

Review

# Ability Tests Measure Personality, Personality Tests Measure Ability: Disentangling Construct and Method in Evaluating the Relationship between Personality and Ability

Patrick C. Kyllonen \* and Harrison Kell

Research & Development, Educational Testing Service, Princeton, NJ 08541, USA; hkell@ets.org

\* Correspondence: pkyllonen@ets.org; Tel.: +1-(609)-734-1056; Fax: +1-(609)-734-1755

Received: 25 May 2018; Accepted: 26 June 2018; Published: 10 July 2018



**Abstract:** Although personality and cognitive ability are separate (sets of) constructs, we argue and demonstrate in this article that their effects are difficult to tease apart, because personality affects performance on cognitive tests and cognitive ability affects item responses on personality assessments. Cognitive ability is typically measured with tests of items with correct answers; personality is typically measured with rating-scale self-reports. Oftentimes conclusions regarding the personality–ability relationship have as much to do with measurement methods as with construct similarities and differences. In this article, we review key issues that touch on the relationship between cognitive ability and personality. These include the construct-method distinction, sources of test score variance, the maximal vs. typical performance distinction, and the special role for motivation in low-stakes testing. We review a general response model for cognitive and personality tests that recognizes those sources of test score variance. We then review approaches for measuring personality through performance (objective personality tests, grit game, coding speed, economic preferences, confidence), test and survey behavior (survey effort, response time, item position effects), and real-world behavior (study time, registration latency, behavior residue, and social media). We also discuss ability effects on personality tests, indicated by age and cognitive ability effects, anchoring vignette rating errors, and instructions to ‘fake good’. We conclude with a discussion of the implications for our understanding of personality and ability differences, and suggestions for integrating the fields.

**Keywords:** personality; ability; constructs; methods; sources of variance; maximal vs. typical performance; high vs. low-stakes testing; objective personality tests; confidence; economic preferences; survey effort; response time; item-position effects; faking

## 1. Introduction

This article reviews evidence for how cognitive ability and personality traits are integrated. There is a substantial literature that examines the correlations between measures of cognitive ability and measures of intelligence, contemporaneously [1,2], and longitudinally [3]. However, this literature almost exclusively treats scores from cognitive abilities and personality measures as pure indicators of cognitive abilities or personality traits, respectively, save for measurement error, and occasionally, as acknowledged by the inclusion of more than one measure, factorial uniqueness.

We do not review that literature.<sup>1</sup> Instead, our point of departure for this article is that personality and cognitive ability are intertwined during item responses on cognitive tests and personality

---

<sup>1</sup> But see [4–6] for recent reviews.

assessments. That is, a response to a cognitive test item typically reflects personality to some extent, and a response to a personality item typically reflects cognitive ability to an extent. The possibility that test scores reflect influences other than ability has long been recognized [7,8]. But the fact that such cross-contamination exists, to the extent that it does, complicates a number of widely held beliefs about both cognitive ability and personality, such as their relative independence, the magnitude of sub-group and country differences in personality and intelligence; the meaning of trend changes, such as maturation and early adolescent ‘storm and stress’ [9]; and the interpretation of predictive validity evidence linking personality and ability measures to educational and workforce outcomes.

The article is organized as follows. First, we review the construct-method distinction—the distinction between cognitive ability and personality constructs (or ‘traits’), and the methods used to measure those constructs. We believe these are almost always confounded, and often conflated, as indicated in, for example, the ‘personality change’ literature, which deals almost exclusively with changes in responses to a very specific kind of assessment, a self-rating Likert scale, rather than to personality *per se* [10,11]. We also review influences other than ability on cognitive test scores, as initially outlined by Thorndike [8], and review a general model to accommodate multiple variance sources, following a proposal by Heckman and Kautz [12]. We argue that the maximal-typical performance and high-stakes, low-stakes distinctions are critical to test score interpretation, and that motivation may be especially important in low-stakes testing. Next, we review studies that measure personality, using measures other than rating scales. These include so-called ‘objective personality tests’, and measures of choosing to put forth effort, such as the grit game and the coding speed test. We also review studies focused on economic preferences, and other studies focused on confidence measures. We review measures of construct irrelevant variance in test and survey behavior, including survey effort, response time, and item position effects. We also review measures of personality obtained in real-world behavior, such as study time, registration latency, and social media. We argue that most of these studies show that what is interpreted as a cognitive ability measure, can often be understood as measuring personality as well, and simultaneously, that it is possible to measure personality outside Likert-scale measures.

Following our review of personality determinants of performance on cognitive measures, we review instances of cognitive influences on traditional personality tests. These include age and cognitive ability effects, anchoring vignette rating errors, and ability to follow instructions to ‘fake good’ on personality tests. We conclude with a discussion of the implications for our understanding of personality and ability differences and suggestions for integrating the fields.

## 2. Construct-Method Distinction

To answer the question, “what is the relationship between personality and intelligence?” it is helpful to start with definitions. A suitable definition of personality comes from the American Psychological Association (APA), as follows: “individual differences in characteristic patterns of thinking, feeling, and behaving”.<sup>2</sup> APA’s 1996 Intelligence Task Force [14] likewise provides a definition of intelligence as “<individual differences in the> ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought”.<sup>3</sup>

For a combination of historical, accidental, and practical reasons, two broad approaches to measuring these two constructs have emerged and come to dominate how we think of them.

---

<sup>2</sup> See also Roberts [13], who defines personality as “relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances”.

<sup>3</sup> It is worth noting that such definitions often include references to “individual differences”. Individual differences should be thought of as a method for identifying the factors discussed here, but it is not the only method. Training effects (of intelligence or personality) are non-individual-differences methods, as are artificial intelligence approaches (e.g., building expert systems).

For personality, the dominant methodology has to do with endorsements of descriptions of characteristic behavior, thoughts, beliefs, or attitudes. Descriptions can be trait terms or statements; endorsements can be ratings, rankings, or preference judgments between two or more descriptions; and the endorsements can be done by the self or by others—peers, teachers, or supervisors. There is a lively body literature in personality psychology on the differences between these methods, but personality is most often measured with self or peer ratings of statements using the Likert scale format [15]. The essence of the method is that it involves evaluating the target’s “characteristic patterns of thinking, feeling, and behaving” represented by descriptions.

For intelligence, the dominant method is the standardized test, with a problem and response format (multiple choice, short answer, and essay), scored as right or wrong, or in some cases, partially right. This characterizes all IQ, achievement, and selection and admissions tests such as the Armed Services Vocational Aptitude Battery (ASVAB) and the College Board’s SAT test.

So, the answer to the question of “what is the relationship between intelligence and personality?” would typically be given by a correlation between a measure of personality using the rating scale method, and a measure of intelligence based on a test. From this literature we find that those who score well on intelligence tests are slightly more likely to say that they “enjoy hearing new ideas” compared to those who score poorly, but are equally likely as their low scoring counterparts to say that they “respect others”, “pay attention to details”, “make friends easily”, or “worry about things”. Studies such as this abound in the literature [2,3,16], varying in the particular tests and in the statement endorsement method used.

However, intelligence can just as easily as personality be evaluated with the statement endorsement methodology. That is, rather than giving a test, we can ask examinees their level of agreement to statements such as “I understand complex ideas”, “I adapt effectively to the environment”, “I learn from experience”, or “I engage in various forms of reasoning to overcome obstacles by taking thought”. This is not the typical way we measure intelligence, and in fact, personality psychologists might recognize or characterize some of these statements (which were taken directly from the APA definition of intelligence, above) as measures of the ‘intellect’ facet of openness. But our point here is to distinguish the constructs (personality and intelligence) from the methods used to measure them (rating scales and tests).

Conversely, personality can be measured with tests. This is arguably more challenging than measuring intelligence with ratings, and there is not a consensus literature on how to do this. In this article, we review a number of different approaches for measuring personality without ratings.

The points here seem obvious, but we believe that there is a tendency to ignore the construct-method distinction when discussing personality and intelligence. We hear too often discussions about personality that really are discussions about examinees’ responses to Likert rating scales of personal descriptions. Similar to Boring’s [17] oft-repeated criticism of intelligence being “what the tests of intelligence test”, personality can equivalently be critiqued as “what it is that personality rating scales assess”.

### 3. Sources of Test-Score Variance

Cronbach [7], following Thorndike [8] (see also [18]), classified the sources of variance in test scores into the dimensions temporary vs. lasting, and general vs. specific individual characteristics. Lasting characteristics include personality, both lasting general (“attitudes, emotional reactions, or habits generally operating in situations like the test situation” [7], and lasting specific (“attitudes, emotional reactions, or habits related to particular test stimuli”, [7]. Temporary-general effects include health, fatigue, emotional strain, mood, and motivation, which may be referred to as state variables [19]. Temporary-specific effects include fluctuations in attention (distraction) and memory, emotions brought on by specific items, specific knowledge pertaining to an item, or item type, perhaps due to coaching and luck.

Table 1 is a modification of the Cronbach–Thorndike table, in which we add several rows (sources of variance) and we also include additional columns on how these sources of variance are treated or

could be treated both in assessment design and in analysis of test scores or item responses. In general, sources of variance other than the intended construct can be labeled as construct irrelevant sources of variance and should be minimized through test design (e.g., clarifying/simplifying instructions, eliminating cues), or statistically controlled for in modeling (e.g., multitrait-multimethod, factor analysis, control variables in regression analysis).

Some of the cognitive test score variance sources are marked as personality (e.g., typical effort, anxiety) or as a state source (e.g., emotional strain and mood/emotion). We do this both to highlight the importance of personality and mood-state variance sources in cognitive testing, which is the main point of this article, but also to show that the role of these other factors has been acknowledged in psychological testing for a long time. We also provide mostly contemporary references for these factors that illustrate how they contribute to performance on tests. In this article, we examine a number of these alternative variance sources in depth.

There has been longstanding awareness of the potential for personality and mood to enter into cognitive test performance. However, the major effect of that awareness has been to be mindful of these potential sources of contamination in test design (third column), and occasionally, for particular studies to model the alternative variance sources (fourth column). For the most part, however, we argue that these potentially confounding influences are ignored.

**Table 1.** Sources of cognitive test score variance, including personality trait and state variance.

Sources of Cognitive Test-Score Variance	Examples	Design Treatment	Analysis Treatment
<b>I. Lasting, general characteristics (lasting person characteristics that pertain to performance on this test and tests like it)</b>			
1. Target construct of the test	general cognitive ability, verbal ability	lengthen test to extent feasible	true score variance
2. Other cognitive factors that might influence test scores	reading, vocabulary, related cognitive factors	minimize role of other factors	factor analysis; MTMM
3. General test-taking skills	comprehend instructions, test-wiseness	test coaching, practice tests	typically ignored
4. Skill with the test's item types	multiple-choice vs. short-answer	vary item type	ignore or model
5. Personality—tendency to typically exert effort	grit game <sup>1</sup> ; picture-number <sup>2</sup> , maximal-typical performance distinction <sup>3</sup>	-	predict important outcomes
6. Personality—lack of anxiety	(lack of) test anxiety <sup>4</sup>	anxiety training	ignored
7. Personality—managing time in a time-limit test	running out of time on a standardized test <sup>5</sup>	provide clocks, warnings	time-accuracy models
<b>II. Lasting, specific characteristics (lasting characteristics that pertain only to this test or item subset)</b>			
1. Skills required by particular item types	mode (PBT, DBT), response format (MC, CR)	multiple methods	ignore or model
2. Skills required by the particular content sample	form differences	create parallel forms	measurement error
3. Personality—Effort inducing states due to test conditions	computer-based assessments, incentives <sup>6</sup>	make tests/items engaging	ignored or researched
4. Personality—Emotional state induced by test stimuli	math anxiety <sup>7</sup> , stereotype threat <sup>8</sup>	minimize inducements	ignored or researched
<b>III. Temporary, general characteristics of the individual (pertain to the whole test and tests like it, but only for a short while)</b>			
1. Temporary health, fatigue, emotional strain	poor performance due to being ill, sad, tired	allow retest	discard all but highest score
2. Environment effects	poor performance due to noisy/hot room	allow retest, venue flexibility	discard all but highest score
3. Level of practice on skills required by tests of this type	novel test format/content	provide pretest practice	model growth/dynamic testing
4. Personality—Effort-inducing states	motivation incentives (feedback, payments) <sup>9</sup>	provide incentives to all	typically ignored
5. Personality—Emotional states	stressors (high stakes, fear of failure) <sup>10</sup>	anxiety training	typically ignored
<b>IV. Temporary, specific characteristics of the individual (pertain only to this test or item subset, and only this time)</b>			
1. Personality—Changes in fatigue/motivation over the course of a test	Item position effects <sup>11</sup>	minimize test length, make test more engaging	error or model
2. Personality—Emotional reaction to item response/feedback	discouragement/slowdown after item failure <sup>12</sup> , "entity" theory of intelligence <sup>13</sup>	Content, sensitivity, fairness reviews	error
3. Fluctuations in attention and memory	mind wandering <sup>14</sup>	make test/items more engaging	error
4. Unique skill or knowledge of these particular items	effects of special coaching <sup>15</sup> , prior exposure <sup>16</sup>	test coaching, practice tests	error or part of construct
5. Mood/emotion State induced by item(s)	test item invokes a negative emotion <sup>17</sup>	content, fairness reviews	error or part of construct
6. Luck in the selection of answers by guessing	guess correct answer	avoid MC or provide many options	error, guessing correction

NOTES: Table adapted from Thorndike ([8], p. 73), Cronbach ([7], p. 175), Stanley ([18], p. 364). MTMM—multitrait multimethod model; PBT—paper-based test; DBT—digital-based test; MC—multiple-choice test; CR—constructed response (short answer) test; SEE—standard error of equating; RT—response time. <sup>1</sup> Alan, Boneva, and Ertac [20]; <sup>2</sup> Segal [21]; <sup>3</sup> Beus and Whitman [22]; Sackett, Zedeck, and Fogli [23]; <sup>4</sup> Hembree [24]; <sup>5</sup> van der Linden [25]; <sup>6</sup> Liu, Bridgeman, and Adler [26], Finn [27]; <sup>7</sup> Hembree [28]; <sup>8</sup> Steele and Aronson [29]; <sup>9</sup> Liu, et al. [26], Finn [27]; <sup>10</sup> Beilock [30]; <sup>11</sup> DeBeer and Janssen [31]; <sup>12</sup> Rabbit [32]; <sup>13</sup> Mueller and Dweck [33]; <sup>14</sup> Kane and McVay [34]; Terhune, Croucher, Marcusson-Clavertz, and Macdonald [35]; <sup>15</sup> Powers and Rock [36]; <sup>16</sup> Irvine [37]; <sup>17</sup> Eich [38].

#### 4. Response Model for Cognitive and Personality Tests

Borghans, Duckworth, Heckman, and ter Weel [39] provide a general framework for measuring performance in any situation. They proposed viewing performance through the framework of the standard factor analysis model, but pointed out some of its limitations, including the arbitrary location of factors, the lack of concern for causality, the fixed nature of factors, and, particularly in personality, the problem of faking. They proposed an alternative based on predicting real world outcomes, which addresses these traditional limitations.

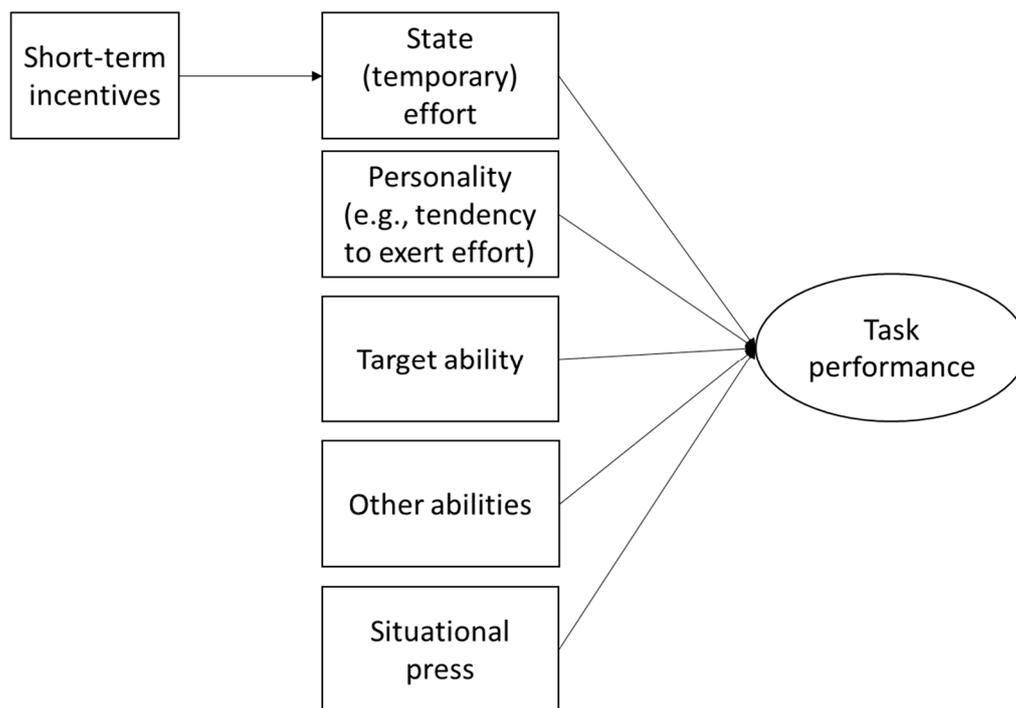
They also explicitly proposed that the measured traits ( $Y_l$ ) in particular situations or occasions ( $Y_l^n$ ), are only imperfect proxies for true traits ( $f_l$ ), with other influences on measured traits being other related traits ( $f_{\sim l}$ ), specific situational incentives associated with the measurement of the target trait ( $R_l^n$ ) (e.g., high-stakes vs. low-stakes testing; rewards for performance), and the context for measuring the target trait ( $W_l^n$ ) (e.g., contexts varying in the appropriateness for expressing the trait, situational press<sup>4</sup> [41,46,47]). They argued from the model, that to measure the desired trait ( $f_l$ ), it is necessary to set benchmark levels for the other influences, for example, setting common incentives,  $R_l^n = \bar{R}_l$ , and contexts,  $W_l^n = \bar{W}_l$ , across respondents. They pointed out, as we do here, that psychologists have been negligent in setting benchmark states, with the consequence of drawing inappropriate conclusions about the generalizability of trait measures across contexts and situations. Table 1 can be seen as a list of categories and examples of traits, incentives, and contexts that influence measured performance in situations ( $Y_l^n$ ), and the design and analysis columns represent some attempts that have been made to either set benchmark states (i.e.,  $\bar{R}_l$ ,  $\bar{W}_l$ ) or adjust for the lack of them afterwards.

From psychology, there are related frameworks for capturing the effects of testing contexts (occasions), incentives, and other influences on test scores. For example, generalizability theory [48,49] specifically identifies a universe score as an expected observed score over all observations in a universe of generalization, where the universe is defined through a set of fixed and random facets (e.g., across all raters and occasions, given some fixed incentives). Latent trait–state theory (LST) [50,51] specifically addresses the importance of latent states as well as traits on performance, by decomposing measurement error into separate latent trait and state residuals from situations and person–situation interactions. The framework also accommodates change over time.

Heckman and Kautz [12] provided a useful graphical depiction of their model, which emphasizes the point that any performance (e.g., test performance) will be a function of (a) abilities, (b) personality, and (c) motivation, and that motivation, in turn, will be a function of the incentives provided.<sup>5</sup> We generalize this idea slightly to propose a range of temporary and lasting influences on task performance, including both the target ability and other abilities, state effort (influenced by short-term incentives), the general tendency to exert effort, and situational press (see Figure 1). Although just one personality trait is listed (i.e., tendency to exert effort), others could also be included (e.g., trait anxiety). This diagram can be viewed as a simplification, reflecting causal directionality, but glossing over some issues, such as multilevel relationships and interactions between factors (e.g., type of short-term incentive  $\times$  personality) [53,54].

<sup>4</sup> Situational press refers to the reduction in trait variance due to situational constraints [40]. H. A. Murray [41] distinguished alpha (consensual, objective) and beta (subjective) press. Related concepts include situational strength [42], personality–situation congruence [43], situational construal [44], and moderated-mediation models [45], with situations as a moderator, and personality as a mediator.

<sup>5</sup> An anonymous reviewer pointed out that motivation itself may be multidimensional [52].



**Figure 1.** Sources of test score (task performance) variance (following Heckman and Kautz [12]).

### 5. Maximal-Typical vs. High-Stakes Low-Stakes Distinction

Personality traits and intelligence are normally conceptualized in different ways. Personality traits are often defined in terms of typicality—stable patterns of behavior over an extended period of time [55–57]. If person A frequently acts in an assertive, talkative manner across a wide variety of everyday situations, she would be considered more extraverted overall than person B, who is only moderately talkative and assertive on average. However, person B, if properly motivated, may be able to act in ways more extraverted than usual, and the upper limit of person B’s extraversion may even exceed person A’s, because of situational press. Conceptualizing and measuring personality traits by their maximal expression has occasionally been considered and attempted [58–63], but the vast majority of research and theorizing treats personality traits as the average expression of a person’s behavior [64]. Consequently, personality traits are usually treated as summaries of what individuals typically do [65].

Intelligence is implicitly (and sometimes explicitly) conceptualized and measured as what people are able to do and is defined as the limit of a person’s intellectual repertoire, which can be expressed when that person is exerting maximum effort [66,67]. Cognitive ability tests are often administered under high-stakes conditions (e.g., personnel selection, university admissions), which are presumed to induce individuals to be motivated to do as well as possible on those tests and, as a consequence, demonstrate their current degree of intelligence to its fullest extent.

Just as people are capable of expressing personality traits to greater (or lesser) extents than they ordinarily do, people are also not usually motivated to express the utmost limit of their intellectual skills on an everyday basis. Consequently, there is no guarantee that individuals will demonstrate the full extent of their intelligence across the situations they encounter in their daily lives (Ackerman, 2018). For example, a person with a Ph.D. in engineering (or in English literature) will be capable of solving highly complex mathematical problems (or writing a thought-provoking essay), but may not feel motivated to do so if those problems (or essay prompts) are presented under low-stakes conditions without adequate incentives. Although the extent to which individuals demonstrate their intelligence in everyday situations has been explicitly studied in terms of dispositions [68], typical intellectual

engagement [69] and through the application of ‘user-friendly’ cognitive tests [70], the majority of research and theory concerned with intelligence treats the construct as what people can maximally do intellectually [58,65,71,72].

Because intelligence is treated as the upper limit of individuals’ cognitive skills, it is amenable to being directly measured, as it is not necessary that people maintain the expression of this upper limit beyond a relatively short period of time (e.g., while taking a high-stakes standardized test). Intelligence tests can thus be conceptualized as *samples*—actual performances that directly demonstrate the construct [73,74].<sup>6</sup> In contrast, because personality traits are defined as typical behavior over a long period of time, there is a view that they cannot be directly measured in conventional assessment settings. Consequently, the most frequent method of personality assessment is self-report; people are asked to complete a questionnaire about themselves, with the idea that personality traits can be indirectly measured via individuals’ self-perceptions, which are partly based on their own observations about trends in their behavior over long periods. Consequently, self-report personality surveys are actually measures of self-concept [79,80] and thus *signs*—indirect indicators of the constructs of interest [73,74]. These self-reports capture some of the shared reality of people’s actual behavior, as correlations between self-reports of personality traits and observers’ reports of personality traits range from  $r = 0.29$  to  $r = 0.41$  [81].<sup>7</sup>

The meta-analytic correlation between typical and maximal performance in the workplace has been estimated to be  $r = 0.42$  [22]. Behavior on-the-job is influenced by both cognitive skills and personality (along with other constructs), and this meta-analytic correlation cannot be considered indicative of the relationship between maximal and typical behavior within either of those domains individually. Nonetheless, the relatively low correlations between behaviors in the same domain, carried out under different conditions, is intriguing and suggests the need for additional research examining the interrelations between maximal and typical expressions of intelligence and personality within and between the two domains.

## 6. Cognitive Test Performance under Low-Stakes Conditions

There is a corollary to the fact that many people do not enact the full extent of their cognitive skills on an everyday basis because they lack the incentive to do so: In testing situations where the stakes are low (e.g., laboratory experiments, nationally-sponsored learning assessments) many test-takers also lack the incentive to exert the effort necessary to perform as well as possible. Although the potential for high-stakes conditions to lead examinees to distort their responses to personality measures has been noted for decades [84] less attention has been made to the potential for low-stakes conditions to introduce construct-irrelevant variance into cognitive test scores (for exceptions see [72,85,86]). When intelligence tests are administered under high-stakes conditions, all individuals are expected to be maximally motivated and, as a consequence, cognitive ability is assumed to be the primary (and perhaps only) source of test score variance [58]. When performance on cognitive tests has little to no consequences for test-takers it is naïve to assume that all test-takers are exerting maximal effort and that subjects do not vary in whatever degree of effort they do put forth [87]: “A common assumption when studying human performance is that subjects are alert and optimally motivated. It is also assumed that the experimenter’s task at hand is by far the most important thing the subject has to do at that time. Thus, although individual differences in cognitive ability are assumed to exist, differences in motivation are ignored”.

---

<sup>6</sup> Although this often how high-stakes intelligence tests are conceptualized, some have argued that they are indirect, as the scores are only the results of unobserved cognitive processes [75–78].

<sup>7</sup> Self-reports of intelligence are also somewhat accurate, as a meta-analytic estimate of the correlation between self-estimates of intelligence and intelligence test scores was  $r = 0.33$  [82]. Similarly, self-reports of typical intellectual engagement and scores on intelligence measures have been found to correlate from  $r = 0.43$  to  $r = 0.50$  [83].

The implications of differences in motivation for the construct validity of intelligence tests administered under low-stakes conditions have occasionally been explored over the past 70 years [88–92] but research in this area has intensified in the last 20 years [27,93]. Some lines of contemporary investigation have sought to demonstrate the influence of effort on test performance when the stakes are low by experimentally inducing motivation. Means of inducing effort have varied across studies but included manipulating motivational frames (e.g., “scores will be made available to employers”; [26]), offering monetary incentives [94], publicly recognizing students for their test performance [95], and providing feedback about performance [96]. Other studies have used nonexperimental procedures to study effort, such as measuring motivation via self-report [97], observational coding [98], filtering out subjects with extreme response times [99], and using person-fit statistics to detect unusual response patterns [100]. The general conclusions from these lines of research is that effort matters: Two meta-analyses have estimated a mean performance difference of 0.59 to 0.64 standard deviations between motivated and unmotivated students [98,101].

Being dispositionally motivated to achieve is related to task persistence and engagement [102], strongly related to conscientiousness [103], and even treated as an element of conscientiousness in some personality taxonomies [104–106]. Taken together with findings that individuals differ in their motivation to do well on tests in the absence of adequate incentives [21], this suggests that variance in scores on assessments administered under low-stakes conditions can be attributed to both intelligence and personality. When scores on such tests are judged to be “pure” indicators of cognitive skills their construct validity is compromised, as personality contributes construct-irrelevant variance [107]. However, if variance in these scores is judged to be attributable to intelligence and personality their construct validity is considerably strengthened. Indeed, test scores under low-stakes conditions can be treated as partially being measures of personality.

That a substantial portion of the variance in cognitive test scores may be attributable to personality in low-stakes settings but not high-stakes settings implies that assessment conditions (and incentives) may be an important moderator of observed associations between personality and intelligence. For instance, given that conscientious people are more likely to exert effort in general, it might be expected that the correlation between conscientiousness and intelligence test scores will be higher in non-incentivized, low-stakes conditions than in high-stakes conditions. A cursory review of studies reporting correlations between conscientiousness and ACT/SAT scores, and conscientiousness and low-stakes test scores in different samples supports this hypothesis in a preliminary way. Richardson, Abraham, and Bond’s [108] meta-analysis reports a sample-weighted correlation of  $-0.05$  between conscientiousness and ACT/SAT, while Poropat’s [109] meta-analysis records a correlation of  $-0.03$ ; the sample-weighted correlation derived from Noftle and Robins’ [110] primary study is  $-0.04$ . These values contrast with correlations reported in some studies, where intelligence tests were administered under low-stakes conditions, such as  $0.29$  [111] and  $0.20$  [112].<sup>8</sup> Any attempt to understand the relationship between personality traits and intelligence must take into careful consideration the circumstances in which assessments were administered. Just as the relationship between intelligence scores and scores on a personality test administered for hiring should not be taken at face value, nor should the association between personality scores and scores on an intelligence test administered under low-stakes conditions.

---

<sup>8</sup> Other studies report much smaller relations between conscientiousness and scores on cognitive tests given under non-incentivized conditions [83,113]. Further complicating matters, a meta-analysis showed a correlation of  $0.14$  between achievement motivation and ACT/SAT [114]. Moreover, using ACT and SAT as the “gold standard” for high-stakes cognitive ability tests likely introduces complications due to range restriction and selection bias [115].

## 7. Personality Measured through Performance Tests

### 7.1. Objective Personality Tests

Cattell and Warburton [116] distinguished three kinds of personality assessments, namely questionnaires (Q-data), biographical data (L-data), and tests (T-data). As noted above, almost all the research in personality has been concerned with Q-data. But in his *Essentials of Psychological Testing*, Cronbach [7] devoted an entire chapter to 'Performance Tests of Personality', tracing the history back to the Character Education Inquiry [117], which included performance tests of honesty (failing to cheat when the opportunity presented itself) and persistence (reading and marking a string of letters that formed sentences). He also reviewed cognitive style tests, in-basket tests, leaderless groups, projective tests, and other methods, which have not had a serious impact on personality testing per se (although such measures are used in applied workforce personnel selection). A problem with many of the early efforts was insufficient reliability.

A more recent treatment was provided in a special issue on Objective Personality Tests in the *European Journal of Psychological Assessment* [118]. Ortner and Proyer [119] provided a comprehensive review of objective personality tests (OPTs), distinguishing between three kinds. One is OPTs masked as achievement tests. An example is the 'time pressure task', in which examinees use dragging and dropping to categorize letters. The time limit gradually decreases, and the score is based on whether the examinee's performance increases or decreases as the time limit goes down. James' [120] conditional reasoning test (CRT) is a very different measure, but it might also be considered a measure of this type. It presents five alternative multiple-choice reading comprehension problems with two correct answers. The two correct answers reflect different world views, which are presumed to be revealed by one's selection.

A second category is OPTs that represent real-life situations, particularly risk propensity. An example is the Balloon Analogue Risk Task (BART; [121]), in which the test taker gets more points for blowing up a balloon; the larger it gets the more points the test taker gets, until it pops, in which case all of the points are lost (test takers decide when to stop blowing up the balloon). Similar risk-taking tests have been made from decisions to cross a road [122,123].

The third category is questionnaire-type OPTs that ask for decisions. An example is one in which a problem is presented (e.g., "You are a mile away from the nearest station when the car breaks down. What would you do? If you know, make a checkmark"), which is scored for assertiveness/confidence as the latency to respond, regardless of the response.<sup>9</sup>

### 7.2. Grit Game

Alan, Boneva, and Ertac [20] evaluated the effects of a 10-session, after-hours educational intervention (referred to as a 'grit intervention'), designed to promote students' ability to set appropriate goals, and to attribute success and failure to effort rather than to factors outside their control (e.g., intelligence). They evaluated the intervention with a real effort mathematical task (a 'grit game'), which was to find pairs of numbers that add up to 100 from a grid; they are given a target number of pairs to find (three) and a time limit (1.5 min). They could choose between a more and less difficult task (varying in the grid size), where the more difficult task paid out more (four gifts for winning vs. one gift for winning; failing to achieve the goal resulted in no gifts in either task). The intervention was successful in that it led to students seeking the more challenging version of the task, apparently to accumulate skills, which in turn led to an increase ( $d = 0.28$ ) in performance on a standardized test. This study illustrates a couple of principles. One is that the relationship

<sup>9</sup> Another type of "objective measure" that could be included here are personality measures based on the behaviors individuals who are high or low on a trait report doing. An example is the Behavioral Indicators of Conscientiousness (BIC) measure [124]. However, these items end up being almost indistinguishable from typical Likert-rating personality items.

between personality (in this case, grit, or the tendency to select challenging goals and exert effort) and intelligence (performance on the math test) is not fixed, but it can be modified by an intervention focusing on beliefs about the importance of effort. Second, it is possible to measure a personality construct by way of a decision behavior related to a game-like task.

### 7.3. Coding Speed Test as a Measure of Personality

Segal [21] argued that what a test measures depends on stakes; high-stakes tests measure cognitive skills, but low-stakes (i.e., unincentived) tests can measure both personality (intrinsic motivation and the tendency to exert effort) and cognitive skills. A particularly good indicator of personality for a test would be one in which the knowledge requirements are minimal, reducing the confounding effects of knowledge on performance. She argued that the Armed Service Vocational Aptitude Battery's (ASVAB) coding speed test, which requires examinees to match common words with four-digit numbers by scanning a test form, satisfies the low knowledge requirement. The coding speed test (along with other ASVAB tests) was administered in the National Longitudinal Study of Youth (NLSY), with no incentives (therefore, low stakes) and it was found that scores on the test were correlated with earnings 23 years later, whether or not controlling (through regression) for the non-speeded portion of the ASVAB (which would measure cognitive ability), and also controlling for educational attainment, suggesting that it is personality and not cognitive ability component of coding speed that relates to workforce success. To buttress this claim, Segal also showed that (a) recruits, who take coding speed under incentivized conditions (military entrance), scored higher than NLSY participants, despite having less education; (b) in an experiment, about a third of the participants responded to incentives by increasing their performance (call this the unmotivated group, in that they need incentives to respond well), the others did not (call this the intrinsically motivated group, since they perform well regardless); (c) both groups (unmotivated and intrinsically motivated) had equal SAT scores; (d) there were more females in the intrinsically motivated group; and (e) males in the intrinsically motivated group had higher conscientiousness scores than the remaining males.

### 7.4. Economic Preference Games

Almlund, Duckworth, Heckman, and Kautz ([125], Table 6) proposed a set of tasks from behavioral economics research measuring time (delay discounting), risk (aversion), and social (leisure, altruism, trust, reciprocity) preferences. Such preferences are assumed to be fairly general, and lasting, therefore by our definition, can be thought of as personality factors. These have tended to correlate, but only weakly, with survey measures of the Big 5<sup>10</sup>. Big 5 items that would appear to reflect preferences include risk tolerance (e.g., "I take risks", "I avoid dangerous situations"), typically considered a facet of extraversion; time preferences (e.g., "I put off unpleasant tasks", "I avoid responsibilities", "I get chores done right away") typically considered a facet of conscientiousness; and social preferences (e.g., "I love to help others" and "I trust others") typically considered a facet of agreeableness.

In a series of studies, Falk and colleagues [127,128] developed a battery of tasks designed for surveys for measuring risk, time, and social preferences—specifically, risk aversion, future time discounting, trust, altruism, and positive and negative reciprocity—which they called the preference survey module. They implemented this module as part of the Gallup World Poll, which was administered to 80,000 individuals in 76 countries.

The module was developed as follows. First, an experimental measure for each of the six preferences was administered; these measures involved real money and payouts. For example, a risk-taking measure asked whether a respondent preferred a lottery or a safe option, with varying

---

<sup>10</sup> The Big 5 is a prominent dimensional analysis of personality, positing that five orthogonal factors—Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness—account for item responses on personality surveys and predict real-world outcomes [126].

amounts of money. A time preference measure asked whether a respondent preferred an immediate or delayed payout with varying amounts of money. A trust measure asked how much money a respondent would give to another in an investment game (i.e., anticipating that the other would return some of that money). A negative reciprocity measure determined what the minimal amount of money a respondent would accept from another would be, before rejecting the offer (in which case both receive no money, i.e., an ultimatum game).

Next, numerous survey items were administered to the same respondents, and regression analyses were used to select the two survey items that best predicted performance on the experimental measures. The survey items were of two types, qualitative and quantitative. Qualitative items were typical Likert style personality items, such as “Are you a person who is generally willing to take risks, or do you try to avoid taking risks?” (risk-taking), “How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?” (time preference), “As long as I am not convinced otherwise, I assume that people only have the best intentions” (trust), and “Are you a person who is generally willing to punish unfair behavior even if this is costly?” (negative reciprocity). The quantitative measures were typically survey versions of the experimental measure, which asked respondents what they would do in a situation. For example, a respondent would be given a set of choices of payments, they would prefer “today” vs. “in 12 months” (e.g., 100 dollars today vs. 120 dollars in 12 months) (time preference). Or they would be told “Suppose you won a lottery for \$1000, how much would you give to charity?” (altruism).

Falk et al. [128] estimated the correlations between performance on the experimental tasks, and a best composite of the two survey items, for each dimension. The correlations ranged from 0.38 (negative reciprocity) to 0.58 (time preference), which are reasonably high given that each was measured by only two survey items, and the test–retest correlations between the game tasks ranged from 0.35 to 0.67.<sup>11</sup> Administering the two-survey-item-per-dimension module as the Global Preference Survey (within the Gallup World Poll), resulted in many interpretable findings, such as women being more risk averse than men, with stronger social dispositions; risk taking being lower with age; cognitive skills correlating with time preference and risk taking; time preference (future time orientation) being related to educational attainment and savings; risk-taking being associated with smoking and self-employment; and social preferences being related to donating, volunteering, and helping others.

### 7.5. Confidence

In a series of studies, Stankov and colleagues [129–132] have argued that confidence is one of the most powerful non-cognitive predictors of academic achievement, as well as other outcomes. In their approach, confidence is measured following the completion of an item response, for example, after a vocabulary, mathematics, and cognitive reflection test [133], or a progressive matrices item. Then, a respondent is presented with a confidence scale (“How confident are you that your answer is correct? Choose one—0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100%”). Up to half the variance in performance is captured by this measure [130]. Several cycles of the Program for International Student Assessment (PISA) have also used a confidence measure (called ‘mathematics self-efficacy’) and similarly, it is among the highest correlates of achievement (average within country correlation,  $r = 0.43$ ; [134]). There is evidence that confidence, measured this way, is somewhat independent of the ability measured by the particular test item [135], which supports the hypothesis that there is a general, lasting trait of self-confidence, which can be measured in the context of specific cognitive ability test items. Although no study, to our knowledge, has directly compared a survey approach to measuring confidence (e.g., a Big 5 facet of extroversion), the achievement correlations given by the

<sup>11</sup> Using the standard correction for disattenuation,  $r_{x'y'} = r_{xy} / (r_{xx}r_{yy})^{1/2}$ , and numbers supplied in Falk et al. [128], we estimate the partly disattenuated correlation to be 0.53, 0.65, 0.77, 0.53, 0.71, 0.46, for risking taking, time, trust, altruism, positive and negative reciprocity, respectively, assuming perfect reliability of the composite.

experimental measure are higher than meta-analytic estimates of achievement correlations given by survey measures [109].

## 8. Personality Measured through Test and Survey Behavior

### 8.1. Survey Effort

The effort one puts forward in responding to a survey, both by returning it, and by completing all items on that survey, might be understood as an indicator of either a temporary or lasting characteristic of that individual [136–138]. Item response rates were examined by Hitt, Trivitt, and Cheng [139], who examined six large-scale longitudinal surveys of adolescents. They found significant relationships between item response rates and educational attainment (a one standard deviation [SD] increase in item response rate was associated with 0.11 to 0.33 additional years of education); this relationship held as significant in four of the six datasets, even after controlling for a number of other factors, including cognitive ability. The fact that cognitive ability, as measured by tests, attenuated the relationship could be seen as evidence that test scores already include variance associated with effort.

A question is whether lack of effort, indicated by skipping background questions, is related to conscientiousness or other Big 5 personality factors, measured by surveys. Zamarro, Cheng, Shakeel, and Hitt [140] examined behavior within the Understanding America Study (UAS), a 30 min panel survey of 6000 households in which respondents are paid \$20 per survey. Specifically, they examined item nonresponse rates and careless answering. Careless answering was defined as a kind of person misfit statistic, in which each item response on a scale was regressed on the average score from the remaining items on a scale, and a standardized residual was taken to represent misfit. Misfit was averaged across several scales to form a composite careless answering variable. They found significant but low correlations between nonresponse rates and careless answering with Big 5 variables. More importantly, they found that careless answering and item nonresponse independently were associated with educational attainment, more so than Big 5 personality traits measured by a survey. In addition, careless answering was associated with labor market outcomes (earnings), although item nonresponse was not.

Besides careless answering, returning the survey itself may be taken as an indication of effort. Cheng, Zamarro, and Orriens [141] examined return rates on the same Understanding America Study (UAS). They found that those returning surveys were more conscientious and less open, after controlling for a wide variety of demographic characteristics,<sup>12</sup> supporting the typical-effort interpretation of the conscientiousness factor.

### 8.2. Item Position Effects

It is commonly assumed that an item's difficulty (e.g., the percentage of test takers who get the item right) is not affected by whether that item is administered early or late in a test (this assumption is implicit in the expression "item difficulty", which is not conditioned on item position). That assumption is sometimes relaxed, in acknowledgement of warm-up and fatigue effects, which may increase item difficulty [142]. A common remedy is creating two or more test forms that vary item position, thereby averaging out the item position effects. Nevertheless, the existence of item position effects is a reflection of noncognitive influence (warming up, being mentally fatigued) on a cognitive test score.

One question concerns the severity of item position effects. This seems to vary depending on circumstances, but there certainly is considerable evidence for them. Albano [143] found that items in the middle of a first-grade reading achievement test were 8% more difficult than they were at the beginning of the test ( $N = 93,000+$ ). He found similar effects for the Graduate Record Exam (GRE)

---

<sup>12</sup> Survey returners also were higher cognitive ability, more likely female, native born, not employed, and African American, controlling for many background factors.

( $N = 5000+$ ). In both studies there was item heterogeneity in the sense that items varied in their susceptibility to position effects; on the GRE the range was from a proportion correct decreasing by 0.17 to increasing by 0.03 in early vs. late item positions.

Another question is whether there are group or individual differences in the susceptibility to item position effects. Debeer, Buchholz, Hartig, and Janssen [144] (see also [31,145]) examined data from the low-stakes Program for International Student Assessment (PISA) 2009 reading assessment ( $N = 460,000+$ , 65 countries), using an item-response theory (IRT) model that included item position effects, which enabled modeling test-taker effort (i.e., less susceptibility to item position is assumed to indicate more [consistent] effort). They found both a general decrease in effort across countries, and large individual, school-level, and country-level differences in the decrease of examinee effort over the course of the test. The amount of decrease was associated with the overall performance level. For example, students from Finland, which is a high (average) performing country, showed a relatively small decrease in effort ( $d = -0.09$ ), whereas students from Greece, which is a lower performing country, on average, showed a larger decrease in effort over the course of the test ( $d = -0.28$ ). There were also school effects on persistence, supporting a positive relationship between persistence and ability (schools with higher ability students also are ones with higher-average-persistence students). The large lesson here is that differences in PISA scores between countries at least partly reflect noncognitive (persistence) differences between students in those countries. This result was replicated at the student level in a large-scale German achievement study [146], using a survey approach to measure effort as a predictor of change in item difficulty during the test.

### 8.3. Response Time

Another indicator of effort on a cognitive ability test, particularly one given under low-stakes conditions, is response time. The idea is that if examinees are responding quickly, for example, in less time than it takes to read the question, they are not putting forward adequate effort for solving the problem. Wise and Kong [99] proposed a measure of response-time-effort (RTE) indicated by the proportion of items for which the examinee takes adequate time to respond, that is, more time to respond than a low threshold (which can be determined in various ways [147], but typically below a second or two, depending on the task). This measure has been shown to relate to test performance as well as other outcomes [148]. Students who display less response-time-effort tend to do more poorly on the assessment. Lee and Jia [149] developed a response-time-effort method to investigate student effort on the National Assessment of Educational Progress (NAEP).

Wise and Gao [150] proposed a broader measure of test taking effort on computer tests, which they refer to as response behavior effort (RBE). In addition to rapid guesses (the response measured with RTE), they proposed rapid omits and rapid perfunctory answers on constructed response items. In all of these cases, they used a threshold of 10% of the average time test takers spent on the item, or 10 s (whatever was lower) as the threshold to defined rapid guessing, omitting, or perfunctory answering. They applied this method to the Office of Economic Cooperation and Development's (OECD) 2013 PISA-Based Test for Schools (PBSTS), and found that about 5% of all of the items showed a RBE value of less than 0.90, due to rapid guesses (71% of non-effortful responses), rapid omits (19%), and rapid perfunctory answers (10%). The highest achievers (by quartile) displayed the highest behavior effort (75,157 solution behavior responses/75,216 responses total = 99.9%; vs. the lowest quartile at 95.3%, computed from information in their [150] Table 3). However, it was also true that for those items in which the test taker did display solution behavior, there still was a difference of 27% vs. 72% correct for the lowest vs. highest quartile.

## 9. Personality Measured through Real World Behavior

### 9.1. Study Time

How students choose to spend their out-of-school time can be thought of as a noncognitive factor that is integral with achievement, as indicated by test scores. Time spent on homework for example, is likely to boost achievement [151]. McMullen [152] estimated that one additional hour of homework per week translated to an improvement in mathematical achievement by 0.24 standard deviations, and that this was even higher for low performing students and schools. Being randomly assigned into a homework required (vs. not-required) group was found to boost test scores, grades, and retention [153]. However, the issue here is whether choosing to do homework could be seen as a personality factor (tendency to put forth effort). A study using time-diary data found that an extra hour of homework per night increased the probability of attending college by 5 percentage points (for males) [154]. However, the authors suggested that this homework effect may be due to an omitted variable (e.g., motivation), based on an instrumental variable analysis in which day of the week (e.g., surveyed on a Friday vs. not) and season (e.g., football season or not) were treated as instruments.

### 9.2. Registration Latency

Richardson, Abraham, and Bond [108] identified procrastination, measured by survey items (e.g., “I generally delay before starting on work I have to do”), as among the highest noncognitive correlates of college success ( $\rho = -0.25$ ), as defined by the grade-point average. Novarese and di Giovanni [155] examined a performance measure of procrastination, registration latency for college (law school in Italy), which is defined as the time between when a student was first eligible to enroll and when the student actually did enroll (before the deadline, near the deadline, or after the deadline, which involved paying a late fee). They found that late registering students were more likely not to complete the first year, less likely to graduate, had poorer performance, passed fewer exams, and received fewer credits. They found that this same pattern held for late registration in years two and three. A question is whether this kind of procrastination is a temporary, one-off characteristic, perhaps due to circumstances, or a more lasting one. They found some evidence pointing to it being a lasting characteristic; the correlation between procrastination from year to year (for the first five years, ignoring the first year, which was a bit of an outlier), ranged from 0.33 to 0.47, suggesting a relatively stable indicator. Interestingly, promptness averaged over years two and three, which was found to be correlated to a performance measures (number of exams passed,  $r = 0.42$ ), was only modestly related to a self-report prompt, “I procrastinate” with a 0–10 response scale ( $r = -0.22$ ), which itself had a lower correlation with the number of exams passed.

### 9.3. Word Use, Office Appearance, and Facebook Likes as Personality Measures

That personality does not have to be measured by Likert scale self-reports has been explored in various studies. Fast and Funder [156] examined the words used in one hour life history interviews, and found a number of moderately high, interpretable correlations between particular kinds of words used (e.g., certainty words, such as ‘absolutely’), and responses on a personality measure (e.g., “is facially and/or gesturally expressive”, “is verbally fluent”). Gosling, Ko, Mannarelli, and Morris [157] had observers view people’s offices and workspaces (when the occupants were not there), and then completed personality ratings of the offices’ occupants. A separate group of coders coded 43 offices’ features, such as their neatness. A number of suggestive correlations were found, such as conscientiousness being related to neatness, openness related to distinctiveness and unconventionalness, and so on. A similar analysis, with similar findings, was conducted on bedrooms [156].

Kosinski, Stillwell, and Graepel [158] explored the relationship between Facebook likes and various characteristics of Facebook users. They reduced a binary matrix of 55,000 Facebook likes (1 = the user indicated a ‘like’ for photos, friends’ status updates, sports, books, web sites) to 100 components using singular value decomposition, then used the resulting components to predict

a variety of user characteristics, such as age, gender, personality, intelligence, relationship status, political views, and religion, using linear or logistic regression analysis (they obtained personality and intelligence through a special app). They found that for the continuous variables, age ( $r = 0.75$ ), size of Facebook friendship network ( $r = 0.47$ ), openness ( $r = 0.43$ ), extraversion ( $r = 0.40$ ), and intelligence ( $r = 0.39$ ), were fairly well predicted; for the categorical variables, gender (AUC = 0.93), sexual orientation (AUC = 0.88, 0.75), political party (AUC = 0.85), and race (AUC = 0.96) were well predicted, with other outcomes (drug, alcohol, cigarette use, relationship status), moderately predicted.

Youyou, Kosinski, and Stillwell [159] followed up this study with the administration of a longer personality (self-ratings) survey to 86,000 Facebook users who completed the survey on themselves and on several peers. They found that Facebook likes correlated more highly with personality self-reports ( $r = 0.56$ ) than others' ratings did ( $r = 0.49$ ). Further, they found that Facebook likes correlated more highly with a variety of life outcomes (e.g., substance use, political attitudes, health) than self-ratings did.

## 10. Ability Effects on Personality Measures

To this point, we have focused on the influences of personality state and trait variables on cognitive test score performance, to make the point that cognitive test scores cannot be unambiguously attributed to cognitive skills. Instead, they represent a mixture of cognitive and noncognitive influences. In this section, we address the opposite question, "To what extent do cognitive factors influence responses to personality assessments?"

### 10.1. Age Effects

There is an established literature on personality change, suggesting that personality generally gets better with age. That is, based on rating scale responses, from young to later adulthood (ages 21 to 60), conscientiousness and agreeableness increase over time and neuroticism declines [160]. However, during the earlier years, from age 10 to 20, as young adults grow in sophistication in understanding language and human nature, the picture is more complex. Responses to rating scale items become more reliable from 12 to 18 years old [161], the Big 5 factor structure becomes more differentiated and closer to the adult structure over that period<sup>13</sup> [162], and below age 13 (from age 10), the alignment of items to their appropriate factors (with respect to the adult structure) deteriorates substantially. In general, coherence (mean inter-item correlation of items measuring a factor) increases substantially from age 10 to 20, and differentiation (mean inter-scale correlations, controlling for unreliability) goes down. This pattern is exactly what would be expected if cognitive ability played a role in responding to personality rating-scale items. In this way, personality scores partly reflect cognitive ability differences.

### 10.2. Cognitive Ability Effects

The differentiation hypothesis of cognitive ability is based on the idea that the role of the general factor diminishes during development (or with increases in ability), dominating early childhood (or, at low ability levels), but becoming increasingly less important with development (or at high ability levels), as the role of specialized abilities (e.g., verbal vs. spatial) become relatively more important. The support for this hypothesis is mixed [163], but a question here is whether there is a similar differentiation in personality. An interpretation of age or ability-related personality differentiation is that a lack of differentiation could be due to a person's lack of cognitive ability to comprehend personality descriptions and properly differentiate levels of agreement with those descriptions.

Möttus, Allik, and Pullman [164] found that out of 35 personality scales (Big 5 plus facets), reliability was significantly higher for high ability groups than for low ability groups for seven of

<sup>13</sup> This is true regardless of whether acquiescence is controlled for or not; acquiescence ("Yea-saying") can be controlled by within-person standardization, or ipsatizing.

those scales, and nominally higher for 30 of the 35 scales<sup>14</sup>. Correlations between scales were also higher for the low ability group than for the high ability group. Of the 10 Big 5 intercorrelations, 8 were nominally higher in the low ability group; this was also seen in the size of the first principal components (22% vs. 27% of the variance for high vs. low ability groups, respectively). Based on findings from a similar study, Allik, Laidra, Realo, and Pullman [165] concluded that some younger children lacked the “developed abilities required for observing one’s own personality dispositions and for giving reliable self-reports on the basis of these observations”. It is certainly possible and desirable to reduce the complexity of items in order to reduce the effects of cognitive ability [166], but our point here is simply that, in general, cognitive ability plays a role in responding to personality items.

In another manifestation of the importance of ability in responding to personality surveys, adults are quite capable of ‘faking’ responses to personality scales (particularly, Likert type scales) to present a favorable impression. The degree to which they are able to do so is related to cognitive ability [167].

### 10.3. *Faking on Personality Tests*

The typical rating scale format of a personality test enables ‘faking’. That is, to convey a positive image of oneself, it is possible to “strongly agree” with positive statements (e.g., “I work hard”) and “strongly disagree” with negative ones (e.g., “I am lazy”), regardless of one’s personality. In fact, there is mixed evidence on the extent to which respondents actually do this, some suggesting respondents do not often fake [168], others suggesting they do [169,170].<sup>15</sup> But for our purposes here, we focus on a category of studies in which respondents are asked to “fake good”. Being able to fake good indicates a sophistication about how responses are interpreted by potential decision-makers (e.g., hiring authorities, admissions committees). If there is differential ability to fake good, and that ability is related to cognitive ability, then that suggests that cognitive ability can have a direct influence on the responses to personality tests. A meta-analysis of studies that instructed respondents to fake good (or bad) suggests that indeed respondents are quite capable of doing so [172]. Effect size estimates for the Big 5 factors ranged from 0.48 (with Agreeableness) to 0.65 (with Openness), in between-subjects designs, slightly larger in within-subjects designs (instructions to fake bad resulted in effect sizes two to three times greater, in the other direction). However, respondents high in cognitive ability are particularly able to fake good, by roughly one half of a standard deviation, compared to low cognitive ability respondents [173]. This suggests that personality tests at least partially measure cognitive ability, depending on conditions, particularly incentives, and according to Griffith and Converse [170], personality tests do so.

### 10.4. *Anchoring Vignettes as a Window into Psychological Understanding*

The anchoring vignettes technique is a method for increasing comparability between respondents on rating scale measures, by having respondents rate both themselves and others, described in vignettes, on the same items [174]. Anchoring vignettes were included in PISA 2012 to address the problem of response style differences between countries in international comparisons [175]. The key for including anchoring vignettes as part of the discussion on the role of cognitive ability in personality testing is related to the task of rating vignettes. In PISA 2012, two sets of vignettes were included, one concerned with the dimension of classroom management, one with teacher support. Each had vignettes designed to be either low, medium, or high on the targeted dimension. For example, the high, medium, and low teacher support vignettes were, (a) “Ms. <a> sets mathematics homework every other day. She always gets the answers back to students before examinations”; (b) “Mr. <b> sets mathematics

<sup>14</sup> The authors argued that although there were differences as we note here, they did not support the conclusion that “personality differs substantially across ability groups” ([163], p. 155). However, our argument is simply that cognitive ability difference contributed to differences in personality responses, which is what the authors found.

<sup>15</sup> Most of this literature is based on a rating-scale response format. Forced-choice methods appear to limit faking susceptibility [171].

homework once a week. He always gets the answers back to students before examinations"; and (c) "Ms. <c> sets mathematics homework once a week. She never gets the answers back to students before examinations", respectively. Following each vignette, students were asked how much they agree with the statement, "Mr./Ms. <x> is concerned about his/her students' learning". They could answer "strongly agree", "agree", "disagree", or "strongly disagree"? Typically, students' responses to the vignettes align with the vignettes intended trait location, so that they are more likely to agree that the high vignette teacher is "concerned about students' learning" than the low vignette teacher is. But student responses are not always aligned with the intended vignette location. In fact, cognitive ability (whether measured by mathematics, reading, or problem solving) was (negatively) associated with either assigning two vignettes the same rating ('ties') or rating the intended higher vignette lower than the intended lower vignettes ('misorderings'), with effects sizes ranging from about a half a standard deviation to about 0.8 standard deviations (see [176], Tables 10 and 11, pp. 29–30;  $k = 52$  countries,  $N \sim 250,000$ ).

## 11. Discussion

This article is a contribution to a special issue of the *Journal of Intelligence* on the integration of personality and intelligence, which invited contributions to "bring these two traditions (personality and intelligence) back to the discussion table and to underscore the relevance of an integrative perspective for both individual differences and developmental research" [177]. A central theme of this article is that, wittingly or unwittingly, intelligence researchers are already studying personality, and personality researchers are studying intelligence.

Another theme is that, while construct and method are typically confounded—intelligence is measured with tests and personality is measured with Likert scale self-reports—in principle, they are separable, and failing to acknowledge the construct-method distinction results in dubious conclusions, such as the highest personality correlate of intelligence is the openness/intellect factor. That conclusion could nearly as justifiably be restated as the method effect for measuring intelligence is observed directly as the correlation between the openness/intellect factor (i.e., self-reports of one's intelligence) and intelligence test scores.

We have several suggestions for how personality and intelligence research and researchers can move forward together. First, we believe that the constructs of intelligence and personality are viable. Carroll's [66] cognitive ability taxonomy, the Big 5 framework [160], and economic preferences [125], should be thought of as useful delineations of skills that people develop with schooling and experience, and apply when making decisions and acting on them. The view of personality as a skill strikes some psychologists as odd, but it has been embraced by policy makers. Consider the title of an op-ed in the *New York Times* by a prominent U.S. Senator, "We need immigrants with skills. But working hard is a skill" [178]. The mistake is to conflate abilities, personality factors, and preferences with the methods used to measure them. Researchers should acknowledge, or attempt to control for ancillary sources of variance in these measures.

Second, as psychology and measurement psychology in particular has been recognized for a long time, it is useful to include multiple measures of a construct to unconfound the construct from its measurement. For new surveys, this might include supplementing Likert scale measures with additional behavioral, or 'objective' measures. And fortunately, as the literature review above shows, there may already exist multiple measures of constructs in extant survey datasets. For example, large-scale achievement surveys, such as NAEP and PISA, have been and can continue to be analyzed for indications of personality traits, such as tendency to put forth effort, by analyzing achievement item response times, item position effects, and the like. There are a large number of potential datasets that could be mined for personality indicators other than Likert scale measures.

Finally, we suggest expanding the editors' call to bring personality and intelligence researchers back to the discussion table to include behavioral, labor, and education economists. As we hope this review demonstrates, economists have already made significant contributions to our understanding of

the integration of personality and intelligence. There are the beginnings of integration with occasional special conferences [179], National Academies reports [180], and papers published in economics journals and handbooks [125]. Collaborations published in psychology journals, such as the *Journal of Intelligence* would be useful moving forward.

**Author Contributions:** Both authors (P.C.K., and H.K.) contributed to Conceptualization, Writing-Original Draft Preparation; Writing-Review & Editing, and Funding Acquisition.

**Funding:** This research was funded through Educational Testing Services' Research Allocation funding.

**Acknowledgments:** We thank two anonymous reviewers for helpful comments.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the writing of the manuscript, nor in the decision to publish the review.

## References

1. Ackerman, P.L. A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence* **1996**, *22*, 229–259. [[CrossRef](#)]
2. Ackerman, P.L.; Heggestad, E.D. Intelligence, personality, and interests: Evidence for overlapping traits. *Psychol. Bull.* **1997**, *121*, 219–245. [[CrossRef](#)] [[PubMed](#)]
3. Ziegler, M.; Danay, E.; Heene, M.; Asendorpf, J.B.; Buehner, M. Openness, fluid intelligence, and crystallized intelligence. *J. Res. Personal.* **2012**, *46*, 173–183. [[CrossRef](#)]
4. Ackerman, P.L. The search for personality–intelligence relations: Methodological and conceptual Issues. *J. Intell.* **2018**, *6*, 2. [[CrossRef](#)]
5. Rammstedt, B.; Lechner, C.M.; Danner, D. Relationships between personality and cognitive ability: A facet-level analysis. *J. Intell.* **2018**, *6*, 28. [[CrossRef](#)]
6. Stankov, L. Low correlations between intelligence and big five personality traits: Need to broaden the domain of personality. *J. Intell.* **2018**, *6*, 26. [[CrossRef](#)]
7. Cronbach, L.J. *Essentials of Psychological Testing*, 3rd ed.; Harper & Row: New York, NY, USA, 1970.
8. Thorndike, R.L. *Personnel Selection: Test and Measurement Techniques*; John Wiley & Sons, Inc.: New York, NY, USA, 1949.
9. Göllner, R.; Roberts, B.W.; Damian, R.I.; Lüdtke, O.; Jonkmann, K.; Trautwein, U. Whose “storm and stress” is it? Parent and child reports of personality development in the transition to early adolescence. *J. Personal.* **2016**, *85*, 376–387. [[CrossRef](#)] [[PubMed](#)]
10. Roberts, B.W.; Walton, K.E.; Viechtbauer, W. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychol. Bull.* **2006**, *132*, 1–25. [[CrossRef](#)] [[PubMed](#)]
11. Roberts, B.W.; Luo, J.; Briley, D.A.; Chow, P.I.; Su, R.; Hill, P.L. A systematic review of personality trait change through intervention. *Psychol. Bull.* **2017**, *143*, 117–141. [[CrossRef](#)] [[PubMed](#)]
12. Heckman, J.J.; Kautz, T. Fostering and measuring skills interventions that improve character and cognition. In *The Myth of Achievement Tests: The GED and the Role of Character in American Life*; Heckman, J.J., Humphries, J.E., Kautz, T., Eds.; The University of Chicago Press: Chicago, IL, USA, 2014; pp. 341–430.
13. Roberts, B.W. Back to the future: Personality and assessment and personality development. *J. Res. Personal.* **2009**, *43*, 137–145. [[CrossRef](#)] [[PubMed](#)]
14. Neisser, U.; Boodoo, G.; Bouchard, T.J.; Boykin, A.W.; Brody, N.; Ceci, S.; Halpern, D.; Loehlin, J.C.; Perloff, R.; Sternberg, R.J.; et al. Intelligence: Knowns and unknowns. *Am. Psychol.* **1996**, *51*, 77–101. [[CrossRef](#)]
15. Likert, R. A technique for measurement of attitudes. *Arch. Psychol.* **1932**, *140*, 5–55.
16. Rammstedt, B.; Danner, D.; Martin, S. The association between personality and cognitive ability: Going beyond simple effects. *J. Res. Personal.* **2016**, *62*, 39–44. [[CrossRef](#)]
17. Boring, E. *Intelligence as the Tests Test it*; New Republic: Washington, DC, USA, 1923; pp. 35–37.
18. Stanley, J.C. Reliability. In *Educational Measurement*, 2nd ed.; Thorndike, R.L., Ed.; American Council on Education: Washington, DC, USA, 1971; pp. 356–442.
19. Chaplin, W.F.; John, O.P.; Goldberg, L.R. Conceptions of states and traits: Dimensional attributes with ideas as prototypes. *J. Personal. Soc. Psychol.* **1988**, *54*, 541–557. [[CrossRef](#)]

20. Alan, S.; Boneva, T.; Ertac, S. Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. HCEO Working Paper. 2016. Available online: <https://ssrn.com/abstract=2761390> (accessed on 13 April 2016). [CrossRef]
21. Segal, C. Working when no one is watching: Motivation, test scores, and economic success. *Manag. Sci.* **2012**, *58*, 1438–1457. [CrossRef]
22. Beus, J.M.; Whitman, D.S. The relationship between typical and maximum performance: A meta-analytic examination. *J. Hum. Perform.* **2012**, *25*, 355–376. [CrossRef]
23. Sackett, P.R.; Zedeck, S.; Fogli, L. Relations between measures of typical and maximum job performance. *J. Appl. Psychol.* **1988**, *73*, 482–486. [CrossRef]
24. Hembree, R. Correlates, causes, effects, and treatment of test anxiety. *Rev. Educ. Res.* **1988**, *58*, 47–77. [CrossRef]
25. Van der Linden, W.J. Setting time limits on tests. *Appl. Psychol. Meas.* **2011**, *35*, 183–199. [CrossRef]
26. Liu, O.L.; Bridgeman, B.; Adler, R.M. Measuring learning outcomes in higher education: Motivation matters. *Educ. Res.* **2012**, *41*, 352–362. [CrossRef]
27. Finn, B. Measuring motivation in low-stakes assessment. *ETS Res. Rep. Ser.* **2015**, *2015*, 1–17. [CrossRef]
28. Hembree, R. The Nature, Effects, and Relief of Mathematics Anxiety. *J. Res. Math. Educ.* **1990**, *21*, 33–46. [CrossRef]
29. Steele, C.M.; Aronson, J. Stereotype threat and the intellectual test performance of African-Americans. *J. Personal. Soc. Psychol.* **1995**, *69*, 797–811. [CrossRef]
30. Beilock, S.L. *Choke: What the Secrets of the Brain Reveal about Getting It Right When You Have to*; Simon & Schuster: New York, NY, USA; Free Press: New York, NY, USA, 2010.
31. Debeer, D.; Janssen, R. Modeling Item-Position Effects within an IRT Framework. *J. Educ. Meas.* **2013**, *50*, 164–185. [CrossRef]
32. Rabbit, P. Error and error correction in choice-response tasks. *J. Exp. Psychol.* **1966**, *71*, 264–272. [CrossRef]
33. Mueller, C.M.; Dweck, C.S. Intelligence praise can undermine motivation and performance. *J. Personal. Soc. Psychol.* **1998**, *75*, 33–52. [CrossRef]
34. Kane, M.J.; McVay, J.C. What Mind Wandering Reveals About Executive-Control Abilities and Failures. *Curr. Dir. Psychol. Sci.* **2012**, *21*, 348–354. [CrossRef]
35. Terhune, D.B.; Croucher, M.; Marcusson-Clavertz, D.; Macdonald, J.S.P. Time contracts and temporal precision declines when the mind wanders. *J. Exp. Psychol.* **2017**, *43*, 1864–1871. [CrossRef] [PubMed]
36. Powers, D.E.; Rock, D.A. Effects of coaching on SAT I: Reasoning Test scores. *J. Educ. Meas.* **1999**, *36*, 93–118. [CrossRef]
37. Irvine, S.H. *Computerised Test Generation for Cross-National Military Recruitment: A Handbook*; IOS Press: Amsterdam, The Netherlands, 2014.
38. Eich, E. Searching for mood dependent memory. *Psychol. Sci.* **1995**, *6*, 67–75. [CrossRef]
39. Borghans, L.; Duckworth, A.L.; Heckman, J.J.; Weel, B.T. The economics and psychology of personality traits. *J. Hum. Resour.* **2008**, *43*, 972–1059.
40. Ackerman, P.L.; Chamorro-Premuzic, T.; Furnham, A. Trait complexes and academic achievement: Old and new ways of examining personality in educational contexts. *Br. J. Educ. Psychol.* **2011**, *81*, 27–40. [CrossRef] [PubMed]
41. Murray, H.A. *Explorations in Personality*; Oxford University Press: Oxford, UK, 1938.
42. Mischel, W. On the future of personality measurement. *Am. Psychol.* **1977**, *32*, 246–254. [CrossRef]
43. Sherman, R.A.; Nave, C.S.; Funder, D.C. Properties of persons and situations related to overall and distinctive personality-behavior congruence. *J. Res. Personal.* **2012**, *46*, 87–101. [CrossRef]
44. Sherman, R.A.; Nave, C.S.; Funder, D.C. Situational construal is related to personality and gender. *J. Res. Personal.* **2013**, *47*, 1–14. [CrossRef]
45. Ng, K.-Y.; Ang, S.; Chan, K.-Y. Personality and leadership effectiveness: A moderated mediation model of leadership self-efficacy, job demands, and job autonomy. *J. Appl. Psychol.* **2008**, *93*, 733–743. [CrossRef] [PubMed]
46. Kelly, R.T.; Rawson, H.E.; Terry, R.L. Interaction effects of achievement need and situational press on performance. *J. Soc. Psychol.* **1973**, *89*, 141–145. [CrossRef] [PubMed]
47. Sherman, R.A.; Nave, C.S.; Funder, D.C. Situational similarity and personality predict behavioral consistency. *J. Personal. Soc. Psychol.* **2010**, *99*, 330–343. [CrossRef] [PubMed]

48. Brennan, R.L. *Generalizability Theory*; Springer: New York, NY, USA, 2001.
49. Cronbach, L.J.; Gleser, G.C.; Nanda, H.; Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*; Wiley: New York, NY, USA, 1972.
50. Steyer, R.; Mayer, A.; Geiser, C.; Cole, D.A. A theory of states and traits—Revised. *Ann. Rev. Clin. Psychol.* **2015**, *11*, 71–98. [[CrossRef](#)] [[PubMed](#)]
51. Steyer, R.; Schmitt, M.; Eid, M. Latent state-trait theory and research in personality and individual differences. *Eur. J. Personal.* **1999**, *13*, 389–408. [[CrossRef](#)]
52. Vansteenkiste, M.; Sierens, E.; Soenens, B.; Luyckx, K.; Lens, W. Motivational profiles from a self-determination perspective: The quality of motivation matters. *J. Educ. Psychol.* **2009**, *101*, 671–688. [[CrossRef](#)]
53. Curran, P.J.; Bauer, D.J. Building path diagrams for multilevel models. *Psychol. Methods* **2007**, *12*, 283–297. [[CrossRef](#)] [[PubMed](#)]
54. McArdle, J.J.; Boker, S.M. *RAMpath: Automatic Path Diagram Software*; Erlbaum: Hillsdale, NJ, USA, 1991.
55. Stagner, R. *Psychology of Personality*, 3rd ed.; McGraw-Hill Book Company, Inc.: New York, NY, USA, 1961.
56. Matsumoto, D. (Ed.) *The Cambridge Dictionary of Psychology*; Cambridge University Press: Cambridge, UK, 2009.
57. VandenBos, G.R. (Ed.) *American Psychological Association Dictionary of Psychology*, 2nd ed.; American Psychological Association: Washington, DC, USA, 2015.
58. Fiske, D.W.; Butler, J.M. The experimental conditions for measuring individual differences. *Educ. Psychol. Meas.* **1963**, *23*, 249–266. [[CrossRef](#)]
59. Klesges, R.C.; Mcginley, H.; Jurkovic, G.J.; Morgan, T.J. The predictive validity of typical and maximal personality measures in self-reports and peer reports. *Bull. Psychon. Soc.* **1979**, *13*, 401–404. [[CrossRef](#)]
60. Paulhus, D.L.; Martin, C.L. The structure of personality capabilities. *J. Personal. Soc. Psychol.* **1987**, *52*, 354–365. [[CrossRef](#)]
61. Turner, R.G. Consistency, self-consciousness, and the predictive validity of typical and maximal personality measures. *J. Res. Personal.* **1978**, *12*, 117–132. [[CrossRef](#)]
62. Wallace, J. An abilities conception of personality: Some implications for personality measurement. *Am. Psychol.* **1966**, *21*, 132–138. [[CrossRef](#)]
63. Willerman, L.; Turner, R.G.; Peterson, M. A comparison of the predictive validity of typical and maximal personality measures. *J. Res. Personal.* **1976**, *10*, 482–492. [[CrossRef](#)]
64. Craik, K.H. Accentuated, revealed, and quotidian personalities. *Psychol. Inq.* **1993**, *4*, 278–280. [[CrossRef](#)]
65. Cronbach, L.J. *Essentials of Psychological Testing*, 5th ed.; Harper & Row: New York, NY, USA, 1990.
66. Carroll, J.B. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*; Cambridge University Press: Cambridge, UK, 1993.
67. Humphreys, L.G. Intelligence from the standpoint of a (pragmatic) behaviorist. *Psychol. Inq.* **1994**, *5*, 179–192. [[CrossRef](#)]
68. Perkins, D.; Tishman, S.; Ritchhart, R.; Donis, K.; Andrade, A. Intelligence in the wild: A dispositional view of intellectual traits. *Educ. Psychol. Rev.* **2000**, *12*, 269–293. [[CrossRef](#)]
69. Goff, M.; Ackerman, P.L. Personality-intelligence relations: Assessment of typical intellectual engagement. *J. Educ. Psychol.* **1992**, *84*, 537–552. [[CrossRef](#)]
70. Dennis, M.J.; Sternberg, R.J.; Beatty, P. The construction of “user-friendly” tests of cognitive functioning: A synthesis of maximal-and typical-performance measurement philosophies. *Intelligence* **2000**, *28*, 193–211. [[CrossRef](#)]
71. Ackerman, P.L. Adult intelligence: The construct and the criterion problem. *Perspect. Psychol. Sci.* **2017**, *12*, 987–998. [[CrossRef](#)] [[PubMed](#)]
72. Cattell, R.B. Personality traits associated with abilities. II: With verbal and mathematical abilities. *J. Educ. Psychol.* **1945**, *36*, 475–486. [[CrossRef](#)] [[PubMed](#)]
73. Goodenough, F.L. *Mental Testing*; Holt, Rinehart, & Winston: New York, NY, USA, 1949.
74. Wernimont, P.F.; Campbell, J.P. Signs, samples, and criteria. *J. Appl. Psychol.* **1968**, *52*, 372–376. [[CrossRef](#)] [[PubMed](#)]
75. Aftanas, M.S. Theories, models, and standard systems of measurement. *Appl. Psychol. Meas.* **1988**, *12*, 325–338. [[CrossRef](#)]
76. Hunt, E. On the nature of intelligence. *Science* **1983**, *219*, 141–146. [[CrossRef](#)] [[PubMed](#)]

77. Hunt, E. Science, technology, and intelligence. In *The Influence of Cognitive Psychology on Testing*; Ronning, R.R., Glover, J.A., Conoley, J.C., Witt, J.C., Eds.; Erlbaum: Hillsdale, NJ, USA, 1987; pp. 11–40.
78. Reeve, C.L.; Scherbaum, C.; Goldstein, H. Manifestations of intelligence: Expanding the measurement space to reconsider specific cognitive abilities. *Hum. Resour. Manag. Rev.* **2015**, *25*, 28–37. [[CrossRef](#)]
79. Cramer, A.O.; Sluis, S.; Noordhof, A.; Wichers, M.; Geschwind, N.; Aggen, S.H.; Kendler, S.H.; Borsboom, D. Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *Eur. J. Personal.* **2012**, *26*, 414–431. [[CrossRef](#)]
80. Möttus, R.; Kandler, C.; Bleidorn, W.; Riemann, R.; McCrae, R.R. Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *J. Personal. Soc. Psychol.* **2017**, *112*, 474–490. [[CrossRef](#)] [[PubMed](#)]
81. Connelly, B.S.; Ones, D.S. Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychol. Bull.* **2010**, *136*, 1092–1122. [[CrossRef](#)] [[PubMed](#)]
82. Freund, P.A.; Kasten, N. How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychol. Bull.* **2012**, *138*, 296–321. [[CrossRef](#)] [[PubMed](#)]
83. Furnham, A.; Monsen, J.; Ahmetoglu, G. Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *Br. J. Educ. Psychol.* **2009**, *79*, 769–782. [[CrossRef](#)] [[PubMed](#)]
84. Ellis, A. The validity of personality questionnaires. *Psychol. Bull.* **1946**, *43*, 385–440. [[CrossRef](#)] [[PubMed](#)]
85. Terman, L.M. *The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale*; Houghton Mifflin: New York, NY, USA, 1916.
86. Thorndike, E.L. *An Introduction to the Theory of Mental and Social Measurements*; Teachers College, Columbia University: New York, NY, USA, 1904.
87. Revelle, W. Individual differences in personality and motivation: 'Non-cognitive' determinants of cognitive performance. In *Attention: Selection, Awareness, and Control*; Baddeley, A., Weiskrantz, L., Eds.; Oxford University Press: Oxford, UK, 1993; pp. 346–373.
88. Benton, A.L. Influence of incentives upon intelligence test scores of school children. *Pedagog. Semin. J. Genet. Psychol.* **1936**, *49*, 494–497. [[CrossRef](#)]
89. Ferguson, H.H. Incentives and an intelligence tests. *Aust. J. Psychol. Philos.* **1937**, *15*, 39–53. [[CrossRef](#)]
90. Klugman, S.F. The effect of money incentive versus praise upon the reliability and obtained scores of the Revised Stanford-Binet Test. *J. Gen. Psychol.* **1944**, *30*, 255–269. [[CrossRef](#)]
91. Knight, F.B.; Remmers, H.H. Fluctuations in mental production when motivation is the main variable. *J. Appl. Psychol.* **1923**, *7*, 209–223. [[CrossRef](#)]
92. Maller, J.B.; Zubin, J. The effect of motivation upon intelligence test scores. *Pedagog. Semin. J. Genet. Psychol.* **1932**, *41*, 135–151. [[CrossRef](#)]
93. Wise, S.L.; Smith, L.F. The validity of assessment when students don't give good effort. In *Handbook of Human and Social Conditions in Assessment*; Brown, G.T.L., Harris, L.R., Eds.; Routledge: New York, NY, USA, 2016; pp. 204–220.
94. Braun, H.; Kirsch, I.; Yamamoto, K. An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teach. Coll. Rec.* **2011**, *113*, 2309–2344.
95. Pedulla, J.; Abrams, L.; Madaus, G.; Russell, M.; Ramos, M.; Miao, J. *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers*; National Board on Educational Testing and Public Policy: Boston, MA, USA, 2003.
96. Zilberberg, A.; Anderson, R.D.; Finney, S.J.; Marsh, K.R. American college students' attitudes toward institutional accountability testing: Developing measures. *Educ. Assess.* **2013**, *18*, 208–234. [[CrossRef](#)]
97. Sundre, D.L.; Moore, D.L. The student opinion scale: A measure of examinee motivation. *Assessment Update* **2002**, *14*, 8–9.
98. Duckworth, A.L.; Quinn, P.D.; Lynam, D.R.; Loeber, R.; Stouthamer-Loeber, M. Role of test motivation in intelligence testing. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7716–7720. [[CrossRef](#)] [[PubMed](#)]
99. Wise, S.L.; Kong, X. Response time effort: A new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* **2005**, *18*, 163–183. [[CrossRef](#)]
100. Meijer, R.R. Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychol. Methods* **2003**, *8*, 72–87. [[CrossRef](#)] [[PubMed](#)]

101. Wise, S.L.; DeMars, C.E. Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educ. Assess.* **2005**, *10*, 1–17. [[CrossRef](#)]
102. Revelle, W. Personality, motivation, and cognitive performance. In *Abilities, Motivation and Methodology*; Kanfer, R., Ackerman, P.L., Cudeck, R., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1989; pp. 297–343.
103. Richardson, M.; Abraham, C. Conscientiousness and achievement motivation predict performance. *Eur. J. Personal.* **2009**, *23*, 589–605. [[CrossRef](#)]
104. Costa, P.T., Jr.; McCrae, R.R.; Dye, D.A. Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personal. Individ. Differ.* **1991**, *12*, 887–898. [[CrossRef](#)]
105. Dudley, N.M.; Orvis, K.A.; Lebiecki, J.E.; Cortina, J.M. A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *J. Appl. Psychol.* **2006**, *91*, 40–57. [[CrossRef](#)] [[PubMed](#)]
106. Hough, L.M.; Johnson, J.W. Use and importance of personality variables in work settings. In *Handbook of Psychology, Volume 12: Industrial and Organizational Psychology*, 2nd ed.; Schmitt, N.W., Highhouse, S., Weiner, I.B., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2013; pp. 211–243.
107. Messick, S. Validity. In *Educational Measurement*, 3rd ed.; Linn, R.L., Ed.; Macmillan: Old Tappan, NJ, USA, 1989; pp. 13–103.
108. Richardson, M.; Abraham, C.; Bond, R. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychol. Bull.* **2012**, *138*, 353–387. [[CrossRef](#)] [[PubMed](#)]
109. Poropat, A.E. A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* **2009**, *135*, 322–338. [[CrossRef](#)] [[PubMed](#)]
110. Noffle, E.E.; Robins, R.W. Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *J. Personal. Soc. Psychol.* **2007**, *93*, 116–130. [[CrossRef](#)] [[PubMed](#)]
111. Judge, T.A.; Higgins, C.A.; Thoresen, C.J.; Barrick, M.R. The big five personality traits, general mental ability, and career success across the life span. *Pers. Psychol.* **1999**, *52*, 621–652. [[CrossRef](#)]
112. Batey, M.; Chamorro-Premuzic, T.; Furnham, A. Individual differences in ideational behavior: Can the big five and psychometric intelligence predict creativity scores? *Creat. Res. J.* **2010**, *22*, 90–97. [[CrossRef](#)]
113. Lounsbury, J.W.; Sundstrom, E.; Loveland, J.M.; Gibson, L.W. Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. *Personal. Individ. Differ.* **2003**, *35*, 1231–1239. [[CrossRef](#)]
114. Robbins, S.B.; Lauver, K.; Le, H.; Davis, D.; Langley, R.; Carlstrom, A. Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychol. Bull.* **2004**, *130*, 261–288. [[CrossRef](#)] [[PubMed](#)]
115. Murray, A.L.; Johnson, W.; McGue, M.; Iacono, W.G. How are conscientiousness and cognitive ability related to one another? A re-examination of the intelligence compensation hypothesis. *Personal. Individ. Differ.* **2014**, *70*, 17–22. [[CrossRef](#)]
116. Cattell, R.B.; Warburton, F.W. *Objective Personality and Motivation Tests*; University of Illinois Press: Urbana, IL, USA, 1967.
117. Hartshorne, H.; May, M.A. *Studies in the Nature of Character: Studies in Deceit*; MacMillan: New York, NY, USA, 1928.
118. Ortner, T.M.; Schmitt, M. Advances and continuing challenges in objective personality testing. *Eur. J. Psychol. Assess.* **2014**, *30*, 163–168. [[CrossRef](#)]
119. Ortner, T.; Proyer, R. Objective personality tests. In *Behavior-Based Assessment in Psychology*; Ortner, T.M., van de Vijver, F.J.R., Eds.; Hogrefe & Huber: Seattle, DC, USA, 2015; pp. 133–149, ISBN 978-0-88937-437-9.
120. James, L. Measurement of personality via conditional reasoning. *Organ. Res. Methods* **1998**, *1*, 131–163. [[CrossRef](#)]
121. Lejuez, C.W.; Read, J.P.; Kahler, C.W.; Richards, J.B.; Ramsey, S.E.; Stuart, G.L.; Brown, R.A. Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *J. Exp. Psychol.* **2002**, *8*, 75–84. [[CrossRef](#)]
122. Gugerty, L. Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. *J. Exp. Psychol.* **1997**, *3*, 42–66. [[CrossRef](#)]
123. Hurwitz, J.B. Assessing a perceptual model of risky real-time decision making. *Proc. Hum. Factors Ergon. Soc. Ann. Meet.* **1996**, *40*, 223–227. [[CrossRef](#)]

124. Jackson, J.J.; Wood, D.; Bogg, T.; Walton, K.E.; Harms, P.D.; Roberts, B.W. What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *J. Res. Personal.* **2010**, *44*, 501–511. [[CrossRef](#)] [[PubMed](#)]
125. Almlund, M.; Duckworth, A.L.; Heckman, J.J.; Kautz, T.D. Personality psychology and economics. (No. w16822). In *Handbook of the Economics of Education*; Hanushek, E.A., Machin, S., Woessmann, L., Eds.; Elsevier: Amsterdam, The Netherlands, 2011; pp. 1–181.
126. Goldberg, L.R. The structure of phenotypic personality traits. *Am. Psychol.* **1993**, *48*, 26–34. [[CrossRef](#)] [[PubMed](#)]
127. Falk, A.; Becker, A.; Dohmen, T.; Enke, B.; Huffman, D.; Sunde, U. *The Nature and Predictive Power of Preferences: Global Evidence*; Institute for the Study of Labor (IZA), Discussion Paper (DP) No. 9504; IZA: Bonn, Germany, 2015.
128. Falk, A.; Becker, A.; Dohmen, T.; Huffman, D.; Sunde, U. *The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences*; Institute for the Study of Labor (IZA), Discussion Paper (DP) No. 9674; IZA: Bonn, Germany, 2016.
129. Stankov, L.; Crawford, J.D. Confidence judgments in studies of individual differences. *Personal. Individ. Differ.* **1996**, *21*, 971–986. [[CrossRef](#)]
130. Stankov, L.; Morony, S.; Lee, Y.-P. Confidence: The best non-cognitive predictor of academic achievement? *Educ. Psychol.* **2013**, *34*, 9–28. [[CrossRef](#)]
131. Stankov, L.; Kleitman, S.; Jackson, S.A. Measures of the trait of confidence. In *Measures of Personality and Social Psychological Constructs*; Boyle, G.J., Saklofske, D.H., Matthews, G., Eds.; Academic Press: London, UK, 2014; pp. 158–189.
132. Stankov, L.; Lee, J.; Luo, W.; Hogan, D.J. Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learn. Individ. Differ.* **2012**, *22*, 747–758. [[CrossRef](#)]
133. Jackson, S.A.; Kleitman, S.; Howie, P.; Stankov, L. Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Front. Psychol.* **2016**, *7*, 1559. [[CrossRef](#)] [[PubMed](#)]
134. Lee, J. Universals and specifics of math self-concept, math-self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learn. Individ. Differ.* **2009**, *19*, 355–365. [[CrossRef](#)]
135. Kleitman, S.; Stankov, L. Self-confidence and metacognitive processes. *Learn. Individ. Differ.* **2007**, *17*, 161–173. [[CrossRef](#)]
136. Huang, J.L.; Curran, P.G.; Keeney, J.; Poposki, E.M.; DeShon, R.P. Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* **2012**, *27*, 99–114. [[CrossRef](#)]
137. Malhotra, N. Completion time and response order effects in web surveys. *Public Opin. Q.* **2008**, *72*, 914–934. [[CrossRef](#)]
138. Meade, A.W.; Bartholomew, C.S. Identifying careless responses survey data. *Psychol. Methods* **2012**, *17*, 437–455. [[CrossRef](#)] [[PubMed](#)]
139. Hitt, C.; Trivitt, J.; Cheng, A. When you say nothing at all: The predictive power of student effort on surveys. *Econ. Educ. Rev.* **2016**, *52*, 105–119. [[CrossRef](#)]
140. Zamarro, G.; Cheng, A.; Shakeel, M.D.; Hitt, C. Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *J. Behav. Exp. Econ.* **2018**, *72*, 51–60. [[CrossRef](#)]
141. Cheng, A.; Zamarro, G.; Orriens, B. Personality as a predictor of unit nonresponse in an internet panel. *Sociol. Methods Res.* **2018**. [[CrossRef](#)]
142. Leary, L.F.; Dorans, N.J. Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Rev. Educ. Res.* **1985**, *55*, 387–413. [[CrossRef](#)]
143. Albano, A.D. Multilevel modeling of item position effects. *J. Educ. Meas.* **2014**, *50*, 408–426. [[CrossRef](#)]
144. Debeer, D.; Buchholz, J.; Hartig, J.; Janssen, R. Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading assessment. *J. Educ. Behav. Stat.* **2014**, *39*, 502–523. [[CrossRef](#)]
145. Borgonovi, F.; Biecek, P. An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learn. Individ. Differ.* **2016**, *49*, 128–137. [[CrossRef](#)]
146. Weirich, S.; Hecht, M.; Penk, C.; Roppelt, A.; Bohme, K. Item position effects are moderated by changes in test-taking effort. *Appl. Psychol. Meas.* **2017**, *41*, 115–129. [[CrossRef](#)] [[PubMed](#)]

147. Kong, X.J.; Wise, S.L.; Bhola, D.S. Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educ. Psychol. Meas.* **2007**, *67*, 606–619. [[CrossRef](#)]
148. Wise, S.L.; Pastor, D.A.; Kong, X.J. Understanding correlates of rapid-guessing behavior in low stakes testing: Implications for test development and measurement practice. *Appl. Meas. Educ.* **2009**, *22*, 185–205. [[CrossRef](#)]
149. Lee, Y.-H.; Jia, Y. Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assess. Educ.* **2014**, *2*, 1–24. [[CrossRef](#)]
150. Wise, S.L.; Gao, L. A general approach to measuring test-taking effort on computer-based tests. *Appl. Meas. Educ.* **2014**, *30*, 343–354. [[CrossRef](#)]
151. Cooper, H.; Robinson, J.C.; Patall, E.A. Does homework improve academic achievement? A synthesis of research, 1987–2003. *Rev. Educ. Res.* **2006**, *76*, 1–62. [[CrossRef](#)]
152. McMullen, S. The Impact of Homework Time on Academic Achievement. Available online: <https://www.unc.edu/courses/2007fall/econ/994/040/2007fall/mcmullen.pdf> (accessed on 9 July 2018).
153. Grodner, A.; Rupp, N.G. The role of homework in student learning outcomes: Evidence from a field experiment. *J. Econ. Educ.* **2013**, *44*, 93–109. [[CrossRef](#)]
154. Kalenkoski, C.M.; Pabilonia, S.W. Does High School Homework Increase Academic Achievement? Available online: <https://www.bls.gov/osmr/pdf/ec150050.pdf> (accessed on 9 July 2018).
155. Novarese, M.; Di Giovinazzo, V. Promptness and Academic Performance. MPRA Paper No. 49746, Posted 11. 2013. Available online: <http://mpra.ub.uni-muenchen.de/49746/> (accessed on 9 July 2018).
156. Fast, L.A.; Funder, D.C. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *J. Personal. Soc. Psychol.* **2008**, *94*, 334–346. [[CrossRef](#)] [[PubMed](#)]
157. Gosling, S.D.; Ko, S.J.; Mannarelli, T.; Morris, M.E. A room with a cue: Personality judgments based on offices and bedrooms. *J. Personal. Soc. Psychol.* **2002**, *82*, 379–398. [[CrossRef](#)]
158. Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [[CrossRef](#)] [[PubMed](#)]
159. Youyou, W.; Kosinski, M.; Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 1036–1040. [[CrossRef](#)] [[PubMed](#)]
160. Srivastava, S.; John, O.P.; Gosling, S.D.; Potter, J. Development of personality in early and middle adulthood: Set like plaster or persistent change? *J. Personal. Soc. Psychol.* **2003**, *84*, 1041–1053. [[CrossRef](#)]
161. Costa, P.T., Jr.; McCrae, R.R. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*; Psychological Assessment Center: Odesa, FL, USA, 1992.
162. Soto, C.J.; John, O.P.; Gosling, S.D.; Potter, J. The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *J. Personal. Soc. Psychol.* **2008**, *94*, 718–737. [[CrossRef](#)] [[PubMed](#)]
163. Tucker-Drob, E.M. Differentiation of cognitive abilities across the lifespan. *Dev. Psychol.* **2010**, *45*, 1097–1118. [[CrossRef](#)] [[PubMed](#)]
164. Möttus, R.; Allik, J.; Pullman, H. Does personality vary across ability levels? A study using self and other ratings. *J. Res. Personal.* **2007**, *41*, 155–170. [[CrossRef](#)]
165. Allik, J.; Laidra, K.; Realo, A.; Pullman, H. Personality development from 12 to 18 years of age: Changes in mean levels and structure of traits. *Eur. J. Personal.* **2004**, *18*, 445–462. [[CrossRef](#)]
166. De Fruyt, F.; Mervielde, I.; Hoekstra, H.A.; Rolland, J.P. Assessing adolescents' personality with the NEO PI-R. *Assessment* **2000**, *7*, 329–345. [[CrossRef](#)] [[PubMed](#)]
167. MacCann, C.; Pearce, N.; Jiang, Y. The general factor of personality is stronger and more strongly correlated with cognitive ability under instructed faking. *J. Individ. Differ.* **2017**, *38*, 46–54. [[CrossRef](#)]
168. Ellingson, J.E.; Sackett, P.R.; Connelly, B.S. Personality assessment across selection and development contexts: Insights into response distortion. *J. Appl. Psychol.* **2007**, *92*, 386–395. [[CrossRef](#)] [[PubMed](#)]
169. Birkeland, S.A.; Manson, T.M.; Kisamore, J.L.; Brannick, M.T.; Smith, M.A. A meta-analytic investigation of job applicant faking on personality measures. *Int. J. Sel. Assess.* **2006**, *14*, 317–335. [[CrossRef](#)]
170. Griffith, R.L.; Converse, P.D. The rules of evidence and the prevalence of applicant faking. In *New Perspectives on Faking in Personality Assessment*; Ziegler, M., MacCann, C., Roberts, R., Eds.; Oxford University Press: New York, NY, USA, 2011; pp. 34–52. [[CrossRef](#)]
171. Cao, M. Examining the Fakability of Forced-Choice Individual Differences Measures. Ph.D. Thesis, Psychology. University of Illinois, Urbana, IL, USA, 2016. Available online: <http://hdl.handle.net/2142/93064> (accessed on 9 July 2018).

172. Viswevaran, C.; Ones, D.S. Meta-analyses of fakability estimates: Implications for personality measurement. *Educ. Psychol. Meas.* **1999**, *59*, 197–210. [CrossRef]
173. Tett, R.P.; Freund, K.A.; Christiansen, N.D.; Fox, K.E. Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personal. Individ. Differ.* **2012**, *52*, 195–201. [CrossRef]
174. King, G.; Wand, J. Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Anal.* **2007**, *15*, 46–66. [CrossRef]
175. Kyllonen, P.C.; Bertling, J.P. Innovative questionnaire assessment methods to increase cross-country comparability. In *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; CRC Press: Boca Raton, FL, USA, 2014; pp. 277–285.
176. Bertling, J.; Kyllonen, P.C. Anchoring adjustments of student questionnaire indexes: Possible scenarios for the PISA 2012 international database. In Proceedings of the OECD PISA Technical Advisory Group (TAG) Meeting, Melbourne, Australia, 4–5 April 2013.
177. Ziegler, M.; Colom, R.; Horstmann, K.T.; Wehner, C.; Bensch, D. Special Issue: “The Ability-Personality Integration”. Available online: [http://www.mdpi.com/journal/jintelligence/special\\_issues/ability\\_personality\\_integration](http://www.mdpi.com/journal/jintelligence/special_issues/ability_personality_integration) (accessed on 4 June 2018).
178. Flake, J. Jeff Flake: We Need Immigrants with Skills. But Working Hard Is a Skill. Available online: <https://curranberger.com/nyt-we-need-immigrants-with-skills-but-working-hard-is-a-skill/> (accessed on 7 June 2018).
179. Heckman, J.J. Measuring and Assessing Skills: Real-Time Measurement of Cognition, Personality, and Behavior. Available online: [https://cehd.uchicago.edu/?page\\_id=1198](https://cehd.uchicago.edu/?page_id=1198) (accessed on 7 June 2018).
180. National Research Council. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*; The National Academies Press: Washington, DC, USA, 2012. [CrossRef]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).