

Age Differences in Strategic Planning as Indexed by the Tower of London

Dustin Albert and Laurence Steinberg

Temple University

The present study examined age differences in performance on the Tower of London, a measure of strategic planning, in a diverse sample of 890 individuals between the ages of 10 and 30. Although mature performance was attained by age 17 on relatively easy problems, performance on the hardest problems showed improvements into the early 20s. Furthermore, whereas age-related performance gains by children and adolescents (ages 10–17) on the hardest problems were partially mediated by maturational improvements in both working memory and impulse control, improved performance in adulthood (ages 18+) was fully mediated by late gains in impulse control. Findings support an emerging picture of late adolescence as a time of continuing improvement in planned, goal-directed behavior.

There is broad consensus that the protracted maturation of the prefrontal cortex (PFC) and its reciprocal connections to other regions contributes to the relatively late gains in efficiency of cognitive control processes in adolescence (e.g., Casey, Getz, & Galvan, 2008). Cognitive control processes, often referred to as *executive functions*, include a suite of abilities enabling successful planning and enactment of goal-directed behavior. Recent research suggests that developmental gains in cognitive control in late adolescence contribute to the progressive capacity to make mature decisions, particularly in risky contexts (Steinberg, 2008). Whereas young adolescents evince an adult-like capability to reason logically about the costs and benefits of decision alternatives, aspects of psychosocial maturity reflecting self-regulatory control (e.g., resistance to peer influence, impulse control, future orientation) show improvements across the course of adolescence and often into the early 20s (Albert & Steinberg, 2011).

Importantly, research charting age differences in such capacities is increasingly consulted as a source of guidance for social and legal policies concerning adolescents. For instance, in a landmark Supreme

Court decision overturning the juvenile death penalty for individuals under 18 years old (*Roper v. Simmons*, 2005), the majority opinion specifically cited evidence from developmental science to argue that adolescents lacked the psychosocial maturity to be held culpable for their crimes to the same degree as adults (Steinberg, Cauffman, Woolard, Graham, & Banich, 2009). Furthermore, interventions aimed at reducing adolescent risk behavior, to the degree that they are successful, must capitalize on an understanding of adolescents' strengths and weaknesses in the self-regulatory domain (Albert & Steinberg, 2011). A fuller understanding of the normative course of self-regulatory development is therefore critical for informing legal and social policies relevant to the health, well-being, and judicial treatment of adolescents.

The planning and enactment of future-directed behavior is likely to involve a variety of higher order control processes, including inhibition of attention to distracting stimuli, sustained suppression of impulsive responding, and the capacity to maintain goal representations in working memory (WM; Unterrainer & Owen, 2006). Luna, Garver, Urban, Lazar, and Sweeney (2004) found that oculomotor indexes of processing speed, response inhibition (RI), and WM did not reach adult levels of maturity until middle to late adolescence, with WM the last to reach asymptote, at age 19. Each process showed a steep increase in performance from childhood through adolescence, followed by a plateau

This research was supported by the John D. and Catherine T. MacArthur Foundation Research Network on Adolescent Development and Juvenile Justice. We are grateful to our project collaborators, including Marie Banich, Elizabeth Cauffman, Sandra Graham, and Jennifer Woolard, for insightful critiques of drafts of this manuscript. Special thanks go to Keith Berg for his skillful tutelage on the nuances of the Tower of London, and to Jason Chein and Brad Conner for their generous conceptual and statistical support.

Correspondence concerning this article should be addressed to Dustin Albert, Department of Psychology, Temple University, Philadelphia, PA 19122. Electronic mail may be sent to dustin.albert@temple.edu.

© 2011 The Authors

Child Development © 2011 Society for Research in Child Development, Inc.
All rights reserved. 0009-3920/2011/8205-0013

DOI: 10.1111/j.1467-8624.2011.01613.x

from adolescence through adulthood, similar to the age function identified for synaptic pruning in the PFC (Huttenlocher, 1990)

Luciana and colleagues found that adolescents reached adult-level performance at progressively later ages for nonverbal WM tasks requiring increasing levels of executive control, with a measure of strategic self-organization showing the latest gains, through ages 16–17 (Luciana, Conklin, Hooper, & Yarger, 2005). The authors interpreted the findings as evidence for a functional dissociation between performance on tasks requiring only the maintenance of information in WM, which relies primarily on ventrolateral PFC, versus tasks requiring strategic self-monitoring and executive control of the contents of WM, which heavily recruit dorsolateral PFC (DLPFC), one of the last brain regions to reach structural and functional maturity in late adolescence (Giedd, 2008).

Perhaps the most informative research on controlled problem solving utilizes the Tower of London (TOL) task, created by Shallice (1982) to study deficits in goal-directed behavior commonly observed among patients with frontal-lobe lesions. On the classic version of the TOL, participants are presented with a test instrument that consists of three differently colored balls placed in a variable configuration on three rods of progressively smaller size. The object of each trial is to rearrange the balls, one at a time and in the minimum number of moves, to match a separately presented goal configuration. Trials are designed to vary in difficulty based on the minimum number of moves required to achieve the solution. Whereas simpler trials may be solved directly through a perceptual match-to-sample strategy, trials with a high minimum number of moves require a sequence of intermediate moves to reach the final goal state. Thus, difficult trials assess complex, integrative problem solving by requiring goal directedness, strategic planning of subgoals (including removal of obstacles), inhibition of prepotent responses, and recall-guided action (Berg & Byrd, 2002). In short, the difficult TOL trials require planning and sustained cognitive control of behavior toward a goal.

Evidence from lesion, pathology, and neuroimaging studies employing the TOL clearly demonstrates the importance of the PFC, and specifically dorsolateral and rostral portions of the PFC, for complex problem solving. Deficits in efficient problem solving have been reported among patients with a variety of unilateral or bilateral frontal-lobe lesions (Unterrainer & Owen, 2006), as well as patients with diverse pathologies of the frontal

lobes or frontostriatal system, including schizophrenia (Morris, Rushe, Woodruffe, & Murray, 1995), Parkinson's disease (Owen et al., 1992), Huntington's disease (Watkins et al., 2000), and attention deficit hyperactivity disorder (ADHD; Young, Morris, Toone, & Tyson, 2007). Furthermore, fMRI findings suggest that specific activation of DLPFC and rostralateral PFC support performance on the most difficult TOL problems (van den Heuvel et al., 2003; Wagner, Koch, Reichenbach, Sauer, & Schlösser, 2006). Given evidence that these brain regions are among the last to reach structural and functional maturity in adolescence (Giedd, 2008), it is reasonable to predict that optimal performance on the TOL will not be attained until late adolescence or early adulthood.

In the largest study to date of age differences in TOL performance, significant age gains were seen in the ability to perfectly solve TOL problems across a sample of 800 children aged 5–12 (Korkman, Kemp, & Kirk, 2001). However, because these data derived from a NEPSY standardization study that did not include an adolescent or adult comparison group, they do not address the question of when TOL problem solving reaches developmental maturity. Several studies utilizing broader age ranges have shown TOL performance gains through the adolescent years. For example, a TOL study of four different age groups (7-, 11-, 15-, and 21-year-olds) found continued gains in problem-solving efficiency between ages 15 and 21 (Huizenga, Dolan, & van der Molen, 2006). Asato, Sweeney, and Luna (2006) reported similar findings in a study of individuals between 8 and 30, with increasing age predicting better performance on more difficult trials of the TOL (> 3 minimum moves). Other cross-sectional studies of age differences in TOL performance have identified performance plateaus as occurring somewhere between ages 15 and 30, depending on the variation of the TOL task employed and the construction of age comparison groups (Anderson, Anderson, Northam, Jacobs, & Catroppa, 2001; De Luca et al., 2003; Luciana & Nelson, 2002; Raizner, Song, & Levin, 2002).

Because of the relatively small sample sizes and restricted age ranges that are common to these studies, it is difficult to conclude whether problem-solving performance is largely mature by middle adolescence (i.e., closer to 15), or whether subtle performance gains continue into early adulthood. Furthermore, most studies utilized TOL problems with a restricted range of difficulty (typically between two and five minimum moves), decreasing

the likelihood of identifying performance differences between adolescents and young adults, which are typically observed for only the most difficult problems. In order to fully describe the developmental course of cognitive processes undergirding TOL performance, it is necessary to have (a) a large enough sample to differentiate among chronological ages, (b) an age range that spans the teens and 20s, and (c) sufficient variability in problem difficulty.

A second question concerns the degree to which developmental improvements in planning and problem solving on the TOL are dependent upon advances in cognitive abilities like WM capacity, manipulation of information in WM (i.e., WM updating), and inhibitory processing, all of which are known to mature between middle and late adolescence (Steinberg, 2008). To plan a solution on a difficult TOL trial, one must mentally represent a path from the start state to the goal state, requiring multiple intermediate steps organized as subgoal operations. As one subgoal operation is mentally enacted, the problem state representation must be updated and new alternative operations evaluated, a process that must be repeated until the goal state is reached and the sequence of operations can be behaviorally reproduced. Clearly, there are basic processing efficiency and capacity prerequisites for mastering such a complex cognitive task, including but not necessarily limited to visuospatial (VS) reasoning skills. For more difficult problems, the ability to maintain and update the problem state representation across successive subgoal operations may require a relatively advanced capacity to select what information goes into WM, hold that information in WM across a delay, and accurately update the contents of WM as new information is processed. Likewise, optimal performance on the TOL may require a mature inhibitory processing system, including the ability to inhibit attention to distracting stimuli, to avoid making seemingly obvious but actually counterproductive “trap” moves, and more generally to delay immediate responding while completing the full planning phase.

The few studies that have investigated the contributions of WM and inhibitory processing to age-related gains on the TOL have reported inconsistent results. Asato et al. (2006) found that oculomotor measures of RI (antisaccades) and WM capacity (i.e., memory guided saccades) predicted TOL solution efficiency across their 8- to 30-year-old sample. In contrast, Huizenga et al. (2006) found WM capacity and WM manipulation did not significantly predict age differences in TOL performance

across four age groups: 7, 11, 15, and 21; only a measure of RI on a modified Stroop task was a significant predictor, and only among 21-year-olds. Adult studies are also mixed. Whereas some studies have predicted TOL problem solving from VS WM (Gilhooly, Wynn, Phillips, Logie, & Della Sala, 2002; Welsh, Satterlee-Cartmell, & Stine, 1999) and RI (Miyake et al., 2000; Welsh et al., 1999), others have failed to replicate these associations and instead point to contributions from basic measures of fluid, VS intelligence (Unterrainer et al., 2004; Zook, Davalos, DeLosh, & Davis, 2004).

Although these divergent findings can be explained in large part by differences in the structure and scoring of TOL and TOH task variants (see Berg & Byrd, 2002), the inconsistent findings for RI merit closer analysis. The studies described above typically assessed RI using either the Stroop (i.e., the degree to which an individual can resist attentional interference from salient stimuli), or one of several tasks assessing the capacity to inhibit an automatic motor response (e.g., Go/No-Go, Stop Signal, or oculomotor antisaccade). Friedman and Miyake (2004) aptly described these and similar tasks as measures of *prepotent response inhibition*, a latent subclass of RI representing the capacity to inhibit automatic (i.e., prepotent) attention or motor responses. Given that the TOL does not establish a strong prepotent response to inhibit, it is not surprising that performance on prepotent RI tasks is inconsistently related to age and individual differences in TOL performance.

Then what aspect of inhibitory processing is important for planned problem solving on the TOL? One possible alternative is suggested by a measure derived from the TOL itself—the capacity to delay responding until planning is complete. Assessed as the length of time that an individual waits before initiating a solution for a given problem, *first-move latency* is consistently predictive of TOL performance, such that individuals with longer average first-move latency correctly solve more trials in fewer moves (Mitchell & Poston, 2001), and instructing participants to fully plan a solution before acting is associated with both increased first-move latency and problem-solving efficiency (Unterrainer, Rahm, Leonhart, Ruff, & Halsband, 2003). This association between planning time and performance has been cited as one possible explanation for the TOL deficits seen in children and adults with ADHD. For instance, a recent study found that whereas healthy adults progressively increased their planning time as problems grew in difficulty, adults with ADHD (matched on age, IQ, and social

class) waited no longer to act on difficult than on easy problems, and consequently performed worse on difficult problems relative to healthy controls (Young et al., 2007). It is also plausible that developmental changes in the tendency to wait before responding on the TOL could account for age-related variance in problem-solving performance. In a previous report that utilized response time (RT; but no other TOL outcome) data from the present data set, we demonstrated linear age gains in first-move latency from 10 through 30 years, with 26- to 30-year-olds waiting significantly longer to act than all other age groups (Steinberg et al., 2008).

In sum, it is likely that longer waiting times on the TOL reflect not only the inhibition of immediate, automatic responding, but also the capacity to sustain this inhibition in support of planning an optimal solution; as such, we consider TOL first-move latency as an index of *impulse control*. This distinction between prepotent RI and a higher order capacity for impulse control is consistent with evidence that children show a dissociation between performance on tasks requiring management of conflicting attentional demands (i.e., a subclass of RI) and tasks requiring sustained delay of behavior, which the authors also referred to as measures of impulse control (Carlson & Meltzoff, 2008). At a broader level, we view impulse control as a higher order self-regulatory capacity, similar to Nigg's (2000) "executive inhibition" construct, which he associates with relatively late developmental gains in conscientiousness, agreeableness, and planful control, alongside a parallel decline in impulsivity. Whereas aspects of VS reasoning, WM, and prepotent RI are likely essential prerequisites for planned problem solving on the TOL, we predict that an additional component contributes to optimal TOL performance—the development of *impulse control*, reflected by prolonged planning of behavior prior to action.

The present study examines age differences in strategic planning on the TOL in a large and ethnically diverse sample ranging from 10 to 30 years old. Consistent with lesion and neuroimaging evidence that optimal performance on the TOL recruits anterior portions of the PFC, and longitudinal MRI findings suggesting that these brain regions are among the latest to mature, we hypothesize that developmental improvements on the TOL will be evident well into the late adolescent years. Furthermore, we hypothesize that age differences in TOL problem solving will be partially mediated by developmental improvements in VS reasoning, WM capacity, WM updating, and prepotent RI. In addition,

we predict that unique age-related variance in performance will be accounted for by age gains in impulse control, operationalized as the amount of time an individual waits before acting.

Method

Participants

Data for the present study were collected from five sites: Denver, Colorado; Irvine, California; Los Angeles, California; Philadelphia, Pennsylvania; and Washington, DC. The combined sample includes 935 individuals ranging in age from 10 to 30 years old. This age range was selected to allow for examination of age differences within the adolescent decade, as well as comparison of adolescent subgroups to individuals in their late teens and early to mid-20s, an age period when the PFC is still maturing, and to individuals in their late 20s, a period in which PFC maturation is, presumably, largely complete (Giedd, 2008). Due to data recording errors associated with the computerized version of the TOL, 39 participants had invalid data on primary TOL outcome measures, and were therefore dropped from the analysis. In addition, 6 individuals were dropped due to missing demographic data, resulting in an analytic sample of 890 individuals. Because all missing cases resulted from technical or administrative error, such cases can be considered missing at random; no significant differences were found between age groups in the proportion of valid cases, $\chi^2(6, N = 935) = 8.67, ns$. To facilitate analysis of age differences with sufficient statistical power, we created the following age groups: 10–11 years ($n = 109$; 56 females), 12–13 years ($n = 130$; 63 females), 14–15 years ($n = 122$; 55 females), 16–17 years ($n = 140$; 74 females), 18–21 years ($n = 141$; 68 females), 22–25 years ($n = 133$; 73 females), and 26–30 years ($n = 115$; 64 females).

The sample was evenly split between males (49%) and females (51%) and was ethnically diverse, with 29% African Americans, 15% Asian Americans, 22% Latino(a)s, 24% Whites, and 10% Others. Participants were predominantly working and middle class. Each site contributed an approximately equal number of participants, although site contributions to ethnic groups were disproportionate, reflecting the demographics of each locale.

Procedure

Prior to data collection, all site project directors and research assistants met at one location for

several days of training to ensure consistent task administration across data collection sites. The project coordinators and research assistants conducted on-site practice protocol administrations prior to enrolling participants.

Participants were recruited via newspaper advertisements and flyers posted at community organizations, Boys and Girls Clubs, churches, community colleges, and local places of business in neighborhoods targeted to have an average household education level of "some college" according to 2000 U.S. census data. Individuals who were interested in the study were asked to call the research office listed on the flyer. Members of the research team described the nature of the study to the participant over the telephone and invited those interested to participate. Given this recruitment strategy, it was not possible to know how many participants saw the advertisements, what proportion responded, and whether those who responded are different from those who did not.

Data collection took place at an office at a participating university or a location in the community where it was possible to administer the test battery in a quiet and private location. Before beginning, participants were provided verbal and written explanations of the study, their confidentiality was assured, and their written consent or assent was obtained. For participants who were under the age of 18, informed consent was obtained from either a parent or guardian.

Participants completed a 2-hr assessment that consisted of a series of computerized tasks, a set of computer-administered self-report measures, a demographic questionnaire, and an assessment of Intelligence Quotient (IQ). The tasks were administered in individual interviews. Research assistants were present to monitor the participant's progress, reading aloud the instructions as each new task was presented and providing assistance as needed. To keep participants engaged in the assessment, participants were told that they would receive \$35 for participating in the study and that they could obtain up to a total of \$50 (or, for the participants under 14, an additional prize of approximately \$15 in value) based on their performance on the computer tasks. In actuality, we paid all participants ages 14–30 the full \$50, and all participants ages 10–13 received \$35 plus the prize. This strategy was used to increase the motivation to perform well on the tasks but ensure that no participants were penalized for their performance. All procedures were approved by the Institutional Review Board of the university associated with each data collection site.

Measures

The present analyses utilize data from the demographic questionnaire, a self-report measure of impulsivity, a standardized assessment of intelligence, two computerized tests of WM, a computerized test of prepotent RI, and a computerized version of the TOL task.

Demographics. Participants reported their age, gender, ethnicity, and household education. The age groups did not differ with respect to gender, although they did marginally differ with respect to ethnicity, $\chi^2(24, N = 890) = 36.62, p = .048$. To locate the source of this difference, we examined standardized residuals for each cell, which represent the degree to which an ethnic group was over- or underrepresented in a given age group; residuals greater than $z = \pm 1.96$ were considered significant at $p < .05$. Among 16- to 17-year-olds, there were more African American ($z = 3.3$) and fewer European American ($z = -3.5$) participants than expected, and among 18- to 19-year-olds, there were fewer African American ($z = -2.0$) and more Asian American ($z = 2.4$) participants than expected. No other significant age differences in ethnic composition were found. To index household education, we utilized reports of parents' highest education level for individuals under 18, and individuals' own educational attainment for those 18 and older. We recognize that using respondents' current level of attainment as a household educational index among college-enrolled individuals aged 18 and older may misrepresent these individuals' actual background, because college students who are adults are coded as having attained "some college" when in fact their parents may have attained more or less than this. However, a strength of the present study is that our young adult sample is not exclusively composed of college undergraduates but rather includes both students and nonstudents. There is no consensus, when studying young adults, about how best to characterize their household education. Although an omnibus analysis of variance showed a marginally significant difference between the age groups in household education, $F(6, 874) = 2.11, p = .05$, further examination utilizing the Tukey post hoc procedure for multiple comparisons revealed no significant differences between any two age groups. Gender, ethnicity, and household education are specified as covariates in all subsequent analyses.

Impulsivity. A widely used self-report measure of impulsivity, the Barratt Impulsiveness Scale, Version 11 (Patton, Stanford, & Barratt, 1995), was part

of the questionnaire battery and has been shown to have good construct, convergent, and discriminant validity. The scale has 30 items comprising six subscales, including motor impulsivity (e.g., "I act on the spur of the moment"), inability to delay gratification (e.g., "I spend more money than I should"), lack of perseverance (e.g., "It's hard for me to think about two different things at the same time"), attention (e.g., "I am restless at movies or when I have to listen to people"), cognitive complexity ("I am a great thinker"), and self-control ("I plan for my future"). Each item is scored on a 4-point scale (*rarely/never, occasionally, often, almost always/always*), with higher scores indicative of greater impulsivity. Item responses were averaged to form a total impulsivity score. Interitem reliability of the full scale was acceptable in the current sample ($\alpha = .80$). Because each of the individual subscales demonstrated unacceptably low reliability (all α coefficients $< .7$), we only utilized full-scale impulsivity scores in the present analyses.

Intelligence. The Wechsler Abbreviated Scale of Intelligence (WASI) Full-Scale IQ Two-Subtest (FSIQ-2; Psychological Corporation, 1999) was used to produce an estimate of general intellectual ability based on two subtests corresponding to crystallized and fluid intelligence (Vocabulary and Matrix Reasoning, respectively). The WASI can be administered in approximately 15 min and is correlated with the Wechsler Intelligence Scale for Children ($r = .81$) and the Wechsler Adult Intelligence Scale ($r = .87$). It has been normed for individuals between the ages of 6–89 years. Small but significant differences were found between the age groups in Full-Scale IQ, $F(6, 882) = 4.17, p < .001$, such that 10- to 11-year-olds scored higher than 14- to 15- or 16- to 17-year-olds, and 16- to 17-year-olds also scored lower than 22- to 25-year-olds. To account for the possibility that potential age differences in TOL performance result from IQ differences between the age groups, IQ was specified as a covariate in analyses examining age effects on the TOL.

Because a second set of analyses specifically examines the cognitive capacities that potentially mediate age differences in TOL performance, we also derived a measure of *visuospatial reasoning* based on raw scores from the WASI's Matrix Reasoning subtest. For each Matrix Reasoning problem, the participant is instructed to examine a VS pattern with a missing component, and identify which of five possible choices best completes the pattern. Raw scores are recorded as the number of problems answered correctly. In contrast to the Full-Scale IQ score, which is standardized based on age norms,

this measure of VS reasoning preserves age differences in performance.

Prepotent response inhibition. A computerized version of the classic Stroop Color-Word task was administered to assess prepotent RI (Stroop, 1935). On each trial, the participant was presented a color word (e.g., *blue, yellow*) and instructed to identify the color in which the word was printed (while ignoring the semantic meaning of the word) by pressing a corresponding key as quickly as possible. Trials varied on whether the color word and the printed color of the word were congruent or incongruent. Participants completed two 48-trial experimental blocks. The first block included an equal mix of congruent and incongruent trials, and the second included a greater number of congruent than incongruent trials. Utilizing all trials, we calculated interference effects for RT and accuracy as the difference in average RT and ratio of accurate responses, respectively, on incongruent versus congruent trials. In order to allow interpretation of Stroop results as a *capacity*, we reverse scored the interference effects for RT and accuracy, such that higher scores represent stronger inhibition of attention to distracting stimuli, and fewer inaccurate responses (i.e., resistance to interference). Descriptive analysis of Stroop resistance to interference scores based on RT revealed a small but counterintuitive relation with age ($r = -.083, p < .05$), suggesting a weak age trend toward *less effective* inhibition. Because inaccurate response trials do not contribute to the index of RT interference, we suspect that this negative correlation with age is an artifact reflecting the greater frequency of inaccurate responses in younger age groups. We therefore opted to utilize Stroop resistance to interference scores based on accuracy as our index of prepotent RI; this measure showed a modest but positive correlation with age ($r = .167, p < .001$).

Working memory. The test battery included two measures of WM. The first, which we refer to as Working Memory Capacity, was based on a standard test of Forward Digit Span. Participants heard a series of 13 sequences of digits (beginning with two digits and increasing to eight) that they were asked to recall. A Working Memory Capacity score was computed based on the highest number of digits correctly recalled within the 13 trials.

In addition, we derived a measure of Working Memory Updating using an item recognition memory task (Thompson-Schill et al., 2002). On each trial of this task, participants saw four probe letters on the screen, followed by a brief delay. They were then presented a single target letter and asked whether

the target was among the four probes. In half of the trials, the probe item was a member of the target data set (i.e., "positive" trials); in the other half of the trials, the probe item was not a member of the target data set (i.e., "negative" trials). To respond accurately, participants pressed a key corresponding to *yes* for positive and *no* for negative trials.

Each subject completed four blocks of experimental trials and one block of control trials; each block included 40 trials. For experimental trials, the trial sequence was manipulated to vary the degree to which items from previous trials would interfere with accurate recognition of target items on current trials. "Recent" trials used probe letters that appeared in the previous target set (not the one against which participants are currently comparing). Thus, recent trials introduce interference to the task; if participants fail to effectively update the WM buffer by clearing items from previous trials and adding items from the current trial, they might inaccurately identify the probe as a member of the target set of letters. "Nonrecent" trials used probe letters that did not appear in either of the previous two target sets, and thus are not as subject to interference effects. This resulted in an equal number of four different types of trials (recent-positive, recent-negative, nonrecent-positive, nonrecent-negative), pseudorandomly distributed throughout the experimental blocks. Following the analytical strategy outlined by Thompson-Schill et al. (2002), we calculated an interference effect as the difference in response accuracy for recent-negative versus nonrecent-negative trials. In order to reduce overlap in terminology with the Stroop interference effect, we refer to the WM interference effect as *working memory updating*; to the degree that subjects effectively update the contents of WM on a trial-by-trial basis, interference created by recent target items is diminished. Analogous to our transformation of Stroop results, we reverse scored the WM updating data, such that higher scores represent greater competence.

Tower of London. A computerized version of the classic TOL task was administered to assess planning and problem solving (Berg & Byrd, 2002; Shallice, 1982). Similar to the physical instrument used in the classic TOL, the computerized TOL instrument consists of three rods, the first of which can hold three balls, the second only two balls, and the last, one ball. On each trial, the subject is shown (a) a goal instrument and (b) a game instrument, each of which displays three differently colored balls distributed across the three rods. The subject is asked to move the balls on the game instrument

from their starting arrangement to match the goal in as few moves as necessary, using the computer cursor to "drag" and "drop" each ball. Five sets of four problems are presented, beginning with those that can be solved in three moves and progressing to those that require a minimum of seven moves.

In the administration of the task, the starting and goal positions are displayed, and the subject takes as much (or as little) time as necessary before making each move. The subject is instructed to click a button indicating completion of the trial when the solution picture matches the goal picture. The trial is considered successfully solved if the solution is correctly submitted within a time limit of 160 s. If the submitted solution does not match the goal presentation, or if the participant does not submit a solution within 160 s, the trial is considered unsolved. After each trial, feedback is presented indicating whether the trial was solved in the minimum number of moves, solved with extra moves, or incorrectly solved.

For each level of problem difficulty, our primary outcome variable is the *percent of trials with perfect solutions* (i.e., trials solved in the minimum number of moves), a measure of optimal planning and execution of the task. We also computed a measure of relative performance on problems that were not perfectly solved (i.e., accurate solutions achieved with a varying number of extra moves), but because the resulting index was highly correlated with average perfect solutions ($r = .89$), and showed a nearly identical pattern of age differences, we henceforth limit our analyses to the percent of perfect solutions. This decision reflects our greater interest in the development of strategic planning, as opposed to "on-line" trial-and-error problem solving.

In addition, we utilized a measure of *first-move latency*, calculated as the length of time between the problem presentation and the participant's first move. As described in the introduction, first-move latency significantly predicts overall performance on the TOL and is commonly interpreted as a measure of the extent to which an individual plans before acting. Because first-move latency data from the present study (but no performance data) were presented in detail in a prior report (Steinberg et al., 2008), we refer the reader to that paper for a full discussion of age differences and their implications for understanding the maturation of impulse control. In the present report, first-move latency is utilized as an index of impulse control, which we examine alongside measures of VS reasoning, WM, and prepotent RI as potential mediators of age differences in problem-solving performance.

Results

Raw correlations between all study variables are presented in Table 1. Means and standard deviations for TOL outcomes, across the age groups and for the sample as a whole, are presented in Table 2. Descriptive statistics for VS reasoning, WM capacity, WM updating, prepotent RI, and self-reported impulsivity are presented in Table 3.

Age Differences in Performance on the TOL

Age differences in perfect solutions were examined using a repeated measures analysis of covariance (ANCOVA). Performance at each level of difficulty (from three to seven minimum move trials) was specified as a five-level within-subject factor. Age group, sex, and ethnicity were entered as independent variables, and household education and IQ were entered as covariates. An alpha level of .01 was applied for the repeated measures ANCOVA, as well as follow-up univariate ANCOVAs at each level of problem difficulty when the omnibus ANCOVA was significant. Significant univariate effects were further examined using post hoc comparisons of pairwise differences between age groups, using a Bonferroni adjustment for a groupwise alpha of .05 (i.e., the cumulative alpha of all post hoc pairwise comparisons between age groups for a given univariate test was equal to .05).

As predicted, analyses revealed a significant main effect for age, with older subjects achieving proportionately more perfect solutions than younger subjects, $F(6, 802) = 19.49, p < .001$. In addition, a significant within-subjects effect of problem difficulty confirmed that, across the age groups, fewer

trials were solved perfectly as problems became more difficult, $F(4, 799) = 10.44, p < .001$. Importantly, we also found a significant interaction between age and problem difficulty, such that age differences in performance varied at different levels of problem difficulty, $F(24, 3208) = 2.96, p < .001$ (Figure 1). To further examine this interaction, we conducted univariate ANCOVAs at each level of problem difficulty, again controlling for IQ and household education. Although the main effect for age was significant at each level of problem difficulty, age accounted for more variance in performance on the hardest problems, reflected in the larger effect size for seven-move problems relative to all other difficulty levels (three-move $\eta^2 = .05$, four-move $\eta^2 = .03$, five-move $\eta^2 = .02$, six-move $\eta^2 = .02$, seven-move $\eta^2 = .11$).

Post hoc age comparisons revealed that the age of mature performance (assessed as the youngest age group that was not significantly different from the highest performing age group) also varied according to problem difficulty. Whereas adult performance was demonstrated by 16- to 17-year-olds on three-move problems, the youngest to perform at a mature level on the seven-move problems was the 22- to 25-year-old group. A polynomial contrast analysis of age differences (adjusting for unequal age intervals) in proportion of perfect solutions averaged across all levels of difficulty showed strong linear improvement by age, $F(1, 883) = 116.5, p < .001$. In addition, a significant quadratic effect, $F(1, 883) = 15.78, p < .001$, reflected the leveling off of performance in the late teens (i.e., no age gains in performance were seen beyond ages 16–17 on perfect solutions averaged across all levels of difficulty). These linear and quadratic trends were significant at each level of problem difficulty, with the

Table 1
Raw Correlations Among Study Variables

	1	2	3	4	5	6	7	8	9	10
Age		.34***	.05	.10**	.05	.24***	.17***	.17***	.35**	-.20***
TOL perfect solutions			.32***	.10**	.26***	.21***	.13***	.12***	.47***	-.21***
Full-scale IQ				.29***	.69***	.21***	.13***	.10**	.23***	-.19***
Household education					.15***	.12***	.10**	.06	.13***	-.05
Visuospatial reasoning						.15***	.11**	.07*	.19***	-.15***
Working memory capacity							.17***	.12***	.13***	-.13***
Working memory updating								.07*	.06	-.05
Prepotent response inhibition									.06	-.08*
Impulse control (TOL latency)										-.13***
Self-reported impulsivity										

Note. Asterisks represent significance level of bivariate Pearson's correlation coefficients. TOL = Tower of London.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 2
Means and Standard Errors of Tower of London Outcomes, Averaged Across All Trials

Age group	First-move latency		% perfect solutions	
	Mean	SE	Mean	SE
10–11	5.23***	0.44	38.9***	1.4
12–13	5.67***	0.39	43.1***	1.3
14–15	5.82***	0.53	45.3***	1.7
16–17	6.73***	0.36	51.6	1.2
18–21	7.35***	0.36	50.6	1.2
22–25	8.39*	0.35	52.8	1.1
26–30	9.98	0.39	55.1	1.3
Total	7.03	0.15	48.2	0.5

Note. All values adjusted for IQ and household education. *Ns* are as follows: 10–11 years (*N* = 109), 12–13 years (*N* = 130), 14–15 years (*N* = 122), 16–17 years (*N* = 140), 18–21 years (*N* = 141), 22–25 years (*N* = 133), and 26–30 years (*N* = 115). Asterisks indicate a significant difference between the mean for a given age group and the mean for the highest performing age group. All significance tests were Bonferroni corrected for groupwise $\alpha = .05$. Means and standard deviations at each level of problem difficulty are available in table form on request from the author. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Means and Standard Errors of Variables Examined for Mediation

Age group	VSR	WMC	WMU	PRI	S-R I
	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
10–11	25.34 (0.54)	6.58*** (0.10)	-0.69** (0.09)	-0.097*** (0.010)	2.29*** (0.03)
12–13	24.53 (0.45)	6.69*** (0.09)	-0.86*** (0.09)	-0.072 (0.008)	2.26** (0.02)
14–15	25.09 (0.38)	7.02 (0.09)	-0.56 (0.07)	-0.075* (0.009)	2.27** (0.02)
16–17	25.03 (0.45)	7.26 (0.08)	-0.49 (0.05)	-0.060 (0.007)	2.19 (0.02)
18–21	25.73 (0.38)	7.21 (0.08)	-0.44 (0.06)	-0.067 (0.010)	2.19 (0.02)
22–25	25.55 (0.44)	7.29 (0.08)	-0.46 (0.05)	-0.049 (0.007)	2.16 (0.03)
26–30	25.55 (0.50)	7.37 (0.08)	-0.32 (0.07)	-0.040 (0.006)	2.12 (0.02)
Total	25.26 (0.17)	7.07 (0.03)	-0.54 (0.03)	-0.065 (0.003)	2.21 (0.01)

Note. Asterisks indicate a significant difference between the mean for a given age group and the mean for the highest performing age group. All significance tests were Bonferroni corrected for groupwise $\alpha = .05$. VSR = visuospatial reasoning; WMC = working memory capacity; WMU = working memory updating; PRI = prepotent response inhibition; S-R I = self-reported impulsivity. * $p < .05$. ** $p < .01$. *** $p < .001$.

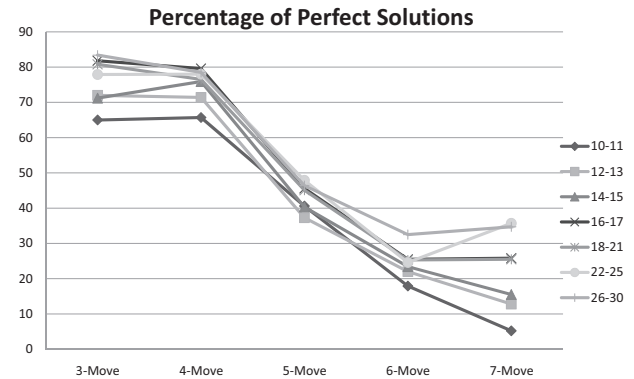


Figure 1. Age differences in percentage of perfect solutions by problem difficulty.

exception of six-move problems, which showed a linear but not a quadratic effect. No higher order polynomial contrasts were significant, nor were significant ANCOVA main effects or interactions found for gender, ethnicity, or household education, although a main effect was found for IQ, $F(1, 802) = 100.91, p < .001$. As expected, a follow-up regression showed that IQ significantly predicted percentage of perfect solutions, averaged across difficulty levels ($\beta = .31, t = 9.33, p < .001, \Delta R^2 = .08$). Together, these data suggest that optimal problem solving, which presumably requires strategic planning to attain a solution without any error, continues to mature through middle adolescence and only reaches adult maturity on the most challenging problems between ages 22 and 25.

Mediation of Age Differences in Problem Solving by WM Capacity, WM Updating, VS Reasoning, Prepotent RI, and Impulse Control

To assess the degree to which developmental gains in WM capacity, WM updating, VS reasoning, prepotent RI, and impulse control account for age-related variance in planned problem solving, we conducted a series of hierarchical regressions on the average percentage of perfect solutions on the TOL, across levels of problem difficulty. In all analyses, we specified ethnicity, gender, and household education as covariates in the first step of the hierarchical regression. We did not include full-scale IQ as a covariate but instead utilized raw scores from the Matrix Reasoning subtest of the WASI as an index of VS reasoning, which we examine as a potential mediator. Our rationale for this decision is twofold: (a) Because IQ scores are standardized to age norms, they do not represent age-related variance in intelligence, and thus are not well suited

for a test of whether intellectual development mediates age gains in TOL performance; and (b) in contrast to the full-scale IQ score (which is jointly determined by Vocabulary and Matrix Reasoning scores on the two-subtest WASI), the Matrix Reasoning subtest provides a specific assessment of VS reasoning capacity, a plausible predictor of performance on the TOL, a visual problem-solving task.

Mediation analyses proceeded as follows. After confirming that age (entered as a continuous variable in Step 2 of the hierarchical regression) accounted for unique variance in TOL performance ($\beta = .32$, $t = 9.95$, $p < .001$, $\Delta R^2 = .10$, total $R^2 = .125$), we examined whether the simultaneous introduction of all potential mediators eliminated or reduced this age effect. As expected, the full model significantly predicted TOL performance, $F(12, 830) = 30.9$, $p < .001$, $R^2 = .31$, and the introduction of the mediator variables significantly improved model fit, F -change(5, 830) = 41.3, $p < .001$, $\Delta R^2 = .17$. Furthermore, although age remained a significant predictor in the full model ($\beta = .17$, $t = 5.21$, $p < .001$), indicating that age differences in TOL performance were not *fully* mediated by developmental gains in any of the examined mediators, the amount of unique variance predicted by age dropped from 10% (i.e., $\Delta R^2 = .10$) to 2.3% ($\Delta R^2 = .023$). These results suggest that one or more of the hypothesized mediator variables indeed contributed to the age differences observed on the TOL.

To formally examine which specific mediator variables were responsible for this effect, we conducted a series of regressions following the procedure outlined by Baron and Kenny (1986). First, we tested whether age significantly predicted the potential mediator variable, controlling for gender,

ethnicity, household education, and all other potential mediator variables. We then tested the degree to which the mediator predicted perfect solutions, using the same covariates, and whether its introduction into the model reduced the predictive effect of age. If these first two criteria for mediation were satisfied, we then conducted a Sobel test to determine the degree of significance of the mediation effect (Sobel, 1982). For ease of interpretation, all regression coefficients are reported as standardized betas; Sobel tests were conducted using unstandardized beta coefficients and corresponding error terms. The regression and Sobel test results for each hypothesized mediator are presented in Table 4.

Working memory capacity. Consistent with its potential role as a mediator, we found that WM capacity (i.e., forward digit span) was significantly predicted by our full covariate model, $F(11, 831) = 9.43$, $p < .001$, and was specifically predicted by age ($\beta = .18$, $t = 4.96$, $p < .001$, $\Delta R^2 = .026$). Furthermore, WM capacity accounted for unique variance in TOL perfect solutions ($\beta = .09$, $t = 2.89$, $p < .005$, $\Delta R^2 = .007$), and a Sobel test revealed that WM capacity partially mediated the age effect ($z' = 2.4$, $p = .02$). Thus, results suggest that WM capacity partially—but not fully—mediates age differences in problem-solving performance on the TOL.

Working memory updating. WM updating was also predicted by the full covariate model, $F(11, 831) = 5.53$, $p < .001$, with unique prediction by age ($\beta = .14$, $t = 3.8$, $p < .001$, $\Delta R^2 = .016$). However, WM updating did not predict unique variance in TOL performance ($\beta = .05$, $t = 1.72$, *ns*), and there was no evidence for mediation by WM updating of the relation between age and TOL performance (Sobel $z' = 1.54$, $p = .12$).

Table 4

Summary of Regressions Testing Mediation of Age Differences in TOL Perfect Solutions Averaged Across All Difficulty Levels

Mediator	Age prediction of mediator			Mediator prediction of TOL solutions			Sobel
	Std. β	Unstd. β (SE)	ΔR^2	Std. β	Unstd. β (SE)	ΔR^2	z test
VSR	-.067	-.060 (0.033)	.004	.135***	.015 (0.003)	.016	-1.71
WMC	.179***	.032 (0.007)	.026	.089**	.048 (0.017)	.007	2.40*
WMU	.141***	.020 (0.005)	.016	.051	.035 (0.021)	.002	1.54
PRI	.146***	.002 (0.001)	.017	.037	.214 (0.172)	.001	1.06
Impulse control	.348***	.253 (0.024)	.110	.378***	.051 (0.004)	.118	8.05***
S-R I	-.149***	-.007 (0.002)	.020	-.109**	-.217 (0.065)	.011	2.59**

Note. Significance levels are indicated by corresponding asterisks. Regressions controlled for gender, ethnicity, household education, and competing mediators (except S-R I, which was examined as a "substitute" for impulse control). VSR = visuospatial reasoning; WMC = working memory capacity; WMU = working memory updating; PRI = prepotent response inhibition; S-R I = self-reported impulsivity; TOL = Tower of London.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Visuospatial reasoning. Utilizing raw scores on the Matrix Reasoning subtest of the WASI, we also examined whether developmental gains in VS reasoning capacity mediated age differences in TOL performance. Although VS reasoning was significantly predicted by the full covariate model, $F(11, 831) = 10.39, p < .001$, the unique age prediction was only marginally significant, and in the direction opposite to what was expected ($\beta = -.07, t = -1.84, p = .07, \Delta R^2 = .004$). Given the lack of a significant bivariate correlation between age and VS reasoning ($r = .05, p = .14$), this trend-level effect is likely an artifact of residual variance from competing covariates representing cognitive processing. VS reasoning did predict unique variance on the TOL ($\beta = .14, t = 4.38, p < .001, \Delta R^2 = .016$), confirming its role as a contributor to problem solving, although not a mediator of age differences in TOL performance.

Prepotent RI. Prepotent RI—operationalized as resistance to the interference effect on Stroop accuracy—was significantly predicted by the full covariate model, $F(11, 831) = 3.48, p < .001$, and was specifically predicted by age ($\beta = .15, t = 3.87, p < .001, \Delta R^2 = .017$). However, controlling for age, demographic variables, and all other potential mediators, prepotent RI did not predict unique variance on the TOL ($\beta = .04, t = 1.21, p = .23, \Delta R^2 = .001$), and the Sobel test of mediation was nonsignificant ($z' = 1.03, ns$). To insure that the absence of a mediation effect did not result from our choice to examine the Stroop interference effect in terms of accuracy (rather than RT), we reconducted the analyses with RT interference as the hypothesized mediator. Consistent with findings for accuracy, the RT interference effect did not predict unique variance (or mediate the age effect) in TOL performance. Furthermore, these results for accuracy and RT interference held whether we examined Stroop outcomes separately or combined across trial blocks (equal vs. unequal proportions of congruent vs. incongruent trials).

Impulse control. Finally, we examined whether the tendency to take more time to consider solutions before acting would partially mediate the relation between age and perfect solutions on the TOL. As predicted, average first-move latency was significantly predicted by the full covariate model, $F(11, 829) = 16.72, p < .001$, with unique prediction by age ($\beta = .34, t = 10.33, p < .001, \Delta R^2 = .104$). Furthermore, first-move latency significantly predicted unique variance in TOL performance ($\beta = .38, t = 11.9, p < .001, \Delta R^2 = .118$) above and beyond all other hypothesized mediators. Finally, a Sobel test

supported our prediction that first-move latency partially mediates the relation between age and TOL performance ($z' = 8.05, p < .001$).

Given the possibility that the strong relation between first-move latency and perfect solutions is due to shared method variance between the two measures, we ran parallel analyses using scores on a self-report measure of impulsivity as the mediation variable, in place of first-move latency. (Recall that first-move latency is significantly negatively correlated with self-reported impulsivity.) Significant results for self-reported impulsivity would further confirm the role of impulse control as a mediator, unconfounded by common method variance. Consistent with findings for first-move latency, self-reported impulsivity was significantly predicted by the full covariate model, $F(11, 831) = 7.54, p < .001$, including a unique effect for age ($\beta = -.15, t = -4.43, p < .001, \Delta R^2 = .02$). In addition, self-reported impulsivity significantly predicted unique variance in TOL performance ($\beta = -.11, t = -3.41, p = .001, \Delta R^2 = .011$), and a Sobel test confirmed that impulsivity partially mediated the relation between age and problem solving on the TOL ($z' = 2.59, p < .01$). In sum, although self-reported impulsivity did not explain as much variance in TOL performance as first-move latency, the parallel findings for these two measures suggest that developmental gains in impulse control contribute to age-related maturation of planned problem solving.

Differential Predictors of Early and Late Gains in TOL Performance

The analyses presented thus far have broadly examined the hypothesis that TOL problem solving gradually matures over the course of adolescence, while investigating the contributions of developmental gains in core cognitive competencies to this general improvement in problem-solving capacity. However, it is also plausible that the specific cognitive advances supporting age-related gains in problem solving differ at distinct stages of development and at varying levels of cognitive challenge. Based on our finding that performance plateaus were attained at a much earlier age (16–17) for relatively easy three-move problems than for the hardest seven-move problems (ages 22–25), we hypothesized that the ability to solve easier problems (i.e., those requiring less extensive planning and manipulation of items in WM) may depend on relatively early maturation of VS reasoning, whereas perfectly solving the most difficult problems may

require more advanced WM and impulse control, competencies that continue to mature through adolescence.

To test these hypotheses, we repeated the mediation analyses described above, with two adjustments. First, we ran separate analyses examining age gains in performance on three- versus seven-move problems, allowing us to test the prediction that age mediation effects seen for higher order cognitive competencies are specific to the most challenging problems. In addition, we examined the mediation effects separately for two subsamples: a child and adolescent group (ages 10–17; $n = 501$), and an adult group (18–30; $n = 389$). This framework allowed us to examine whether age gains in performance were dependent upon the same or different cognitive advances during each broadly defined stage of development. Our rationale for splitting the sample at age 17 followed from the finding that most gains in TOL performance reached asymptote by this age, whereas performance on the most difficult problems extended into early adulthood. Thus, for the older group, we expected age to predict significant variance in performance only on the seven-move problems. By splitting the sample, we were able to isolate the cognitive gains that contributed to this late development in performance. Consistent with prior analyses, regressions controlled for all demographic variables (i.e., gender, ethnicity, household education). After first examining whether age significantly predicted performance, we then added the cognitive competency variables to examine poten-

tial mediation effects. We here report mediation findings only for the variables that significantly predicted TOL performance. Comprehensive results organized by problem difficulty and age range are presented in Table 5.

Results from the regression predicting perfect solutions on three-move problems confirmed the pattern of age differences reported in the ANCOVA analyses, such that age significantly predicted performance for the child-adolescent group ($\beta = .19$, $t = 4.09$, $p < .001$, $\Delta R^2 = .033$), but not for adults ($\beta = -.033$, $t = -.68$, ns). Within the younger group, the only significant predictor beyond age was VS reasoning ($\beta = .15$, $t = 3.24$, $p < .001$, $\Delta R^2 = .02$). However, VS reasoning was not significantly predicted by age in the younger group ($\beta = -.04$, $t = -.78$, ns), and therefore did not satisfy the criteria for mediation of the age effect. Neither WM, prepotent RI, nor impulse control were significant predictors of performance on 3-move problems in the younger group. Within the older group, significant predictors included VS reasoning ($\beta = .13$, $t = 2.37$, $p < .05$, $\Delta R^2 = .011$) and impulse control ($\beta = .21$, $t = 3.87$, $p < .001$, $\Delta R^2 = .036$). Because age was not predictive in this subsample, we did not conduct mediation analyses.

In contrast to the three-move problem regressions, results from analyses predicting perfect solutions on seven-move problems demonstrated significant age effects for both the child-adolescent ($\beta = .29$, $t = 6.72$, $p < .001$, $\Delta R^2 = .085$) and adult ($\beta = .11$, $t = 2.22$, $p < .05$, $\Delta R^2 = .012$) subsamples. Within the younger group, the addition of the

Table 5
Differential Prediction of TOL Perfect Solutions by Age and Problem Difficulty

Variable	3-Move problems		7-Move problems	
	Ages 10–17	Ages 18–30	Ages 10–17	Ages 18–30
	Std. β	Std. β	Std. β	Std. β
Age (premediators)	.219***	.004	.293***	.128*
Age (postmediators)	.194***	-.033	.196***	.060
VSR	.153**	.129*	.044	.029
WMC	.031	.005	.100**,+	.035
WMU	-.048	.015	.057	-.016
PRI	.029	.014	.016	.010
Impulse control	.081	.210***	.324***,+	.395***,+

Note. All regressions controlled for gender, ethnicity, household education, and all other hypothesized mediating variables. Asterisks indicate significant beta coefficients. Plus signs indicate that the variable significantly mediated a portion of the age effect, as confirmed by a Sobel test. VSR = visuospatial reasoning; WMC = working memory capacity; WMU = working memory updating; PRI = prepotent response inhibition; TOL = Tower of London.

* $p < .05$. ** $p < .01$. *** $p < .001$.

+ $p < .01$.

cognitive competency variables to the model reduced the unique variance explained by age (from 8.5% to 3.3%), but did not eliminate its significance ($\beta = .196$, $t = 4.44$, $p < .001$, $\Delta R^2 = .033$). Beyond age, the younger group's performance on seven-move problems was significantly predicted by WM capacity ($\beta = .1$, $t = 2.27$, $p < .05$, $\Delta R^2 = .009$) and impulse control ($\beta = .32$, $t = 7.51$, $p < .001$, $\Delta R^2 = .095$). Consistent with partial mediation, age significantly predicted WM capacity ($\beta = .23$, $t = 5.06$, $p < .001$, $\Delta R^2 = .048$), and a Sobel test was significant ($z' = 2.11$, $p < .05$). Similarly, impulse control was significantly predicted by age ($\beta = .16$, $t = 3.48$, $p = .001$, $\Delta R^2 = .024$), and the mediation effect was supported by a significant Sobel test ($z' = 3.07$, $p < .01$).

Within the adult group, the introduction of the mediator variables to the regression completely eliminated the significance of the age effect ($\beta = .06$, $t = 1.23$, $p = .22$, $\Delta R^2 = .003$), and only impulse control significantly predicted solutions on the hardest problems ($\beta = .40$, $t = 7.97$, $p < .001$, $\Delta R^2 = .138$). Furthermore, age significantly predicted impulse control ($\beta = .18$, $t = 3.44$, $p < .001$, $\Delta R^2 = .029$), and a Sobel test ($z' = 3.17$, $p < .01$) confirmed that impulse control significantly mediated the age gains in performance seen after age 17 on difficult TOL problems.

In sum, on the easiest TOL problems—those presumably requiring little planning of successive subgoal operations to attain the final goal state—age gains were only seen prior to age 18, and these age differences were not explained by any of the cognitive competence variables measured in this study. In contrast, on the harder problems, age gains were seen in both the child-adolescent and adult subsamples (although the age effect was considerably smaller among adults). Age gains in performance within the child and adolescent period were partially mediated by both WM capacity and impulse control, consistent with results reported for regressions predicting average perfect solutions in the full sample. In contrast, age gains in performance on seven-move problems observed in the adult subsample were unrelated to WM capacity, and were instead fully mediated by protracted maturation of impulse control.

Discussion

Consistent with neurobiological evidence showing gradual and prolonged maturation of brain systems implicated in cognitive control, we found that per-

formance on a standard measure of strategic planning and problem solving—the TOL—continues to improve well into late adolescence and early adulthood. Although developmental gains were evinced across our sample of 10- to 30-year-olds for problem solving at all difficulty levels, the greatest age differences in performance were seen on the hardest problems, which require planning and execution of multiple intermediate subgoals to reach a correct solution. Using the most stringent criterion—the ability to plan and enact perfect solutions on seven-move problems—adult performance was not attained until between ages 22 and 25, suggesting that the ability to plan a perfect solution to a difficult problem continues to develop into early adulthood.

Many researchers have questioned whether developmental improvements on the TOL represent advances in strategic planning and control of behavior, or whether age differences in performance can be accounted for by maturation of basic cognitive processing abilities like VS reasoning, WM, and prepotent RI. Although our findings confirmed the importance of WM for TOL performance, they also demonstrated an important and unique role for impulse control in support of optimal problem solving. Specifically, whereas individual differences in WM capacity (but not WM updating, VS reasoning, or prepotent RI) partially mediated age-related gains in the ability to perfectly solve problems, an additional, unique mediation effect was found for average first-move latency, the amount of time the individual inhibited responding (and presumably engaged in planning) before initiating a solution. Furthermore, results demonstrating a parallel mediation effect for self-reported impulsivity suggest that this effect is not a mere artifact of common task variance. In sum, we find age-related gains in WM capacity and impulse control in a large and diverse sample spanning two decades of development, and these gains predict stronger performance on the TOL.

Our results also suggest that not all TOL problems are created equally. The ability to execute perfect solutions on the easiest TOL problems—those that presumably require little planning and execution of intermediate subgoal operations—does not show improvement beyond age 17. Furthermore, the gains in simple problem solving that are apparent before this age are not mediated by parallel gains in any of the cognitive capacities we measured. In contrast, age gains in perfectly solving the hardest TOL problems extend well into the 20s. Across the child and adolescent period, advances in

WM capacity and impulse control both partially mediate these age gains in performance, whereas only impulse control mediates age gains among adults on the hardest problems.

At first glance, it may appear puzzling that impulse control mediates age gains in problem solving but prepotent RI does not, given that both constructs reflect aspects of inhibitory processing. Upon closer examination, the contrasting findings suggest an interesting distinction between cognitive control mechanisms at two different levels of processing. We operationalized prepotent RI in terms of the Stroop interference effect on accuracy—that is, the degree to which individuals were able to resist interference from salient but irrelevant stimuli and maintain control over goal-directed behavior. Inaccurate responses represent a failure to suppress automatic behavioral reactions to irrelevant stimuli, and thus poor cognitive control over prepotent behavior. In light of the TOL task structure, which does not “train” a strong prepotent response or directly challenge the participant with distracting stimuli, it is not surprising that Stroop did not predict TOL performance. In contrast, the measure of impulse control used in this study—the amount of time an individual waited before attempting to solve a problem—was a robust predictor of TOL performance, a finding that has been consistently reported in the TOL literature. Again, this relation is not very surprising, given that perfectly solving a multistep problem is likely to benefit from increased planning. In our view, impulse control represents not only the successful inhibition of immediate responding, but a further *sustained* delay of responding in support of effortful planning of future behavior. This conceptual distinction is consistent with current formulations suggesting that impulse control is part of a higher order “executive inhibition” factor representing mature self-regulation (Nigg, 2000). Although early childhood growth in analogous capacities like “effortful control” has been well studied (e.g., Rothbart, Ahadi, Hershey, & Fisher, 2001), research on the continued maturation of self-regulatory capacities in adolescence and adulthood remains sparse, and constitutes an important goal for future research.

The finding that TOL performance was predicted by a basic measure of WM storage capacity but not WM updating is more surprising. Given that WM updating requires greater *control* of information in WM than simple storage capacity (i.e., holding information across a delay), we expected that maturation of WM updating would support age gains in TOL performance. However, it is possible that plan-

ning solutions on TOL problems does not require this specific level of control over WM. Whereas our WM updating task included a rapid series of trials requiring clearing the WM buffer of old information and introducing new information, TOL trials were self-paced (within a 160-s limit) and lacked an explicit “interference” challenge. Furthermore, our study did not explicitly examine VS WM, which may be particularly important for solving inherently visual problems on the TOL (e.g., Gilhooly et al., 2002). In sum, our findings demonstrate that WM, at a minimum, plays a support role in planning solutions on the TOL, consistent with significant mediation by WM capacity of age gains in problem solving. We suspect that some form of “executive” WM also contributes to TOL performance; future research should utilize measures of WM requiring concurrent storage and manipulation of VS information to identify this active WM component.

Evidence that performance gains in planned problem solving extend well into late adolescence and early adulthood is consistent with previous studies of age differences on the TOL (Anderson et al., 2001; Asato et al., 2006; De Luca et al., 2003; Huizenga et al., 2006; Luciana & Nelson, 2002). However, to our knowledge, the present study is the first to utilize a large, diverse, and broad enough sample to confidently identify an approximate age of maturation. That strategic planning continues to improve until ages 22–25 adds to a growing body of evidence that cognitive control processes are only gradually consolidated over the course of adolescence, coincident with ongoing structural and functional maturation of the PFC (Casey et al., 2008). In particular, adolescent improvements in cognitive control are evidenced by performance gains on tasks known to activate the dorsolateral PFC, including relatively difficult tests of RI (Luna & Sweeney, 2004), spatial WM (Conklin, Luciana, Hooper, & Yarger, 2007), flexible rule use (Crone, Donohue, Honomichl, Wendelken, & Bunge, 2006), and strategic self-organization (Luciana et al., 2005). Given the strong evidence for the role of the DLPFC in optimal problem solving on the TOL (e.g., van den Heuvel et al., 2003), our finding of TOL performance gains into the early to mid-20s provides further indirect support for the link between DLPFC maturation and improvements in cognitive control.

Limitations

Several limitations to the current study must be acknowledged when drawing conclusions about the

developmental course of strategic planning. Although the TOL is widely used to study planning, the task cannot be considered a “pure” planning measure, given the diverse array of cognitive abilities likely required for optimal problem solving. However, our finding that first-move latency (widely considered an indicator of planning time) strongly predicted performance and partially mediated the impact of age on performance suggests that strategic planning plays a key role in solving TOL problems. Because first-move latency in this study is derived from the TOL task itself, it is possible that its strong correlation with performance relies to some degree on shared method variance. Although we addressed this concern by replicating findings for first-move latency with a self-report measure of impulsivity, future research should incorporate an independent behavioral measure of impulse control to further confirm its importance to TOL problem solving. Finally, our suggestions concerning the role of structural and functional brain maturation in the development of TOL performance are necessarily speculative. Although research examining TOL performance in children and adults with PFC lesions has generally demonstrated deficits in planned problem solving (Unterrainer & Owen, 2006), several studies suggest that the relation between frontal functioning and problem solving varies depending on the specific location of the lesion and the aspect of problem solving under investigation (Morris, Kotitsa, & Bramham, 2005).

Conclusions

Despite these limitations, the research reported here provides important new evidence that strategic planning and problem solving undergo continued refinement well into late adolescence and, in some respects, early adulthood. Adolescents may evince adult-like competence in basic cognitive capabilities by the time they are 15 or 16 (Steinberg et al., 2009), but the ability to marshal these abilities in the implementation of a plan may not fully mature until some years later. Although steady gains in core cognitive processes clearly contribute to this achievement, the present study suggests that a higher order process is also at work—namely, the emergence of mature self-regulation. Indeed, late adolescence increasingly looks like a time for consolidation of gains in a variety of self-regulatory domains, including impulse control (Steinberg et al., 2008), future orientation (Steinberg et al., 2009), reward and punishment learning (Cauuffman et al., 2010), emotion regulation (Dahl, 2001), and

resistance to peer influence (Steinberg & Monahan, 2007). Perhaps it is this consolidation of self-regulatory competence that best distinguishes the passage from adolescence to adulthood.

References

- Albert, D., & Steinberg, L. (2011). Adolescent judgment and decision making. *Journal of Research on Adolescence, 21*, 211–224.
- Anderson, V., Anderson, P., Northam, E., Jacobs, R., & Catroppa, C. (2001). Development of executive functions through late childhood and adolescence in an Australian sample. *Developmental Neuropsychology, 20*, 385–406.
- Asato, M. R., Sweeney, J. A., & Luna, B. (2006). Cognitive processes in the development of TOL performance. *Neuropsychologia, 44*, 2259–2269.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.
- Berg, K. W., & Byrd, D. L. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology, 24*, 586–604.
- Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science, 11*, 282–298.
- Casey, B. J., Getz, S., & Galvan, A. (2008). The adolescent brain. *Developmental Review, 28*, 62–77.
- Cauuffman, E., Shulman, E., Steinberg, L., Claus, E., Banich, M., Graham, S., et al. (2010). Age differences in affective decision making as indexed by performance on the Iowa Gambling Task. *Developmental Psychology, 46*, 193–207.
- Conklin, H., Luciana, M., Hooper, C., & Yarger, R. (2007). Working memory performance in typically developing children and adolescents: Behavioral evidence of protracted frontal lobe development. *Developmental Neuropsychology, 31*, 103–128.
- Crone, E., Donohue, S., Honomichl, R., Wendelken, C., & Bunge, S. (2006). Brain regions mediating flexible rule use during development. *Journal of Neuroscience, 26*, 11239–11247.
- Dahl, R. (2001). Affect regulation, brain development, and behavioral/emotional health in adolescence. *CNS Spectrums, 6*, 1–12.
- De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J., Proffitt, T. M., Mahony, K., et al. (2003). Normative data from the CANTAB. I: Development of executive function over the lifespan. *Journal of Clinical and Experimental Neuropsychology, 25*, 242–254.
- Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General, 133*, 101–135.

- Giedd, J. N. (2008). The teen brain: Insights from neuroimaging. *Journal of Adolescent Health, 42*, 335–343.
- Gilhooly, K. J., Wynn, V., Phillips, L. H., Logie, R. H., & Della Sala, S. (2002). Visuo-spatial and verbal working memory in the five-disc Tower of London task: An individual differences approach. *Thinking and Reasoning, 8*, 165–178.
- Huizenga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia, 44*, 2017–2036.
- Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia, 28*, 517–527.
- Korkman, M., Kemp, S. L., & Kirk, U. (2001). Effects of age on neurocognitive measures of children ages 5 to 12: A cross-sectional study on 800 children from the United States. *Developmental Neuropsychology, 20*, 331–354.
- Luciana, M., Conklin, H. M., Hooper, C. J., & Yarger, R. S. (2005). The development of nonverbal working memory and executive control processes in adolescents. *Child Development, 76*, 697–712.
- Luciana, M., & Nelson, C. A. (2002). Assessment of neuropsychological function through use of the Cambridge Neuropsychological Testing Automated Battery: Performance in 4- to 12-year-old children. *Developmental Neuropsychology, 22*, 595–624.
- Luna, B., Garver, K., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development, 75*, 1357–1372.
- Luna, B., & Sweeney, J. A. (2004). The emergence of collaborative brain function: fMRI studies of the development of response inhibition. *Annals of the New York Academy of Science, 1021*, 296–309.
- Mitchell, C. L., & Poston, C. L. (2001). Effects of inhibiting of response on Tower of London performance. *Current Psychology, 20*, 164–168.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100.
- Morris, R. G., Kotitsa, M., & Bramham, J. (2007). Planning in patients with focal brain damage: From simple to complex task performance. In R. G. Morris & G. Ward (Eds.), *The psychology of planning: Cognitive and neuropsychological perspectives* (pp. 153–180). Abingdon, UK: Taylor and Francis.
- Morris, R. G., Rushe, T., Woodruffe, P. W., & Murray, R. M. (1995). Problem solving in schizophrenia: A specific deficit in planning. *Schizophrenia Research, 14*, 235–246.
- Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin, 126*, 220–246.
- Owen, A. M., James, M., Leigh, P. N., Summers, B. A., Quinn, N. P., Marsden, C. D., et al. (1992). Fronto-striatal cognitive deficits at different stages of Parkinson's disease. *Brain, 115*, 1727–1751.
- Patton, J., Stanford, M., & Barratt, E. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*, 768–774.
- Psychological Corporation (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Author.
- Raizner, R. D., Song, J., & Levin, H. S. (2002). Raising the ceiling: The Tower of London—Extended version. *Developmental Neuropsychology, 21*, 1–14.
- Roper v. Simmons, 541 U.S. 1040 (2005).
- Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*, 1394–1408.
- Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London, B, 298*, 199–209.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp. 290–312). Washington, DC: American Sociological Association.
- Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental Review, 28*, 78–106.
- Steinberg, L., Albert, D., Cauffman, E., Banich, M., Graham, S., & Woolard, J. (2008). Age differences in sensation seeking and impulsivity as indexed by behavior and self-report: Evidence for a dual systems model. *Developmental Psychology, 44*, 1764–1777.
- Steinberg, L., Cauffman, E., Woolard, J., Graham, S., & Banich, M. (2009). Are adolescents less mature than adults? Minors' access to abortion, the juvenile death penalty, and the alleged APA “flip-flop.” *American Psychologist, 64*, 583–594.
- Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., & Banich, M. (2009). Age differences in future orientation and delay discounting. *Child Development, 80*, 28–44.
- Steinberg, L., & Monahan, K. (2007). Age differences in resistance to peer influence. *Developmental Psychology, 43*, 1531–1543.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.
- Thompson-Schill, S., Jonides, J., Marshuetz, C., Smith, E. E., D'Esposito, M., Kan, I. P., et al. (2002). Effects of frontal lobe damage on interference effects in working memory. *Cognitive, Affective, & Behavioral Neuroscience, 2*, 109–120.
- Unterrainer, J. M., & Owen, A. M. (2006). Planning and problem solving: From neuropsychology to functional neuroimaging. *Journal of Physiology—Paris, 99*, 308–317.
- Unterrainer, J. M., Rahm, B., Kaller, C. P., Leonhart, R., Quiske, K., Hoppe-Seyler, K., et al. (2004). Planning abilities and the Tower of London: Is this task measuring a

- discrete cognitive function? *Journal of Clinical and Experimental Neuropsychology*, 26, 846–856.
- Unterrainer, J. M., Rahm, B., Leonhart, R., Ruff, C. C., & Halsband, U. (2003). The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Cognitive Brain Research*, 17, 675–683.
- van den Heuvel, O. A., Groenewegen, H. J., Barkhof, F., Lazeron, R., van Dyck, R., & Veltman, D. J. (2003). Frontostriatal system in planning complexity: A parametric functional magnetic resonance version of Tower of London task. *Neuroimage*, 18, 367–374.
- Wagner, G., Koch, K., Reichenbach, J. R., Sauer, H., & Schlösser, R. G. M. (2006). The special involvement of the rostralateral prefrontal cortex in planning abilities: An event-related fMRI study with the Tower of London paradigm. *Neuropsychologia*, 44, 2337–2347.
- Watkins, L. H., Rogers, R. D., Lawrence, A. D., Sahakian, B. J., Rosser, A. E., & Robbins, T. W. (2000). Impaired planning but intact decision making in early Huntington's disease: Implications for specific frontostriatal pathology. *Neuropsychologia*, 38, 1112–1125.
- Welsh, M. C., Satterlee-Cartmell, T., & Stine, M. (1999). Towers of Hanoi and London: Contributions of working memory and inhibition to performance. *Brain and Cognition*, 41, 231–242.
- Young, S., Morris, R., Toone, B., & Tyson, C. (2007). Planning ability in adults with attention-deficit/hyperactivity disorder. *Neuropsychology*, 21, 581–589.
- Zook, N. A., Davalos, D. B., DeLosh, E. L., & Davis, H. P. (2004). Working memory, inhibition, and fluid intelligence as predictors of performance on Tower of Hanoi and London tasks. *Brain and Cognition*, 56, 286–292.