Published in final edited form as:
J Exp Psychol Gen. 2024 January ; 153(1): 70–85. doi:10.1037/xge0001471.

Associations between executive functions assessed in different contexts in a genetically informative sample

Samantha M. Freis[1,2], **Jordan D. Alexander**[3], **Jacob E. Anderson**[3], **Robin P. Corley**[1,2], **Alejandro I. De La Vega**[4], **Daniel E. Gustavson**[1], **Scott I. Vrieze**[1,3], **Naomi P. Friedman**[1,2], **University of Colorado Boulder**

[1]Institute for Behavioral Genetics, University of Colorado Boulder

[2]Department of Psychology and Neuroscience, University of Colorado Boulder

[3]Department of Psychology, University of Minnesota

[4]Department of Psychology, The University of Texas at Austin

Abstract

Executive functions (EFs) are cognitive functions that help direct goal-related behavior. EFs are usually measured via behavioral tasks assessed in highly controlled laboratory settings under the supervision of a research assistant. Online versions of EF tasks are an increasingly popular alternative to in-lab testing. However, researchers do not have the same control over the testing environment during online EF assessments. To assess the extent to which EFs assessed in-lab and online are related, we used data from the Colorado Online Twin Study (CoTwins; 887 individual twins aged 13.98–19.05) and constructed an Lab Common EF factor and an Online Common EF factor from four EF tasks assessed in-lab and online. The In-lab Common and Online Common EF factors were genetically identical (rA = 1.00) but phenotypically separable (r = .77, 95% CI [.59, .94]) indicating that these EF factors have the same genetic underpinnings but may be differentially influenced by environmental factors. We examined phenotypic, genetic, and environmental correlations between the EF factors and a general cognitive ability factor (g) assessed in the lab and found similar relationships between In-lab Common EF and g and Online Common EF and g. Overall, these results suggest that Common EF factors assessed in different contexts are highly related to each other and similarly related to other cognitive outcomes. These findings indicate that online task-based EF assessments could be a viable strategy for increasing sample sizes in large-scale studies, particularly genetically informed studies.

Keywords

Cognitive control; executive control; heritability; environmental contexts

[*]Correspondence: Samantha M. Freis, Samantha.freis@colorado.edu.

The authors have no conflicts of interest to declare. Deidentified CoTwins data is in the process of being deposited on the Inter-University Consortium for Political and Social Research/ National Addiction and HIV Data Archive Program (ICPSR/NAHDAP) at the University of Michigan and will be available to qualified investigators. All stimulus materials are available upon request. This study was not preregistered. Additionally, de-identified cognitive task data after trimming and transformation and Mplus analysis code for the phenotypic and genetic models presented in the results is available at the following OSF web page: https://osf.io/vsxb2/.

Executive functions (EFs) are cognitive functions that help individuals control their thoughts and actions during goal-directed behavior (Friedman & Miyake, 2017). EFs relate to numerous important psychological and behavioral outcomes across the lifespan, including other cognitive functions, academic performance, and mental health (Diamond, 2013; Freis, et al., 2022; Friedman et al., 2020; Malanchini et al., 2019). They are typically assessed with behavioral tasks administered in carefully controlled laboratory settings, which are relatively time-consuming to collect, limiting sample sizes. Not surprisingly, online EF tasks have become increasingly popular (e.g., Enkavi et al., 2019, Bycroft et al., 2018), dramatically expanding the availability of EF task data. However, in-lab and online EF tasks can differ substantially, most notably because participants may complete online tasks in a wide range of contexts over which the experimenter has little control. Thus, it is unclear how closely they are related, although researchers using online EF tasks may often assume they tap the same variance as in-lab tasks. Although correlations between in-lab and online EF tasks have been reported in small clinical samples (Moore et al., 2020; Shvetz et al., 2021), such correlations can be attenuated by unreliability and task impurity, meaning that they do not permit clear tests of whether online tasks may assess different variance. Using data from a large twin sample ($N = 887$), we examined relationships between in-lab and online tasks at the level of latent variables ("Common EF" factors), which reduce the influence of measurement error and task impurity (Friedman & Miyake, 2017). Moreover, because the sample is genetically informative, we have the unique ability to evaluate the hypothesis that in-lab and online EF tap similar genetic variance but differ in environmental influences.

Technological Advances in EF Assessments

EFs are complex cognitive functions that are related to numerous psychological and behavioral outcomes, ranging from other cognitive functions and academic achievement to behavioral problems and psychopathology across the lifespan (Friedman & Miyake, 2017). Therefore, studying EFs is of interest to multiple areas of psychology, and making EF assessments more accessible and easier to administer could advance our understanding of numerous psychological outcomes. Assessing EFs remotely with online tasks continues to grow in popularity, as technological advances can increase speed, efficiency, and sample sizes in data collection (Ahmed et al., 2022; Miller & Barr, 2017). For example, EF tasks are frequently used as part of neuropsychological testing in clinical settings, and some researchers have emphasized the promise of remote EF and cognitive assessments (Bilder & Reise, 2019; Lavigne et al., 2022; Miller & Barr, 2017). Additionally, the COVID-19 pandemic has emphasized the importance of remote psychological assessments and tasks. Remote EF tasks could be useful not only in clinical settings but also in large cohort studies or psychological studies more broadly. Finally, the speed and ease of assessment of online testing could provide researchers with the opportunity to collect multiple EF assesments instead of relying on individual tasks.

Though some traditional EF tasks adapted to be administered online replicate expected effects (like the Stroop effect) and display similar reliability to their in-lab counterparts (Crump et al., 2013), it remains unclear if EFs assessed remotely via online tasks and EFs assessed in-lab are strongly correlated. Though research has shown that performance on online general cognitive ability tasks are related to performance on in-lab tasks (Malanchini et al., 2021), research specific to relationships between EF tasks assessed in different contexts is limited. There are several methods of collecting EF task data remotely. Some researchers have argued that remote assessment of EFs using traditional tasks administered by a research assistant via video conferencing are useful and valid (Ahmed et al., 2022). This supervised method may be particularly useful for younger samples (Ahmed et al., 2022). However, computerized unsupervised remote EF assessments are less labor intensive to collect. In two small samples, researchers reported moderate correlations between traditional in-lab EF tasks and EF tasks administered online and unsupervised at home in their clinical and control groups (Moore et al., 2020; Shvetz et al., 2021). However, due to their unreliability and task impurity, relying exclusively on relationships between individual EF tasks does not allow clear tests of whether tasks assessed in different contexts may measure different abilities.

Latent Variable Models of EFs

Many studies of EFs have exclusively used single EF tasks, which can be problematic for the interpretation of individual differences (Miyake et al., 2000). Theory-driven methods like confirmatory factor analysis extract commonalities across different EF tasks, such as latent variables of response inhibition, updating working memory, and shifting between task sets (Miyake et al., 2000). One well replicated framework of EFs that uses latent variable modeling is the unity/diversity framework (Friedman & Miyake, 2017). This framework decomposes variance in EF tasks into a Common EF factor that influences performance on all tasks, and orthogonal specific factors (Updating-Specific and Shifting-Specific) that additionally influence performance on updating and shifting tasks. We use this framework in the current study to focus on the similarity of the Common EF factor across assessment contexts. This factor, which is proposed to capture individual differences in goal construction and maintenance and the use of those goals to bias ongoing processing (Friedman & Miyake, 2017), is the EF factor that is generally most closely related to self-regulation and behavior problems (Friedman & Miyake, 2017).

Latent variables are unmeasured variables that are estimated by extracting shared variance between measured variables that are theoretically influenced by an underlying construct (Kline, 2016). Latent variable modeling is advantageous when studying EFs due to the task impurity problem: EFs involve controlling other cognitive processes, so EF tasks must include non-executive processes that may influence individual differences in task performance (Friedman & Miyake, 2017; Miyake et al., 2000). For example, the classic color/word Stroop task requires controlling prepotent responses (word reading) but also requires visually processing colors and vocally or manually responding, so individual differences in these other processes may contribute to variance in performance in addition to the control processes of interest. The task impurity problem, along with low reliability, contributes to low correlations between individual EF tasks (Miyake et al., 2000). Since latent variable modeling extracts shared variance across measures, when those measures vary in non-executive processes, the shared variance excludes these measure-specific factors as well as random measurement error (Miyake et al., 2000). Thus, if latent variables for in-lab and online EF factors assess the same ability, like Common EF, we would expect them to correlate highly (i.e., not significantly lower than 1.0), even if correlations between individual EF tasks assessed in different contexts are lower than unity. To our knowledge, no study has examined latent variable models of EFs assessed both in-lab and online to examine how related they are after eliminating measure-specific factors and random measurement error.

Using Twin Data to Inform Etiological Differences between EFs Assessed in Different Contexts

As reviewed in previous sections, it is unclear the extent to which EFs assessed in different contexts are related, and there are several possible reasons why EFs assessed in different contexts may not be perfectly correlated. Twin and family studies provide the unique opportunity to examine genetic and environmental influences on an outcome of interest and to examine why two outcomes are related: i.e, through genetic and environmental correlations (Rijsdijk & Sham, 2002). Twin studies decompose the phenotypic variance in an outcome into variation attributable to additive genetic (also known as heritability), shared environmental (environmental factors that make siblings more alike), and nonshared environmental (environmental factors unique to each sibling) influences. Notably, estimates of genetic and environmental influences from twin models represent *proportions of variance* accounted for by genetic and environmental differences in a population rather than specific genes or environmental factors. Therefore, these estimates are specific to the population being studied at a particular age, at a particular time, and under particular environmental circumstances.

Heritability estimates can change over time. For example, the heritability of height increased across Finnish birth cohorts born in the first half of the 20th century (Silventoinen et al., 2000). It would be improbable that the specific genes influencing height changed in this short period of time, so the increased proportion of variance in height explained by genes is likely due to environmental changes like improvements in nutrition and the standard of living. While that example covers a range of generations, similar changes can be seen within one generation across time: The same trait measured in the same sample at different points in development can display different patterns of genetic and environmental influence. For example, Common EF is developmentally stable, and in one twin study, that stability was due almost entirely to high genetic correlations across late adolescence and early adulthood (Friedman et al., 2016). However, Common EF factors at ages 17 and 23 were not perfectly phenotypically correlated, and this change in Common EF was explained by small but significant nonshared environmental influences that were unique to age 23. Finally, two traits, like depression and anxiety, can be perfectly genetically correlated but be influenced by different environmental factors (Kendler, 1996). Overall, these examples demonstrate the utility of using twin studies to examine relationships between the same or different traits at multiple levels of analysis. However, no genetically informed study has examined latent variables of EFs assessed both in-lab and online. Here we review several possible explanations for correlations less than unity between EFs assessed in different contexts and how the twin method would inform the etiology of those differences.

First, EF tasks assessed in different contexts could show low to moderate correlations due to minor measurement differences and random noise. If that is the case, latent variables of those EF factors should be perfectly correlated for both genetic and nonshared environmental influences, as latent variables eliminate random measurement error.

Second, EF factors derived from tasks assessed in different contexts could be separable because of differential influence from environmental factors. For example, lab-based EF tasks are typically administered in highly controlled settings under the supervision of research staff. Therefore, EF tasks assessed in the lab may assess performance under optimal conditions with minimal environmental variation (Friedman & Gustavson, 2022; Toplak et al., 2013; Wennerhold & Friese, 2020). Somewhat consistent with the low environmental variation hypothesis, genetic models of latent variables of EFs (measured in the lab) typically show low estimates of environmental influence at multiple developmental stages (Engelhardt et al., 2015; Friedman et al., 2016; Friedman & Miyake, 2017). This feature may limit the generalizability of in-lab EF performance to settings in the real world, where control is required in the face of distraction (for example, remembering to stop at the grocery store on the way home from work). Meanwhile, online EF tasks could measure performance in more distracting and diverse settings, as researchers do not control the environment in which the participants complete them. If EF tasks administered online are substantially influenced by home environmental differences, one would expect to observe higher estimates of environmental variance and lower estimates of genetic variance for EF assessed online, nonshared environmental influences to be uncorrelated, and phenotypic correlations between EFs assessed in different contexts less than unity. However, if environmental influences primarily explain the lack of unity, we would still expect the genetic variance for online EF tasks to correlate with the genetic variance for in-lab EF tasks. That is, the same genetic influences might affect in-lab and online EFs, but they do not correlate perfectly phenotypically because online EF has environmental influences that are uncorrelated with the environmental influences for in-lab EF.

Finally, there may be fundamental differences beyond environmental variation between EF assessed in the lab and EF in action in a more real-world setting. This possibility, along with other issues like ratings being influenced by social desirability bias and poor insight into one’s own control abilities, has been raised as one potential explanation for the low correlations observed between task-based and ratings-based EF, cognitive control, and self-regulation assessments (see Friedman & Gustavson, 2022, for more discussion). Though we are focusing only on task-based measures here, this hypothesis is still relevant. If there is something fundamentally different about how people execute EF in the lab versus real-world settings, and these tasks are picking up on those differences, one would expect to observe divergent genetic correlations, nonshared environmental influences to be uncorrelated, and phenotypic correlations between EFs assessed in different contexts to be less than unity.

If online and in-lab EF factors are separable, a natural question is whether one is more valid than the other. One way to address this question is with a measure that can indicate criterion validity. Latent variables of EFs assessed in-lab demonstrate a range of relationships with other cognitive functions, like general cognitive ability (*g*) or intelligence (IQ). In adulthood, EFs are phenotypically correlated with, but distinct from, *g* or IQ, and these correlations vary between EF factors (Blair, 2006; Friedman et al., 2006; Gustavson et al., 2022; Kane & Engle, 2002). Genetically informed studies have identified different patterns of genetic and environmental influence on EFs and *g* or IQ (Engelhardt et al., 2015; Freis et al., 2022; Friedman et al., 2016; Gustavson et al., 2022; Haworth et al., 2010; Tucker-Drob et al., 2013). This consistent pattern of commonality but separability between EFs and *g* makes *g* a useful criterion variable in methodological studies of EFs. Therefore, to the extent that online and in-lab EF factors differ, it will be important to compare both factors with measures of *g* or IQ, as this may shed light on the (potentially) different sets of individual differences captured by both sets of measures.

Present Study

Here, we analyze data (*N* = 887 individual twins) from the Colorado Online Twin Study (CoTwins) to address these issues. CoTwins is an adolescent twin study primarily focused on answering research questions related to substance use, mental-health, cognition, personality, and behavior by using traditional in-lab assessments followed by remote psychological assessments in an intensive longitudinal design. The CoTwins study protocol began with participants coming into the lab for a baseline appointment to complete several tasks and surveys. During the appointment, participants also provided their emails and signed up for the CoTwins App so we could deploy follow up remote online tasks and surveys longitudinally.

The present study took advantage of our unique adolescent twin sample with both in-lab and online remote EF assessments to examine three main questions: 1) How correlated are latent variables derived from in-lab and online EF tasks? 2) If latent variables derived from in-lab and online EF tasks are not perfectly correlated, is their separability due to different genetic and/or environmental influences? and 3) If we find that EFs assessed in different contexts are separable, are they also differentially related to a criterion variable, *g*?

Method

Sample

Data were obtained as part of the CoTwins study, which is a joint project led by investigators at the University of Colorado Boulder and the University of Minnesota. Participants were identified from birth records provided by the Colorado Department of Health and Human Services as part of the Colorado Twin Registry (Rhea et al., 2012). CoTwins included 887 adolescent twins recruited while they were in high school in the state of Colorado (ages 13.98–19.05, *M* age 16.28, *SD* = 1.22, 478 female, 409 male, and 0 other, as indicated by the multiple-choice self-report question “I identify my sex as...”). 63.02% had at least one parent with a bachelor’s degree or higher and 48.82% had an annual household income of $100,000 or more. Participants self-reported race and ethnicity (Hispanic/Latino, Non-Hispanic/Latino, White, American Indian/Alaskan Native, Black or African American, Asian, and Native Hawaiian or other Pacific Islander) and were able to select multiple identities. 15.56% of the sample identified as Hispanic/Latino. 80.61% of the sample identified as White, 1.13% of the sample identified as American Indian/Alaskan Native, 1.13% of the sample identified as Black or African American, 0.45% of the sample identified as Asian, 0.23% of the sample identified as Native Hawaiian or other Pacific Islander, 10.15% of the sample identified as more than one race, and 6.31% of the sample declined to report race.

CoTwins recruited participants for the in-lab baseline protocol in 2 waves. Wave 1 (*n* = 670 adolescents; *M* age at lab visit = 16.03, *SD* = 1.13) ran from 2015–2016 (see Alexander et al., 2022). Wave 2 ran from 2018–2022 and included 217 new participants. The second wave of data collection focused on recruiting older adolescents (*M* age at lab visit = 17.57, *SD* = .65). Additionally, wave 2 included 510 participants re-enrolled from wave 1 for the online data collection protocol, but these participants did not come back into the lab for re-assessment. Figure 1 displays the in-lab and online data collection protocols and the age-by-wave distributions in the CoTwins sample.

Twin pairs were categorized as either monozygotic (MZ, identical) or dizygotic (DZ, fraternal). Zygosity classification was determined by two independent testers who assessed the twins on using a five-point similarity scale across six physical traits, including characteristics like hair color and texture, eye color, and complexion. Testers discussed any major discrepancies in similarity ratings and reached a consensus on zygosity classification. The sample contained 167 MZ pairs and 277 DZ pairs (189 same-sex DZ and 88 opposite sex DZ). There were 444 total pairs; however, one MZ co-twin did not participate, leading to 887 total participants.

All research protocols were reviewed and approved by the University of Minnesota and Colorado’s Institutional Review Board. Parental permission and informed consent or adolescent assent were obtained from each participant or parent.

Measures

The CoTwins EF task battery included three in-lab tasks (antisaccade, keep track, and category-switch) and three online tasks (Stroop, keep track, and category-switch), the latter two of which were administered multiple times. All tasks were computerized. Task stimuli were counterbalanced and randomized, and the stimuli order within each task was identical for all participants. All tasks contained practice trials to confirm that the participants understood the task instructions, as well as warm-up trials (not analyzed) at the beginning of blocks.

Antisaccade (adapted from Roberts et al., 1994) was only administered in the lab. This response inhibition task requires participants to restrain a prepotent eye movement to a visual cue and instead look in the opposite direction to see a quickly presented target stimulus. The timing of the task was speeded so that it would be impossible for the participant to see the target if they first looked at the cue. Participants were seated 18 inches from the computer monitor. Each trial started with a centered fixation cross that appeared for one of nine durations from 1,500 to 3,500 millisecond (ms) in 250-ms increments. When the fixation cross disappeared from the screen a cue (a black 1/8 in. square, inner edge 3.375 in. from the center), appeared on the right or left side of the screen with equal probability for 250 ms. Once the cue disappeared, a numeric target (a digit 1–9, 26-point Helvetica font, presented in a 7/16-in. square with its inner edge 3.25 in. from the fixation) appeared for

150 ms before being masked, on the same side as the cue for the prosaccade practice block and the opposite side for the antisaccade blocks. The mask remained on the screen until the participant verbalized the target number (or guessed). The experimenter then entered the response, which initiated the subsequent trial. The participants practiced on 12 prosaccade trials to familiarize them with the procedure, followed by 12 practice antisaccade trials. They then received the target 36 antisaccade trials (preceded by 2 warm-up antisaccade trials that were not analyzed). The dependent measure was the proportion of correct responses out of 36 antisaccade trials.

Stroop (based on Stroop, 1935) was only administered online. This interference control/ response inhibition task requires participants to avoid the prepotent tendency to read color words (RED, GREEN, or BLUE) and instead indicate the color the word is printed in (red, green, or blue). The task was presented on a white background. Each trial began with a black fixation cross that appeared in the center of the screen for 500 ms, followed by a 500 ms blank screen, followed by the colored stimulus, which remained on the screen until the participant indicated its color by pressing the r, g, or b key on the keyboard, or until 1,500 ms. If the participant did not respond within 1,500 ms, the stimulus disappeared and “Respond Faster!” (black font) appeared in its place for 250 ms before the next trial started. The task began with a block of 42 neutral trials (colored strings of 3-to-5 dollar signs), followed by 3 blocks of 54 mixed trials (18 neutral trials and color words printed in the congruent [18 trials] or incongruent [18 trials] colors). Each target block incorporated 2 additional warm-up trials. At least 10 neutral practice trials preceded the first neutral block, and at least 12 mixed practice trials preceded the first mixed block (participants completed as many practice blocks as necessary to reach 80% accuracy in the practice trials before moving on to the target trials). In the practice blocks only, participants received feedback on each trial: “Correct!” or “Incorrect!” appeared below the stimulus for 250ms after each response. The dependent measure was the difference in mean RTs for correct responses on incongruent trials versus neutral trials in the mixed blocks.

Before the Stroop task, participants completed an online colorblind screening, using 10 plates from the Ishihara test (Ishihara, 1917). Each plate consisted of a number formed from colored dots that was embedded within a circular background of contrasting colored dots (e.g., red number and green background). On each trial, one plate appeared on the screen and participants had to type the number in an answer box. There were 7 red/green plates and 3 red/gray (control) plates. If participants responded incorrectly on more than 2 red/green plates, they were flagged as color-blind, and the Stroop task was not collected.

Keep track (based on Yntema, 1963) was administered both in the lab and online. This working memory updating task requires participants to keep track of a series of words in 2–5 relevant categories from a stream of words in 6 categories (animals, colors, countries, distances, metals, and relatives). In each trial of the lab version, participants first saw a screen instructing them “Recall the last item from the following sets:” in the center of the screen and “Press the space bar to continue” at the bottom of the screen. They pressed the space bar when ready to view the categories, at which point the central instruction disappeared and the relevant category names appeared horizontally arranged on a single line about 2/3 down on the screen, in the same location they remained throughout the trial. They pressed the space bar when they were ready to start the trial, after which they saw “GET READY” in 30pt pink font in the center of the screen for 2,000 ms, then saw a list of 15 words that appeared in 45pt font in the center of the screen for 2,000 ms each. At the end of the trial, the category list disappeared from the bottom of the screen and a pink “???” appeared in the center of the screen, signaling the participants to tell the experimenter the last word in each of the target categories. The words were drawn from a list of 6 from each category, and the complete list of 36 possible words and which categories they belonged to were included in the initial task instructions. The in-lab keep track contained 2 practice trials with 3 categories each, then 10 target trials: 4 trials with 3 categories and 6 trials with 4 categories.

Keep track was administered twice online. Slight changes were made between the online keep track assessments to improve the task. The first online keep track (administered during wave 1) was shortened considerably compared to the in-lab version to minimize participant burden. It contained 1 practice trial with 2 categories, followed by 5 target trials: 1 trial with 3 categories followed by interleaved 2 trials with 4 categories, and 2 trials with 5 categories. Participants indicated their answers by selecting them with the mouse from the list of 36 possible words on the screen. The dependent measure was the proportion of 21 total words recalled across all trials. The second online keep track (administered during wave 2) was lengthened and adjusted to require recall instead of recognition of the answers, bringing it more in line with the lab version. It contained 2 practice trials with 3 categories and 9 target trials: 3 3-word trials, 3 4-word trials, and 3 5-word trials. Participants indicated their responses by typing them in a blank entry box at the end (answers were screened for spelling errors and typos before scoring). The dependent measure for the in-lab and online keep track tasks was the proportion of total words remembered across all trials (out of 36 for the in-lab and second online versions, and out of 21 for the first online version).

Category-switch (Mayr & Kliegl, 2000) was administered both in the lab and online. This mental set shifting task requires participants to switch between two word categorization rules based on a random cue. In each trial in the lab version, participants first saw a cue symbol (heart or crossed arrows), which indicated the categorization rule (living/nonliving vs. smaller/larger than a soccer ball, respectively). After 350 ms a word printed in 36pt courier font appeared under the cue. The word remained on the screen until the participant categorized it with one of two button box buttons, and an error buzz sounded after incorrect responses. There was a 350 ms inter-trial interval. The complete list of words (alligator, snowflake, bicycle, mushroom, cloud, goldfish, lizard, table, marble, shark, knob, lion, pebble, sparrow, coat, or oak) and their correct categorizations were provided in the instructions. The task started with 2 practice 12-trial single-task blocks (living/nonliving then size), a practice 24-trial mixed block, then 2 mixed blocks comprised of 64 trials (plus 4 warm-up trials) each, in which the two rules were mixed and half the trials required switching rules.

As with the keep track task, category switch was administered twice online (once in wave 1 and once in wave 2), and between the online category-switch assessments, slight changes were made to improve the task. In the first online category-switch task, the word remained on the screen until the participant categorized it with one of two keys (f or j) on the keyboard, and errors were indicated with a red circle. In the second online category switch task, there was no error feedback, but if participants did not make a response within 2,500 ms, the task proceeded to the next trial. Additionally, for the second version of the online task, participants re-did the practice blocks until they got 80% correct before completing the target blocks. Both online category switch versions started with a practice 12-trial single-task living/nonliving block, then a 32-trial (plus 2 warm-up trials) living/nonliving block. Participants then completed a practice 12-trial size block and a 32-trial (plus 2 warm-up trials) size block. (The additional single-task blocks were included in the online versions to give participants additional practice with the response mappings, given that no experimenter was present to monitor whether they had mastered them.) Finally, participants completed a practice 24-trial mixed block, then 2 mixed blocks comprised of 54 trials (plus 2 warm-up trials) each in the first online version, or 64 trials (plus 4 warm-up trials) each in the second online version. The dependent measure for all the in-lab and online category switch tasks was the difference between average RTs for switch trials and repeat trials in the mixed blocks (the local switch-cost).

General Cognitive Ability was measured during the in-person lab visit with a block design task and a vocabulary task from the Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II) (Wechsler, 1999). *Block design* requires participants to recreate a pattern displayed in front of them with a set of blocks at quickly as possible. Block design consisted of a maximum of 13 trials and participants discontinued after 2 consecutive responses that were either incorrect or over the time limit. *Vocabulary* requires participants to define a series of increasingly difficult words in a verbal free response format. Vocabulary consisted of a maximum of 31 trials and participants discontinued after 3 consecutive incorrect responses.

Procedure

In-Lab.—When they were recruited into the study, participants first came into the lab to complete the in-lab EF and general cognitive ability tasks and a larger battery of individual differences measures including interviews, questionnaires, and the Peabody Individual Achievement Test (PIAT; Markwardt, 1997) to check reading ability and ensure they could understand the written task instructions. A small number (n = 55) of participants came into the lab after completing some online assessments because their in-lab session was delayed by the COVID-19 pandemic.

Each twin was tested individually. The appointment began with the completion of parental consent and adolescent assent. Participants then provided their email addresses, registered their smart phones with the CoTwins App, and provided salivary samples for later genotyping efforts. Participants then completed surveys and interviews to assess substance use and psychopathology. During the first half of the appointment, participants completed the category-switch, antisaccade, and keep track tasks in that order. The in-lab tasks were presented via PsyScope X B51 (Cohen et al., 1993) on a Macintosh laptop with a 15-inch screen size and 1680×1050 screen resolution, and participants responded verbally (antisaccade and keep track), or via a ms-accurate button box (category-switch). In addition to written instructions presented during the tasks, the experimenter verbally explained the tasks and instructions, checked that subjects understood the instructions, and monitored participants to ensure that they were completing the tasks correctly.

Online.—Participants provided their email addresses to complete follow-up online surveys and EF tasks. The CoTwins App was used for longitudinal surveys, which are not included in the present study. Participants were later sent a link to complete the online EF tasks at home. The participants were instructed to complete the online tasks on a computer with a physical keyboard in their home in a quiet room. The online tasks were programmed to not administer if the user attempted to complete them on an AndroidOS or iOS phone device. The experimental design was to assess EFs every 9–12 months; however, the second online assessment we use here was collected during Wave 2 and took some time to be implemented, so the average interval between the online assessments was longer than 1 year. The mean time between the in-lab assessments and the first online assessment was 1.12 years for keep track, 1.17 years for category-switch, and 2.8 years for Stroop. The mean time between the repeated online assessments was 3.08 years for keep track and 3.02 years for category-switch. The first versions of the online tasks were programmed in JavaScript, and the second versions of the online tasks were programmed with the jsPsych framework (de Leeuw, 2015), which is a JavaScript library for creating and administering behavioral experiments via web browsers.

Here, we analyze the in-lab data for all participants and include the online EF data from the wave 1 versions of category-switch and keep track, as well as the wave 2 versions of category-switch and keep track plus the Stroop task which was implemented in wave 2. Therefore, some participants only received the first or second version of the online tasks, but the wave 2 re-recruits received both versions of the online tasks. For both waves, there were no changes to the in-lab EF tasks or the general cognitive ability tasks.

Statistical Procedures

Data trimming and transformation.—Results from the CoTwins EF data have not previously been published; however, data trimming and transformation procedures were matched to those used in Friedman et al. (2016). We screened completed task data for chance-level accuracy. Chance performance for each task was calculated with an accuracy cut-off based on the binomial probability ($p < .01$) that the participant would have achieved that score by chance. We removed 13 lab antisaccade scores, 4 lab category-switch scores, 0 lab keep track scores, 24 scores from the first version of the online keep track task, 7 scores from the second version of the online keep track task, 0 scores from the first version of the online category-switch, 8 scores from the second version of the online category-switch, and 0 scores from the online Stroop for falling below their respective task chance-level accuracy cut-offs. In addition, for the online category-switch tasks, 3 scores were removed for not understanding or following the task instructions (as indicated by accuracy patterns suggesting that the participants did not switch task sets). For the online Stroop task, colorblind participants were disqualified from completing the task ($n = 15$, 40% male).

For the RT-based tasks, we applied a within-subject trimming procedure (Wilcox & Keselman, 2003) within conditions prior to averaging RTs; this procedure yields measures of central tendency that are robust to non-normality of the RT distributions. For all EF tasks, we also implemented between-subjects trimming to reduce the influence of extreme scores while maintaining them in the distribution: We replaced scores more than 3 standard deviations above or below the mean with a value exactly 3 standard deviations above or below the mean (36 scores replaced across all tasks; *n* trimmed for each task: lab antisaccade = 0; lab category-switch = 9, lab keep track = 12, first version of online category-switch = 8, second version of online category-switch = 4, first version of online keep track = 0, second version of online keep track = 2, online Stroop = 1). Finally, for the keep track and antisaccade task accuracy scores, we used the arcsine of the proportion correct because the overall accuracy showed evidence of ceiling effects (Friedman et al., 2016).

Model estimation.—The CoTwins EF battery was designed to examine the unity and diversity of EFs (Friedman & Miyake, 2017), a framework that decomposes variance in EF tasks into specific factors (typically Updating-Specific and Shifting-Specific) that influence ability on updating and shifting tasks, as well as a Common EF factor that influences ability on all tasks. In these models, because we use different versions of the same task in the lab and online, we refer to these specific factors as Keep track-specific and Category switch-specific.

All models were estimated in Mplus version 8 (Muthén & Muthén, 1998–2017). Age and sex were regressed out of all cognitive tasks, and we used the standardized residuals in all models. To account for non-independence of twin pairs we used Mplus' TYPE = COMPLEX for phenotypic analyses; this option provides model chi-squares and standard errors adjusted for nonindependence with a sandwich estimator.

Chi-square tests of model fit are sensitive to sample size, so we also assessed model fit with root mean square error of approximation (RMSEA) < .06 and confirmatory fit index (CFI) > .95 (Hu & Bentler, 1998). The statistical significance of parameter estimates for the phenotypic models was assessed with the Wald tests in the model output. For twin models, these tests are not invariant to model parameterization (Neale et al., 1989), so the significance of variance components and their covariances in the twin models was assessed using chi-square difference tests, where a significant *p*-value ($p < .05$) indicates a significant reduction in model fit (i.e., that a significant parameter was dropped).

Twin analyses.—The classical twin design (Rijsdijk & Sham, 2002) uses information from MZ and DZ twin pairs to partition the phenotypic variance of an outcome into additive genetic (A), shared environmental (C), and nonshared environmental (E) influences. C includes factors that lead siblings to correlate (e.g., neighborhood factors and socioeconomic status). In contrast, E includes factors that lead siblings not to correlate (e.g., differential responses to parental treatment, different peer groups, different extracurricular activities). MZ and DZ twins share their familial environments, but MZ twins share 100% of their segregating genes, while DZ twins, like non-twin siblings, share on average 50% of their segregating genes. The structural equation twin model formalizes these biometrical genetic principles to partition the variance of a phenotype into latent A, C, and E components. In this model, the A components correlate at 1.0 in MZ twins, and they correlate at 0.5 in DZ twins (consistent with the average proportion of shared segregating genes for MZ and DZ twin pairs). The C components correlate at 1.0 in MZ and DZ twins (because both types of twin pairs are reared together). Finally, the E components do not correlate (because they represent environmental influences unique to each twin).

Genetic influences (A) are suggested when MZ twin pairs display a higher correlation for a phenotype compared to DZ twin pairs. Shared environmental (C) influences are suggested when the DZ correlation is greater than half the MZ correlation. Nonshared environmental influences (E) are those that lead twins to differ; so, they are suggested when the MZ correlation is less than 1. E also includes random measurement error for manifest variables, as error would lead twins to be uncorrelated; however, E for latent variables does not include error, because latent variables include only reliable (correlated across tasks) variance. The twin model can be extended to multiple traits to estimate genetic (*r*A) and environmental (*r*C and *r*E) correlations (e.g., between Common EF and *g*) by allowing the A, C, and E variables to correlate across phenotypes.

Transparency and Openness

Deidentified CoTwins data is in the process of being deposited on the Inter-University Consortium for Political and Social Research/ National Addiction and HIV Data Archive Program (ICPSR/NAHDAP) at the University of Michigan and will be available to qualified investigators. Stimulus materials are available upon request. Additionally, de-identified cognitive task data after trimming and transformation and Mplus analysis code for the phenotypic and genetic models presented in the results is available at the following OSF web page: https://osf.io/vsxb2/ (Freis, et al., 2023). This study was not pre-registered.

Results and Discussion

Preliminary Analyses

Descriptive statistics, internal reliability, and test-retest reliability, when applicable, for the lab and online tasks are reported in Tables 1 and 2. Zero-order correlations between the tasks can be found in Table 3, and univariate twin results can be found in Table 4. The phenotypic correlations between the individual EF tasks were consistent with the typical pattern that individual EF tasks often weakly correlate due to task-specific variance (Friedman & Miyake, 2017), and the correlations observed were similar to existing research in this age group (Friedman et al., 2016).

Because we had variable amounts of time between the EF assessments, we examined whether the relationships between the different versions of the EF tasks were moderated by the amount of time between assessments. For each task pair (e.g., in-lab keep track and online keep track), we tested models where performance on the online task was predicted by performance on the in-lab task, the amount of time between assessments, and the interaction between performance on the in-lab task and the amount of time between assessments. None of these interactions were significant (all interaction betas < 0.30, $p > .094$), indicating that differences in the intervals between assessments did not influence the associations between the lab and online tasks.

Are EFs Assessed in Different Contexts Separable?

To examine relationships between in-lab and online EF factors, we created a phenotypic model distinguishing Common EF assessed in the lab and Common EF assessed online. This model had five factors: Lab Common EF, Online Common EF, Keep track-specific, Category switch-specific, and *g*. In this study, we focus on relationships between the Common EF factors and *g*; however, all results are presented in Figure 2. This model fit the data well, $\chi^2(25) = 51.10$, $p = .002$, RMSEA = 0.034, CFI = .969. Lab and Online Common EF were strongly correlated ($r = .77$; $p < 0.001$; 95% CI [.59, .94]). We compared this full model with the four EF factors and *g* factor to a model with the correlation between the Common EF factors constrained to 1.0, and a model that collapsed the two Common EF factors into one Common EF factor. Both reduced models significantly hurt model fit ($\Delta\chi^2(1) = 5.80$ $p = .016$ and $\Delta\chi^2(2) = 7.43$ $p = .024$), indicating that these two EF factors are highly overlapping but separable.

Why are EFs Assessed in Different Contexts Separable?

We constructed a genetic model of our context-specific EF model with *g* and estimated A, C, and E variance components for Lab Common EF, Online Common EF, and Keep track-specific, Category switch-specific, and *g*. We also calculated genetic and environmental correlations between the Lab and Online Common EF factors. When we first estimated the genetic correlation between Lab and Online Common EF, the estimate was greater than 1, so we implemented a boundary constraint of <1 in Mplus, which minimally changed model fit ($\Delta\chi^2 = 0.05$). This model (Figure 3) fit the data adequately according to the RMSEA, although the CFI was less than .95, the typical criterion for good fit[1] ($\chi^2(388) = 528.161$, $p < .001$, RMSEA = 0.04, CFI = .902).

Results of the genetic model with and without *g* were highly similar, so here we present the variance component estimates and model comparisons from the model of EFs without *g,* which enables more focused chi-square difference tests. However, we still could only conduct 2-df tests for the A and E components; e.g., we could not set the A variance for Lab Common EF to zero without also setting its covariance with the A variance for the Online Common EF factor to zero, or else the resulting model would not be identified. We observed significant estimates of additive genetic variance for all latent EF variables. Additive genetic factors explained 89.5% of the variance in Lab Common EF ($\Delta\chi^2(2) = 27.71$, $p < 0.001$), but only 44.4% of the variance in Online Common EF ($\Delta\chi^2(2) = 20.69$, $p < 0.001$). For shared environmental factors, we only observed a C estimate above

[1] The genetic models estimate the A, C, and E parameters from the variance/covariance matrix for each twin group. Therefore, the genetic models require splitting the sample into twin 1 and twin 2 variables for MZ and DZ groups, reducing sample sizes that contribute to each element of the covariance matrix and increasing sampling error. The twin models assume strict model invariance across twins and zygosity groups, and examination of the model residuals suggested that the misfit was largely due to differences in variances/means of the online tasks across twins and zygosity groups. We attribute these differences to sampling error due to the smaller *n*s for some of the online tasks, as different random assignments led to different variables that showed this pattern of different variances across twins, with no systematic patterns. Given that the invariance assumption is necessary for the classic twin model and that the RMSEA indicated good fit, we proceeded with the models unaltered.

0 for Online Common EF: Shared environmental factors explained 18.2% of the variance in Online Common EF although this variance component was not significant ($\Delta\chi^2(1) = 0.846$, $p = 0.358$). Non-shared environmental factors explained 10.4% of the variance in Lab Common EF ($\Delta\chi^2(2) = 2.22$, $p = 0.330$) and 37.5% of the variance in Online Common EF ($\Delta\chi^2(2) = 4.53$, $p = 0.104$), although these variance components were not statistically significant.

Additive genetic factors explained 90.6% of the variance in Keep track-specific ($\Delta\chi^2(1) = 18.40$, $p < 0.001$) and 87.2% of the variance in Category switch-specific ($\Delta\chi^2(1) = 5.78$, $p = 0.016$). Non-shared environmental factors 9.3% of the variance in Keep track-specific ability ($\Delta\chi^2(1) = 0.96$, $p = 0.328$), and 12.8% of the variance in Category switch-specific ability ($\Delta\chi^2(1) = 0.859$, $p = 0.354$).

We then implemented a series of model constraints to examine if Lab and Online Common EF displayed the same patterns of additive genetic, shared environmental, and non-shared environmental influences. Though our estimate of A was higher for Lab Common EF, constraining the standardized A variance components for Lab and Online Common EF to be equal did not hurt model fit ($\Delta\chi^2(1) = 1.34$, $p = 0.247$), nor did constraining the standardized C and E variance components for Lab and Online Common EF to be equal (C constraint: $\Delta\chi^2(1) = 0.44$, $p = 0.507$; E constraint: $\Delta\chi^2(1) = 1.658$, $p = 0.197$). A 2-df test constraining the C and E variance components for Lab and Online Common EF to be equal also did not significantly hurt model fit ($\Delta\chi^2(2) = 1.97$, $p = 0.373$). Though the estimates of C and E were higher while the A estimate was lower for Online Common EF compared to Lab Common EF we did not observe statistically significant differences in the estimates of C and E for these two factors. Potentially due to the non-significant estimates of C and E, we could not distinguish if the C and E estimates were different for Lab and Online Common EF. The results of this model were consistent with a full genetic model of EFs and *g* (Figure 3).

Tests of whether the A, C, and E components could explain similar or different proportions of variance in the EF factors does not tell us if those variance components are tapping the same underlying influences. To examine the amount of genetic or environmental overlap between these factors, we estimated genetic and environmental correlations between variance components that were greater than 0. We observed a significant genetic correlation between Lab and Online Common EF (rA = 1.00, $\Delta\chi^2(1) = 20.69$, $p < 0.001$). This estimate suggests that the same genes are influencing the Lab and Online Common EF factors. Additionally, we observed a non-significant non-shared environmental correlation between Lab and Online Common EF (rE = .63, $\Delta\chi^2(1) = 1.45$, $p = 0.229$). The rA of 1 between Lab and Online EF combined with the lower estimate of rE and nominally larger estimates of C and E for Online Common suggests that these EF factors may differ at the environmental level. Although it is difficult to draw conclusions given the lack of statistically significant differences, given the estimated genetic correlation of 1.0, the phenotypic separability can only be attributed to environmental differences.

Are EFs Assessed in Different Contexts Differentially Related to *g*?

In the phenotypic full model with the four EF factors and *g*, Lab Common EF and Online Common EF were significantly correlated with *g*: $r = .54$ ($p < 0.001$; 95% CI [.31, .78]) and *r* = .57 (*p* < 0.001; 95% CI [.38, .76]), respectively. Additionally, in the full genetic models, we found that Lab and Online Common EF were similarly genetically correlated with *g* (Lab EF with *g* *r*A = 0.68, $\Delta\chi^2(1) = 52.32$, *p* < 0.001; Online EF with *g* *r*A = 0.67, $\Delta\chi^2(1) = 4.00$, *p* = 0.045). Finally, we observed a non-significant shared environmental correlation between Online Common EF and *g* (*r*C = 0.67, $\Delta\chi^2(1) = 0.81$, *p* = 0.368). These results indicate that even though Common EF factors assessed in different contexts are separable, they are similarly related at the phenotypic and genetic levels to the criterion variable *g* (assessed in lab). These patterns indicate that the two Common EF factors both display criterion validity and one factor is not a demonstrably more valid measure of Common EF than the other.

General Discussion

We examined phenotypic, genetic, and environmental relationships between in-lab and online task-based latent variables of EFs and *g* in an adolescent twin sample to evaluate how strongly Common EF factors derived from tasks assessed in different contexts are related. First, we examined phenotypic relationships between EFs assessed in different contexts and found that although latent variables of Common EF assessed in-lab and online were moderately to strongly related, they were separable. Second, we examined genetic and environmental correlations between Lab and Online Common EF factors and found that they were genetically identical but phenotypically separable, suggesting that environmental factors explain differences in Common EF across task settings. Third, we examined phenotypic, genetic, and environmental correlations between the Common EF factors and *g* and observed similar relationships between the two Common EF factors and *g*, suggesting Common EF factors show similar criterion validity across task settings. This study is the first to examine latent variables of Common EF assessed in different contexts. It points to the validity and usefulness of Common EF assessed in different contexts, but also suggests that Common EF factors assessed in-lab and online display different patterns of genetic and environmental influences.

Some researchers have proposed dropping EF tasks altogether for particular research questions due to low test-retest reliability as compared to ratings-based measures of EFs (Enkavi et al., 2019). For example, Enkavi et al. (2019) examined the reliability of self-regulation dependent tasks administered via an online platform and concluded that surveys might be more appropriate in self-regulation studies due to higher reliability. However, due to the use of individual online EF tasks, it is possible that correlations were attenuated due to task and testing-specific environmental variation (like a distracting testing environment). Importantly, latent variables of EFs derived from in-lab behavioral tasks are reliable and demonstrate developmental stability (Friedman et al., 2016). Furthermore, removing task-specific nonexecutive influences and random measurement error frequently leads to stronger correlations of behavior with latent EF variables compared to correlations of behavior with individual EF tasks (Friedman & Miyake, 2017).

Similar improvements in reliability have been observed when using factors for the online tasks used in the Enkavi et al. (2019) study, although the resulting factors did not predict substance use behaviors, whereas rating-based measures did (Eisenberg et al., 2019). However, other studies have found that, when examined together, latent EF variables derived from tasks and ratings-based measures of self-regulation are both independently predictive of real-world outcomes like psychopathology and academic achievement (Freis, et al., 2022; Friedman et al., 2020; Friedman & Gustavson, 2022; Malanchini et al., 2019; Shields et al., 2022). Moreover, it is primarily the genetic influences on EF that are related to outcomes such as psychopathology (Freis, et al., 2022; Friedman et al., 2020; Malanchini et al., 2019; Shields et al., 2022). Thus, our finding that EFs assessed in-lab and online show identical genetic influences suggest that both in-lab and online EF measures may be equally useful predictors of these outcomes, just as we found that they were equally related to our in-lab measure of *g* in the current study.

When interpreting the genetic and environmental variance components observed in our model, it is important to caution against common misinterpretations of behavioral genetics results. One common misinterpretation is that high A estimates indicate that the outcome is immutable while higher C and E estimates indicate that the outcome is malleable or a good intervention target. For example, we found a high estimate of A for In-lab Common EF. This high estimate of A for Common EF is consistent with prior twin studies on EF assessed with in-lab task batteries in children, adolescents, and adults (Engelhardt et al., 2015; Freis, et al., 2022; Friedman et al., 2016, 2020). However, our current findings and the existing twin literature on Common EF *does not* indicate that Common EF is insensitive to environmental influences or genetically determined and immutable. Furthermore, our nominally higher estimates of environmental influences for Online Common EF *do not* indicate that Common EF assessed online is more malleable. These variance components (A, C, and E) do not characterize specific genes or environmental factors but denote the proportion of variance in a particular population at a particular time and in a particular context that can be attributed to additive genetic, shared environmental, or nonshared environmental influences (Harden, 2021). Because variance component estimates derived from the classical twin design are context-specific, a traditional cross-sectional twin design alone is uninformative for intervention research. These estimates can change across development and under different environmental conditions (Tucker-Drob et al., 2013). Therefore, an intervention that changes the specific environmental conditions in which the outcomes of interest in the twin model are measured could also lead to different variance component estimates. Indeed, our finding that the heritability estimate of Common EF nominally changes in different contexts empirically demonstrates that heritability depends on environmental context (see Friedman & Miyake, 2017, for more discussion).

Still, it is somewhat perplexing why twin studies of Common EF observe little to no influence of C, considering the extensive literature on the importance of family-level environmental factors like socioeconomic status and early childhood adversity on numerous aspects of cognitive functioning and development (Farah, 2017; Merz et al., 2019; Zelazo, 2020). For example, existing research on Common EF indicates that even when C and E estimates are low, Common EF is still significantly phenotypically correlated with parental socioeconomic status (Friedman et al., 2011). One possible explanation for low estimates of C for Common EF in the twin literature is the presence of dominance (nonadditive) genetic influences (D). Traditional twin studies cannot estimate more than three variance components in the same model as that model would not be identified, and this limitation can bias variance component estimates (Friedman et al., 2021). For example, in a twin study, if both D and C influence a trait, and only one of the variance components is estimated, the D and C influences can cancel each other out (Gustavson et al., 2022). This canceling out arises because C typically increases DZ twin correlations, while D typically decreases DZ twin correlations relative to MZ twin correlations (Gustavson et al., 2022). However, extended designs incorporating twins, siblings, and adopted siblings can minimize this bias by simultaneously estimating A, D, C, and E (Friedman et al., 2021; Gustavson et al., 2022). Indeed, in a combined twin and adoption study of EFs, Gustavson et al. found evidence suggesting that D masks the contribution of C on EFs in the traditional twin design (2022).

Our findings also suggest that environmental influences may contribute to the phenotypic separability of Common EF assessed in different contexts. These environmental influences may be specific to the testing environment. For example, the laboratory is likely a more controlled setting than participants' homes, so in-lab EF tasks are potentially picking up on performance in an optimal setting that minimizes environmental variation. Due to this highly controlled laboratory environment, in-lab EF tasks have been proposed to have low ecological validity and may not always translate to real-world situations that require control (Friedman & Gustavson, 2022). However, this highly controlled laboratory setting is beneficial as it standardizes the testing experience for all participants and minimizes distractions that may make already complex tasks harder. As is typical with online cognitive assessments, we attempted to recreate some of the controlled aspects of the lab environment. For example, our online testing instructions prompted participants to complete their tasks at home in a quiet room, potentially minimizing environmental variation. However, we did not have control over the testing environment and the lab setting involved a research assistant present to read task instructions and note any potential issues with data collection. Additionally, technological differences, like using different keyboards (Schubert et al., 2013), can introduce additionally testing variability in at home settings. These factors could contribute to the home testing environment differing from the lab environment. This lack of researcher control could increase environmental variation but also jeopardize the validity of these online EF assessments. However, we found that the Lab and Online Common EF factors correlated highly, supporting the viability of collecting EF data online. These results suggest that latent variables derived from online EF assessments are valid and could be used to increase sample sizes in psychological research or potentially reach participants who may otherwise not have been able to participate in a lab setting.

Moreover, environmental influences like socioeconomic status may be more influential during online testing because socioeconomic status will impact the environmental variability of the home environment, while the lab environment is designed to be consistent across participants. While standardized testing protocols minimize environmental variation in the lab, it is important to note that existing research in developmental and cognitive psychology points to relationships between socioeconomic status, brain development, and cognitive functions, including EFs (Farah, 2017; Friedman et al., 2011; Merz et al., 2019), which would influence performance on lab tasks.

Overall, there are benefits to both in-lab and online EF assessments. In-lab assessments may be more standardized and measure performance under optimal conditions for all participants. In contrast, online assessments may be more accessible and could increase ecological validity and better represent "real-world" performance, as individuals must exercise control in diverse and variable environments.

Limitations and Future Directions

Because the online EF data collection was remote, not all research participants completed all the online EF tasks. So, the average *n* for the in-lab tasks was higher than the online tasks, potentially contributing to higher sampling error for the online measures. Additionally, the smaller sample sizes of the online tasks may have impacted our ability to detect the significance of the variance components in our genetically informed model. Moreover, we only had in-lab measures of general cognitive ability, so we could not examine method-specific relationships in our analyses of EFs assessed in different contexts and *g*. Additionally, as described in the method, we made minor alterations between the two data collection waves to improve the online category-switch and keep track tasks, so the online tasks were not identical. However, the correlations across these versions of the same tasks remained high.

We focus on Common EF assessed in different contexts due to its relevance to other behavioral outcomes (Friedman & Miyake, 2017) and our limited number of tasks. Due to the limited number of tasks, we could not construct context-specific Keep track-specific and Category switch-specific factors, so we cannot comment on if the patterns we observed of nominally different genetic and environmental influence hold for EF factors beyond Common EF. Additionally, in this study, because we used the same tasks in the lab and online, the Updating-Specific (Keep track-specific) and Shifting-Specific (Category switch-specific) factors are confounded with task-specific variance. Finally, since we did not have identical task batteries for the in-lab and online protocols we cannot directly compare the means and variances for the in-lab and online factors.

Constraints on generality.

CoTwins participants were recruited from birth records from the entire state of Colorado to be representative of the state rather than a specific area. As a result, the CoTwins sample is reasonably representative of the state of Colorado. However, the CoTwins sample is over-represented for White and higher socioeconomic status participants compared to the adolescent population of the United States (United States Census Bureau, 2020). Additionally, though this study did not use the phone-based data, having a smart phone was a study requirement because of the CoTwins phone application. Though smart phones are quite common, this requirement could have inflated the number of high socioeconomic status participants in our sample. When research lacks socioeconomic, racial, and ethnic diversity, the findings are less generalizable and translatable, which can exacerbate existing disparities in many different psychological and behavioral outcomes (Holden et al., 2022). This lack of generalizability is often discussed as a significant concern in scientific research. Another equally important consideration, particularly in the behavioral genetics field, is that when samples lack diversity, we also miss out on the full range of environmental factors relevant to the outcomes of interest (Holden et al., 2022). In this study, for example, a more diverse sample could have increased the environmental variability in the at-home testing environment. Therefore, these findings need to be replicated in a more diverse sample.

Future directions.

Our finding that Common EF factors assessed in different contexts are highly correlated, particularly at the genetic level, has implications for large-scale data collection efforts attempting to employ deep phenotyping like biobanks (Bycroft et al., 2018). For example, using at-home online EF tasks could allow large scale biobanks to gather more EF data with more diverse task batteries. The result that EFs assessed in different contexts are genetically identical indicates that online tasks would be appropriate measures in genetically informed studies. Similarly, online EF tasks could reduce the burden of bringing participants into the lab for assessments and allow researchers to administer more tasks and not rely on small single-task in-lab data to study EF and its correlates. Finally, our findings reinforce the important existing efforts to standardize and validate online behavioral experiments (Sochat et al., 2016) by demonstrating that online EF tasks are highly related to traditional in-lab tasks.

One important future direction we are pursuing is to examine if EFs assessed in different contexts are similarly related to other outcomes beyond a cognitive variable like *g*. For example, it could be particularly valuable to examine if the Common EF factors assessed in different contexts show similar patterns of relationships with behavioral outcomes commonly associated with behavioral control, like impulsivity or psychopathology. For example, one hypothesis for the observation that EF factors derived from in-lab behavioral tasks display modest correlations with self-reported control abilities is that in-lab tasks measure performance under optimal conditions (minimal environmental variation), while self-report measures may measure average control across a wider range of conditions (Friedman & Gustavson, 2022). Online tasks could provide an intermediate measure that is more objective than self-report, but also taps more environmental variation than in-lab tasks, and hence may be more related to real-world behavioral outcomes that reflect individual differences in control abilities. Future studies that examine genetic and environmental correlations between Common EF assessed online and real-world outcomes may be able to elucidate if the environmental variance in Online Common EF is meaningful.

Summary and Conclusions

We found that Common EF factors assessed in different contexts are highly related, but not identical. This separability appears to be due to environmental factors, as Common EF factors assessed in different contexts were identical at the genetic level, but phenotypically separable. Overall, our results indicate that researchers can use both in-lab and online EF tasks to study EFs, and the use of EF tasks assessed in different contexts could allow researchers to examine different questions.

Context

We took advantage of our unique sample that collected both in-lab and online EF tasks to examine how closely related these EFs assessed in different contexts are. This question is important given the growing popularity of online EF assessments; yet, to our knowledge, no study has employed latent variable models to examine relationships between EFs assessed in different contexts. Our group has extensively studied the unity/diversity model of EFs (Friedman & Miyake, 2017); however, that research has primarily relied on in-lab EF tasks, and we wanted to examine this well replicated model with both in-lab and online tasks. This study replicated the unity/diversity model of EFs in a previously unpublished sample with our unique task battery. Finally, twin studies are sometimes underestimated as mere studies of genetic influences/heritability. However, twin studies are invaluable for examining questions about environmental influences (Friedman et al., 2021). This project exemplifies how twin studies can be used to get at environmental research questions and examine relationships between variables at different levels of analysis.

Acknowledgments

The authors gratefully thank all participants and members of the Colorado Online Twin Study (CoTwins). Special thanks to Amy Ledbetter, Rio Hindami, Rachel Hawley, and Enning Tang for their excellent work on participant recruitment, data collection, and data management. This research was supported by the National Institutes of Health (NIH) grants DA046413, MH016880 and AA023974.

References

Ahmed SF, Skibbe LE, McRoy K, Tatar BH, & Scharphorn L (2022). Strategies, recommendations, and validation of remote executive function tasks for use with young children. Early Childhood Research Quarterly, 60, 336–347. 10.1016/j.ecresq.2022.03.002

Alexander JD, Zhou Y, Freis SM, Friedman NP, & Vrieze SI (2022). Individual differences in adolescent and young adult daily mobility patterns and their relationships to big five personality traits: A behavioral genetic analysis. Journal of research in personality, 100, 104277. [PubMed: 35991708]

Bilder RM, & Reise SP (2019). Neuropsychological tests of the future: How do we get there from here? The Clinical Neuropsychologist, 33(2), 220–245. 10.1080/13854046.2018.1521993 [PubMed: 30422045]

Blair C (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. Behavioral and Brain Sciences, 29(2), 109–125. 10.1017/S0140525X06009034 [PubMed: 16606477]

United States Census Bureau (2020). 2019 Population Estimates by Age, Sex, Race and Hispanic Origin. Census.Gov.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, & Marchini J (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726), 203–209. 10.1038/s41586-018-0579-z [PubMed: 30305743]

Cohen J, MacWhinney B, Flatt M, & Provost J (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behavior Research Methods, Instruments, & Computers, 25(2), 257–271. 10.3758/BF03204507

Crump MJ, McDonnell JV, & Gureckis TM (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one, 8(3), e57410. [PubMed: 23516406]

de Leeuw JR (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behavior Research Methods, 47(1), 1–12. doi:10.3758/s13428-014-0458-y [PubMed: 24683129]

Diamond A (2013). Executive functions. Annual Review of Psychology, 64(1), 135–168. 10.1146/annurev-psych-113011-143750

Eisenberg IW, Bissett PG, Zeynep Enkavi A, Li J, MacKinnon DP, Marsch LA, & Poldrack RA (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. Nature Communications, 10(1), 2319. 10.1038/s41467-019-10301-1

Engelhardt LE, Briley DA, Mann FD, Harden KP, & Tucker-Drob EM (2015). Genes unite executive functions in childhood. Psychological Science, 26(8), 1151–1163. 10.1177/0956797615577209 [PubMed: 26246520]

Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, & Poldrack RA (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. Proceedings of the National Academy of Sciences, 116(12), 5472–5477. 10.1073/pnas.1818430116

Farah MJ (2017). The neuroscience of socioeconomic status: Correlates, causes, and consequences. Neuron, 96(1), 56–71. 10.1016/j.neuron.2017.08.034 [PubMed: 28957676]

Freis SM, (2023). Associations between executive functions assessed in different contexts in a genetically informative sample. https://osf.io/vsxb2/

Freis SM, Morrison CL, Lessem JM, Hewitt JK, & Friedman NP (2022). Genetic and environmental influences on executive functions and intelligence in middle childhood. Developmental Science, 25(1), e13150. 10.1111/desc.13150 [PubMed: 34288270]

Freis SM, Morrison CL, Smolker HR, Banich MT, Kaiser RH, Hewitt JK, & Friedman NP (2022). Executive functions and impulsivity as transdiagnostic correlates of psychopathology in childhood: A behavioral genetic analysis. Frontiers in Human Neuroscience, 16. 10.3389/fnhum.2022.863235

Friedman NP, Banich MT, & Keller MC (2021). Twin studies to GWAS: There and back again. Trends in Cognitive Sciences. 10.1016/j.tics.2021.06.007

Friedman NP, & Gustavson DE (2022). Do rating and task measures of control abilities assess the same thing? Current Directions in Psychological Science, 31(3), 262–271. 10.1177/09637214221091824 [PubMed: 35928929]

Friedman NP, Hatoum AS, Gustavson DE, Corley RP, Hewitt JK, & Young SE (2020). Executive functions and impulsivity are genetically distinct and independently predict psychopathology: Results from two adult twin studies. Clinical Psychological Science, 8(3), 519–538. 10.1177/2167702619898814 [PubMed: 33758683]

Friedman NP, & Miyake A (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex, 86, 186–204. 10.1016/j.cortex.2016.04.023 [PubMed: 27251123]

Friedman NP, Miyake A, Altamirano LJ, Corley RP, Young SE, Rhea SA, & Hewitt JK (2016). Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. Developmental Psychology, 52(2), 326–340. 10.1037/dev0000075 [PubMed: 26619323]

Friedman NP, Miyake A, Corley RP, Young SE, DeFries JC, & Hewitt JK (2006). Not all executive functions are related to intelligence. Psychological Science, 17(2), 172–179. 10.1111/j.1467-9280.2006.01681.x. [PubMed: 16466426]

Friedman NP, Miyake A, Robinson JL, & Hewitt JK (2011). Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: A behavioral genetic analysis. Developmental Psychology, 47(5), 1410–1430. 10.1037/a0023750 [PubMed: 21668099]

Gustavson DE, Reynolds CA, Corley RP, Wadsworth SJ, Hewitt JK, & Friedman NP (2022). Genetic associations between executive functions and intelligence: A combined twin and adoption study. Journal of Experimental Psychology: General. 10.1037/xge0001168

Harden KP (2021). "Reports of My Death Were Greatly Exaggerated": Behavior genetics in the postgenomic era. Annual Review of Psychology, 72(1), 37–60. 10.1146/annurev-psych-052220-103822

Haworth CMA, Wright MJ, Luciano M, Martin NG, De Geus EJC, Van Beijsterveldt CEM, Bartels M, Posthuma D, Boomsma DI, Davis OSP, Kovas Y, Corley RP, Defries JC, Hewitt JK, Olson RK, Rhea SA, Wadsworth SJ, Iacono WG, McGue M, … Plomin R (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. Molecular Psychiatry, 15(11), 1112–1120. 10.1038/mp.2009.55 [PubMed: 19488046]

Holden LR, Haughbrook R, & Hart SA (2022). Developmental behavioral genetics research on school achievement is missing vulnerable children, to our detriment. New Directions for Child and Adolescent Development, 2022(183–184), 47–55. 10.1002/cad.20485 [PubMed: 36162231]

Hu LT, & Bentler PM (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3(4), 424–453. 10.1037/1082-989X.3.4.424

Ishihara S (1917). Tests for color-blindness, Handaya, Hongo Harukicho, Tokyo.

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 and Review, 9(4), 637–671. 10.3758/BF03196323 [PubMed: 12613671]

Kendler KS (1996). Major depression and generalised anxiety disorder: same genes,(partly) different environments–revisited. The British Journal of Psychiatry, 168(S30), 68–75. [PubMed: 8770431]

Kline RB (2016). Principles and practice of structural equation modeling (4th ed). Guilford Press.

Lavigne KM, Sauvé G, Raucher-Chéné D, Guimond S, Lecomte T, Bowie CR, Menon M, Lal S, Woodward TS, Bodnar MD, & Lepage M (2022). Remote cognitive assessment in severe mental illness: A scoping review. Schizophrenia, 8(1), 1–12. 10.1038/s41537-022-00219-x [PubMed: 35132080]

Malanchini M, Engelhardt LE, Grotzinger AD, Harden KP, & Tucker-Drob EM (2019. “Same But Different”: Associations between multiple aspects of self-regulation, cognition, and academic abilities. Journal of Personality and Social Psychology, 117(6), 1164–1188. 10.1037/pspp0000224 [PubMed: 30550329]

Malanchini M, Rimfeld K, Gidziela A, Cheesman R, Allegrini AG, Shakeshaft N, Schofield K, Packer A, Ogden R, McMillan A, Ritchie SJ, Dale PS, Eley TC, von Stumm S, & Plomin R (2021). Pathfinder: A gamified measure to integrate general cognitive ability into the biological, medical, and behavioural sciences. Molecular Psychiatry, 26(12), 7823–7837. 10.1038/s41380-021-01300-0 [PubMed: 34599278]

Markwardt FC Jr (1997). Peabody Individual Achievement Test--Revised--Normative Update.

Mayr U, & Kliegl R (2000). Task-set switching and long-term memory retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(5), 1124–1140. 10.1037/0278-7393.26.5.1124 [PubMed: 11009248]

Merz EC, Wiltshire CA, & Noble KG (2019). Socioeconomic inequality and the developing brain: Spotlight on language and executive function. Child Development Perspectives, 13(1), 15–20. 10.1111/cdep.12305

Miller JB, & Barr WB (2017). The technology crisis in neuropsychology. Archives of Clinical Neuropsychology, 32(5), 541–554. 10.1093/arclin/acx050 [PubMed: 28541383]

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, & Wager TD (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. Cognitive Psychology, 41(1), 49–100. 10.1006/cogp.1999.0734 [PubMed: 10945922]

Moore RC, Campbell LM, Delgadillo JD, Paolillo EW, Sundermann EE, Holden J, Schweitzer P, Heaton RK, & Swendsen J (2020). Smartphone-Based measurement of executive function in older adults with and without HIV. Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 35(4), 347–357. 10.1093/arclin/acz084 [PubMed: 31942632]

Muthén LK, & Muthén BO (1998-2017). Mplus: Statistical Analysis with Latent Variables: User’s Guide (Version 8).

Neale MC, Heath AC, Hewitt JK, Eaves LJ, & Fulker DW (1989). Fitting genetic models with LISREL: Hypothesis testing. Behavior genetics, 19(1), 37–49. [PubMed: 2712812]

Rhea SA, Gross AA, Haberstick BC, & Corley RP (2006). Colorado twin registry. Twin Research and Human Genetics, 9(6), 941–949. [PubMed: 17254434]

Rijsdijk FV, & Sham PC (2002). Analytic approaches to twin data using structural equation models. Briefings in Bioinformatics, 3(2), 119–133. 10.1093/bib/3.2.119 [PubMed: 12139432]

Roberts RJ, Hager LD, & Heron C (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. Journal of Experimental Psychology: General, 123(4), 374–393. 10.1037/0096-3445.123.4.374

Shields AN, Malanchini M, Vinnik L, Tucker-Drob EM, Harden KP, & Tackett JL (2022). Genetic variance in conscientiousness relates to youth psychopathology beyond executive functions. Journal of psychopathology and clinical science, 131(8), 830. [PubMed: 36326625]

Schubert TW, D'Ausilio A, & Canto R (2013). Using Arduino microcontroller boards to measure response latencies. Behavior Research Methods, 45(4), 1332–1346. 10.3758/s13428-013-0336-z [PubMed: 23585023]

Shvetz C, Gu F, Drodge J, Torous J, & Guimond S (2021). Validation of an ecological momentary assessment to measure processing speed and executive function in schizophrenia. Npj Schizophrenia, 7(1), 1–9. 10.1038/s41537-021-00194-9 [PubMed: 33479257]

Silventoinen K, Kaprio J, Lahelma E, & Koskenvuo M (2000). Relative effect of genetic and environmental factors on body height: differences across birth cohorts among Finnish men and women. American journal of public health, 90(4), 627. [PubMed: 10754982]

Sochat VV, Eisenberg IW, Enkavi AZ, Li J, Bissett PG, & Poldrack RA (2016). The experiment factory: Standardizing behavioral experiments. Frontiers in psychology, 7, 610. [PubMed: 27199843]

Stroop JR (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643–662. 10.1037/h0054651

Toplak ME, West RF, & Stanovich KE (2013). Practitioner Review: Do performance-based measures and ratings of executive function assess the same construct? Journal of Child Psychology and Psychiatry, 54(2), 131–143. 10.1111/jcpp.12001 [PubMed: 23057693]

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]

Wechsler D (1999). Wechsler Abbreviated Scale of Intelligence-Second Edition. 10.1037/t15171-000

Wennerhold L, & Friese M (2020). Why self-report measures of self-control and inhibition tasks do not substantially correlate. Collabra: Psychology, 6(1), 9. 10.1525/collabra.276

Wilcox RR, & Keselman HJ (2003). Modern robust data analysis methods: Measures of central tendency. Psychological Methods, 8(3), 254–274. 10.1037/1082-989X.8.3.254 [PubMed: 14596490]

Yntema DB (1963). Keeping track of several things at once. Human Factors, 5(1), 7–17. 10.1177/001872086300500102 [PubMed: 14002586]

Zelazo PD (2020). Executive function and psychopathology: A neurodevelopmental perspective. Annual Review of Clinical Psychology, 16(1), 431–454. 10.1146/annurev-clinpsy-072319-024242

Public significance statement:

This study highlights that performance on online cognitive tasks show a close correspondence to cognitive tasks administered in highly controlled laboratory settings, suggesting that at home online cognitive tasks are valid. Online cognitive tasks could be more accessible to administer and incorporate into clinical, academic, and research settings.

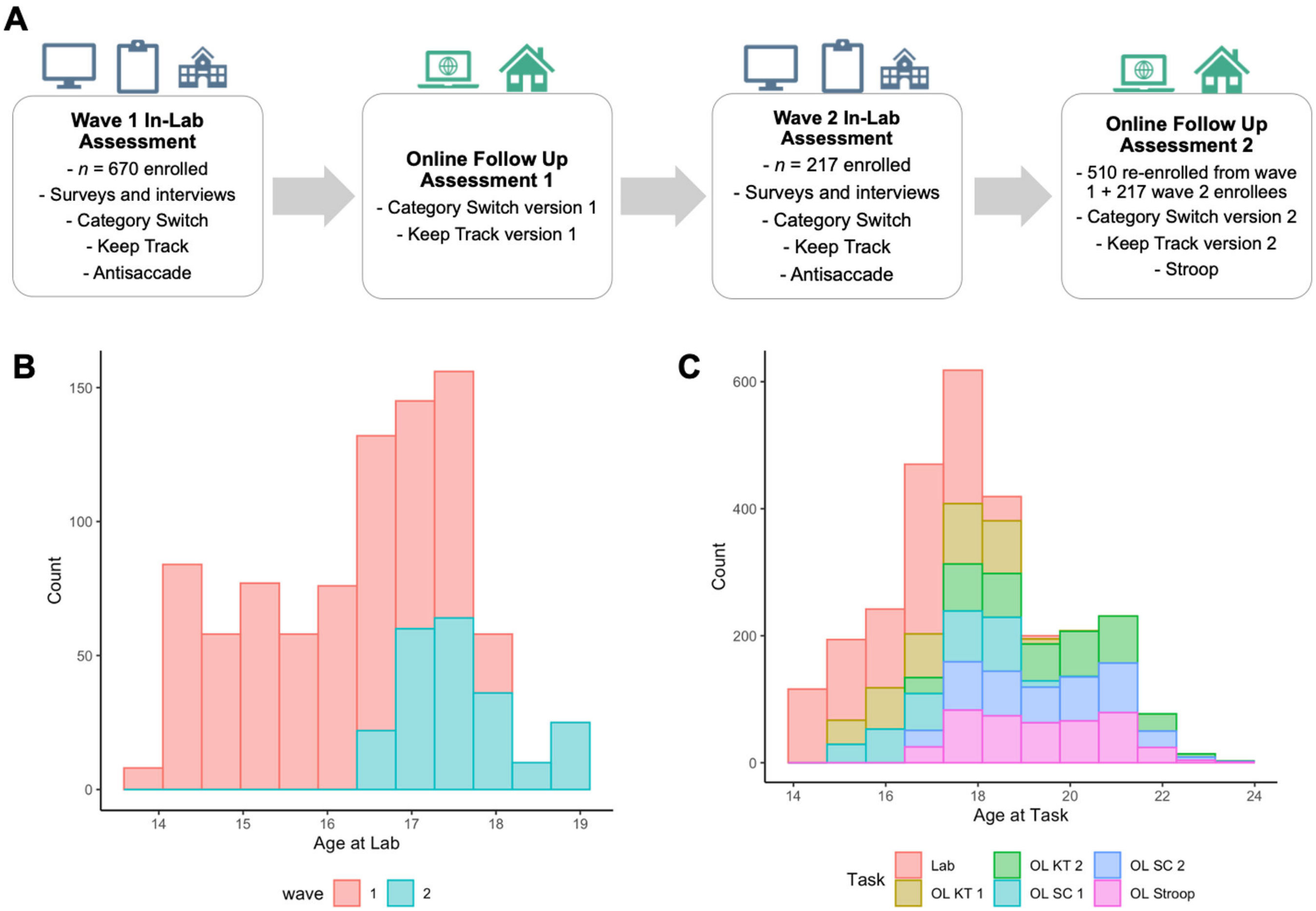

Figure 1.

Visualization of CoTwins in-lab and remote online protocols (A), distributions of age at lab visit by study wave (B), and distribution of age at each task (C). Total *N* = 887. Wave 1 in-lab *n* = 670, *M* age = 16.03, *SD* age = 1.13, min age = 13.98, max age = 17.95. Wave 2 in-lab *n* = 217, *M* age = 17.57, *SD* = .65, min age = 16.48, max age = 19.05. OL SC = online category-switch; OL KT = online keep track; OL Stroop = online Stroop; 1 and 2 refer to the version of the online category-switch and keep track tasks.

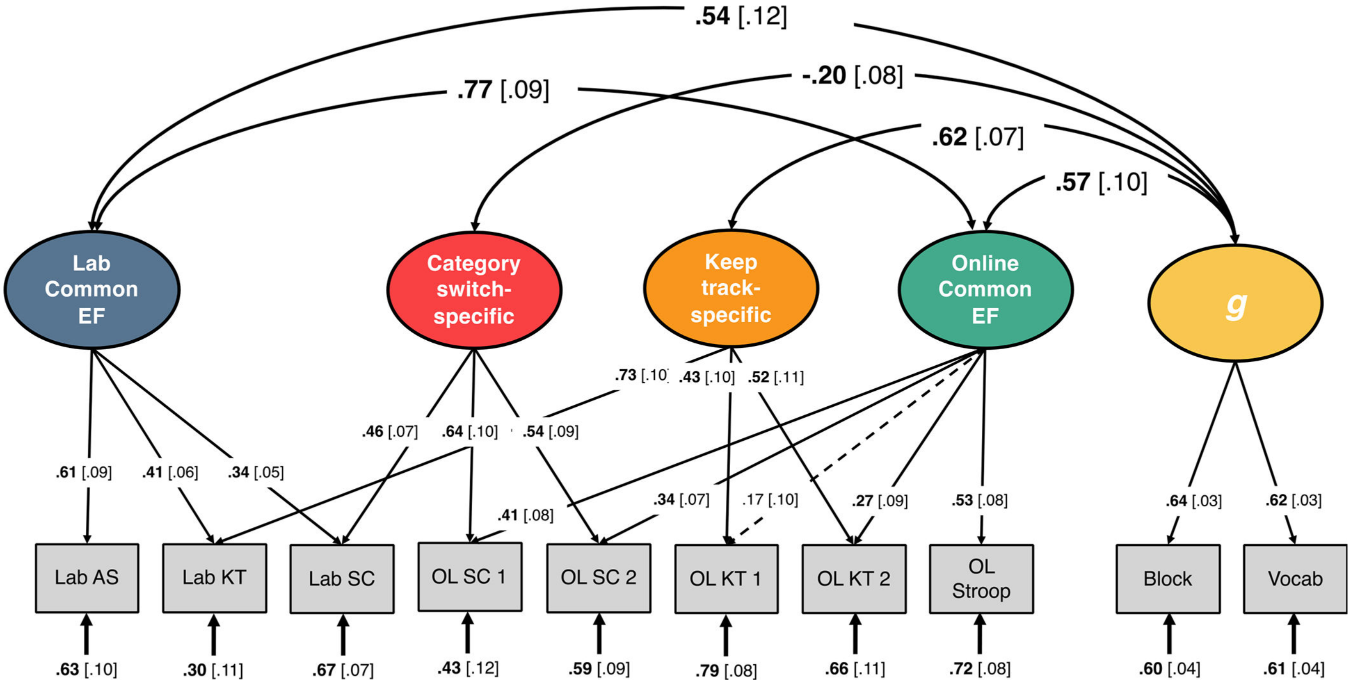

Figure 2.

Phenotypic factor model of the context-specific executive function (EF) unity/diversity model and *g*. Lab SC = in-lab category-switch; Lab AS = in-lab antisaccade; Lab KT = in-lab keep track; OL SC = online category-switch; OL KT = online keep track; OL Stroop = online Stroop; 1 and 2 refer to the version of the online category-switch and keep track tasks. Block = block design; Vocab = Vocabulary. Ellipses indicate latent variables; rectangles indicate observed variables. Numbers on single-headed arrows indicate standardized factor loadings. Numbers in brackets are standard errors. Numbers at the ends of arrows are residual variances. OL = online; Vocab = vocabulary. Double headed arrows indicate correlations. Solid lines and boldface type signify $p < .05$.

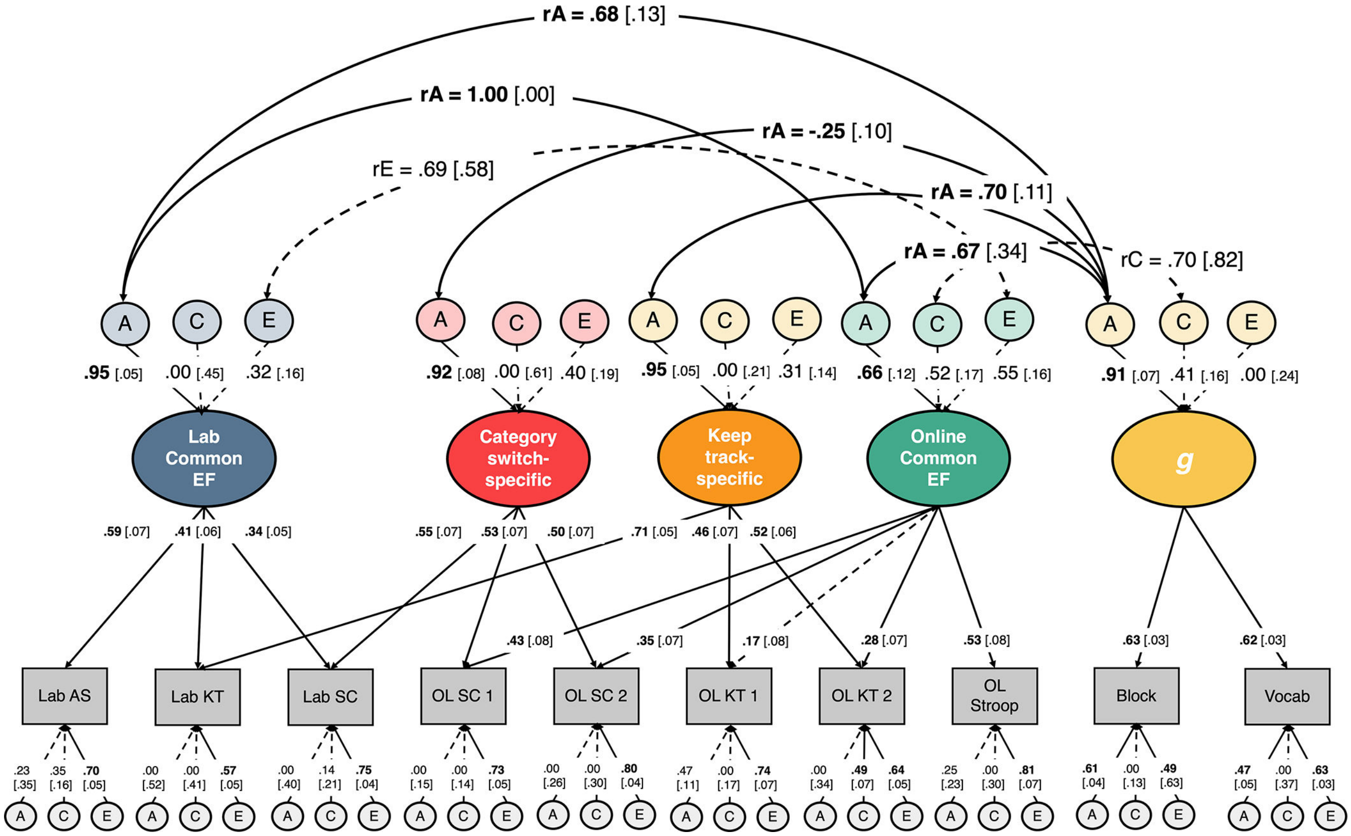

Figure 3.

EF = executive function; A = additive genetic variance; C = shared environmental variance; E = nonshared environmental variance; rA = genetic correlation; rC = shared environmental correlation; rE = nonshared environmental correlation; Lab SC = in-lab category switch; Lab AS = in-lab antisaccade; Lab KT = in-lab keep track; OL SC = online category switch; OL KT = online keep track; OL Stroop = online Stroop; Block = block design; Vocab = vocabulary; 1 and 2 refer to the version of the online category-switch and keep track tasks. Ellipses indicate latent variables; rectangles indicate observed variables. Numbers on single-headed arrows indicate standardized factor loadings. Numbers in brackets are standard errors. Double-headed arrows indicate correlations. Solid lines and boldface type indicate $p < .05$, whereas dashed lines and regular text indicate $p > .05$, determined with chi-square difference tests for the A, C, and E effects and their correlations.

Table 1

Task Descriptive Statistics

Measure	*n*	mean	sd	min	max	range	skew	kurtosis
Lab Category-switch	881	212.26	166.2	−216.94	746.05	962.99	0.94	0.9
Lab Antisaccade	872	0.96	0.28	0.31	1.57	1.26	−0.08	−0.25
Lab Keep track	886	0.85	0.22	0.2	1.52	1.33	0.19	0.28
OL Category-switch 1	316	146.88	122.96	−58.82	523.96	582.78	1.1	0.89
OL Keep track 1	359	0.88	0.2	0.5	1.57	1.07	0.82	1.33
OL Stroop	419	147.41	93.87	−81.94	422.76	504.7	0.28	−0.41
OL Category-switch 2	407	129.71	103.05	−53	462.89	515.89	0.79	0.43
OL Keep track 2	404	0.92	0.19	0.45	1.57	1.12	0.14	0.18
Block Design Raw	885	44.02	12.26	8	70	62	−0.44	−0.17
Vocab Raw	886	37.62	4.49	16	52	36	−0.64	0.92

Note. Descriptive statistics for cognitive tasks after data screening. OL = online.

Table 2

Task Reliabilities

Task	Method	Within-task Reliability	Test-retest with subsequent tasks
Category-switch	Lab	.64[a]	.34–.42
Antisaccade	Lab	.87[a]	–
Keep track	Lab	.76[b]	.33–.40
Category-switch	Online	.85–.92[a]	.51–.56
Keep track	Online	.50–.72[b]	.41–.57
Stroop	Online	.89–.90[a]	-

Note. Dashes indicate the correlation was not available because the tasks were not administered twice.

[a] Calculated by adjusting odd-even correlations with the Spearman-Brown prophecy formula.

[b] Calculated using Cronbach's alpha.

Table 3

Phenotypic Correlations

	1	2	3	4	5	6	7	8	9	10
1. LabSC	1									
2. LabAS	0.21*	1								
3. LabKT	0.12*	0.26*	1							
4. olSC 1	0.39*	0.25*	0.09	1						
5. olKT 1	0.06	0.10	0.34*	0.06	1					
6. olStroop	0.16*	0.21*	0.10	0.24*	0.07	1				
7. olSC 2	0.35*	0.17*	0.00	0.50*	0.06	0.19*	1			
8. olKT 2	0.07	0.11*	0.42*	0.00	0.45*	0.08	0.09	1		
9. Block	0.05	0.23*	0.44*	0.04	0.20*	0.24*	0.01	0.23*	1	
10. Vocab	0.07	0.15*	0.42*	0.06	0.12*	0.12*	0.00	0.21*	0.39*	1

Note. Correlation matrix of cognitive tasks after data screening. LabSC = in-lab category-switch; LabAS = in-lab antisaccade; LabKT = in-lab keep track; olSC = online category-switch; olKT = online keep track; olStroop = online Stroop; 1 and 2 refer to the version of the online category-switch and keep track tasks. Block = block design; Vocab = Vocabulary. In the full models, category-switch and Stroop task scores were multiplied by −1 so that so that for all tasks, higher numbers indicate better performance

*$p < .05$.

Table 4

Univariate Twin Models

	Twin Correlations		Variance Components			Model Fit				
Task	MZ	DZ	A	C	E	χ^2	df	*p*	RMSEA	CFI
LabAS	.49	.28	.38*	.10	.52*	4.09	6	.664	.000	1
LabSC	.44	.17	.40*	.00	.60*	14.86	6	.021	.082	.790
LabKT	.65	.29	.64*	.00	.36*	6.35	6	.385	.016	.997
olSC 1	.15	.20	.00	.18	.82*	4.51	6	.609	.000	1
olKT 1	.58	−.12	.40*	.00	.60*	23.32	6	.001	.163	.329
olSC 2	.29	.17	.20	.07	.73*	5.45	6	.488	.000	1
olKT 2	.52	.41	.28	.24	.48*	9.96	6	.127	.072	.856
olStroop	.38	.04	.30	.00	.70*	4.22	6	.647	.000	1
Block	.78	.38	.79*	.00	.21*	7.53	6	.275	.034	.992
Vocab	.61	.35	.51*	.10	.39*	5.99	6	.424	.000	1

Note. LabSC = in-lab category-switch; LabAS = in-lab antisaccade; LabKT = in-lab keep track; olSC = online category-switch; olKT = online keep track; olStroop = online Stroop; 1 and 2 refer to the version of the online category-switch and keep track tasks. Block = block design; Vocab = Vocabulary.

* *p* < .05, determined with chi-square difference tests.