Contemporary Educational Psychology 39 (2014) 314–325

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

The Academic Diligence Task (ADT): assessing individual differences in effort on tedious but important schoolwork ☆

Brian M. Galla [a,*], Benjamin D. Plummer [a], Rachel E. White [a], David Meketon [a], Sidney K. D'Mello [b], Angela L. Duckworth [c]

[a] University of Pennsylvania, 3701 Market St., Suite 214, Philadelphia, PA 19104
[b] University of Notre Dame, 106 Haggar Hall, Notre Dame, IN 46556
[c] University of Pennsylvania, 3701 Market St., Suite 215, Philadelphia, PA 19104

ARTICLE INFO

Article history:
Available online 7 August 2014

Keywords:
Academic diligence
Self-control
Grit
Assessment
Academic performance
Validation

ABSTRACT

The current study reports on the development and validation of the Academic Diligence Task (ADT), designed to assess the tendency to expend effort on academic tasks which are tedious in the moment but valued in the long-term. In this novel online task, students allocate their time between solving simple math problems (framed as beneficial for problem solving skills) and, alternatively, playing Tetris or watching entertaining videos. Using a large sample of high school seniors ($N = 921$), the ADT demonstrated convergent validity with self-report ratings of Big Five conscientiousness and its facets, self-control and grit, as well as discriminant validity from theoretically unrelated constructs, such as Big Five extraversion, openness, and emotional stability, test anxiety, life satisfaction, and positive and negative affect. The ADT also demonstrated incremental predictive validity for objectively measured GPA, standardized math and reading achievement test scores, high school graduation, and college enrollment, over and beyond demographics and intelligence. Collectively, findings suggest the feasibility of online behavioral measures to assess noncognitive individual differences that predict academic outcomes.

> Diligence is the mother of good fortune, and idleness—its opposite—never brought a man to the goal of any of his best wishes.
>
> ~Miguel de Cervantes

> The expectations of life depend upon diligence; the mechanic that would perfect his work must first sharpen his tools.
>
> ~Confucius

1. Introduction

Calvin and Susie are doing their homework. Tonight's assignment: multiplication tables. Despite their teacher's repeated appeals that "math is important," Calvin and Susie find this activity to be extremely tedious and would much prefer to be doing something else. Calvin, barely a few problems into the assignment, disengages, deciding instead to watch YouTube videos. Susie, on the other hand, ignores the urge to do something easier and more fun and stays focused on the task. Thirty minutes later, Calvin is now fully consumed in his search to find obscure cereal commercials from the 1980s, while Susie is just finishing the final problem on her assignment.

At least some of the time, as the example above illustrates, even the most valuable academic tasks can be tedious. Of the tasks faced daily by students, William James (1899) once observed, in a lecture delivered to local school teachers, there is by necessity "a large mass of material that must be dull and unexciting . . . It is certain that most schoolroom work, till it has become habitual and automatic, is repulsive, and cannot be done without voluntarily jerking back the attention to it every now and then" (pp. 104–105, 109). Disengaging from a task may bring about short-term relief—but at the sacrifice of long-term gains, particularly skill improvement. The importance of diligence to success in and beyond the classroom accords with common wisdom and, further, has been affirmed in several longitudinal studies. For instance, Oliver, Guerin, and Gottfried (2007) showed that adolescents rated higher by their parents on items similar to "stays at a task until it's done" prospectively predicted their high school GPA, college GPA, and total years of education at levels comparable to general intelligence. Likewise, Vaillant and

☆ This research was supported by funding from the National Institutes of Health (5-K01-AG033182-02), the Bill and Melinda Gates Foundation (OPP1046454), and the National Science Foundation (HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. The authors are grateful to the students, teachers, and school administrators who participated in this research.

* Corresponding author. Department of Psychology, University of Pennsylvania, 3701 Market St. Suite 214, Philadelphia, PA 19104, USA. Fax: (215) 573-2188.
E-mail address: gallabri@upenn.edu (B.M. Galla).

http://dx.doi.org/10.1016/j.cedpsych.2014.08.001
0361-476X/

Vaillant (1981) found that industriousness in early adolescence, based on school records, teacher evaluations, and child and parent interviews, predicted employment, income, and mental health better than either childhood socioeconomic status or intelligence.

Given the vital importance of diligence to success both in and beyond the classroom, the absence of a validated behavioral assessment thereof is surprising. In the current study, we report on the creation and validation of the Academic Diligence Task (ADT), a novel behavioral measure in which time and effort are voluntarily allocated to a tedious but "good for you" math skill-building activity versus entertaining but frivolous distractions (e.g., Tetris, music videos, and movie trailers). Performance is measured by either percent time on task or by productivity (i.e., number of problems correctly solved). A large sample ($N = 921$) of high school seniors completed this task along with a comprehensive battery of self-report questionnaires and a standard measure of intelligence. We assess evidence for the task's convergent and discriminant validity. Finally, we examine the incremental predictive power of this task for five objectively measured academic outcomes: GPA, math and reading standardized achievement test scores, high school graduation, and full-time college enrollment.

1.1. What is academic diligence?

Academic diligence can be defined as working assiduously on academic tasks which are beneficial in the long-run but tedious in the moment, especially in comparison to more enjoyable, less effortful diversions. We conceptualize academic diligence as a domain-specific facet of self-control, which refers more broadly to the regulation of thoughts, emotions, and behaviors in the face of momentary temptations and distractions (Tangney, Baumeister, & Boone, 2004). Of course, self-control is also called for in non-academic domains (e.g., exercise, diet, interpersonal relationships), but "studenting" is the primary occupation of adolescents (Corno & Xu, 2004) and universally understood by students as vitally important to their futures (Galla, Duckworth, Rikoon, & Haimm, 2014). Exercising self-control depends in part upon executive functions, a suite of higher-level cognitive processes, including working memory and response inhibition, which collectively enable top-down, goal-directed control over lower-level impulses (Diamond, 2012).

It was Freud Equation (1916–1917) who originally proposed that the chief developmental task of childhood is to learn how to control impulses toward immediate gratification. Today's students face a world in which putting up with what Freud called "a little unpleasure" is exceedingly difficult (p. 357). While incontrovertible evidence is difficult to muster on this point, it seems obvious that modern technology and an endlessly inventive consumer marketing sector have created a world in which short-term pleasures are increasingly salient and convenient (Akst, 2011; Moffitt, Poulton, & Caspi, 2013). It has never been easier to indulge in diversions that feel good right away, even if they do not bring about long-term benefit. Exercising self-control, according to both Freud and modern theorists (Eisenberg, Duckworth, Spinrad, & Valiente, 2014; Mischel, Shoda, & Rodriguez, 1989) can be distinguished from complying with external authority insofar as it is in service of the individual's own personally valued—rather than externally imposed—goals.

Academic diligence is also conceptually related to grit, defined as perseverance and passion for long-term and especially challenging goals. In contrast to self-control, grit is distinguished by its emphasis on effort and interest sustained over months and years rather than minutes and hours (Duckworth, Peterson, Matthews, & Kelly, 2007). In terms of the Big Five taxonomy of personality, grit and self-control belong to the conscientiousness family, which also encompasses the traits of dependability, punctuality, and orderliness, among other facets (MacCann, Duckworth, & Roberts, 2009) and is the Big Five factor most robustly associated with academic performance (Poropat, 2009). Accordingly, we expect measures of academic diligence to relate more strongly to conscientiousness than to other Big Five factors, including extraversion, emotional stability, and openness to experience. Big Five agreeableness, which encompasses traits related to empathy and warmth, as well as helpfulness and compliance (DeYoung, Quilty, & Peterson, 2007), may be related to academic diligence but is distinct in its motivational underpinnings: Agreeableness is primarily motivated by the desire to follow externally imposed rules and to get along with others.

1.2. Why develop a behavioral measure of academic diligence?

There are questionnaires for academic diligence (e.g., Bernard & Schuttenberg, 1995), but questionnaire measures suffer from numerous well-known limitations (Lucas & Baird, 2005). For instance, when endorsing items, respondents must compare the target (e.g., a teacher rating a student, a student rating himself) to some implicit standard, and standards may vary from respondent to respondent. To one student, "I am a hard worker" may be exemplified by doing five hours of homework each day; for another, the same statement may be exemplified by simply showing up for class. For both informant-report and self-report questionnaires, bias that arises from heterogeneous frames of reference reduces their validity (though, not necessarily, their reliability) (Heine, Lehman, Peng, & Greenholtz, 2002). For self-report questionnaires, social desirability bias and faking are other important limitations, particularly when there are incentives to appear admirable. Likewise, memory recall limitations and acquiescence bias can influence self-report responses, and halo effects can influence informant-report responses (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Finally, questionnaires may be relatively insensitive to subtle changes in behavior over time, either because of reference bias (i.e., standards for judging behavior changing over time in tandem with behavior itself) or because judgments that integrate behavioral observations over time may be less sensitive to changes therein (Duckworth, Yeager, & Bryk, 2014).

Limitations of questionnaire measures prompted us to develop a behavioral measure of academic diligence. Behavioral tasks, which do not rely upon subjective judgments but instead assay behavior directly, obviate the limitations related to reference bias, social desirability bias, and faking. Moreover, because tasks do not rely upon individuals to integrate a large number of observations over time into a summary judgment, they may also be more sensitive to subtle changes over time. While no measure is perfect, a behavioral measure of academic diligence may be better suited than questionnaires for certain purposes, including assessing the effects of interventions, evaluating students in high-stakes situations in which there are incentives for social desirability and faking, and assessing age-related changes in academic diligence.

1.3. Our approach to measurement

As a starting point in our design process, we reviewed tasks developed for preschool children about four decades ago by Patterson and Mischel (1975, 1976). In these studies, children aged 3 to 6 years were asked to work on a facile, repetitive activity (e.g., filling a large pegboard with pegs, copying Xs and Os into a grid) in the presence of a distracting "Mr. Clown Box," a large wooden box painted in a clown's likeness, complete with flashing lights, a voice, and a see-through belly containing toys. Performance was quantified both as the percentage of time each child spent working and the amount of work completed during the 10 to 12 minute testing period. More recent attempts to assess diligence using behavioral measures have been rare, but those that we found required participants to solve simple math problems while being tempted by a variety of distractions (e.g., TV commercials, magazines, video games) (Gollwitzer

& Schaal, 1998; Magen & Gross, 2007; Parks-Stamm, Gollwitzer, & Oettingen, 2010; Vohs et al., 2008). Like the early Patterson and Mischel work, diligence was quantified either as percentage of time on task (versus off task) or number of problems solved. Without exception, older and more recent studies were experimental in aim, designed to assess the effect of metacognitive interventions or situational manipulations. Not surprisingly, therefore, no attempt was made to present evidence for the convergent, discriminant, and predictive validity of diligence tasks as individual difference measures.

Unlike diligence on facile but tedious work tasks, perseverance in the face of extreme difficulty has inspired many attempts at behavioral operationalizations. Such measures typically present individuals with a difficult, impossible, and/or physically uncomfortable tasks and assess how long they are willing or able to continue (Feather, 1962). In such paradigms, there are typically no diversions, and total time before quitting is recorded rather than productivity or time on task. For example, Hartshorne and May (1929) designed a persistence task for school-age children in which a wooden puzzle was placed on each child's desk and the time they elected to solve it covertly recorded. Over the last century, these early efforts to assess persistence have since been expanded upon by others using an array of activities (e.g., solving difficult or impossible puzzles, squeezing a hand grip) (e.g., Battle, 1965; Crutcher, 1934; Feather, 1962; Ryans, 1938; Sansone, Weir, Harpster, & Morgan, 1992; Sansone, Wiebe, & Morgan, 1999; Ventura, Shute, & Zhao, 2013; Vohs et al., 2008).

1.4. The current study

While persistence in the face of high challenge is obviously important to achievement, working diligently while resisting diversions is at least as important—and perhaps more common in the daily lives of students. Indeed, since James made his observations of "schoolroom drudgery," entertaining distractions have arguably multiplied in number and potency (Moffitt et al., 2013). We therefore aimed to create an ecologically valid task (cf., Burgess et al., 2006) simulating the contemporary real-world conflict between homework and distractions of the digital age. Specifically, our task pits completing an important but tedious math skill-building exercise against playing a video game or watching entertaining video clips. While drilling basic math skills has high long-term utility (i.e., basic skills prepare students to learn more complex math operations), it is clear that most students do not enjoy such activities. For example, a recent national survey found that 56% of middle school students would rather eat broccoli than do their math homework, and 44% would rather take out the trash (Raytheon Company, 2012). At the same time, both middle school and high school students identify homework, studying, and other academic work as more important to their long-term goals than any other type of daily activity (Galla et al., 2014).

In a large sample of high school seniors, we examined the convergent validity of the ADT with self-report questionnaire measures of Big Five conscientiousness, and in particular, two of its facets: self-control and grit. We assessed evidence for discriminant validity using self-report questionnaire measures of theoretically unrelated personality traits, such as Big Five extraversion and emotional stability, as well as test anxiety, life satisfaction, and positive and negative affect. Given its relationship to compliance with externally-imposed rules (DeYoung et al., 2007), we expected Big Five agreeableness to positively correlate with the ADT, but importantly, we did not expect agreeableness to explain the association between the ADT and conscientiousness, self-control, or grit. Finally, we examined evidence for incremental predictive validity for five objective indicators of academic performance, including senior year grades, performance on two standardized tests, graduation from high school, and subsequent college enrollment.

2. Method

2.1. Participants

The sample included $N = 921$ high school seniors (mean age = 17.90 years, $SD = 0.51$) drawn from two large public high schools in the Northeast United States. These students were drawn from a larger study on college persistence. Both schools were comprehensive public high schools. All students who were sufficiently able for inclusion in mainstream classes according to official school records were eligible to participate. In other words, only students with severe mental or physical disabilities, no English language comprehension, or both were excluded. According to school records, approximately 36% of the students were Black, 33% were White, 21% were Asian, 8% were Latino, and 2% were of other ethnic backgrounds; 49% were female. About 55% of students were from low-income families, as indicated by their qualification for free or reduced-price lunch. About 4% of students were designated as English-language learners, and about 7% received special education services. Students were recruited through an information letter from the principal that also contained an opt-out parent consent form.

2.2. Procedure

Students completed the 20-minute Academic Diligence Task, a battery of self-report questionnaires, and a battery of cognitive tests during regular school hours on school computers. Students at School 1 ($n = 419$) completed the study measures during 3 separate 50-minute sessions (ADT, cognitive tests, and self-report questionnaires, respectively), and students at School 2 ($n = 502$) completed the study measures during a single 2.5 hour session (cognitive tests, self-report questionnaires, and ADT, respectively). Note that while students were given 2.5 hours to complete all the study measures at School 2, very few students required the full period to complete all the measures (this was usually the case for students who arrived late to the testing session, or for students who may have initially experienced technological issues, such as no Wi-Fi connectivity). Similarly, not every student at School 1 required the entire 50 minute period to complete the various testing batteries. For example, students at both schools completed the self-report questionnaire battery on average in fewer than 20 minutes.

2.3. Measures

2.3.1. The Academic Diligence Task (ADT)

The chief goal of this paper was to develop and validate a novel behavioral measure to assess individual differences in academic diligence. As alluded to above, the ADT was designed to mirror a real-world choice students must confront when completing homework: the choice to remain engaged in tedious, but important assignments, or browse the Internet to watch entertaining videos or play video games. Specifically, the task consisted of a split-screen interface with the choice to either complete single-digit subtraction problems ("Do Math") or watch YouTube video clips (90 second clips of popular music videos, movie trailers, sports highlights, or general interest videos) or play Tetris ("Play game or watch movie"). See Fig. 1 for a graphical depiction of the main user interface. At any point during the activity students were free to either productively focus on the skill-building task or pass the time by engaging with the distractions, although the software restricted engagement to one activity at a time.

Before starting the task, each student was provided a pair of ear buds to wear during the entire testing session. The ear buds ensured that students were able to interact with the task independently, and they also reduced distractions for students sitting close to one

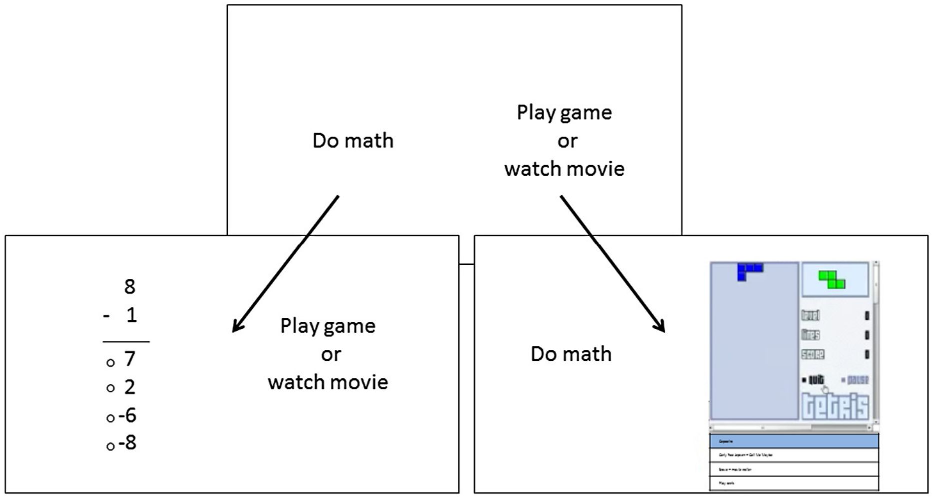

Fig. 1. This figure depicts the main user interface of the Academic Diligence Task. Students are given the choice to "Do math" or "Play game or watch movie." If they click "Do math," they solve single digit subtraction problems. If they click "Play game or watch movie," a pull down menu is displayed that contains various video clips or the option to play the video game Tetris. At any point during the activity students are free to either productively focus on the skill-building task or pass the time by engaging with the distractions, although the program restricts engagement to one activity at a time.

another. To get acquainted with the software, students first completed a one-minute practice block of single-digit subtraction problems but without the option to watch videos or play Tetris. Students read: "First we are going to do some math problems for one minute. Try to solve as many problems as quickly and accurately as you can. To choose an answer, click on the bubble next to the answer you would like to select."

After completing the practice block, students read an FAQ-style cover story that described specifically what they would be doing. Students first read that they would be solving math problems similar to the practice block. They were again encouraged to try to solve as many problems as quickly and accurately as possible.

Next, the general utility of completing the subtraction problems was emphasized to students. Students read that practicing basic math skills, like single-digit subtraction, can help make them better at problem solving ("You are doing this activity because it can make you smarter. Research shows that practicing basic math skills, like simple math, makes you a better problem solver . . . The more problems that you do, the better you will be at solving problems in the future.").[1] Thus, if they desired, students could reasonably see completing the math problems as useful to their academic skills.

Students then read that if they felt like it they were free to click on the opposite side of the screen to watch fun videos or play Tetris, but were also reminded that the more problems they completed, the more likely it is that their problem solving ability would improve. Thus, the instructions presented students with a choice: they could either spend their time solving simple, but "good for you" math problems, or alternatively, watching frivolous, but entertaining videos and playing video games. It is important to note that students were free to do whatever they preferred; students were in no way obligated to do the math if they did not want to, and nor would they be punished if they decided to watch the videos or play Tetris (indeed, about 5% of students exclusively watched videos or played Tetris).

After reading the cover story, students answered, "Why are you going to do math problems?," by selecting one of the following response options: (a) Because I'm in school, (b) Practicing math can make me smarter, and (c) My teacher told me to. We also included this question to rule out the alternative hypothesis that performance on the Academic Diligence Task merely involved compliance with school rules and authority (as indicated by response options (a) and (c)). About 79% ($n = 726$) of students selected the expected response (Practicing math can make me smarter), suggesting that they both understood the message of the cover story and that it was credible.

Next, the basic cover story was reiterated to students a final time:

> You are about to begin the activity! Remember, you will be able to watch videos or play Tetris whenever you feel like it. To watch videos or play Tetris, click on the right side of the screen. To switch back to math, click on the left side of the screen. If you have any final questions before you begin please raise your hand now and an experimenter will assist you.

Students then began the 20-minute test phase of the ADT. The test phase consisted of five, four-minute blocks where the student could toggle between completing the single-digit subtraction problems or watching videos or playing Tetris. Math problems were presented one at a time, and students selected the correct answer among four response options. Once a response was selected, another math problem was immediately displayed; students did not receive performance feedback during the task. After each task block students were asked, "How bored were you by the math during the last session?." They responded using a 5-point scale from 1 = *not at all bored* to 5 = *very bored*. From start to finish (including the practice block, reading the cover story, answering subjective experience questions, completing the 20-minute test phase, etc.), the entire ADT lasted for about 30 minutes.

Unknown to the students, the software recorded information regarding their interaction with the task, from which we derived two indices of academic diligence: (a) productivity and (b) time on task. Productivity represents the total number of math problems solved correctly, summed across all five task blocks. Time on task represents the total percentage of time students spent solving the math problems, averaged across all five task blocks. Time on task excludes time students spent on the videos/Tetris or time spent idling, which was coded as the amount of time it took for the student to

[1] This cover story was based on research suggesting that the retrieval of basic math facts can impact performance on standardized math achievement tests (Royer, Tronsky, Chan, Jackson, & Marchant, 1999).

Table 1
Descriptive statistics.

Variable	*M*	*SD*	Alpha	Range
Academic Diligence Task				
Productivity	244	180	.89	0–966
Time on task	54%	33%	.85	0–100%
Boredom	3.33	1.07	.84	1.00–5.00
Intelligence	93.64	17.01	–	40–132
Self-Report Questionnaires				
Big Five extraversion	3.56	0.84	.75	1.00–5.00
Big Five conscientiousness	3.79	0.66	.66	1.50–5.00
Big Five agreeableness	4.08	0.62	.66	1.00–5.00
Big Five openness	3.94	0.70	.68	1.50–5.00
Big Five emotional stability	2.94	0.90	.77	1.00–5.00
Grit	3.83	0.73	.77	1.00–5.00
Self-control	3.76	0.58	.76	1.50–5.00
Life satisfaction	5.17	1.28	–	1.00–7.00
Positive affect	3.65	0.74	.79	1.00–5.00
Negative affect	2.97	0.70	.75	1.00–5.00
Test anxiety	2.84	1.04	.79	1.00–5.00
Attitudes toward math	4.39	2.14	–	1.00–7.00
Academic outcomes				
Graduated high school on time	95%	–	–	–
Full-time college enrollment	53%	–	–	–
Standardized math achievement test	1398	265	–	816–2587
Standardized reading achievement test	1350	224	–	700–2214

Note: Internal reliability consistency estimates were not calculated for life satisfaction and attitudes toward math because these were single-item scales.

make his/her first selection of "Do Math" or "Play game or watch movie" following the start of each task block.[2]

2.3.2. Self-report questionnaires

Unless stated otherwise, students rated all questionnaire items on a 5-point scale from 1 = *not at all like me* to 5 = *very much like me*. Some items were modified slightly to improve comprehension (for example, the emotional stability item, "Is depressed, blue," was modified to, "Is sad or unhappy."). For each questionnaire measure, items were averaged to create a composite score in which higher scores indicated more agreement with the construct. The order of the surveys was fixed across schools, although items to each survey were presented in randomized order. Descriptive statistics and internal reliability consistency estimates for each measure are shown in Table 1.

2.3.2.1. Big Five personality. Students completed items adapted from the short form version of the Big Five Inventory of dimensions of personality (John & Srivastava, 1999). Using four items per dimension, we assessed extraversion (e.g., "I am outgoing. I like to meet new people"), agreeableness (e.g., "I am helpful and I cooperate with other people"), conscientiousness (e.g., "I am organized and neat"), emotional stability (e.g., "I get stressed out easily"), and openness to experience (e.g., "I am curious. I am very interested in learning new things.").

2.3.2.2. Grit. Students completed four items adapted from the Grit scale (Duckworth et al., 2007), which measures passion and perseverance for long-term goals (e.g., "I work very hard. I keep working when others stop to take a break").

2.3.2.3. Self-control. Students completed the Domain-Specific Impulsivity Scale for children (Tsukayama, Duckworth, & Kim, 2013). Using four items per dimension, we assessed self-control in the domains of schoolwork (e.g., "I pay attention and resist distractions in class") and interpersonal relationships (e.g., "I can remain calm even when criticized or otherwise provoked."). A self-control composite score was calculated as the mean of all eight items, with higher scores indicating higher self-control.

2.3.2.4. Test anxiety. Students answered four questions about their test anxiety: "Even when I'm well prepared for a test, I feel very nervous about it;" "During a test I often get so nervous I forget the answers that I know;" "As soon as an exam is over I try to stop worrying about it, but I just can't;" and "During a test I often think about what will happen if I fail."

2.3.2.5. Positive and negative affect. Students rated from 1 = *never* to 5 = *always* how often they experience five positive emotions (i.e., happy, relaxed, excited, healthy and awake, full of energy) and five negative emotions (i.e., worried, sad, bored, angry, frustrated).

2.3.2.6. Life satisfaction. Students answered from 1 = *extremely unsatisfied* to 7 = *extremely satisfied* the following question about their life satisfaction, "Overall, how satisfied or unsatisfied are you with your life?" Similar single-item measures of life satisfaction have been used in large-scale national panel studies and show adequate reliability estimates (Lucas & Donnellan, 2012).

2.3.2.7. Attitudes toward math. To control for the possibility that performance on the Academic Diligence Task was merely a reflection of existing preferences for math, we included the covariate of attitudes toward math to analyses reported below. Students answered from 1 = *strongly dislike* to 7 = *strongly like* the following question regarding their preferences for math, "Please rate your attitude toward math." Using single-item scales to measure attitudes is common practice as they demonstrate equivalent predictive validity to multiple-item scales (Bergkvist & Rossiter, 2007).

2.3.3. Intelligence

Intelligence was assessed with the matrix reasoning subtest of the Kaufmann Brief Intelligence Test (Kaufman & Kaufman, 1990). Students are shown a series of patterns in which one portion of the pattern is missing. From a set of response options, students must determine the shape/pattern that completes the pattern. The current version of the task included a total of 36 matrix reasoning problems, and the task ended after four consecutive incorrect responses or completion of all problems. The number of correct answers before a ceiling of four incorrect trials in a row constituted the raw score, which was converted to an age-normed scaled score in accordance with the scoring manual.

2.3.4. Academic outcomes

2.3.4.1. Grade point average (GPA). From school records, we recorded senior year GPA. At School 1, GPA was coded on a 4.0 scale ($M = 2.51$, $SD = 0.97$), while at School 2, GPA was coded on a 100 point scale ($M = 84.97$, $SD = 7.82$). To accommodate the different grading scales between schools, we z-standardized the two sets of GPAs within each school before combining them into one GPA variable ($M = 0$, $SD = 1$). Outlier values ($n = 3$)—those more than 3.29 standard deviations below the mean—were set to −3.29 (Erceg-Hurn & Mirosevich, 2008).

[2] Independent samples t-tests indicated a significant effect of school on ADT performance, but not boredom. Specifically, students at School 1 had a higher average time on task (School 1: $M = 0.59$, $SD = 0.30$; School 2: $M = 0.50$, $SD = 0.35$; $t = -4.13$, $p < .001$) and higher average productivity (School 1: $M = 278$, $SD = 180$; School 2: $M = 216$, $SD = 175$; $t = -5.22$, $p < .001$). Schools did not differ in average levels of self-reported boredom during the task (School 1: $M = 3.36$, $SD = 1.10$; School 2: $M = 3.31$, $SD = 1.05$; $t = -0.66$, $p = .512$). Performance differences between schools likely stem from the fact that students at School 1 completed the ADT during a single session, while students at School 2 completed it during a larger testing battery. The general pattern of convergent, discriminant, and incremental predictive validity reported below was mostly similar when running analyses separately for each school. As such, and to improve statistical power, we combined the two schools into one large sample. To control for school differences in performance on the ADT, we included a dummy code for school in every analysis reported in the paper.

2.3.4.2. Standardized achievement test scores. From school records we also recorded scores on state-mandated, standardized achievement tests for math and reading taken during the students' junior year. We used scaled scores for math and reading, which are equal-interval, normed conversions of the raw scores.

2.3.4.3. High school graduation. From school records we coded a binary indicator of whether each student graduated high school on time (0 = *no*, 1 = *yes*). About 5% of the sample (aggregated across both schools) did not graduate high school on time.

2.3.4.4. Full-time college enrollment. We collected college enrollment data for the fall semester following high school graduation using the National Student Clearinghouse (NSC, http://www.studentclearinghouse.org) database. The NSC is a non-profit organization created in connection with the student financial aid lending industry to gather enrollment verification data for student borrowers (Schoenecker & Reeves, 2008). In the years since its inception, the NSC database has become a valuable tool for researchers interested in tracking college enrollment because of its extremely high coverage rate (Dynarski, Hemelt, & Hyman, 2013), and because objective student enrollment records can be collected without the need to contact individual schools or students. Using the NSC data, we created a binary indicator (0 = *no*, 1 = *yes*) of whether or not each student was enrolled in college full-time (at either a 2-year or 4-year institution) in fall semester immediately following high school graduation. At the time of our data request (April 2014), 53% of students had enrolled full-time in college during the previous fall semester.

3. Results

3.1. Descriptive statistics for the Academic Diligence Task

As intended, the ADT was easy in the sense of presenting math problems students could do: The mean accuracy rate on math problems was 96%. Also as expected, the two indices of diligence, productivity (the total number of math problems solved correctly) and time on task varied widely. As shown in Table 1, some students solved no problems at all, choosing instead to exclusively watch the videos or play Tetris, whereas other students spent the entire 20 minutes on the math skill-building activity. In fact, the most diligent student solved 966 problems—more than three times as many problems as the average student. As might be expected, productivity and time on task were also tightly yoked ($r = .85$, $p < .001$).

What evidence is there to suggest that the ADT measured academic diligence and not just math ability? We posit that if the ADT were simply a measure of math ability then productivity would not decline over time. That is, if math ability were mostly driving performance on the task then a student should be equally productive across task blocks (we might also expect productivity to increase over time as students get better at solving the problems). And yet, the opposite pattern was found in the data. As shown in Fig. 2, both productivity and time on task each declined over time, suggesting that a reduction in focused engagement, not math ability, was primarily driving change in performance on the task. During the first four-minute block, students solved an average of 63 math problems. By the final block this number decreased by 38%, such that

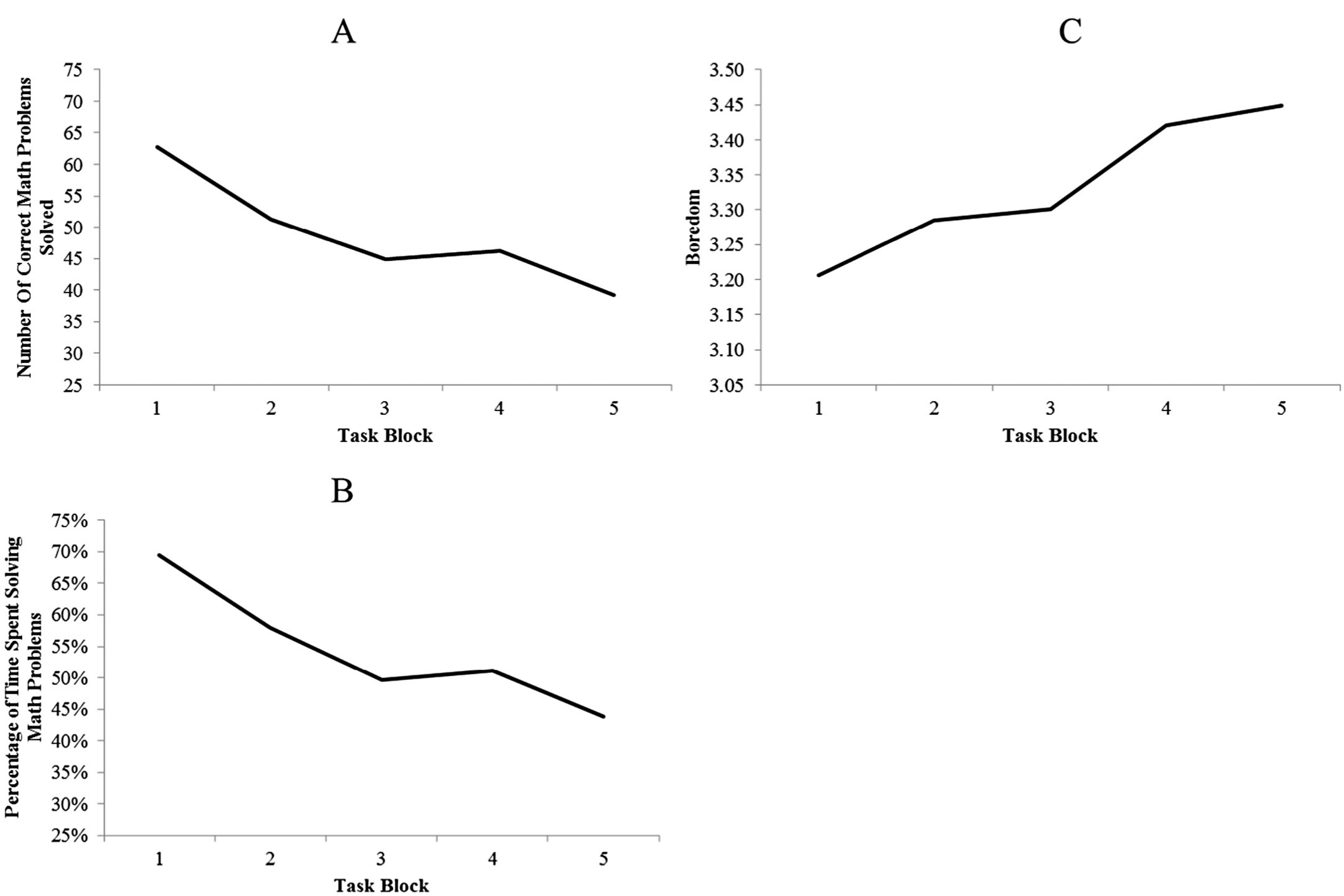

Fig. 2. This figure depicts the mean number of math problems solved (A), the mean percentage of time spent on task (B), and mean boredom (C) during each task block.

students only completed an average of 39 math problems. In other words, the average rate of correctly solved problems decreased from roughly 16 problems per minute in Block 1 to roughly 10 problems per minute in Block 5. Similarly, the amount of time students spent solving the math problems declined from 70% in the first block to 44% in the final block.

Multilevel growth curve analyses confirmed that performance on the ADT declined over time. Specifically, we fit two multilevel growth curve models in which productivity and time on task (measured as observed variables) were modeled as a function of an intercept and a linear growth term (task block). We permitted the intercept and the slope to vary randomly between students, and we used restricted maximum likelihood estimator to model regression parameters. We also included a school dummy variable as a level 2 covariate of the intercept and slope (see Model 2 in Appendix for a complete description of the model; unconditional means models with intra-class correlations are shown under Model 1). The number of problems solved significantly decreased with each task block (b = –5.21, t = –16.32, p < .001, Pseudo-R^2 = .16), as did the percentage of time spent solving the math problems (b = –0.06, t = –17.99, p < .001, Pseudo-R^2 = .14). These results indicate that productivity dropped by about 5 math problems per block, and time on task dropped by about 6% per block. Pseudo-R^2 values indicate that 16% of the within-person variance in Productivity and 14% of the within-person variance in Time on Task was explained by the linear growth term.

Also as intended, students found the math skill-building activity to be increasingly boring: Self-reported boredom increased by an average of 7%, from 3.21 in Block 1 to 3.45 in Block 5, a significant change (b = 0.06, t = 5.01, p < .001, Pseudo-R^2 = .20; Fig. 2). Growth curve analyses also indicated that boredom and diligence were linked within each student across time. Specifically, we fit two additional multilevel growth curve models predicting within-student change in boredom as a function of within-student change in either productivity or time on task (see Model 3 in Appendix). A one standard deviation (*SD*) change in the number of problems students solved (relative to their own mean level) predicted a 0.20 *SD* change in how bored they felt (b = 0.20, t = 8.43, p < .001). Likewise, a one *SD* change in the time spent on the math activity also predicted a 0.20 *SD* change in boredom (b = 0.20, t = 9.82, p < .001).

Students who spent more time on task also showed steeper increases in boredom during the task compared to students who spent less time on task. To illustrate these trajectories, we fit a final multilevel growth curve analysis in which boredom was modeled as a function of the intercept and time (i.e., task block) at level 1, and total time on task as a z-standardized level 2 (between-student) predictor of the intercept and slope (see Model 4 in Appendix). Total time on task did not predict the initial levels of boredom (b = 0.00, t = 0.01, p = .99), but it did moderate the effect of task block on boredom (b = .05, t = 4.27, p < .001). As shown in Fig. 3, simple slopes analyses indicated that students one *SD* above the mean in total time on task showed significant increases in boredom across task block (b = .12, t = 6.59, p < .001). Contrariwise, students who spent most of their time watching videos and playing Tetris (i.e., students one *SD* below mean in total time on task) did not show increases in boredom across task block (b = 0.01, t = 0.52, p = .61).

In sum, results suggest that performance on the ADT required focused engagement despite sharp increases in boredom and the immediate availability of more pleasurable distractions.

3.2. Reliability of the Academic Diligence Task

To test whether the ADT was internally reliable, we examined the intercorrelations between performance and boredom across blocks. The number of problems solved across blocks showed medium-to-large correlations (rs = .50 to .66, ps < .001; $r_{average}$ = .60), and similarly, time on task was also correlated across blocks (rs = .41 to .59, ps < .001; $r_{average}$ = .52). Boredom was correlated across blocks (rs = .35 to .66, ps < .001; $r_{average}$ = .52). In other words, students who solved more problems in a given block, and who spent more time on task, were more likely to do so in other blocks. Likewise, feeling bored in any given block was related to feeling bored in the other blocks. We also examined alpha coefficients for the task by treating performance and boredom in each block as individual items. Performance and boredom exhibited high internal consistency estimates: α = .89 (productivity), α = .85 (time on task), and α = .84 (boredom).

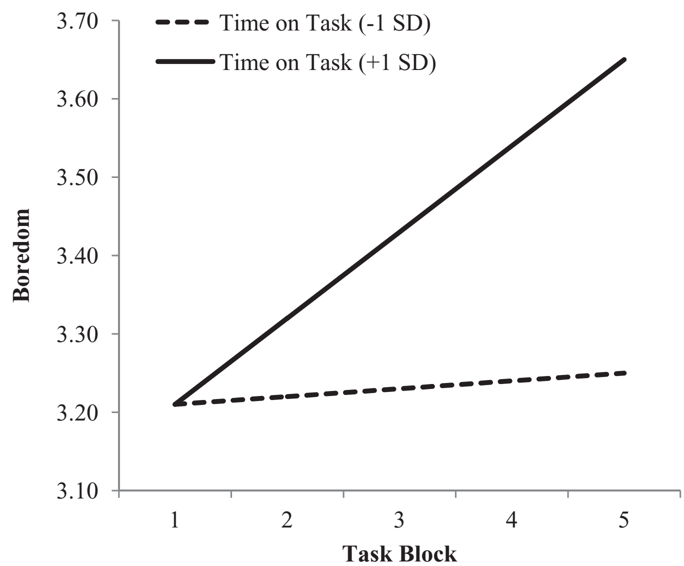

Fig. 3. Estimated change in boredom across task block as a function of total amount of time on task. Students who spent more time on task (1 *SD* above the mean) showed significant increases in boredom across task block (b = 0.12, t = 6.59, p < .001), while students who spent less time on task (1 *SD* below the mean) showed no increase in boredom across task block (b = 0.01, t = 0.52, p = .61).

3.3. Convergent and discriminant validity of the Academic Diligence Task

Convergent and discriminant validity were examined using a series of partial correlations (*pr*). In each analysis, we included dummy codes for school, gender, race/ethnicity (with White as the reference category), and free or reduced price lunch status; additional controls for intelligence and attitudes toward math were also added to each model. We included these control variables to address potential concerns that performance on the task was confounded by membership in certain demographic categories (e.g., girls versus boys), intelligence, or positive preferences for math. Missing data were handled using full information maximum likelihood which produces less biased and more efficient results than other methods, such as listwise or casewise deletion (Baraldi & Enders, 2010; Schafer & Graham, 2002).

As shown in Table 2, performance on the ADT, measured by either productivity or time on task, converged with Big Five conscientiousness (pr = .08 and .09, ps < .05, respectively) and its facets, self-control (pr = .11 and .15, ps < .005) and grit (pr = .16 and .17, ps < .001). These correlations, while small in size, are comparable to meta-analytically derived estimates of the correlation size between questionnaire measures and behavior measures of self-control (rs = .10 to .21; Duckworth & Kern, 2011).

Are Big Five conscientiousness, self-control, and grit equally strong predictors of performance on the ADT? Paunonen and Ashton (2001) showed that facet-level scores on personality scales were

Table 2
Partial correlations between academic diligence task performance and self-report measures.

Variable	Productivity	Time on task
Convergent validity		
Big Five conscientiousness	.08*	.09**
Grit	.16**	.17**
Self-control	.11**	.15**
Discriminant validity		
Big Five extraversion	–.03	–.04
Big Five emotional stability	.00	.03
Big Five openness	.04	.05
Big Five agreeableness	.07*	.11**
Positive affect	.02	.05
Negative affect	–.03	–.06
Life satisfaction	–.01	.02
Test anxiety	–.01	.00

Note: * $p < .05$, ** $p < .01$. All correlations controlled for school, gender, race/ethnicity, free or reduced price lunch status, intelligence, and attitudes toward math.

Table 3
Regression analyses examining incremental predictive validity of productivity during the academic diligence task on academic outcomes.

Variable	Senior year GPA	Math test scores	Reading test scores	Graduated high school on time	Full-time college enrollment
	β	β	β	OR	OR
School	.00	–.05†	.14**	0.17**	0.90
Female	.22**	–.01	.08*	2.75**	1.45*
F/R lunch	–.03	–.09**	–.12**	1.81†	0.84
Ethnicity					
Black	–.19**	–.15**	–.14**	0.42†	0.89
Asian	.04	.07*	–.05	0.43	1.77**
Mixed race	–.10**	–.06*	–.06†	0.29*	1.06
Attitudes toward math	.19**	.29**	–.01	1.21	1.14†
Intelligence	.20**	.36**	.31**	1.22	1.35**
Productivity	.17**	.23**	.22**	1.84**	1.44**

Note: † $p < .10$, * $p < .05$, ** $p < .01$. *OR* = odds ratio. School was entered as a dummy variable, with students at School 2 ($n = 502$) serving as the reference category. Race/ethnicity was entered as a series of dummy variables, with White as the reference category. The mixed ethnicity category consists of Latino youth as well as those categorized as multiracial by school records. F/R lunch = Free or reduced price lunch status.

stronger predictors of theoretically-related outcomes than were broad, factor-level scores. On this view, we would expect the facets of self-control and grit to "win out" over Big Five conscientiousness when predicting ADT performance. To test this possibility, we fit a series of simultaneous regression models predicting productivity and time on task, respectively. In each model, conscientiousness was entered as a predictor along with either self-control or grit, as well as the covariates mentioned previously. In these models, self-control and grit each predicted unique variance in ADT performance (βs = .09 to .17, *ps* < .05), but conscientiousness did not (βs = –.02 to .02, *ns*).

As predicted, performance on the ADT was unrelated to Big Five extraversion, openness, or emotional stability (*prs* = –.04 to .05, *ns*; $pr_{average} = .03$). Moreover, performance was unrelated to negative affect, positive affect, life satisfaction, and test anxiety (*prs* = –.06 to .05, *ns*, $pr_{average} = .02$). Consistent with our prediction that performance on the ADT might be weakly associated with traits associated with compliance, both productivity and time on task were correlated with Big Five agreeableness ($pr = .07$, $p < .05$; $pr = .11$, $p < .001$, respectively).

Since Big Five agreeableness was also associated with grit ($r = .41$, $p < .001$) and self-control ($r = .53$, $p < .001$), we fit another series of simultaneous regression models similar to those described above to rule out the possibility that performance on the ADT was solely a function of Big Five agreeableness. This time we included agreeableness as a predictor of ADT performance along with either self-control or grit (and covariates). In each model, Big Five agreeableness failed to predict unique variance in ADT performance (βs = .01 to .04, *ns*). Contrariwise, grit and self-control each predicted unique variance in ADT performance when controlling for agreeableness (βs = .09 to .15, *ps* < .05).

In sum, results indicate that performance on the Academic Diligence Task—measured by either productivity or time on task—was reliably associated with individual differences in self-control and grit above and beyond demographics, intelligence, and attitudes toward math. Performance on the ADT was not reliably associated with Big Five extraversion, openness, and emotional stability, or with other theoretically unrelated constructs (e.g., life satisfaction). Finally, results indicate that Big Five agreeableness, a personality trait that encompasses compliance, does not confound the association between performance on the ADT and self-control or grit. As another check to rule out the effect of compliance to school authority on the observed results, we reran tests of convergent and discriminant validity while also controlling for students' responses to the question, "Why are you going to do math problems?" Results were virtually unchanged when we included this variable in our analyses.

3.4. Incremental predictive validity of the Academic Diligence Task on academic outcomes

We next tested the incremental predictive validity of productivity and time on task for five objectively measured academic outcomes: senior year GPA, standardized achievement test scores in math and reading, high school graduation, and full-time college enrollment. Predictive validity for GPA and achievement test scores were tested using ordinary least-squares regressions, and predictive validity for high school graduation and full-time college enrollment were tested using logistic regressions. In all models, we controlled for school, free and reduced price lunch status, gender, race/ethnicity, intelligence, and attitudes toward math. Again, we used full information maximum likelihood to handle missing data. To facilitate interpretation of odds ratios (*OR*) in logistic regressions, we *z*-standardized continuous variables (i.e., intelligence, attitudes toward math, ADT performance) prior to analyses.

As shown in Table 3, productivity on the ADT predicted senior year GPA (β = .17, $p < .001$) and junior year standardized math scores (β = .23, $p < .001$) and reading scores (β = .22, $p < .001$). Productivity also prospectively predicted the odds of graduating high school on time (*OR* = 1.84, $p = .004$) and full-time college enrollment (*OR* = 1.44, $p < .001$). Specifically, students 1 *SD* above the mean in productivity had 84% greater odds of graduating high school on time, and 44% greater odds of being enrolled in college full-time, compared to students with mean levels of productivity.

In similar models, regression coefficients for time on task were smaller in magnitude but were nevertheless statistically significant. As shown in Table 4, time on task predicted senior year GPA (β = .07, $p = .018$) and junior year standardized math scores (β = .07, $p = .012$) and reading scores (β = .09, $p = .003$). Time on task also prospectively predicted the odds of graduating high school on time (*OR* = 1.65, $p = .006$) and full-time college enrollment (*OR* = 1.17, $p = .03$). Specifically, students 1 *SD* above the mean in time on task had 65% greater odds of graduating high school on time, and 17% greater odds of being enrolled in college full-time, compared to students whose time on task was exactly average.[3]

[3] Again, to rule out the effect of compliance to school authority on observed results, we reran tests of incremental predictive validity while also controlling for Big Five agreeableness or students' answers to the question, "Why are you going to do math problems?" Incremental predictive validity of productivity and time on task on academic outcomes was largely the same.

Table 4
Regression analyses examining incremental predictive validity of time on task during the academic diligence task on academic outcomes.

Variable	Senior year GPA	Math test scores	Reading test scores	Graduated high school on time	Full-time college enrollment
	β	β	β	OR	OR
School	.03	–.02	.17**	0.17**	0.99
Female	.23**	.01	.09**	2.62**	1.50**
F/R lunch	–.03	–.10**	–.13**	1.75†	0.82
Ethnicity					
Black	–.20**	–.18**	–.16**	0.39*	0.83
Asian	.04	.07*	–.05	0.47	1.80**
Mixed race	–.11**	–.07*	–.07*	0.27*	1.00
Attitudes toward math	.22**	.33**	.02	1.27	1.19*
Intelligence	.22**	.40**	.34**	1.23	1.42**
Time on task	.07*	.07*	.09**	1.65**	1.17*

Note: † $p < .10$, * $p < .05$, ** $p < .01$. *OR* = odds ratio. School was entered as a dummy variable, with students at School 2 ($n = 502$) serving as the reference category. Race/ethnicity was entered as a series of dummy variables, with White as the reference category. The mixed ethnicity category consists of Latino youth as well as those categorized as multiracial by school records. F/R lunch = Free or reduced price lunch status.

4. Discussion

In this paper, we summarized the design and validation of a novel behavioral measure of academic diligence. During this computerized task, students must allocate attention and effort between a tedious, but important math skill-building task, or alternatively, entertaining videos or Tetris. Using a large sample of high school seniors, results indicated that completing more math problems, and spending more time solving problems, were associated with increases in boredom. Moreover, students typically disengaged from the primary skill-building activity as the task progressed, choosing instead to focus on the more pleasant media distractions. Ultimately, the total amount of time students spent on task versus the amount of time they spent on media distractors was approximately even (54% on task).

Performance on the ADT, measured either by productivity (i.e., the total number of math problems solved) or time on task (versus watching videos/playing Tetris), demonstrated convergent validity with individual differences in Big Five conscientiousness and even more so for its facets, self-control and grit. Overall, the observed correlations with self-control and grit were small in size (prs = .11 to .17; $pr_{\text{average}} = .15$), but compare favorably to meta-analytic estimates of the association between questionnaire and behavioral measures of self-control (rs = .10 to .21; Duckworth & Kern, 2011). Performance on the ADT also showed a clear pattern of discriminant validity: it was unrelated to Big Five extraversion, openness, and emotional stability, and furthermore, showed no reliable association with life satisfaction, positive and negative affect, or test anxiety. While ADT performance was also associated with Big Five agreeableness, this association was not significant when controlling for either self-control or grit. Finally, both productivity and time on task showed incremental predictive validity over and above demographics, intelligence, and attitudes toward math for five objectively measured academic outcomes: senior year GPA, scores on standardized math and reading achievement tests, graduation from high school, and full-time college enrollment (measured nearly a full year after completion of the ADT). In sum, results provide evidence for the reliability and validity of the Academic Diligence Task as a measure of individual differences in academic diligence among high school seniors.

4.1. Implications for research and practice

Results of the current study affirm the critical role of diligence, as captured through the Academic Diligence Task, for succeeding academically. To our knowledge, the Academic Diligence Task is the first behavioral measure to demonstrate validity as an individual differences measure. This novel task has the potential to catalyze research on academic diligence in several ways. Importantly, the Academic Diligence Task was designed for group administration, is relatively brief, and requires no special materials or trained experimenters. These advantages stand in contrast to the behavioral measures of persistence designed by Hartshorne and May (1929), complete evidence for whose validity was never demonstrated. The design features of the ADT should therefore increase the usefulness of the task for school-based research where time and resources may be limited.

Moreover, recent studies have shown that the ADT appears to be sensitive to subtle changes in behavior over time. For example, randomized controlled trials show that performance on the ADT can be improved through interventions that target a self-transcendent purpose for learning (Yeager et al., 2014) and a psychologically distanced perspective toward work (White et al., 2014). These emerging findings suggest that the Academic Diligence Task may be a useful tool—combined with questionnaire methods—in which to evaluate the effect of school-based interventions designed to improve academic self-regulation.

4.2. Potential criticisms of the Academic Diligence Task

We see five potential criticisms of the Academic Diligence Task. The first is that it merely taps math ability or an interest in math. There are several reasons to doubt this possibility. Accuracy on the task was extremely high—the mean accuracy rate was 96%—indicating that students had little difficulty completing the math problems if they chose to do so. Additionally, productivity on the math problems declined as students advanced through the task, which is more suggestive of disengagement than differences in ability. Second, productivity and time on task each demonstrated evidence of convergent, discriminant, and predictive validity above and beyond measures of intelligence (i.e., fluid reasoning) and attitudes toward math. Finally, productivity and time on task predicted scores on standardized achievement tests of reading as well as math. Taken together, these results support our contention that the ADT is measuring the underlying construct of diligence rather than just math ability or an interest in math activities. Of course, replication using non-math skill-building activities (e.g., tests of verbal skills, spatial skills) is required to more conclusively rule out the competing argument that math ability and preferences for math are driving observed effects.

The second potential criticism is that performance on the ADT merely taps compliance with school rules and authority. We expected, and indeed showed, that task performance was weakly correlated with Big Five agreeableness (Poropat, 2009), assessed with items including, “I am helpful and I cooperate with other people,” and “I am considerate and kind to almost everyone.” However, there are several reasons why we doubt compliance substantially influenced the observed results. We carefully worded the cover story to ensure students did not feel obligated to complete the math problems if they did not want to. Students were free to choose how to spend their time and there were no negative consequences for them if they chose to watch videos or play Tetris. Also, we explicitly assessed compliance with school authority via students’ responses to the cover story, and controlling for this variable had no impact on tests of convergent, discriminant, and incremental predictive validity. Finally, Big Five agreeableness did not predict ADT performance above and beyond self-reported grit or self-control. Controlling for

Big Five agreeableness (results not shown) also had almost no impact on incremental predictive validity.

Third, one might wonder whether the split-screen interface pitting "Do math" versus "Play game or watch movie" is, in fact, ecologically valid. A recent naturalistic study of homework completion found that students frequently switch between on-task and off-task behaviors when doing their homework, and that digital distractions are most often to blame (Rosen, Carrier, & Cheever, 2013). In this study, middle school, high school, and university students were observed for 15 minutes while studying in their homes. Students of all ages had an average of 2 to 3 computer windows open simultaneously while doing their homework, and switched to a distracting task (e.g., Facebook, texting) every 5 to 6 minutes. This study suggests that our split-screen design, in which a homework activity competes for attention with immediately available digital distractions, does in fact simulate a typical daily experience for teens.

A fourth objection to the Academic Diligence Task is that the motivational salience of watching entertaining YouTube video clips far outweighs the motivational salience of completing a boring math assignment, and therefore biases against students completing the math problems. This was intentional. It is precisely this unbalanced motivation between academic work and entertaining distractions that we were trying to assess with the ADT. Recall that we define academic diligence as working assiduously on academic tasks which are beneficial in the long-run but tedious in the moment, especially in comparison to more enjoyable, less effortful diversions. As such, a critical design element involved simulating a scenario in which students would need to muster continuous effort to remain focused on a task that they knew to be valuable to their long-term goals, but which was far less appealing than an available alternative activity (see also Patterson & Mischel, 1975, 1976).

Fifth, and finally, it is possible that performance on the Academic Diligence Task was confounded by existing attitudes toward watching YouTube videos and Tetris. For example, the student who solved 966 math problems might have done so because he or she was uninterested in watching video clips or playing Tetris. We agree that some students may not be interested in watching YouTube videos or playing Tetris, and that because of this lack of interest in YouTube/Tetris, such students might decide to spend more time solving math problems. However, just as the student who solved 966 problems is not representative of the entire sample, we also do not think that a lack of interest in online videos is representative of the typical student. Students spent, on average, just under 50% of the 20-minute task watching videos and playing Tetris. This suggests to us that the *average* student was at least somewhat interested in watching the video distractors compared to solving the math problems.

While we cannot completely rule out the possibility that attitudes toward online videos influenced ADT performance, we were able to address this concern indirectly. During the Academic Diligence Task, students answered the question "How tempting were the videos/games during the activity?" from *1 = not at all tempting* to *5 = very tempting*. Students at School 1 completed this question after Block 3 of 5, and students at School 2 completed this question after the Block 5 of 5. For the following exploratory analyses, we combined these ratings into a single-item measure of video temptation. Unsurprisingly, students who felt more tempted by the videos/Tetris were less productive ($r = -.33$, $p < .001$) and spent less time on task ($r = -.35$, $p < .001$). Of course, feeling tempted by the videos/Tetris is not the same thing as being *interested* in watching videos, though they may be related. It is also important to remember that video temptation was assessed *after* students had completed most (School 1) or all (School 2) of the Academic Diligence Task, and is therefore confounded by experience with the task. That is, students who watched more videos would be more likely to say they were tempted by them. For these two reasons, we did not include this measure in main analyses. (Incidentally, the student who solved 966 math problems reported a temptation score of 3 out of 5 and a mean boredom score of 4.2 out of 5, suggesting some level of internal conflict between solving math problems and watching videos/playing Tetris).

Keeping in mind the limitations of the temptation rating, we reran tests of convergent validity with grit and self-control and incremental predictive validity for academic outcomes while simultaneously controlling for video temptation. The majority of results were effectively unchanged. In fact, controlling for video temptation substantively influenced only 2 main results: (1) self-control no longer correlated significantly with productivity ($r = .06$, $p = .10$), and (2) time on task no longer significantly predicted graduation from high school ($OR = 1.40$, $p = .08$). We hasten to note that, despite the nonsignificant p-values, the direction of effects was unchanged and the effect sizes were only slightly attenuated. These exploratory analyses show that (as would be expected) video temptation did affect performance, but again, it did not confound the convergent, discriminant, and incremental predictive validity of the Academic Diligence Task.

4.3. Limitations of the current study

There are several limitations of the current study. First, while the current study used an ethnically and socioeconomically diverse sample, replication studies with students from different age groups will strengthen the external validity of the ADT. The use of a very easy single-digit subtraction skill-building activity should make the task suitable even for students in primary grades where math ability may be more of a concern.

Second, due to the limited amount of research contact with each student, we had to rely on short-form versions of personality scales. While all reliability estimates exceeded $\alpha = .65$, future research examining the convergent validity of the ADT might incorporate full-item scales. However, there is little rationale to expect that Type I error was increased since measurement unreliability tends to attenuate the estimates of interest. Moreover, short-form personality measures—despite lower reliability—demonstrate comparable patterns of convergent, discriminant, and predictive validity with long-form Big Five measures (Donnellan, Oswald, Baird, & Lucas, 2006).

Third, and relatedly, we relied on a single-item measure of global attitudes toward math, and therefore our assessment of math attitudes is rather limited. Our decision to use a single-item measure was again pragmatic: given the limited amount of research contact with students we had to be selective in the number and size of the measures we administered. However, we note that single-item measures of attitudes are common and demonstrate equivalent predictive validity to multi-item attitude scales (Bergkvist & Rossiter, 2007). Nevertheless, future research testing the validity of the ADT might incorporate more complete assessments of attitudes toward math (e.g., value of math).

Finally, and as described previously, we did not assess (and control for) students' existing attitudes toward watching YouTube videos and playing Tetris. While controlling for ratings of temptation toward videos/Tetris did not affect major results, this is not the same as controlling for whether students had an initial interest in watching YouTube or playing Tetris. Because the ADT provides students with a choice between two mutually exclusive activities, future research should assess (prior to the start of the task) perceptions about how enjoyable versus how important watching YouTube videos is compared to doing math problems.

5. Conclusion

From time to time, classroom learning activities are unavoidably tedious and boring (D'Mello, 2013; Galla et al., 2014; James, 1899). Despite this unpleasant reality, students also understand the importance of schoolwork for their future success (Galla et al., 2014; Raytheon Company, 2012). Diligently engaging on a tedious assignment despite an overwhelming desire to do something more pleasant is therefore a consequential, albeit, difficult task for most students. And yet, little has been done to validate behavioral measures to assess this important quality. To address this gap in the literature, we designed a scalable, Internet-delivered measure of academic diligence. Our measure, the Academic Diligence Task, simulates real-world challenges students often confront when doing their homework: the choice to stay engaged in a tedious, but valuable assignment or consume online media. In a large sample of high school seniors, performance on the task—measured by productivity and time on task—demonstrated convergent validity with self-report ratings of self-control, grit, and Big Five conscientiousness as well as discriminant validity from theoretically unrelated constructs, such as Big Five extraversion, openness, and emotional stability, life satisfaction, test anxiety, and positive and negative affect. The ADT also demonstrated incremental predictive validity for five objectively measured academic outcomes: senior year GPA, scores on standardized math and reading achievement tests, high school graduation, and college enrollment, over and beyond demographics, intelligence, and attitudes toward math. Collectively, findings further demonstrate the feasibility of online behavioral measures to assess noncognitive individual differences that predict academic outcomes.

Appendix: Multilevel Growth Curve Models

Model 1: Unconditional Means Models

Level 1 (within-student level):

$$Y_{ij} = B_{0j} + e_{ij} \tag{1}$$

Level 2 (between-student level):

$$B_{0j} = \gamma_{00} + u_{0j} \tag{1a}$$

In Model 1, we fit unconditional means models to partition the within-student and between-student variance in Productivity, Time on Task, and Boredom. Intra-class correlations indicated that 58% of the variance in Productivity, 49% of the variance in Time on Task, and 52% of the variance in Boredom was attributable to between-student sources.

Model 2: Within-Student Change in Performance and Boredom across Task Block

Level 1 (within-student level):

$$Y_{ij} = B_{0j} + B_{1j}(\text{task block}) + e_{ij} \tag{2}$$

Level 2 (between-student level):

$$B_{0j} = \gamma_{00} + \gamma_{01}(\text{School}) + u_{0j} \tag{2a}$$

$$B_{1j} = \gamma_{11} + \gamma_{12}(\text{School}) + u_{1j} \tag{2b}$$

In Model 2, we fit a multilevel growth curve model predicting within-student change in productivity, time on task, and boredom as a function of an intercept and a linear growth term (task block). To aid interpretation of the intercept as performance/boredom during Block 1, task block was coded such that Block 1 = 0. In this model, we permitted the intercept and the slope to vary randomly between students (u_{0j} and u_{1j}, respectively), and we used restricted maximum likelihood estimator to model regression parameters. We also included a school dummy variable (School 1 = 1, School 2 = 0) as a level 2 covariate of the intercept and slope.

In addition to fitting linear growth models, we also tested quadratic and cubic growth trajectories for Productivity, Time on Task, and Boredom. In addition to significant linear change, Productivity and Time on Task were also characterized by significant quadratic and cubic trajectories. Boredom, on the other hand, was characterized by significant linear change only. Because we were interested primarily in showing directional trends over time—rather than mapping ideal trajectories—we only report information about linear change as these provide close approximations to the true change in performance over time (as shown in Fig. 2). Results of nonlinear growth trajectories are available upon request.

Model 3: Within-Student Change in Performance Predicts Within-Student Change in Boredom

Level 1 (within-student level):

$$\begin{aligned} \mathit{Boredom}_{ij} = B_{0j} &+ B_{1j}(\text{task block}) \\ &+ B_{2j}(\text{productivity/time on task}) + e_{ij} \end{aligned} \tag{3}$$

Level 2 (between-student level):

$$B_{0j} = \gamma_{00} + \gamma_{01}(\text{School}) + u_{0j} \tag{3a}$$

$$B_{1j} = \gamma_{11} + \gamma_{12}(\text{School}) + u_{1j} \tag{3b}$$

$$B_{2j} = \gamma_{21} + \gamma_{22}(\text{School}) + u_{2j} \tag{3c}$$

In Model 3, we fit a multilevel growth curve model predicting within-student change in boredom as a function of an intercept, a linear growth term (task block), and either productivity or time on task. In these models, productivity and time on task were centered at each student's own mean (or, within-student centered). To aid interpretation of regression parameters in standard deviation units, prior to running the model we *z*-standardized boredom, productivity, and time on task across task block and across participants. All other model specifications were the same as Model 2.

Model 4: Within-Student Change in Boredom as a Function of Between-Student Time on Task

Level 1 (within-student level):

$$\mathit{Boredom}_{ij} = B_{0j} + B_{1j}(\text{task block}) + e_{ij} \tag{4}$$

Level 2 (between-student level):

$$B_{0j} = \gamma_{00} + \gamma_{01}(\text{School}) + \gamma_{02}(\text{Time on Task}) + u_{0j} \tag{4a}$$

$$B_{1j} = \gamma_{11} + \gamma_{12}(\text{School}) + \gamma_{13}(\text{Time on Task}) + u_{1j} \tag{4b}$$

In Model 4, we fit a multilevel growth curve model predicting within-student change in boredom as a function of an intercept and a linear growth term (task block). At level 2, we included average time on task as a *z*-standardized, level 2 (between-student) predictor of the intercept and growth slope. All other model specifications were the same as Model 2.

References

Akst, D. (2011). *We have met the enemy: Self-control in an age of excess*. New York: Penguin Press.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5–37. doi:10.1016/j.jsp.2009.10.001.

Battle, E. S. (1965). Motivational determinants of academic task persistence. *Journal of Personality and Social Psychology*, *2*(2), 209–218.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184. doi:10.2307/30162466.

Bernard, H., & Schuttenberg, E. M. (1995). Development of the Diligence Inventory-higher education form. *Journal of Research & Development in Education*, *28*(2), 91–100.

Burgess, P. W., Alderman, N., Forbes, C., Costello, A. M.-A., Coates, L., Dawson, D. R., et al. (2006). The case for the development and use of “ecologically valid” measures of executive function in experimental and clinical neuropsychology. *Cambridge Journals Online*, *12*, 194–209.

Corno, L., & Xu, J. (2004). Homework as the job of childhood. *Theory Into Practice*, *43*(3), 227–233. doi:10.1207/s15430421tip4303_9.

Crutcher, R. (1934). An experimental study of persistence. *Journal of Applied Psychology*, *18*(3), 409–417. doi:10.1037/h0073473.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. doi:10.1037/0022-3514.93.5.880.

Diamond, A. (2012). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, *18*(2), 192–203.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101.

Duckworth, A. L., Yeager, D. S., & Bryk, A. S. (2014). Measurement matters: Assessing attributes other than cognitive ability. Manuscript in preparation.

Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. (2013). The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes *NBER Working Paper*: National Bureau of Economic Research.

Eisenberg, N., Duckworth, A. L., Spinrad, T. L., & Valiente, C. (2014). Conscientiousness: Origins in childhood? *Developmental Psychology*, *50*, 1331–1349. doi:10.1037/a0030977.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601.

Feather, N. T. (1962). The study of persistence. *Psychological Bulletin*, *59*(2), 94–115.

Freud, S. (1916–1917). *Introductory lectures on psychoanalysis*. New York: W. W. Norton & Company.

Galla, B. M., Duckworth, A. L., Rikoon, S. H., & Haimm, C. A. (2014). Academic work is unpleasant but important: Reports of momentary subjective experience in a national sample of adolescents. Manuscript in preparation.

Gollwitzer, P. M., & Schaal, B. (1998). Metacognition in action: The importance of implementation intentions. *Personality and Social Psychology Review*, *2*(2), 124–136.

Hartshorne, H., & May, M. A. (1929). *Studies in the nature of character: Volume II: Studies in self-control* (Vol. 2). New York: McMillan.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918.

James, W. (1899). *Talks to teachers on psychology and to students on some of life's ideals*. New York: Holt and Company.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford Press.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test (KBIT)*. Circle Pines, MN: American Guidance Service.

Lucas, R., & Donnellan, M. B. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research*, *105*(3), 323–331. doi:10.1007/s11205-011-9783-z.

Lucas, R. E., & Baird, B. M. (2005). Self-report methodologies. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective*. Washington, DC: American Psychological Association.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, *19*, 451–458.

Magen, E., & Gross, J. J. (2007). Harnessing the need for immediate gratification: Cognitive reconstrual modulates the reward value of temptations. *Emotion*, *7*(2), 415–428.

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, *244*(4907), 933–938.

Moffitt, T., Poulton, R., & Caspi, A. (2013). Lifelong impact of early self-control: Childhood self-discipline predicts adult quality of life. *American Scientist*, *101*(5), 352–359.

Oliver, P. H., Guerin, D. W., & Gottfried, A. W. (2007). Temperamental task orientation: Relation to high school and college educational accomplishments. *Learning and Individual Differences*, *17*(3), 220–230.

Parks-Stamm, E. J., Gollwitzer, P. M., & Oettingen, G. (2010). Implementation intentions and test anxiety: Shielding academic performance from distraction. *Learning and Individual Differences*, *20*(1), 30–33.

Patterson, C. J., & Mischel, W. (1975). Plans to resist distraction. *Developmental Psychology*, *11*(3), 369–378.

Patterson, C. J., & Mischel, W. (1976). Effects of temptation-inhibiting and task-facilitating plans on self-control. *Journal of Personality and Social Psychology*, *33*(2), 209–217.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five predictors of academic achievement. *Journal of Research in Personality*, *35*(1), 78–90.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338.

Raytheon Company. (2012). Math relevance to U.S. middle school students. Retrieved from http://www.raytheon.com/newsroom/rtnwcm/groups/corporate/documents/content/rtn12_studentsmth_results.pdf.

Rosen, L. D., Carrier, M. L., & Cheever, N. A. (2013). Facebook and texting made me do it: Media-induced task-switching while studying. *Computers in Human Behavior*, *29*(3), 948–958. doi:10.1016/j.chb.2012.12.001.

Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. J., & Marchant, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology*, *24*, 181–266.

Ryans, D. G. (1938). An experimental attempt to analyze persistent behavior: I. measuring traits presumed to involve “persistence”. *The Journal of General Psychology*, *19*(2), 333–353. doi:10.1080/00221309.1938.9711208.

Sansone, C., Weir, C., Harpster, L., & Morgan, C. (1992). Once a boring task always a boring task? Interest as a self-regulatory mechanism. *Journal of Personality and Social Psychology*, *63*(3), 379–390.

Sansone, C., Wiebe, D. J., & Morgan, C. (1999). Self-regulating interest: The moderating role of hardiness and conscientiousness. *Journal of Personality*, *67*(4), 701–733.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schoenecker, C., & Reeves, R. (2008). The National Student Clearinghouse: The largest current student tracking database. *New Directions for Community Colleges*, *2008*(143), 47–57. doi:10.1002/cc.335.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–322.

Tsukayama, E., Duckworth, A. L., & Kim, B. E. (2013). Domain-specific impulsivity in school-age children. *Developmental Science*, *16*(6), 879–893.

Vaillant, G. E., & Vaillant, C. O. (1981). Natural history of male psychological health: X. Work as a predictor of positive mental health. *American Journal of Psychiatry*, *138*(11), 1433–1440.

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52–58. doi:10.1016/j.compedu.2012.07.003.

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology*, *94*(5), 883–898.

White, R. E., Ostergaard, E., Shablack, H., Orvell, A., Shaefer, C., Kross, E., et al. (2014). Self-distancing promotes perseverance in work tasks. Manuscript in preparation.

Yeager, D. S., Henderson, M., Paunesku, D., Walton, G. M., Spitzer, B., D'Mello, S., et al. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. *Journal of Personality and Social Psychology* (in press).