Within-Individual Variability in Neurocognitive Performance: Age- and Sex-Related Differences in Children and Youths From Ages 8 to 21

David R. Roalf, Raquel E. Gur, Kosha Ruparel, Monica E. Calkins, Theodore D. Satterthwaite, and Warren B. Bilker
University of Pennsylvania

Lauren Julius Harris
Michigan State University

Hakon Hakonarson
Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania

Ruben C. Gur
University of Pennsylvania and Philadelphia Veterans Administration Medical Center, Philadelphia, Pennsylvania

Objective: The transition from childhood to adulthood is characterized by improved motor and cognitive performance in many domains. Developmental studies focus on average performance in single domains but ignore consistency of performance across domains. Within-individual variability (WIV) provides an index of that evenness and is a potential marker of development. Method: We gave a computerized battery of 14 neurocognitive tests to 9138 youths ages 8–21 from the Philadelphia Neurodevelopmental Cohort. Results: As expected, performance improved with age, with both accuracy and speed peaking in adulthood. WIV, however, showed a U-shaped course: highest in childhood, declining yearly into mid-adolescence, and increasing again into adulthood. Young females outperformed and were less variable than males, but by early adulthood male performance matched that of females despite being more variable. Conclusion: We conclude that WIV declines from childhood to adolescence as developmental lags are overcome, and then increases into adulthood reflecting the emergence of cognitive specializations related to skill-honing and brain maturation.

Keywords: age differences, cognitive neuroscience, development, sex differences, within-individual variability

Supplemental materials: http://dx.doi.org/10.1037/neu0000067.supp

The transition from childhood to adulthood is characterized by improvement in motor (Gallahue, Ozmun, & Goodway, 1998), cognitive (Fischer, 1980), and social domains (Blakemore & Choudhury, 2006). Developmental studies of neurocognition, that is, of performance on tasks linked to brain systems, have focused on single domains, including executive-control (e.g., Goldberg, Maurer, & Lewis, 2001), memory (e.g., Dempster, 1985; Luciana, Conkin, Hooper, & Yarger, 2005), language and reasoning (e.g., Friederici & Wartenburger, 2010), and more recently, social cognition (e.g., Burnett, Sebastian, Cohen Kadosh, & Blakemore, 2011; Shaw, Grosbras, Leonard, Pike, & Paus, 2012).

Studies of performance in single domains are useful but are insensitive to potentially important phenomena related to the relative and uneven pace of development of different abilities (In-
absence (Roalf, Gur et al., 2013). Finally, in providing additional
information about typical development (e.g., Van Geert & Van Dijk, 2002). It also could help identify individuals at
risk for certain brain-behavior disorders, because variability is
higher in focal disorders such as infarction, hemorrhage, trauma,
and tumor (Stuss, Murphy, Bins, & Alexander, 2003) and in
developmental disorders such as attention deficit–hyperactivity
disorder (Leth-Steensen, Elbaz, & Douglas, 2000) and schizophre
nia (Roalf, Gur et al., 2013). Finally, in providing additional
information about typical development, the study of WIV across
domains could offer important new information about individual
differences.

Individual differences in performance include sex differences in
specific behavioral domains (e.g., R. C. Gur et al., 1995; Halpern
et al., 2007). Males outperform females on certain spatial and
motor tests (e.g., Harris, 1985; Moreno-Briseño, Diaz, Campos-
Romo, & Fernandez-Ruiz, 2010; Thomas & French, 1985), whereas females do better on certain verbal and memory tests (e.g.,
R. C. Gur et al., 2012; Herlitz, Nilsson, & Backman, 1997). There
also is evidence that WIV is higher in males than females (e.g.,
Deary, Thorpe, Wilson, Starr, & Whalley, 2003; Shea, Lubinski,
& Benbow, 2001). If high WIV is an indication of better perfor-
ance, or “specialization,” in some domains relative to others,
then higher WIV in males could be an indication that more males are
“cognitive specialists” whereas more females are “cognitive
generalists,” performing well across a broad range of functions,
or domains. Measuring WIV during development therefore could be
another, more sensitive way to document the emergence of individ-
ual differences in fundamental patterns of cognitive specializa-
tion. Thus, if part of development involves reaching levels of
performance that are more nearly equal or even across domains,
we would expect WIV to decline with age. However, to our
knowledge, WIV has not been examined in large samples with
sufficiently broad batteries of tests across a wide age range to
permit an evaluation of its development and relevance for overall
performance.

It is helpful to distinguish WIV from two other measures of
variability that have been examined in earlier studies. One refers to
within-individual consistency of performance across time; the
other examines sample variance on single measures (e.g., children
often have high performance variability even when solving the
same problem over short or long intervals; Siegler & McGilley,
1989). Dykiert, Der, Starr, and Deary (2012) found that variability
on one test decreased from childhood through adolescence, followed
by an increase throughout adulthood and into old age. Sample variance has also been studied in different populations,
motivated by Darwin’s observation that evolution is associated
with high variability within species (Darwin, 1871). Thorndike
(1910) proposed that males have greater variability in performance
and that this was is why they have reached greater intellectual
prominence: “if men differ in intelligence and energy by wider
extremes than do women, eminence in and leadership of the
world’s affairs of whatever sort will inevitably belong oftener to
men. They will often deserver it” (p. 10). Although arguments
against this conclusion had already been raised by scholars in
Thorndike’s own time (e.g., Hollingworth, 1914), the debate still
rages; the empirical support for the higher variability in males
seems to differ according to the sample studied and the specific
measures used (Halpern et al., 2007).

In two prior studies, we used a Computerized Neurocognitive
Battery (CNB), a 1-hr-long battery of neuroscience-based tests that
have been validated with neuroimaging studies as recruiting spe-
cific brain systems (R. C. Gur et al., 2001; R. C. Gur et al., 2010). The CNB measures performance in 5 domains: executive-control,
episodic memory, complex cognitive processing, social cognition,
and sensorimotor speed. The battery is designed for large-scale
studies and enables efficient data collection and separate scoring
for accuracy and speed. The range of domains permits reliable
estimates of across-Test WIV. In a prior report (R. C. Gur et al.,
2012), we examined global neurocognitive performance (GNP),
that is, average performance across domains, for the first 3,500
participants from a sample of more than 9,000 children and youth,
ages 8 to 21. We found substantial improvement across the age
range in both accuracy and speed with some domain variability in
effect size. We also found significant sex differences, although
effect sizes were small and moderated by earlier maturation in
females. We did not, however, examine WIV, so its developmental
course in this cohort is unknown. In the current study, we exam-
ined data, including WIV, from the complete sample and tested the
following hypotheses: a) Average neurocognitive performance
will improve with age; b) Females will reach mature performance
(e.g., higher GNP) earlier than males; c) Over the 8- to 21-year
span WIV will decrease, showing increasing evenness of perfor-
ance associated with maturation. Before testing these hypothe-
ses, we examined whether we could confirm in this larger sample
the sex difference found in a previously published study: better
performance by females than males on tests of attention, memory,
and social cognition; better performance by males than females on
tests of spatial processing and motor speed. We also tested
Thorndike’s hypothesis of greater variance among males, although
our single-time measure did not permit evaluation of variability in
performance within individuals over time.

Method

Participants

The sample included children and young adults (ages 8–21)
recruited through a NIMH-funded Grand Opportunity (GO) study
characterizing clinical and neurobehavioral phenotypes in a pro-
spectively acquired community cohort. Participants previously
consented for genomic studies when they came for pediatric ser-
ices at the Children’s Hospital. At that time they provided a blood
sample for genetic studies, authorized access to Electronic Medical
Records (EMRs), and gave written informed consent/assent to be
recontacted for future studies. Of the 50,540 genotyped partici-
pants, 18,344 met criteria for inclusion in the current study and,
from this group, the participants in the current study were ran-
domly selected, with stratification for age, sex and ethnicity.
Overall, 9,138 individuals enrolled in the study and are included in the current analysis. All recruitment and study procedures were completed within a 2-year period from November 2009 to November 2011. The sample included ambulatory children in stable health, proficient in English, physically and cognitively capable of participating in an interview and performing the computerized neurocognitive testing. Youths with disorders that markedly impaired motility or cognition (e.g., significant paresis or palsy, intellectual disability) were excluded. Notably, participants were not recruited from psychiatric clinics, and the sample did not include individuals who sought psychiatric help. Participants were not necessarily receiving ongoing medical care at the Children’s Hospital but had received care at least once from this hospital and provided a genetic sample. Information about ongoing medical care is available but utilization of health services was not analyzed for this study. Participants and their parents provided informed consent (or assent) and agreed to be recontacted for participation in further studies.

### Table 1
Sample Demographics From the Philadelphia Neurodevelopmental Cohort

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Overall (n = 9010)</th>
<th>Females (n = 4685)</th>
<th>Males (n = 4325)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Race (n)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>2859</td>
<td>1572</td>
<td>1287</td>
</tr>
<tr>
<td>American Indian</td>
<td>14</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Asian</td>
<td>89</td>
<td>53</td>
<td>36</td>
</tr>
<tr>
<td>Caucasian</td>
<td>5113</td>
<td>2537</td>
<td>2576</td>
</tr>
<tr>
<td>Hawaiian/Pacific Island</td>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>More than one</td>
<td>926</td>
<td>511</td>
<td>415</td>
</tr>
</tbody>
</table>

*Note.* Means and standard deviations for demographic variables. Minimum and maximum values are included in brackets.

### Table 2
Standardized neurocognitive performance accuracy by age and sex

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
<th>Mean ± SD</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Males</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>444</td>
<td></td>
<td>0.56 ± 1.11</td>
<td>425</td>
<td>-0.85 ± 1.21</td>
<td>420</td>
<td>0.63 ± 1.37</td>
<td>359</td>
<td>-0.22 ± 1.08</td>
<td>361</td>
<td>-0.53 ± 0.87</td>
<td>362</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>361</td>
<td></td>
<td>-0.30 ± 1.05</td>
<td>351</td>
<td>-0.73 ± 1.20</td>
<td>347</td>
<td>-0.26 ± 1.06</td>
<td>344</td>
<td>-0.11 ± 1.10</td>
<td>390</td>
<td>-0.34 ± 0.90</td>
<td>394</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>395</td>
<td></td>
<td>-0.13 ± 1.02</td>
<td>383</td>
<td>-0.54 ± 1.16</td>
<td>377</td>
<td>-0.16 ± 1.03</td>
<td>333</td>
<td>-0.09 ± 0.96</td>
<td>332</td>
<td>-0.26 ± 0.92</td>
<td>332</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>333</td>
<td></td>
<td>-0.08 ± 1.01</td>
<td>321</td>
<td>-0.33 ± 1.05</td>
<td>323</td>
<td>-0.04 ± 0.90</td>
<td>336</td>
<td>-0.13 ± 1.08</td>
<td>333</td>
<td>-0.09 ± 0.92</td>
<td>331</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>335</td>
<td></td>
<td>0.12 ± 0.99</td>
<td>331</td>
<td>-0.15 ± 0.93</td>
<td>328</td>
<td>0.08 ± 0.96</td>
<td>334</td>
<td>0.03 ± 0.92</td>
<td>333</td>
<td>0.00 ± 0.94</td>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>334</td>
<td></td>
<td>0.06 ± 0.96</td>
<td>328</td>
<td>-0.10 ± 1.02</td>
<td>326</td>
<td>0.17 ± 0.77</td>
<td>343</td>
<td>-0.02 ± 1.00</td>
<td>342</td>
<td>0.10 ± 0.93</td>
<td>342</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>342</td>
<td></td>
<td>0.20 ± 0.89</td>
<td>341</td>
<td>0.07 ± 0.83</td>
<td>337</td>
<td>0.15 ± 0.77</td>
<td>366</td>
<td>0.09 ± 0.88</td>
<td>367</td>
<td>0.14 ± 0.94</td>
<td>368</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>368</td>
<td></td>
<td>0.15 ± 0.92</td>
<td>368</td>
<td>0.18 ± 0.88</td>
<td>363</td>
<td>0.29 ± 0.78</td>
<td>353</td>
<td>0.02 ± 0.93</td>
<td>352</td>
<td>0.23 ± 0.97</td>
<td>353</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>352</td>
<td></td>
<td>0.21 ± 0.96</td>
<td>349</td>
<td>0.27 ± 0.76</td>
<td>348</td>
<td>0.33 ± 0.71</td>
<td>305</td>
<td>0.10 ± 0.83</td>
<td>304</td>
<td>0.29 ± 0.93</td>
<td>304</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>303</td>
<td></td>
<td>0.28 ± 0.93</td>
<td>303</td>
<td>0.33 ± 0.75</td>
<td>302</td>
<td>0.33 ± 0.67</td>
<td>333</td>
<td>0.06 ± 0.93</td>
<td>332</td>
<td>0.25 ± 0.90</td>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>332</td>
<td></td>
<td>0.27 ± 0.88</td>
<td>330</td>
<td>0.42 ± 0.73</td>
<td>327</td>
<td>0.33 ± 0.82</td>
<td>223</td>
<td>0.07 ± 1.02</td>
<td>222</td>
<td>0.39 ± 0.96</td>
<td>221</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>223</td>
<td></td>
<td>0.29 ± 0.82</td>
<td>222</td>
<td>0.40 ± 0.76</td>
<td>216</td>
<td>0.41 ± 0.51</td>
<td>128</td>
<td>0.09 ± 0.88</td>
<td>128</td>
<td>0.20 ± 0.94</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>127</td>
<td></td>
<td>0.43 ± 0.87</td>
<td>128</td>
<td>0.55 ± 0.63</td>
<td>124</td>
<td>0.39 ± 0.78</td>
<td>57</td>
<td>0.05 ± 1.00</td>
<td>56</td>
<td>0.21 ± 0.93</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>57</td>
<td></td>
<td>0.23 ± 0.94</td>
<td>57</td>
<td>0.59 ± 0.50</td>
<td>57</td>
<td>0.45 ± 0.95</td>
<td>3</td>
<td>0.06 ± 0.70</td>
<td>3</td>
<td>0.34 ± 0.75</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Overall Inclusion Criteria
Several criteria had to be met for inclusion in the study. Participants ages 18 and over had to be able to provide signed informed consent. For participants under age 18, assent and parental consent were required. All participants had to be proficient in English and physically and cognitively able to participate in computerized neurocognitive testing. In addition to the neurocognitive tests (see Procedures and Measures), participants were questioned about general medical conditions, and electronic medical records were reviewed. Probands (age 11–21) and collaterals (of probands age 8–17) also received a structured interview to screen for major domains of psychopathology, which was adapted from the Kiddie-SADS (Kaufman et al., 1997).
Procedures and Measures

The CNB was administered using a Web-based system (see R. C. Gur et al., 2010). The battery consists of 14 tests in 5 domains:

Executive-Control

Abstraction & mental flexibility. The Conditional Exclusion Test presents four objects at a time, and the participant selects the object that does not belong with the rest based on one of three sorting principles. Sorting principles change, and feedback guides their identification (R. C. Gur et al., 2010; Kurtz, Ragland, Moberg, & Gur, 2004).

Attention. The Continuous Performance Test presents 7-segment displays at a rate of 1/sec. The participant must press the space bar whenever the display forms a digit or a letter (R. C. Gur et al., 2001; R. C. Gur et al., 2010; Kurtz, Moberg, Gur, & Gur, 2001).

Working memory. The Letter N-back test displays sequences of uppercase letters. In the 0-back condition, participants respond to the letter X. In the 1-back condition they respond if the letter is identical to that preceding it. In the 2-back condition, they respond if the letter is identical to that presented two trials back (R. C. Gur et al., 2001; R. C. Gur et al., 2010; Ragland et al., 2002).

Episodic Memory

Verbal. The Word Memory Test presents 20 target words that are then mixed with 20 foils equated for frequency, length, concreteness, and imageability. Participants are asked to memorize the target words as they are presented and then to indicate whether a word presented was included in the target list (Glahn, Gur, Ragland, Cen-sits, & Gur, 1997; R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Facial. The Face Memory Test presents 20 digitized faces subsequently intermixed with 20 foils equated for age, sex, and ethnicity. Participants indicate whether or not they recognize each face (Glahn et al., 1997; R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Spatial. The Visual Object Learning Test uses Euclidean shapes as stimuli with the same paradigm as the word and face memory tests (Glahn et al., 1997; R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Complex Cognition

Language reasoning. The Verbal Reasoning Test consists of verbal analogy problems with simplified instructions and vocabulary (R. C. Gur et al., 1982; R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Nonverbal reasoning. The Matrix Reasoning Test consists of matrices requiring reasoning by geometric analogy and contrast principles (R. C. Gur et al., 2010).

Spatial ability. The Line Orientation Test presents two lines at an angle, and participants click on a button to make one line rotate until it has the same angle as the other. The relative location of the lines and their sizes differ across trials (R. C. Gur et al., 2010).

Social Cognition

Emotion identification. The Emotion Identification Test displays 5 faces, one of which shows an emotionally neutral expres-

<table>
<thead>
<tr>
<th></th>
<th>SMEM_A_z</th>
<th>LAN_A_z</th>
<th>NVR_A_z</th>
<th>SPA_A_z</th>
<th>EID_A_z</th>
<th>EDI_A_z</th>
<th>ADI_A_z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.17</td>
<td>1.06</td>
<td>0.42</td>
<td>0.19</td>
<td>0.59</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>SD</td>
<td>0.02</td>
<td>1.06</td>
<td>0.42</td>
<td>0.19</td>
<td>0.59</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>n</td>
<td>0.11</td>
<td>0.94</td>
<td>0.44</td>
<td>0.17</td>
<td>0.57</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>n</td>
<td>0.08</td>
<td>0.97</td>
<td>0.36</td>
<td>0.15</td>
<td>0.57</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>SD</td>
<td>0.02</td>
<td>1.06</td>
<td>0.42</td>
<td>0.19</td>
<td>0.59</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>Mean</td>
<td>0.17</td>
<td>1.06</td>
<td>0.42</td>
<td>0.19</td>
<td>0.59</td>
<td>0.75</td>
<td>0.10</td>
</tr>
<tr>
<td>SD</td>
<td>0.02</td>
<td>1.06</td>
<td>0.42</td>
<td>0.19</td>
<td>0.59</td>
<td>0.75</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Sensorimotor Speed

Sensorimotor processing speed. The Motor Praxis task requires moving the mouse and clicking on a green square that disappears after the click. With each click, the square gets smaller and changes from one of a number of locations to another location in a random fashion (R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Simple motor speed. The Finger Tapping Test measures how quickly the participant can press the spacebar using only the index finger (R. C. Gur et al., 2001; R. C. Gur et al., 2010).

Each test provides measures of accuracy and speed, except for sensorimotor tests, which measured only speed. A brief standardized reading test from the Wide Range Achievement Test (WRAT4; Wilkinson & Robertson, 2006) was administered to determine participants’ ability to complete the battery and to provide an estimate of IQ.

Participants were given the option of being tested at home or in the laboratory. The option of home testing was given to increase representativeness of the sample by minimizing recruitment obstacles (e.g., transportation, distance, time, scheduling) and to accommodate the large number of participants: as many as 165 participants enrolled in the study per week. Approximately 70% preferred home assessment. To ensure uniformity in procedure, a trained tester sat at the table next to the participant and read aloud the instructions that appeared on the screen. At the time of appointment scheduling, recruiters gauged the appropriateness of a home appointment by inquiring about the availability of a quiet, private space in the home that could be set up for testing. Specifically, they requested a quiet space without distractions, where others (including parents) could not overhear, and a flat surface, like a table or a desk, to set up the laptop, with two chairs that could be pulled up to work at the laptop. If such a space was not available, an in-office appointment was scheduled. Moreover, if assessors arrived at the home and the testing situation could not accommodate the above conditions, the appointment was rescheduled for an in-office appointment. Tests were administered in a fixed order that, based on prior experience, assured novelty to preserve participant interest (R. C. Gur et al., 2012). Brief breaks were offered between tests. Overall, 47 individuals had invalid
VARIABILITY IN NEUROCOGNITIVE PERFORMANCE

Data Analysis

Accuracy and speed values were transformed to their standard equivalents (z scores) based on means and SDs for the entire sample. A global index of neurocognitive performance (GNP) was calculated by averaging the standardized scores across all CNB tests. An index of within-individual variability (WIV) across all tests was calculated for each participant in SD units (see Holtzer, Verghese, Wang, Hall, & Lipton, 2008; Roalf, Ruparel, et al., 2013). Both GNP and WIV were calculated for accuracy and speed separately. These values were entered as dependent measures in ANCOVAs with age, age-squared, and sex as factors. The age-squared factor was used because initial data evaluation using Generalized Additive Models indicated significant nonlinear effects (all ps < .001) with quadratic functions for both GNP and WIV.

Parental education was measured as the average of mother and father education. Equality of variance between sexes was assessed using Levene’s F tests. Post hoc contrasts were used to interpret interactions; Satterthwaite corrections were used when unequal variances could not be assumed; corrected degrees of freedom are reported where appropriate. Cohen’s d values are presented for age-specific WIV comparisons. All ANCOVAs were corrected for race and average parental education. SAS 9.3 was used for all statistical analyses.

Results

Replicating Sex Differences Across Age

To determine whether the current complete sample replicates sex differences, we followed Gur et al.’s (2012) procedure of calculating z scores for each measure. The z scores were calculated in SAS using the PROC STANDARD routine. Age-adjusted scores were used by employing the “by age” (in whole years) option. The sample size was sufficient to generate reliable z scores for each year. When examining age effects, this option was removed so that the z scores would be based on the entire sample. Performance data by age and sex, and the number of participants completing each task are presented in Tables 2 and 3 and online Supplemental Table S2. Overall, performance was better in older (15+) than younger (under 15) participants. The results replicated previous findings for females of better accuracy in attention and better accuracy and speed for word and face memory and all social cognition tests, and for males better accuracy in spatial cognition and greater simple motor speed (see Supplement materials for additional notes for Results; see Figure 1).
Global Neurocognitive Performance: Effects of Age and Sex

For GNP accuracy (Figure 2A), the main effects of age, $F(1, 8921) = 4919.55, p < .0001$, age-square, $F(1, 8921) = 484.99, p < .0001$, and sex, $F(1, 8921) = 14.29, p = .0002$, were all significant. The interaction of age and sex, $F(1, 8921) = 9.65, p = .0019$, but not age-square and sex, was significant. As expected, GNP accuracy increased with age, peaking for females at age 19 and for males at age 20. Variance of GNP accuracy was higher in males, Levene $F(1, 9008) = 21.60, p < .0001$.

For GNP speed (Figure 2B), the main effects of age, $F(1, 8921) = 2189.58, p < .0001$, age-square, $F(1, 8921) = 783.35, p < .0001$, and sex, $F(1, 8921) = 9.55, p = .0020$, were likewise significant. The interaction of age and sex, $F(1, 8921) = 9.47, p = .0021$, but not age-square and sex, was significant. As expected, GNP speed increased with age, males being faster than females at early ages, females drawing even after age 11. Both sexes peaked at age 16 and declined thereafter, more sharply in males. Variance of GNP speed did not differ between the sexes.

Within-Individual Variability: Effects of Age and Sex

WIV for accuracy showed a nonlinear (U-shaped) relationship with age (Figure 3A). ANCOVAs showed main effects of age, $F(1, 8921) = 643.20, p < .0001$, and age-square $F(1) = 257.73, p < .0001$, but not sex. The interaction of age-square and sex, $F(1, 8921) = 9.67, p = .0019$, but not age and sex, was significant. As hypothesized, WIV decreased with age but then, unexpectedly, increased significantly after age 17 and into early adulthood, especially in males. Eight-year-old males, also had
higher WIV for accuracy than their female counterparts, \( t(768.50) = 3.57, p = .0004; \) Cohen’s \( d = 0.26 \). However, during most of adolescence, sex differences were absent until, starting in early adulthood, males again showed higher values, in particular at age 19, \( t(430.58) = 2.61, p = .0094; \) Cohen’s \( d = 0.24 \). Variance of WIV for accuracy was higher in males, \( F(1, 9008) = 29.51, p < .0001 \).

For WIV speed (Figure 3B), the main effects of age, \( F(1, 8921) = 86.24, p < .0001 \), age-square, \( F(1, 8921) = 416.90, p < .0001 \) and sex, \( F(1, 8921) = 8.92, p = .0042 \), were significant. The interaction of age and sex, \( F(1, 8921) = 18.21, p < .0001 \), but not age-square and sex, was significant. As hypothesized, WIV for speed decreased in childhood but here again for males more than for females, it increased significantly in adolescence and into early adulthood. Overall, males had higher WIV for speed than females, which became more evident with increased age. Speed WIV was comparable in males and females during childhood (8–10) and decreased yearly in both sexes until about age 15, when WIV began to increase and became higher in males than females, in particular at age 15 \( t(687.12) = 3.89, p < .0001; \) Cohen’s \( d = 0.27 \).16 [t(671.75) = 2.70, \( p = .0071; \) Cohen’s \( d = 0.20 \)], 17 [t(603.46) = 2.26, \( p = .0239; \) Cohen’s \( d = 0.20 \)], 19 [t(407.97) = 4.13, \( p < .0001; \) Cohen’s \( d = 0.35 \)], and 21 [t(98.38) = 2.33, \( p = .0218; \) Cohen’s \( d = 0.42 \)]. Variance of WIV speed was higher in males, \( F(1, 9008) = 17.39, p < .0001 \). Overall, WIVs for accuracy and speed were highest in childhood, lowest around age 15, and then higher into adulthood. Males had generally higher values than females, most clearly in late adolescence and early adulthood.

**Potential Contribution of Medical and Psychiatric Conditions**

Although the participants were physically and cognitively able to perform the tests using the computerized format, because the sample was drawn from a pediatric hospital, it is possible that, for some participants, certain medical or psychiatric conditions affected their performance. In a follow-up analysis, we excluded all participants (\( n = 2167; 51 % \) female) with significant medical conditions, such as respiratory disease (e.g., cystic fibrosis), gastrointestinal disorders (e.g., inflammatory bowel disease), immune dysfunction (e.g., common variable immunodeficiency), and endocrine/metabolic disorders (e.g., diabetes). To compare differences in health status between younger and older participants, the sample was split into two groups, younger and older than 15. The overall rate of significant medical conditions within this sample is 24.04%. Of participants under the age of 15, 20.30% had a significant medical condition compared to 28.61% of participants age 15 and older. We also generated a composite score from responses on the general mental health screen and entered it as a covariate for all participants. Neither procedure altered the pattern of results (see online Supplemental Figure S1).

**Performance Associations With WRAT4 and Parental Education**

WRAT4 score differed by age, \( F(913,8982) = 39.77, p < .0001 \) (see online Supplemental Table S1). There was a significant difference between ages 8 and 9, 12 and 13, and 15–18. As a group, individuals under 15 [104.20(15.40)] had higher WRAT4 standard
scores than individuals 15 years and older [100.20 (16.42)]. Although statistically significant, the groups differed by only 4 points on the standard scale.

Overall, parental education differed by age, $F(2, 8925) = 32.08$, $p < .0001$ (see online Supplemental Table S1). This was most evident between the parents of the 8-year-olds and 21-year-olds. Individuals under 15 [14.43 (2.34)] had higher parental education than those 15 and older [14.19 (2.31)]. The difference, however, was less than 1 year.

Correlations for each sex for CNB test scores, parental education, and WRAT4 are listed in Table 4. In general, they were similar for males and females. GNP was negatively correlated with WIV for both accuracy and speed. Parental education was positively correlated with GNP accuracy and speed and negatively correlated with WIV accuracy and speed. Estimated IQs based on the WRAT4 were average for both sexes, with an age-adjusted mean scale score of ~103 and $SD$ of ~16 for males and ~102 and $SD$ of ~16 for females, indicating that the IQs for our population-based sample were similar to those for the standardization sample. Overall, the correlations between performance and age-adjusted scale scores on the WRAT4 were in the expected directions: positively correlated with GNP accuracy and speed, negatively correlated with WIV accuracy and speed.

**Discussion**

By examining within-individual variability, we found a heretofore unreported age-related change in within-individual variability.

**Table 4**

<table>
<thead>
<tr>
<th>Correlations</th>
<th>GNP accuracy</th>
<th>GNP speed</th>
<th>WIV accuracy</th>
<th>WIV speed</th>
<th>Parental education</th>
<th>WRAT STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNP accuracy</td>
<td>—</td>
<td>.36</td>
<td>−.54</td>
<td>−.24</td>
<td>.25</td>
<td>.38</td>
</tr>
<tr>
<td>GNP speed</td>
<td>.36</td>
<td>—</td>
<td>−.23</td>
<td>.68</td>
<td>.08</td>
<td>.13</td>
</tr>
<tr>
<td>WIV accuracy</td>
<td>−.55</td>
<td>−.25</td>
<td>—</td>
<td>.21</td>
<td>−.14</td>
<td>−.22</td>
</tr>
<tr>
<td>WIV speed</td>
<td>−.19</td>
<td>−.67</td>
<td>.21</td>
<td>—</td>
<td>−.07</td>
<td>−.13</td>
</tr>
<tr>
<td>Parental education</td>
<td>.21</td>
<td>.04</td>
<td>−.11</td>
<td>.02</td>
<td>—</td>
<td>.33</td>
</tr>
<tr>
<td>WRAT STD</td>
<td>.37</td>
<td>.11</td>
<td>−.19</td>
<td>−.10</td>
<td>.39</td>
<td>—</td>
</tr>
</tbody>
</table>

*Note. Females are displayed above the diagonal, and males below the diagonal. Significant ($p < .001$) correlations are in bold type.*
of neurocognitive performance. At 8 years, the youngest age group sampled, WIV was at its highest level and then decreased every year for both accuracy and speed to its lowest level in early adolescence. This decrease with age was predicted by the hypothesis of uneven development, whereby some domains mature before others. But we also observed, unexpectedly, that after about age 14, WIV increased every year into adulthood, although never reaching the levels shown by young children. This pattern was clearer for speed than accuracy and was stronger for males than for females. We propose that this post-pubescence increase in WIV reflects the unfolding of greater specialization related to skill honing and brain maturation. Such increased specialization in speed, as indexed by increased WIV for speed that continues into adulthood, is consistent with evidence that myelination of brain white matter in some cortical regions is still incomplete at age 21, especially for males, according to Giedd et al., 1999; and myelination of the corpus callosum is still incomplete at age 21, especially for males (Giedd & Rapoport, 2010; Raznahan et al., 2011). This disproportionate development of white matter with age and sex may underlie superior speed performance in some domains but poorer performance in other domains. For example, young adult males are faster than adolescent males in motor performance, but slower during emotion identification. To the extent that certain tasks, such as motor and visual tasks, also require rapid interhemispheric transfer of information, increased speed of performance in some tasks but not others is consistent with evidence that the myelination of the corpus callosum is still incomplete at age 21 (Giedd et al., 1999; Pujol, Vendrell, Junque, Martí-Vilalta, & Capdevila, 1993).

Our results for across-task WIV are similar to those from a previous study of variability in a single reaction time (RT) task, which, like our results, showed decreased variability from childhood through adolescence, followed by an increase throughout adulthood and into old age (Dykiert et al., 2012). In our study, however, this increase began in mid-adolescence. No multitask WIV has been reported previously. Our measure of variability, which has been used before in aging research (MacDonald, Li, & Backman, 2009), is within a single person but occurs across tasks in several domains and therefore provides an index of evenness, or consistency, of performance across domains. Our finding that our youngest participants (8 years) were more uneven than our mid-adolescents (12–14 years) fits with prior reports that children show higher variation in performance (Siegler, 1994). Indeed, as previously noted, children often have high variability when solving the same problem over short or long intervals (Siegler & McGilly, 1989). What has not been reported before and was unexpected was the increase in variability in late adolescence (after age 14) and into early adulthood. We do not believe this pattern in older adolescents and young adults should be seen as a reversion to a childlike pattern of variability reflecting uneven development. Instead, we propose that it reflects increased cognitive specialization, meaning that some individuals are improving more on tasks in certain domains relative to tasks in other domains. If, indeed, this pattern reflects increased cognitive specialization, it is noteworthy that it is more evident in speed than in accuracy, at least within the time parameters of this study. Assuming we can do follow-up studies with the same participants when they are older, it is conceivable that the pattern also will be seen in accuracy.

Our hypothesis of greater across-domain variability in males was strongly supported. Males were more uneven, most notably for speed after age 15. Our results suggest that with maturation there is greater specialization in males for some domains, reflected in differential cognitive processing, whereas in females, specialization more often develops more nearly equally across a range of domains. These results are also consistent with studies indicating, as previously noted, that males are more likely than females to be “cognitive specialists,” performing some functions better than others, whereas females are more likely to become “cognitive generalists,” performing at more nearly equal levels across domains (Halpern et al., 2007). In addition to finding higher WIV in males, our results support Thorndike’s (1910) hypothesis of higher variance in males. Higher variance in males also is consistent with studies of sex differences in IQ and spatial ability (e.g., Deary et al., 2003; Shea et al., 2001), findings that we now have extended over a broader age range. Our findings also align with research showing males to be overrepresented in both extremes of the ability distribution (Thorndike, 1910, but see Hollingworth, 1914), similar to sex differences in IQ in young children (Deary et al., 2003). The sex difference in variability is small, but insofar as the differences are maximized in the tails of the distribution, it may explain why more men both over- and underachieve in comparison with women in the real world. The results support the potential utility of WIV as a dimension or metric related to cognitive development. Initial evaluation of the construct validity of this new dimension suggests that WIV correlates negatively with average performance and with parental education, which is itself positively correlated with average performance. It appears that higher parental education is associated with an environment where both better performance and greater evenness of performance are fostered. These effects also could be genetically mediated, and incorporation of the genomic data available on this sample could help shed light on this issue.

Our GNP data add to the substantial literature showing increases in neurocognitive ability throughout development (e.g., R. C. Gur et al., 2012; Steinberg, 2005). Adolescence and early adulthood are marked by significant increases in brain maturation and subsequent increases in behavioral and cognitive functioning, including improvement in information processing and reasoning, especially in abstract, organized, and logical thinking (Steinberg, 2005). The domains captured in the computerized neurocognitive battery show significant improvement with age, particularly in children and young adults (R. C. Gur et al., 2012) and can be linked to age changes in functional or neuroanatomical systems (e.g., Ragland et al., 2002). In general, males and females showed similar age-related differences in performance. The main difference was that females were more accurate earlier (age 13), whereas males were faster earlier (ages 10–11). The accuracy results fit reports of earlier physical (Hills & Byrne, 2010) and behavioral maturation (e.g., Tiemeier et al., 2010) in females. Perhaps males are more inclined to trade accuracy for speed or they are more impulsive and less able to inhibit their initial responses (Cross, Coping, & Campbell, 2011; Feingold, 1994). So far as children are concerned, any differences in performance also could be related to developmental differences in the pace of brain maturation, in particular prefrontal myelination (Gogtay et al., 2004). Notably, these sex differences do not appear to be modulated by sex hormones or puberty during early adolescence (Herlitz, Reuterskiold, Loven, Thiels, & Rehnman, 2013), although these factors may play a role in late adolescence.
The cross-sectional nature of our study limits our ability to determine developmental effects from cohort effects. However, many of the participants have agreed to be followed longitudinally and these studies are underway. Participant enrollment through pediatric clinics at the Children’s Hospital may affect generalizability as a result of comorbid medical conditions, but we have taken steps to address this potential limitation in our statistical analyses. Other cohort factors may also unintentionally influence our data. For example, our sample may be influenced by the recruitment methods used by the Children’s Hospital. However, children and young adults with physical or mental conditions likely to affect performance were excluded. Moreover, because the study was completed in a short period of time (24 months), within-cohort bias should be reduced. This is especially true in comparison with studies that required a decade or more to complete or that attempt to combine disparate data sets to achieve a large sample.

Although our sample was large and demographically diverse, we included individuals with medical conditions (e.g., neurological or neuropsychological) that could have affected their performance and skewed some of the data. The prevalence of chronic health conditions in our sample (24.04%) is similar to a recent study of approximately 5,000 children (from three cohorts) over the course of 6 years (20.8%; Van Cleave, Gottmaker & Perrin, 2010). Indeed, our overall prevalence is lower than the prevalence reported in one of the cohorts (26.6%) in this study. Furthermore, the presence of significant medical conditions does not imply permanence because medical conditions change over time, more effective treatments and access to health care may become more available, and the natural course of development may affect the condition. Van Cleave et al. also note that the persistence of medical conditions within a given individual in their study was relatively low, approximately 7%. Thus, simply excluding individuals based on significant medical conditions at one time point may ignore natural developmental milestones that help resolve these medical issues and their impact on cognition. Finally, our follow-up analysis that excluded these individuals and controlled for general mental health showed similar results. Notably, the average WRAT4 score for our sample was identical to that obtained in the normative sample, indicating that our sample can provide a valid approximation of typical development. Another limitation is that our final sample included more females than males. Because standardization of CNB scores was based on the entire sample, estimations for some effects in certain age groups could have been imprecise because of disparities in the numbers of males and females, although standardization across the sample should have minimized any systematic distortion of parameter estimates. Given the modest number of tests in each neurobehavioral domain, global scores used for performance and variability may have been biased by the test instruments used. However, because previous studies (R. C. Gur et al., 2012; R. C. Gur et al., 2010) have consistently shown that the CNB is balanced for sex differences, we believe the instrument provides a robust set of tests for measuring global performance and variability. Differences in the parental education and WRAT4 scores of younger and older participants were minimal but could potentially explain some of the differences in performance. However, including parental education and WRAT4 in the statistical model did not change the results for any measure, indicating minimal effects of these moderating variables. Although age-related differences in WIV look systematic, many factors likely to contribute to variability. Differences in WIV therefore may depend to a large extent on the specific constellation of abilities being measured; if so, its generalizability needs further scrutiny. Longitudinal follow-up studies could expand upon the variability findings by examining variability across time within specific tests and assess additional underlying factors that may affect variability such as comorbid health conditions, substance use, or other social or environmental factors.

Notwithstanding these limitations, the current study of more than 9,000 children and youths demonstrates age- and sex-related differences in global neurocognitive performance along with significant within-individual variability. These data provide a framework for the longitudinal study of individual neurocognitive developmental trajectories and potential deviations from those trajectories similar to those found in longitudinal studies of physical growth, intelligence, and personality (e.g., Bayley & Schaefer, 1964; Bradley & Caldwell, 1976; Crano & Mendoza, 1987; Kagan & Moss, 1959). The difference is the new window that WIV offers into behavioral architecture and its development. It could signify a normal variant of degree of cognitive specialization and, when abnormally high, it could index vulnerability or giftedness. Measuring variability of performance also could aid in the identification of critical developmental periods within individuals and provide a benchmark for evaluation of developmental disorders. By screening children for existing neurocognitive problems (e.g., higher than normal WIV for a given sex/age), cognitive weaknesses can be assessed and monitored in a way that can lead to targeted educational or clinical interventions (Sege & De Vos, 2008). In combination with follow-up clinical evaluation and measurement of neurobiological markers (e.g., neuroimaging data, genomics), such data have the potential to unveil specific mechanisms that may underlie developmental traits as well as genetic variants possibly underlying key individual differences, including those related to age and sex. The next step will be to make the neuro-cognitive connection. For example, do individual differences in prefrontal development predict individual differences in variability of performance on cognitive tests that presumably tap into prefrontal functions as well as on tests in other cognitive domains? We have extensive neuroimaging data on a subsample of about 1500 children that could be used to test such hypotheses.

References


