
Gq	Intercept (b_0)		.726	7.98	< .001	[.642, .810]	93.621	.006
	Age linear slope (b_1)		.003	4.47	= .050	[−.000, .006]		
Linear spline age model								
Grw	Intercept (b_0)			< 4				
	Linear spline 1 (b_1)			< 4				
	Linear spline 2 (b_2)			< 4				
	Linear spline 3 (b_3)			< 4				
	Linear spline 4 (b_4)			< 4				
Exponential age model								
Gs	Horizontal asymptote (b_0)		.822	17.19	< .001	[.800, .844]	71.786	.002
	Age scaling factor (b_1)			< 4				
	Age growth rate (b_2)		−.249		.051			

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
Exponential age model								
Gv	Horizontal asymptote (b_0)		.786	44.046	< .001	[.754, .819]	90.588	.009
	Age scaling factor (b_1)		.000	7.355	0.002	[.000, .001]		
	Age growth rate (b_2)		−.404		.004			
Exponential age model								
Gwm	Horizontal asymptote (b_0)		.755	13.508	< .001	[.702, .807]	87.783	.009
	Age scaling factor (b_1)		.012	6.506	0.042	[.001, .023]		
	Age growth rate (b_2)		−.193		.398			
	Step 2. Test-Retest Interval and Age							
	Exponential interval, exponential age, and interval-age							
Complete	Horizontal asymptote (b_0)		.746	58.142	< .001	[.717, .774]	93.398	.007
	Age scaling factor (b_1)		.005	44.483	< .001	[.004, .006]		
	Age growth rate (b_2)		−.223		< .001			
	Interval-age-interaction (b_3)		−.002		< .001			
	Test-retest interval scaling factor (b_4)		−.025	96.836	< .001	[−.037, −.014]		
	Test-retest interval growth rate (b_5)		−.265		.203			
	Exponential interval and exponential age model							
g	Horizontal asymptote (b_0)		.716	30.210	< .001	[.673, .759]	93.387	.004
	Test-retest interval scaling factor (b_1)		−.095	48.855	< .001	[−.122, −.069]		
	Test-retest interval growth rate (b_2)		−.138		.007			
	Age scaling factor (b_3)		.003	28.989	< .001	[.002, .004]		
	Age growth rate (b_4)		−.258		< .001			
	Linear interval and linear age model							
Ga	Intercept (b_0)			< 4				
	Test-retest interval linear slope (b_1)			< 4				
	Age linear slope (b_2)			< 4				
	Exponential interval and exponential age model							
Gc	Horizontal asymptote (b_0)		.780	23.592	< .001	[.737, .823]	92.642	.003
	Test-retest interval scaling factor (b_1)		−.035	37.403	< .001	[−.050, −.021]		
	Test-retest interval growth rate (b_2)		−.281		.083			
	Age scaling factor (b_3)		.004	13.704	< .001	[.003, .005]		
	Age growth rate (b_4)		−.282		< .001			
	Linear interval and exponential age model							
Gf	Horizontal asymptote (b_0)		.780	28.335	< .001	[.746, .814]	89.946	.007
	Test-retest interval linear slope (b_1)			< 4				
	Age scaling factor (b_2)		.041	33.077	< .001	[.023, .058]		
	Age growth rate (b_3)		−.094		.074			
	Linear interval and linear age model							
Gl	Intercept (b_0)			< 4				

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	Test-retest interval linear slope (b_1)			< 4				
	Age linear slope (b_2)			< 4				
	Quadratic interval and linear age model							
Gq	Intercept (b_0)		.699	5.15	< .001	[.596, .803]	88.889	.004
	Test-retest interval linear slope (b_1)			< 4				
	Test-retest interval quadratic slope (b_2)			< 4				
	Age linear slope (b_3)			< 4				
	Linear interval and linear spline age model							
Grw	Intercept (b_0)			< 4				
	Test-retest interval linear slope (b_1)			< 4				
	Linear spline 1 (b_2)			< 4				
	Linear spline 2 (b_3)			< 4				
	Linear spline 3 (b_4)			< 4				
	Linear spline 4 (b_5)			< 4				
Linear interval and exponential age model								
Gs	Horizontal asymptote (b_0)		.822	16.33	< .001	[.800, .845]	71.714	.002
	Test-retest interval linear slope (b_1)			< 4				
	Age scaling factor (b_2)			< 4				
	Age growth rate (b_3)		−.247		< .001			
Exponential interval and exponential age model								
Gv	Horizontal asymptote (b_0)		.692	18.355	< .001	[.621, .763]	87.727	.007
	Test-retest interval scaling factor (b_1)		−.052	30.104	< .001	[−.077, −.026]		
	Test-retest interval growth rate (b_2)		−.250		.278			
	Age scaling factor (b_3)		.001	9.233	< .001	[.001, .002]		
	Age growth rate (b_4)		−.348		.004			
Linear interval and exponential age model								
Gwm	Horizontal asymptote (b_0)		.712	4.729	< .001	[.637, .787]	86.555	.008
	Test-retest interval linear slope (b_1)		−.023	7.811	.115	[−.052, .007]		
	Age scaling factor (b_2)		.045	11.178	.010	[.013, .077]		
	Age growth rate (b_3)		−.095		.074			
H4: Cognitive Ability Captured								
Complete	Intercept		.792	118.604	< .001	[.775, .809]	98.252	.014
	Ga	*g*	−.145	4.034	.043	[−.283, −.007]		
	Gc	*g*	−.022	78.301	.181	[−.055, .010]		
	Gf	*g*	−.109	66.623	< .001	[−.146, −.072]		
	Gl	*g*		< 4				
	Gq	*g*	−.049	7.806	.360	[−.166, .068]		
	Grw	*g*		< 4				
	Gs	*g*	−.050	21.585	.018	[−.091, −.009]		

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	Gv	*g*	−.045	64.584	.007	[−.077, −.013]		
	Gwm	*g*	−.127	20.054	.002	[−.201, −.052]		
H5: General Cognitive Ability Level								
Complete	Intercept		.830	33.936	< .001	[.613, 1.047]	97.990	.013
	Cognitive ability		.000	35.163	.723	[−.002, .002]		
g	Intercept		.907	33.242	< .001	[.705, 1.110]	97.920	.008
	Cognitive ability		−.001	34.290	.326	[−.003, .001]		
Ga			Not calculated because of h < 4 samples					
Gc	Intercept		1.058	9.346	.001	[.529, 1.586]	94.848	.004
	Cognitive ability		−.002	9.608	.316	[−.008, .003]		
Gf	Intercept			< 4			93.630	.002
	Cognitive ability			< 4				
Gl			Not calculated because of h < 4 samples					
Gq			Not calculated because of h < 4 samples					
Grw			Not calculated because of h < 4 samples					
Gs	Intercept			< 4			67.568	.005
	Cognitive ability			< 4				
Gv	Intercept		.892	7.666	.002	[.424, 1.361]	85.624	.009
	Cognitive ability		−.001	7.805	.520	[−.006, .003]		
Gwm	Intercept			< 4			30.927	.001
	Cognitive ability			< 4				
H6: Test Instrument								
Complete	Intercept		.800	63.135	< .001	[.777, .824]	97.626	.011
	CFT	WISC	−.085	5.513	.019	[−.150, −.020]		
	Kuhlmann	WISC	−.004	4.835	.925	[−.110, .102]		
	Raven	WISC	−.169	19.158	< .001	[−.239, −.099]		
	Stanford_Binet	WISC	.032	30.417	.099	[−.006, .071]		
	Woodcock_Johnson	WISC	−.083	7.541	.004	[−.131, −.035]		
	Other instruments	WISC	−.040	70.712	.056	[−.080, .001]		
	Mixed instruments	WISC	−.041	31.636	.044	[−.081, −.001]		
g	Intercept		.800	63.356	< .001	[.776, .823]	97.824	.011
	Kuhlmann	WISC	−.004	4.829	.932	[−.110, .103]		
	Stanford_Binet	WISC	.033	30.337	.095	[−.006, .072]		
	Woodcock_Johnson	WISC	−.083	7.530	.004	[−.130, −.035]		
	Other instruments	WISC	−.034	77.786	.084	[−.073, .005]		
	Mixed instruments	WISC	−.036	34.486	.084	[−.077, .005]		
Ga			WISC	Not calculated because of h < 4 samples				
Gc	Intercept	WISC	.806	33.329	< .001	[.773, .839]	90.905	.003
	Woodcock_Johnson	WISC	−.014	6.951	.692	[−.098, .069]		
	Other instruments	WISC	.004	29.591	.872	[−.046, .054]		
Gf	Intercept	Raven	.634	14.181	< .001	[.575, .694]	94.240	.011

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	CFT	Raven	.085	10.281	.034	[.008, .162]		
	WISC	Raven	.110	7.263	.081	[−.017, .238]		
	Woodcock_Johnson	Raven	.056	7.398	.220	[−.042, .154]		
	Other instruments	Raven	.127	25.716	.003	[.049, .206]		
	Mixed instruments	Raven	.157	19.402	< .001	[.090, .224]		
Gl			Not calculated because of h < 4 samples					
Gq	Intercept			< 4			79.271	.002
	Other instmments	Woodcock_Johnson	−.020	5.3	.742	[−.162, .123]		
Grw			Not calculated because of h < 4 samples					
Gs	Intercept		.725	4.75	< .001	[.605, .845]	79.044	.003
	WISC	Woodcock_Johnson	−.036	8.00	.481	[−.147, .076]		
	Mixed instruments	Woodcock_Johnson	.052	9.39	.303	[−.056, .160]		
Gv	Intercept		.789	27.263	< .001	[.758, .820]	80.497	.004
	Woodcock_Johnson	WISC	−.096	5.734	.036	[−.183, −.009]		
	Other instruments	WISC	−.098	10.463	.031	[−.186, −.011]		
	Mixed instruments	WISC	.018	7.449	.468	[−.037, .074]		
Gwm	Intercept		.706	4.452	< .001	[.659, .753]	84.327	.007
	Woodcock_Johnson	Woodcock_Johnson	.034	8.434	.312	[−.038, .105]		
	Mixed instruments	Mixed instruments	−.029	9.410	.571	[−.139, .082]		
H7: Varying Measurement Instruments								
Complete	Intercept		.772	158.909	< .001	[.756, .787]	98.163	.013
	Different tests	Same test	−.074	39.847	< .001	[−.113, −.035]		
	Same test family	Same test	.007	23.334	.766	[−.043, .057]		
g	Intercept		.820	108.065	< .001	[.804, .835]	97.986	.008
	Different tests	Same test	−.104	37.385	< .001	[−.144, −.063]		
	Same test family	Same test	−.001	19.219	.970	[−.052, .050]		
Ga			Not calculated because of h < 4 samples					
Gc	Intercept		.807	54.747	< .001	[.784, .831]	92.321	.003
	Different tests	Same test	−.161	4.161	.026	[−.290, −.032]		
	Same test family	Same test	−.051	11.446	.299	[−.155, .052]		
Gf	Intercept		.721	51.928	< .001	[.694, .749]	94.556	.011
	Different tests	Same test	−.161	4.275	.002	[−.229, −.092]		
	Same test family	Same test		< 4				
Gl			Not calculated because of h < 4 samples					
Gq			Not calculated because of h < 4 samples					
Grw			Not calculated because of h < 4 samples					
Gs			Not calculated because of h < 4 samples					
Gv	Intercept		.756	47.000	< .001	[.728, .784]	87.971	.008
	Different tests	Same test		< 4				
	Same test family	Same test	.001	6.989	.983	[−.111, .113]		
Gwm	Intercept		.702	19.492	< .001	[.657, .747]	82.974	.006

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	Different tests	Same test	−.283	19.492	< .001	[−.328, −.238]		
	Same test family	Same test		< 4				
H8: Complete Test								
Complete	Intercept		.774	150.494	< .001	[.758, .790]	98.284	.014
	Not complete test	Complete test	−.042	90.905	.009	[−.073, −.010]		
g	Intercept		.813	115.258	< .001	[.797, .829]	98.321	.008
	Not complete test	Complete test	−.046	44.762	.034	[−.088, −.004]		
Ga			Not calculated because of h < 4 samples					
Gc	Intercept		.807	38.051	< .001	[.777, .837]	93.994	.004
	Not complete test	Complete test	−.035	63.596	.171	[−.087, .016]		
Gf	Intercept		.693	36.838	< .001	[.660, .726]	94.193	.012
	Not complete test	Complete test	.037	43.367	.210	[−.022, .097]		
Gl			Not calculated because of h < 4 samples					
Gq	Intercept			< 4			91.402	.004
	Not complete test	Complete test	−.015	5.43	.773	[−.141, .110]		
Grw			Not calculated because of h < 4 samples					
Gs	Intercept		.694	5.36	< .001	[.650, .739]	91.846	.008
	Not complete test	Complete test	.056	8.75	.103	[−.014, .126]		
Gv	Intercept		.754	31.295	< .001	[.713, .796]	89.695	.009
	Not complete test	Complete test	−.016	51.826	.557	[−.071, .039]		
Gwm	Intercept		.654	7.531	< .001	[.540, .767]	86.522	.009
	Not complete test	Complete test	.053	15.398	.350	[−.064, .171]		
H9: Geographic Location								
Complete	Intercept		.780	128.319	< .001	[.763, .796]	98.007	.013
	Asia	North America		< 4				
	Europe	North America	−.050	113.646	.002	[−.082, −.018]		
g	Intercept		.814	102.153	< .001	[.796, .831]	98.317	.011
	Europe	North America	−.046	58.103	.024	[−.086, −.006]		
Ga			Not calculated because of h < 4 samples					
Gc	Intercept		.802	53.404	< .001	[.775, .828]	93.122	.004
	Europe	North America	−.042	19.330	.208	[−.110, .026]		
Gf	Intercept		.724	27.554	< .001	[.686, .763]	95.390	.013
	Europe	North America	−.033	53.672	.237	[−.089, .022]		
Gl			Not calculated because of h < 4 samples					
Gq	Intercept		.781	6.42	< .001	[.714, .847]	91.112	.004
	Europe	North America	−.018	5.89	.785	[−.172, .136]		
Grw			Not calculated because of h < 4 samples					
Gs	Intercept		.716	16.10	< .001	[.661, .771]	91.312	.008
	Europe	North America	.062	11.50	.164	[−.029, .153]		
Gv	Intercept		.744	43.155	< .001	[.712, .777]	89.486	.009
	Europe	North America	.007	14.987	.845	[−.065, .078]		

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
Gwm	Intercept		.713	16.219	< .001	[.669, .757]	85.446	.008
	Europe	North America		< 4				
H10: Reliability Analyses Based on Effect Sizes with Available Reliability								
Exponential age model without adjusting for reliability								
Complete	Horizontal		.801	32.614	< .001	[.768, .834]	95.832	.007
	Age scaling factor		.009	17.639	< .001	[.005, .013]		
	Age growth rate		−.182		< .001			
Exponential age model after adjusting for reliability								
Complete	Horizontal		.899	32.392	< .001	[.864, .934]	95.303	.009
	Age scaling factor		.010	17.663	< .001	[.006, .015]		
	Age growth rate		−.180		.005			
Magnitude of ρ without adjusting for reliability								
Complete	Intercept		.761	56.826	< .001	[.735, .788]	99.289	.013
g	Intercept		.826	31.391	< .001	[.799, .852]	99.499	.008
Ga	Intercept			< 4				
Gc	Intercept		.811	17.345	< .001	[.776, .845]	88.723	.003
Gf	Intercept		.707	26.793	< .001	[.667, .748]	95.875	.028
Gl	Intercept		.639	4.973	< .001	[.568, .709]	67.240	.003
Gq	Intercept			< 4			51.352	.000
Grw	Intercept			< 4			94.549	.005
Gs	Intercept		.733	12.332	< .001	[.666, .800]	92.661	.009
Gv	Intercept		.736	12.201	< .001	[.680, .791]	82.018	.006
Gwm	Intercept		.688	9.620	< .001	[.595, .780]	89.827	.012
Magnitude of ρ after adjusting for reliability								
Complete	Intercept		.855	55.116	< .001	[.827, .884]	98.608	.009
g	Intercept		.891	30.655	< .001	[.861, .920]	99.023	.005
Ga	Intercept			< 4				.016
Gc	Intercept		.879	17.503	< .001	[.846, .913]	89.627	.004
Gf	Intercept		.815	26.522	< .001	[.765, .865]	93.775	.023
Gl	Intercept		.701	4.997	< .001	[.571, .832]	87.076	.013
Gq	Intercept		.828	4.518	< .001	[.782, .873]	73.547	.001
Grw	Intercept			< 4				.005
Gs	Intercept		.840	12.281	< .001	[.761, .919]	92.286	.012
Gv	Intercept		.850	12.156	< .001	[.791, .908]	79.133	.006
Gwm	Intercept		.777	9.562	< .001	[.681, .874]	87.445	.012
Cognitive Ability Captured without adjusting for reliability								
Complete	Intercept		.818	23.255	< .001	[.782, .855]	99.050	.013
	Ga	*g*	−.172	4.169	.029	[−.315, −.029]		
	Gc	*g*	−.006	19.648	.805	[−.058, .045]		
	Gf	*g*	−.124	41.108	< .001	[−.181, −.067]		
	Gl	*g*		< 4				

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	Gq	*g*		< 4				
	Grw	*g*	−.035	4.169	.719	[−.286, .215]		
	Gs	*g*	−.090	11.762	.023	[−.165, −.015]		
	Gv	*g*	−.062	11.977	.117	[−.141, .018]		
	Gwm	*g*	−.182	8.847	.033	[−.345, −.019]		
Cognitive Ability Captured after adjusting for reliability								
	Intercept		.891	22.262	< .001	[.850, .932]	97.297	.006
	Ga	*g*	−.166	4.146	.042	[−.323, −.009]		
	Gc	*g*	.000	18.925	.992	[−.051, .052]		
	Gf	*g*	−.073	38.394	.055	[−.147, .002]		
	Gl	*g*		< 4				
	Gq	*g*		< 4				
	Grw	*g*	−.054	4.146	.595	[−.308, .201]		
	Gs	*g*	−.055	11.322	.200	[−.144, .034]		
	Gv	*g*	−.029	11.854	.503	[−.119, .062]		
	Gwm	*g*	−.168	8.504	.073	[−.356, .020]		
Test instrument without adjusting for reliability								
Complete	Intercept		.791	10.929	< .001	[.738, .844]	98.742	.010
	CFT	WISC	−.072	9.451	.054	[−.146, .002]		
	Raven	WISC	−.177	4.525	.105	[−.409, .055]		
	Stanford_Binet	WISC		< 4				
	Woodcock_Johnson	WISC	−.067	10.591	.045	[−.132, −.002]		
	Other instruments	WISC	.008	23.670	.806	[−.062, .079]		
	Mixed instruments	WISC	−.103	3.199	.207	[−.302, .097]		
Test instrument after adjusting for reliability								
	Intercept		.857	10.830	< .001	[.800, .915]	98.480	.010
	CFT	WISC	−.054	9.422	.259	[−.154, .047]		
	Raven	WISC	−.084	4.371	.581	[−.462, .294]		
	Stanford_Binet	WISC		< 4				
	Woodcock_Johnson	WISC	−.058	10.690	.109	[−.131, .015]		
	Other instruments	WISC	.050	23.420	.141	[−.018, .117]		
	Mixed instruments	WISC	−.051	3.200	.636	[−.355, .252]		
Exploratory Analyses								
1. Simultaneous Inclusion of all Categorical Moderators								
Complete	Intercept		.830	77.592	< .001	[.807, .853]	96.224	.010
	Ga	*g*	−.119	4.346	.107	[−.278, .039]		
	Gc	*g*	−.022	86.966	.186	[−.054, .011]		
	Gf	*g*	−.062	39.666	.004	[−.102, −.021]		
	Gl	*g*	−.040	4.395	.503	[−.188, .108]		
	Gq	*g*	−.039	7.946	.380	[−.136, .058]		
	Grw	*g*		< 4				

Dataset	Predictor	Reference category	ρ	df	p	95% Confidence Interval	I^2	τ^2
	Gs	*g*	−.056	27.911	.024	[−.104, −.008]		
	Gv	*g*	−.053	67.000	.002	[−.086, −.020]		
	Gwm	*g*	−.142	23.888	.001	[−.215, −.069]		
	CFT	WISC	−.046	9.912	.141	[−.109, .018]		
	Kuhlmann	WISC	−.032	4.961	.464	[−.138, .073]		
	Raven	WISC	−.127	36.833	.003	[−.208, −.046]		
	Stanford_Binet	WISC	.019	30.594	.298	[−.017, .054]		
	Woodcock_Johnson	WISC	−.037	11.294	.259	[−.106, .032]		
	Other instruments	WISC	−.030	66.848	.153	[−.072, .011]		
	Mixed instruments	WISC	.017	33.795	.486	[−.032, .065]		
	Different tests	Same test	−.114	32.901	< .001	[−.167, −.061]		
	Same test family	Same test	−.021	26.847	.338	[−.065, .023]		
	Not complete test	Complete test	−.027	61.537	.189	[−.068, .014]		
	Asia	North America		< 4				
	Europe	North America	−.021	98.267	.153	[−.050, .008]		
2. Exponential age model based on age homogenous samples								
Complete	Horizontal		.793	97.100	< .001	[.772, .815]	93.706	.007
	Age scaling factor		.004	45.200	< .001	[.003, .005]		
	Age growth rate		−.230		< .001			

Note. Complete = Complete dataset including *g* and CHC broad abilities. As we subtracted 5 from the test-retest interval, the intercepts in models including test-retest interval represented a 5-year interval instead of a 0-year interval. As we subtracted 20 from the age, the intercepts in models including age represented a 20-year age instead of a 0-year age.

Table 4

Magnitude of Rank-Order Stability in Cognitive Ability for a Sample Age of 20 Years and a Test-Retest Interval of Five Years

Dataset	Predictor	ρ	*df*	*p*	95% CI	I^2	τ^2
Magnitude of ρ							
Complete	Intercept	.762	197.665	< .001	[.748, .776]	98.661	.015
g	Intercept	.801	142.932	< .001	[.786, .817]	98.738	.010
Ga	Intercept	.651	3.996	< .001	[.512, .789]	90.373	.013
Gc	Intercept	.791	67.290	< .001	[.766, .816]	94.154	.003
Gf	Intercept	.708	56.745	< .001	[.680, .735]	95.602	.012
Gl	Intercept	.688	6.990	< .001	[.591, .786]	95.529	.025
Gq	Intercept	.770	10.005	< .001	[.711, .829]	93.647	.004
Grw	Intercept	.776	5.897	< .001	[.636, .915]	96.280	.008
Gs	Intercept	.738	24.135	< .001	[.694, .781]	92.028	.008
Gv	Intercept	.747	55.570	< .001	[.719, .775]	90.031	.009
Gwm	Intercept	.687	20.871	< .001	[.636, .739]	87.504	.009

Note. Parameters were estimated based on random effects intercept-only models. Prior to the analyses, test-retest and age effects were residualized out from the effect sizes. Therefore, in each model, the intercept can be interpreted as the magnitude of rank-order stability for a sample age of 20 years and a test-retest interval of five years. Complete = Complete dataset including *g* and CHC broad abilities.

Table 5.

Comprehensive Meta-Analyses on the Rank-Order Stability of Psychological Constructs

Reference	Construct	ρ	*h*	*M* age in years	*M* interval in years	Statistically Significant Moderators
Roberts & DelVecchio, 2000	Personality Traits	.35-.75[a]	175	17.84	6.75	age, interval, type of trait
Trzesniewski et al., 2003	Self-Esteem	.50	168	14.00	4.90	age, interval, year of assessment
Low et al., 2005	Vocational Interests	.55-.83[a]	148	18.00	7.06	age, interval, type of coefficient, cohort, scale generality, interest classification
Jin & Rounds, 2012	Work Values	.62	28	≈20.50	2.41	generation
Scherrer & Preckel, 2019	Motivation Related Construct	.51	65	school age	≈1.65	N.A.
Mund et al., 2020	Loneliness	.36-.78[a]	76	N.A.	2.65	age, interval, measurement instrument
Bleidorn et al., 2022	Personality Traits	.61	189	22.66	4.88	age, interval, trait, reliability
Breit et al., 2023 (current meta-analysis)	Cognitive Abilities	.77	205	20.00[b]	5.00[b]	age, interval, cognitive ability captured, test instrument, geographic location

Note. ρ = effect size; *h* = number of samples.

[a] ranges are presented when no average effect over all constructs was reported.

[b] these values do not represent the exact averages but the age and interval values for which the effect size ρ was calculated.