A. Constructing Cognitive Skills Scores Methods for Scale Development and Scoring

Prepared by Carly Tubbs, Ph.D. Candidate, New York University; Louise M. Bahry, Ph.D. Candidate, University of Massachusetts Amherst; Robin Audy, World Bank.

Background and Measures

Data for this study come from a 34-item survey module designed for use by the World Bank to assess five different "cognitive" skills. These cognitive skills can be conceptualized as falling into two domains:

(1) *Executive functioning skills,* defined as the cognitive control capacities that enable individuals to "organize their thinking and behavior with flexibility, decrease their reactive responding to contextual cues and contingencies, and engage in self-regulated … behavior" (Welsh et al., 2010). Researchers in developmental psychology and elsewhere propose that such skills are important for school readiness and labor force attainment since they enable individuals to regulate cognitive and emotional responses that in turn allow individuals to engage more effectively in learning activities (Fuchs et al., 2005). We assessed one component of executive functioning – working memory – using a 12-item memory scale adopted from the Skills and Labor Market Survey (ENHAB)[1]. These items tested the short-term recall of increasingly longer number sequences (starting with two numbers and ending with 9 numbers). Enumerators gave respondents three practice examples with two-number sequences to train the respondents on how to answer the questions, and were instructed to read out numbers at a regular pace to avoid grouping.

(2) *Domain-specific skills,* consisting of "knowledge of ideas, facts and definitions, as well as … formulas and rules" (Boekarts, 1997, p. 164) about specific domains such as literacy and numeracy. In turn, each broader domain can be conceptualized as including other branches; mathematics, for example, includes concepts such as number recognition, arithmetic, and graph comprehension (Fuchs et al., 2005; Pinker, 1990). In this study, we assessed various concepts within the domains of literacy and numeracy using multiple-choice questions with four answer choices. Within literacy, these concepts include: (1) *semantics*, assessed using seven items, with five items assessing respondents' familiarity with vocabulary, one item testing understanding of a national idiom, and one item measuring comprehension of the meaning of a complex sentence;[2] (2) *reading comprehension*, assessed by asking respondents to read a 257-word non-technical narrative text and then answering five questions about the text; and (3) *information comprehension,* assessed using four items based on instructions for taking a medicine and reading a timetable describing inter-city bus schedules. Within numeracy, concepts include: (1) *arithmetic*, assessed using three questions about prices in an advertisement; and (2) *graph comprehension*, assessed using three questions based on a graph of Bulgaria's population growth from 1900 to 2011. The items assessing reading

[1] The ENHAB is a recent survey in Peru which gathers data on cognitive and socio-emotional test scores, individual's characteristics, educational trajectory, and wages.

[2] An issue with translation of the items comprising the semantics scale rendered the data from this set of items unusable. The semantics scale was thus not considered for analysis, leaving the total number of assessed skills at five.

comprehension and semantics were taken from existing instruments fielded by the World Bank with Bulgarian students, while the items assessing mathematics and information comprehension were adapted from the Adult Literacy and Lifeskills Survey (Murray, Clermont, & Binkley, 2005).

These domains are not meant to be exhaustive, but to serve as useful heuristics. Moreover, executive functioning skills and domain-specific skills are related: A number of recent studies provide evidence that executive functioning skills such as working memory actually contribute to the development of literacy and numeracy skills (Blair & Razza, 2010; Swanson, Jerman, & Zheng. 2008). From a policy perspective, this suggests that educators should focus on the promotion of *both* executive functioning and domain-specific skills, particularly in the pre-school and elementary school years when such functions are most malleable to intervention (Welsh et al., 2010).

Analysis Strategy

All missing values were recoded as incorrect answers, resulting in a set of 33 dichotomous or binary items.[3] In choosing how to score the items, we were motivated by a primary concern of reducing the measurement error in each score. That is, when we administer a survey measure or test, we want to ensure that the variability in scores is due to what we are trying to measure – in this study, executive functioning or domain-specific skills – as opposed to error or bias. Traditional or unrefined methods of scoring – such as summing the survey items – do not account for this measurement error, leading to bias in future regression analyses (for more information, see Box B1). Refined scoring methods that account for measurement error include the production of factor scores using factor analysis or item response theory (IRT) methods.

Box A1: What is Item Response Theory and When Can We Use It?

Item Response Theory (IRT) is an approach, or family of statistical models, used to analyze assessment item data, such as cognitive skills assessment data. Several IRT models have been developed to estimate ability or person parameters that are scored either dichotomously (i.e. only two response categories) or polytomously (i.e. more than two response categories; Hambleton, Swaminathan, & Rogers, 1991). Traditionally, IRT has been used for educational applications for Computerized Adaptive Testing (CAT), test score equating, item analysis, and test banking. However, due to the advantages of IRT, other disciplines have recently developed an interest in using IRT for scoring, validation, and other psychometric analyses (Reise & Henson, 2003).

There are two over-arching families of item response models which differ greatly in theoretical and mathematical background and analysis. The first of the two families, the logistic models, relate examinee ability (θ) and item parameters using logistic functions. The logistic family of IRT models allow for the estimation of up to three item parameters, or characteristics. The one-parameter (1PL)

[3] Ideally, we would be able to identify four, not two, sets of responses: answered correctly; answered incorrectly; not answered and didn't know; and not answered due to time constraints or motivation but known. While such codes were initially included in the survey instrument, issues with data processing rendered such codes unusable. We were thus forced to collapse the codes into a dichotomous response: correct or incorrect. The implications of this choice are discussed further in the Implications and Future Directions section.

model is the most basic and involves, as the name states, only one item parameter: the *b*-parameter is included in every IRT model and is considered the difficulty parameter (Yen & Fitzpatrick, 2006). The *b*-parameter is at the point on the θ scale where the probability of a correct response is equal to 0.50 and typically varies from -2.00 to 2.00 (Hambleton et al., 1991; Yen & Fitzpatick, 2006), increasing as items become more difficult. The two-parameter model (2PL) includes a second item parameter, the discrimination parameter, *a. a* is the slope of the item characteristic curve (ICC) at the point of inflection and the higher the value of *a*, the more sharp the discrimination (Yen & Fitzpatrick, 2006). Finally, the three-parameter model (3PL) includes the *c*-parameter, called the guessing or pseudo-chance parameter. This parameter was introduced to account for the possibility that even students with low ability have some chance of answering even difficult questions correctly. This parameter is not always necessary, and if set to zero, equates the 3PL with the 2PL (Yen & Fitzpatrick, 2006).

One of the big advantages of using IRT is that the ability or person parameters (θ) are not item or test dependent, and item and test characteristics are not dependent on the ability or person parameters. This is called the *property of invariance* (Hambleton et al., 1991; Lord, 1980). It means that the test and item parameters remain the same regardless of the sample of respondents, and the ability or person parameters do not vary depending on the test items administered or the time of test, provided the items are relevant to and representative of the same domain of interest.

Although there are clear benefits to the invariance property, there are two integral assumptions of IRT. First, there is an assumption regarding the *dimensionality* of the underlying ability or trait. While there are multi-dimensional IRT models (MIRT), the traditional IRT model requires that a single trait or ability accounts for an individual's θ score. When this assumption of the data holds, the examinees can be placed along a single, meaningful scale (Hambleton et al., 1991). Second, there is the assumption of *local item independence*. When the items on an assessment are locally independent, a response to any item is independent of a response to any other item on the same assessment for a given individual. This assumption allows us to determine the probability of an individual response pattern occurring given the individual's ability or trait level (Hambleton et al., 1991; Lord, 1980). If either of these assumptions is not met, item and person parameters will not be properly estimated and thus, indefensible.

In addition to these assumptions, an assessment of model-data fit is also important in IRT. A poorly specified model creates problems with estimating both item parameters and θ scores. Consider the following: An analyst mistakenly specifies a model which only specifies two parameters when in fact the data fit a model consisting of three item parameters. Because the pseudo-guessing parameter has not been specified, the θ values may be over-estimated as the individual's ability to correctly guess the answer has not been taken into consideration. Guessing is not considered to be included in ability and, as such, it should not be allowed to unduly influence scores. While IRT provides distinct advantages to classical methods of analyzing assessment data, these advantages come with several very restrictive assumptions which, if violated, calls into question the validity of the results.

In order to assess whether it was appropriate to employ an IRT model with this data, we decided to first empirically determine the dimensionality of the items by conducting an exploratory factor analysis (EFA) with an oblimax rotation on a randomly selected half of participants stratified by

country (N = 3,965). [4] Should a one-factor model provide a good fit to the data, we would be able to proceed with IRT analyses. Should a multi-factor model provide a good fit to the data, the dimensionality assumption required by IRT methodologies would be violated. In that case, we proceed by examining the results of the EFA and confirming the factor structure using the second half of the sample (N = 3,964). All analyses were conducted in MPlus (Muthén & Muthén, 1998-2012; Version 6.12) and adjusted for any clustering of the data due to sampling design.[5] Responses were treated as ordered categorical data to account for the skewed nature of the data.

Once we determined a factor structure that provided a good fit to the data, we created individual scores on each of these factors using refined factor scoring techniques. As detailed above, factor scoring is preferable in this case to traditional sum scoring methods given that factor scores account for: (1) the weight of individual item loadings; and (2) shared variance between the items and the factors *and* measurement error (DiStephano, Zhu, & Midrila, 2009). Factor scores were created using maximum a posteriori (MAP) estimation in MPLUS, which accounts for the non-normal distribution of item response (Muthén & Muthén, 1998-2012).

Results

The initial EFA indicated that a one-factor model did not provide a good fit to the data (χ^2 (324) = 8981.68, CFI: .888, RMSEA: .082, .081 < 95% CI < .084).[6] Thus we decided that it was not feasible to proceed with an IRT analysis due to the plausibility of violating the dimensionality assumption. In examining the factor loadings, we noted that the 12 items making up the original construct of working memory loaded cleanly onto one factor. This factor was left intact and removed from the exploratory analyses. We then chose a 2-factor solution to model associations between the remaining 15 items. This model provided a good fit to the data (χ^2 (76) = 1261.15, CFI=.951, RMSEA=.063, .060 < 95% CI < .066) while modeling the observed indicators parsimoniously.

A confirmatory factor analysis then confirmed the fit of a 3-factor model for all 27 items in which factors were allowed to correlate (χ^2 (321) = 3128.37, CFI=.981, RMSEA=.033, .032 < 95% CI < .034).[7] The three identified factors described in Table 1, below, were: (1) Working Memory (12 items); (2) Reading Comprehension (5 items); and (3) Informational Numeracy (10 items). In addition, preliminary measurement equivalence analyses indicate that this same factor structure provides a good fit to the data in Uzbekistan, Kyrgyzstan, and Tajikistan (χ^2 (97c3) = 10531.15, CFI=.953, RMSEA=.061, .060 < 95% CI < .062).[8] Finally, given the high correlation between the

[4] An oblimax rotation was chosen to account for the hypothesized correlation between factors.

[5] In Tajikistan – but not in Uzbekistan or Kyrgyzstan – up to two individuals per household were administered the non-cognitive skills module. To account for any non-independence of the data that may occur due to individuals being nested in households, we used the Type=Complex and Cluster=psuid commands in MPlus.

[6] In assessing model goodness of fit, the following criteria are used: A RMSEA < .08 provides an acceptable fit to the data, while an RMSEA < .05 provides a good fit to the data; a CFI > .9 provides an acceptable fit to the data while a CFI > .95 provides a good fit to the data (Kline, 2011).

[7] Factor correlations in the CFA were: Working Memory-Literacy (r=.428, p<.001), Working Memory-Informational Numeracy (r=.480, p<.001), and Literacy-Informational Numeracy (r=.69, p<.001).

[8] Tests of measurement invariance seek to establish whether we are measuring the same construct in the same way across different groups. As of this writing, our preliminary analyses have established *configural invariance:* that the same factor structure (e.g., the same number of factors and the same pattern of loadings) exists in the samples from all three countries.

literacy and informational numeracy items, initial analyses were also conducted to determine whether a higher-order "cognitive" factor may account for the covariation between factors (Cattell, 1978).[9] This model was uninterpretable due to factor loadings above 1.

Table A1. Unstandardized Results from Final CFA of Cognitive Skills Module

		Loading	SE
Working Memory			
1.	Working Memory Item 1	0.974	0.009
2.	Working Memory Item 2	0.985	0.006
3.	Working Memory Item 3	0.987	0.005
4.	Working Memory Item 4	0.962	0.004
5.	Working Memory Item 5	0.926	0.006
6.	Working Memory Item 6	0.904	0.006
7.	Working Memory Item 7	0.862	0.006
8.	Working Memory Item 8	0.866	0.006
9.	Working Memory Item 9	0.816	0.008
10.	Working Memory Item 10	0.795	0.011
11.	Working Memory Item 11	0.861	0.012
12.	Working Memory Item 12	0.900	0.013
Reading Comprehension			
13.	Reading Comprehension Item 13	0.800	0.012
14.	Reading Comprehension Item 14	0.748	0.011
15.	Reading Comprehension Item 15	0.843	0.009
16.	Reading Comprehension Item 16	0.734	0.009
17.	Reading Comprehension Item 17	0.788	0.010
Informational Numeracy			
18.	Information Comprehension Item 18	0.522	0.014
19.	Information Comprehension Item 19	0.553	0.013
20.	Information Comprehension Item 20	0.588	0.013
21.	Information Comprehension Item 21	0.812	0.009
22.	Arithmetic Item 22	0.574	0.013
23.	Arithmetic Item 23	0.741	0.010
24.	Arithmetic Item 24	0.591	0.013
25.	Graph Comprehension Item 25	0.726	0.012
26.	Graph Comprehension Item 26	0.832	0.009

Future analyses will examine other levels of invariance, establishment of which increases our certainty that observed differences between countries is attributable only to true differences in the variability of the scores.

[9] For over a century, researchers have been interested in defining and measuring an overall measure of cognitive ability, or "g" factor (Jensen, 1998; Heckman, Stixrud, & Urzua, 2006). It is beyond the scope of this paper to comment extensively on such research; however, as developmental psychologists with an interest in applying research to policy, we take the position that it is useful to identify and understand the *components* of cognitive ability to better design programs to support the development of such skills.

27.	Graph Comprehension Item 27	0.667	0.011

Interpretation and Future Directions

Our analyses indicated that the data from the cognitive skills module is best represented by three related factors that correspond to some – but not all – of the five cognitive skills described above. For example, our analyses indicated items 1-12 all indexed the hypothesized underlying executive functioning skill of Working Memory, while items 13-17 corresponded to the hypothesized underlying domain-specific skill of Reading Comprehension. Substantively, this indicates that individuals that have higher Working Memory factor scores are better able to temporarily store and manipulate information that is necessary for domain-specific cognitive tasks such as reading comprehension (Baddeley, 1992). Individuals with higher Reading Comprehension scores have a better ability to read and process text and understand its meaning than individuals with lower Reading Comprehension scores (National Reading Panel, 2000).

The other factor represented in the data is a combination of items meant to index facets of both Literacy (items 18-21) and Numeracy (items 22-27). This pattern of relationships can be understood in that the Information Comprehension items all involved number recognition (a component of numeracy), while the Numeracy items all tapped the ability to locate and use information contained in various formats such as advertisements and graphs (a component of information comprehension). Individuals who score highly on Informational Numeracy have the ability to recognize and manipulate numbers contained in and represented by various formats.

There are three things to consider when interpreting the above analysis. First, the factor scores created through the factor analysis procedures described above are not invariant across different tests assessing cognitive ability. While such scores could have resulted from using IRT methodologies, we have evidence that using IRT with this cognitive assessment is not defensible given the likely violation of the assumption of dimensionality and as a result, item dependence. As such, we proceeded with creating refined factor scores that – although they do not inherently have the property of invariance – reduce the amount of measurement error contained in the scores. It should be noted, however, that invariance is a property that can be assessed through the use of factor analytic methods. Second, many of the items included in the cognitive skills assessment are not "clean" items. That is, they assess more than one skill at the same time: Items meant to tap the construct of Arithmetic, for example, also involve elements of reading comprehension and information comprehension. The factors – particularly Reading Comprehension and Information Numeracy – are thus highly correlated, which may be problematic for establishing predictive validity. To address this, we recommend that future analyses with this data
consider a bi-factor analysis in which orthogonal or non-correlated grouping factors are created by allowing a "general" trait to correlate with the items (Reise, Moore, & Haviland, 2010). Finally, as noted in footnote 2, we were limited in our ability to discriminate between correct, incorrect, and missing answers due to issues in data processing. Given that missing answers were all recoded to be incorrect, it is likely that the scores underestimate the cognitive ability level present in the sample population. To address this, we recommend that future data collection activities carefully assess the type and extent of missing data to allow for better sensitivity tests of results to such specifications.

References

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*(2), 161-186.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1-11.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*(1), 29-43.

Hambleton, R. K., Swaminathan, H., & Rogers, H . J. (1991). *Fundamentals of item response theory*. Sage Publications, Newbury Park: CA.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). *The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior* (No. w12006). National Bureau of Economic Research.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Murray, T. S., Clermont, Y., & Binkley, M. (2005). *Measuring adult literacy and life skills: New frameworks for assessment*. Ottawa: Statistics Canada.

Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus user's guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

National Reading Panel, National Institute of Child Health & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.

Pinker, S. 1990. A theory of graph comprehension. In R. Friedle (Ed.), *Artificial intelligence and the future*

of testing. Hillsdale, N.J.: Erlbaum.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*(revised and expanded ed.). Chicago: The University of Chicago Press. (Originial work published 1960)

Reise, S. P. & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81(2),* 93-103. doi:10.1207/S15327752JPA8102_01

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Swanson, H. L., Jerman, O., & Zheng, X. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, *100*, 343–379.

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of educational psychology*, *102*(1), 43-53.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-154). Westport, CT: American Council on Education.

B. Constructing Non-Cognitive Skills Scores Methods for Scale Development and Scoring

Prepared by Carly Tubbs, Ph.D. Candidate, New York University

Background and Measures

Data for this study come from a 33-item survey module designed for use by the World Bank to assess 11 different "non-cognitive" skills (see Table B1, below; Duckworth and Guerra, 2012). These non-cognitive skills can be conceptualized as falling into two domains:

Personality traits, defined as enduring patterns of thinking, feeling, and behaving which are relatively stable across time and situations (Borghans, Duckworth, Heckman, and ter Weel, 2008; Paunonen, 2003). The "Big Five" factors of personality—openness, conscientiousness, extraversion, agreeableness, and neuroticism (or emotional stability)—are the most widely accepted taxonomy of broad personality traits (Goldberg, 1990), having been validated for use across developmental stages (John and Srivastava, 1999) and cultures (Soto, John, Gosling, and Potter, 2008). The survey assessed each of these five factors with three items in the short Big Five Inventory (BFI-S) originally developed by John and Srivastava (1999) and later validated in large-scale panel surveys (Lang et al., 2011). Given its association with important labor market outcomes, assessed grit—a component of conscientiousness—was also assessed, with three items from the Grit Scale (Duckworth et al., 2007).

Socio-emotional skills, defined as the learned knowledge, attitudes, and skills necessary to understand and manage emotions, set and achieve positive goals, establish and maintain positive relationships, and make responsible decisions (CASEL, 2014). Although different cultures may differentially name, conceptualize, and prioritize such skills, socio-emotional skills and learning are of critical importance across all regions of the world (Torrente, Alimchandani, and Aber, in press). There does not currently exist an organization of socio-emotional skills similar to that developed for personality traits; as such, this survey measures socio-emotional skills that are both valued by employers in countries in Europe and Central Asia (World Bank, 2009, 2013) and amenable to intervention efforts (Yeager and Dweck, 2012). These skills include: hostile bias (2 items; Dodge, 2003), decision making (4 items; Mann, Burnett, Radford, and Ford, 1997), achievement striving, and self-control (3 items and 2 items, respectively; Goldberg et al., 2006). In addition, the fixed vs. growth mindset, or the belief that intelligence is fixed versus malleable, was measured (4 items; Yeager and Dweck, 2012).

These domains are not meant to be exhaustive, but to serve as useful heuristics. Moreover, personality traits and socio-emotional skills are related: individuals with certain personality traits may tend to employ certain socio-emotional skills (McAdams, 1995). For program and policy purposes, however, there is a key distinction between personality traits and socio-emotional skills: while personality traits are predictive of labor market outcomes, they are less amenable to direct change via intervention. Socio-emotional skills, however, have been shown to be malleable to various intervention efforts across cultures (e.g., Jones, Brown, and Aber, 2011; Torrente et al., 2014). In turn, building socio-emotional skills can result in changes to enduring patterns of thinking and behaving (Dweck, 2008).

Table B1. Original 33 Items Included in the Non-Cognitive Skills Module[10]

Personality Traits	*Extraversion*
	Are you talkative?
	Do you like to keep your opinions to yourself? Do you prefer to keep quiet when you have an opinion? (R)
	Are you outgoing and sociable, do you make friends easily?
	Conscientiousness
	When you perform a task, are you very careful?
	Do you prefer relaxation more than hard work? (R)
	Do you work very well and quickly?
	Openness
	Do you come up with ideas others haven't thought of before?
	Are you interested in learning new things?
	Do you enjoy beautiful things, like nature, art, and music?
	Emotional Stability
	Are you relaxed during stressful situations?
	Do you tend to worry? (R)
	Do you get nervous easily? (R)
	Agreeableness
	Do you forgive other people easily?
	Are you very polite to other people?
	Are you generous to other people with your time or money?
	Grit
	Do you finish whatever you begin?
	Do you work very hard? For example, do you keep working when others stop to take a break?
	Do you enjoy working on things that take a very long time to complete?
Socio-emotional Skills	*Hostile Bias*
	Do people take advantage of you?
	Are people mean/not nice to you?
	Decision Making
	Do you think about how the things you do will affect your future?
	Do you think carefully before you make an important decision?
	Do you ask for help when you don't understand something?
	Do you think about how the things you do will affect others?

[10] All items except the Fixed Versus Growth Mindset items were scaled using a 4-point Likert scale (1 = Almost always; 4 = Almost never). The Fixed Versus Growth Mindset items employed a 6-point Likert scale (1 = Totally agree; 6 = Strongly disagree). Items that are marked with an (R) were reverse coded so that a low value indicates the same valence of response on every item.

Achievement Striving	
	Do you do more than is expected of you?
	Do you strive to do everything in the best way?
	Do you try to outdo others, to be best?
Self Control	
	Do you spend more than you can afford?
	Do you do crazy things and act wildly?
Fixed Versus Growth Mindset	
	The type of person you are is fundamental, and you cannot change much.
	You can behave in various ways, but your character cannot really be changed.
	As much as I hate to admit it, you cannot teach an old dog new tricks. You cannot change their most basic properties.
	You have a certain personality and not much can be done to change that.

Note: Items and scales in blue are personality trait measures, items and scales in orange are socio-emotional skill measures.

Analysis Strategy

Our initial analyses revealed three main issues with the data. First, correlations between items in the same groupings (e.g., openness, grit) were low—generally ranging from .2 to .4—suggesting that each item is measuring a different facet of the grouping. Second, sum-scoring items according to the 11 hypothesized constructs and computing reliability coefficients indicated the scores were composed of a significant degree of measurement error. Third, the distribution of item responses across the Likert scales deviated substantially from normality, invalidating the assumptions inherent in traditional statistical measurement techniques. To address these issues, factor analyses were conducted in a multi-step process.

Box B1: What is Factor Analysis and Why Do We Use it?

Factor analysis is a statistical technique that can be used to examine the relationship between observed items or *indicators* (see Table B1, above) and unobserved latent constructs or *factors* that are hypothesized to underlie the associations between indicators (in this study, openness, conscientiousness, etc.). There are three primary goals of or reasons to use factor analysis: (1) data reduction; (2) scale structure; and (3) to reduce measurement error. First, survey instruments provide a lot of data—some surveys to assess adult personality factors include over 500 items. Not only is it not practical to analyze that much data, but testing effects on multiple discrete indicators increases the likelihood of having a "false positive," or Type I error. Factor analysis assists with data reduction by establishing a lesser number of factors that account for the variation between indicators. Second, surveys are frequently designed to capture multiple constructs (in our study, various personality traits and socio-emotional skills) using items that may relate more strongly to some constructs than others. For example, in our study, the item "Do you think about how the things you do will affect your future?" may be a better indicator of Decision Making than, "Do you ask for help when you don't

understand something?" Factor analysis allows us to understand the *internal scale structure* by quantifying the number of factors in the data and the extent to which items are related to each factor. Finally, when we administer a survey measure or test, we want to ensure that the variability in scores is due to what we are trying to measure—in this study, personality traits or socio-emotional skills—as opposed to error or bias. Traditional or unrefined methods of scoring—such as summing the survey items—do not account for this measurement error, leading to biases in regression analyses. Factor analysis allows us to adjust for *measurement error* by fitting an underlying model accounting for both variation among observed items and random error variance.

There are two primary types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). While both EFA and CFA attempt to model the relationship between observed items using a smaller set of latent constructs, they differ in the *a priori* restrictions that are placed on the model. EFA is a data-driven technique that is primarily used when the factor structure (e.g., the appropriate number of underlying factors and the relationships of the items to the factors) is unknown, whether because the survey has never been administered before or is being administered in new contexts. In CFA, a researcher uses a strong theoretical foundation to specify at the outset the number of hypothesized factors and the patterns of how the items relate to the factors. This solution is then evaluated with respect to how well it fits the observed data. EFA is used most frequently early in the process of scale development, while CFA is used once the researcher has established the factor structure based on prior empirical and theoretical grounds.

Given that the non-cognitive skills module has never before been administered in the countries of interest in this study, we decided to proceed by first conducting exploratory factor analyses (EFAs) with an oblimax rotation on a randomly selected half of participants stratified by country (N = 3,885).[11] In doing so, we are not making *a priori* assumptions about the factor structure of the module in these new contexts. Then, to support the EFA results, the factor structure was confirmed (in a confirmatory factor analysis, or CFA) using the second half of the sample (N = 3,887). All analyses were conducted in MPlus (Muthén and Muthén, 1998–2012; Version 6.12) and adjusted for any clustering of the data due to sampling design.[12] Responses were treated as ordered categorical data to account for the skewed nature of the data, and full information maximum likelihood (FIML) estimation was employed to handle missing data.[13]

Once we determined a factor structure that provided a good fit to the data, we created individual scores on each of these factors using refined factor scoring techniques. As detailed above, factor scoring is preferable in this case to traditional sum scoring methods given that factor scores account

[11] An oblimax rotation was chosen to account for the hypothesized correlation between factors.

[12] In Tajikistan—but not in Uzbekistan or Kyrgyzstan—up to two individuals per household were administered the non-cognitive skills module. To account for any non-independence of the data that may occur due to individuals being nested in households, we used the Type=Complex and Cluster=psuid commands in MPlus.

[13] FIML utilizes all available data points, even for cases with missing item responses, by assessing during parameter estimation missing data patterns as well as by using information from all available data points. While FIML does not impute missing data, its use of information from all observed data is conceptually similar to missing data imputation, where a missing value is computed conditioned on several other included variables (Muthén, Kaplan and Hollis, 1987). In this sample, 120 cases did not have data on any of the items and were removed from the analysis.

for: (1) the weight of individual item loadings; and (2) shared variance between the items and the factors *and* measurement error (DiStephano, Zhu, and Midrila, 2009). Factor scores were created based on the exploratory factor analysis solution using maximum a posteriori (MAP) estimation in MPLUS, which accounts for the non-normal distribution of item response (Muthén and Muthén, 1998-2012).

Results

The initial EFA revealed two groupings of items: those that loaded well onto one factor, and those that did not. The 4 items making up the original construct of "Fixed Versus Growth Mindset" loaded cleanly onto one factor. This factor was left intact and removed from the exploratory analyses; it was subsequently confirmed to provide a good fit to the data (χ^2 (2) = 27.52, CFI: .996, RMSEA: .057, .039 < 95% CI < .077).[14] Also removed from analyses at this juncture were items that loaded below .2 on any construct and items that were reverse coded due to factor-item correlations in unexpected directions. We then chose a 4-factor solution to model associations between the remaining 18 items; in this solution, items were allowed to cross-load on multiple factors and factors were allowed to correlate.[15] This model provided an excellent fit to the data (χ^2 (87) = 530.89, CFI=.985, RMSEA=.036, .033 < 95% CI < .039) while modeling the observed indicators parsimoniously.

The four identified factors described in Table B2, below, were: (1) Openness to New Ideas and People (5 items; e.g., "Are you outgoing and sociable?"; "Are you interested in learning new things?"); (2) Workplace Attitude and Behavior (5 items; e.g., "Do you enjoy working on things that take a very long time to complete?"; "Are people mean/not nice to you?"); (3) Decision Making (5 items; e.g., "Do you think about how the things you do will affect others?"; "Do you think carefully before making an important decision?"); and (4) Achievement Striving (3 items; "Do you do more than is expected of you?"). As detailed above, confirmatory factor analysis confirmed the fit of this model (χ^2 (129) = 2336.52, CFI=.922, RMSEA=.066, .064 < 95% CI < .069). In addition, preliminary measurement equivalence analyses indicate that this same factor structure provides a good fit to the data in Uzbekistan, Kyrgyzstan, and Tajikistan (χ^2 (459) = 69484.24, CFI=.932, RMSEA=.068, .066 < 95% CI < .070).[16]

---

[14] In assessing model goodness of fit, the following criteria are used: A RMSEA < .08 provides an acceptable fit to the data, while an RMSEA < .05 provides a good fit to the data; a CFI > .9 provides an acceptable fit to the data while a CFI > .95 provides a good fit to the data (Kline, 2011).

[15] Factor correlations in the final EFA ranged from .1 to .65. The highest correlations were: Openness-Decision Making (.535), Openness-Achievement Striving (.556), and Decision Making-Achievement Striving (.65).

[16] Tests of measurement invariance seek to establish whether we are measuring the same construct in the same way across different groups. As of this writing, our preliminary analyses have established *configural invariance:* that the same factor structure (e.g., the same number of factors and the same pattern of loadings) exists in the samples from all three countries. Future analyses will examine other levels of invariance, establishment of which increases our certainty that observed differences between countries is attributable only to true differences in the variability of the scores.

Table C2. Unstandardized Results from Final CFA of Non-Cognitive Skills Module

		Loading	SE
Extraversion			
1.	Are you talkative?	0.502	0.015
2.	Are you outgoing and sociable, do you make friends easily?	0.672	0.012
3.	Are you interested in learning new things?	0.635	0.013
4.	Do you enjoy beautiful things, like nature, art, and music?	0.528	0.015
5.	Are you very polite to other people?	0.648	0.013
Workplace Attitudes and Behaviors			
6.	Do you come up with ideas others haven't thought of before?	0.575	0.019
7.	Do you work very hard? For example, do you keep working when others stop to take a break?	0.693	0.018
8.	Do you enjoy working on things that take a very long time to complete?	0.506	0.019
9.	Do people take advantage of you?	0.360	0.020
10.	Are people mean/not nice to you?	0.207	0.024
Decision Making			
11.	Do you finish whatever you begin?	0.622	0.013
12.	Do you think about how the things you do will affect your future?	0.673	0.011
13.	Do you think carefully before you make an important decision?	0.683	0.011
14.	Do you ask for help when you don't understand something?	0.592	0.013
15.	Do you think about how the things you do will affect others?	0.669	0.011
Achievement Striving			
16.	Do you do more than is expected of you?	0.587	0.014
17.	Do you strive to do everything in the best way?	0.723	0.013
18.	Do you try to outdo others, to be best?	0.463	0.016
Fixed Versus Growth Mindset			
19.	The type of person you are is fundamental, and you cannot change much.	0.678	0.009
20.	You can behave in various ways, but your character can not really be changed.	0.711	0.009
21.	As much as I hate to admit it, you cannot teach an old dog new tricks. You cannot change their most basic properties.	0.697	0.008
22.	You have a certain personality and not much can be done to change that.	0.704	0.008

Interpretation and Future Directions

Our analyses indicated that the data from the non-cognitive skills module is best represented by five factors that correspond to some—but not all—of the 11 personality traits and socio-emotional skills described in Table B1. For example, our analyses indicated that items 19–22 and 16–18 index the hypothesized underlying socio-emotional skills Fixed Versus Growth Mindset and Achievement Striving, respectively. Substantively, this indicates that individuals that have higher Achievement Striving factor scores tend to strive to go "above and beyond" and to do more than is expected of

them, while individuals who have higher Fixed Versus Growth Scores tend to believe new skills can be learned.

The other three factors represented in the data are combinations of items meant to index both personality traits and socio-emotional skills; this pattern of relationships can be understood in that certain personality traits tend to be related to certain learned attitudes and skills. For example, our factor of Decision Making consists of items originally thought to index both decision-making skills and the trait of grit. In this case, individuals who think carefully and thoroughly about the repercussions of their decisions and behaviors (see items 12–15) tend to follow through with their actions (see item 11)—perhaps anticipating the repercussions of not following through. Our factor of Workplace Attitudes and Behaviors consists of items meant to index both Grit and Hostile Bias. Individuals who work very hard when others take a break (see items 6–8) may tend to feel that others take advantage of them or are mean (see items 9–10). Thus individuals who score higher on this construct may be workers who work hard and are innovative but perceive interactions with others as hostile; individuals who score lower on this construct tend to work less hard and on discrete projects, without perceiving workplace interactions as negative. Finally, our construct of Openness to New Ideas and People reflects items thought to index the personality traits of extraversion, agreeableness, and openness. Individuals who score high on this construct are social and open to new ideas, people, and things (see items 1–5).

There are two plausible reasons why the data did not reflect the expected 11 traits and skills. First, only 2–4 items were used to originally index each trait/skill; this may not have been enough to validly and reliably fully "capture" the constructs of interest. Instead, these items appear to reflect weak to moderately related aspects of a trait/skill that co-vary with aspects of other traits/skills. This is unsurprising given demonstrated correlations between: (a) Big Five personality traits (Digman, 1997); and (b) personality traits and socio-emotional skills (McAdams, 1995). To address this issue, future surveys should consider including a broader range of items to represent each trait/skill. A second explanation that we cautiously proffer is that the items do not relate to each other in the same way in Tajikistan, Uzbekistan, and Kyrgyzstan as in the samples from which the items were developed. For example, in the Grit scale in this sample, "finishing what was begun" is not related to "enjoying working on things that take a long time to complete." In ECA contexts, grit might not be well-indexed by such behaviors. To investigate this, future research should: (1) conduct qualitative research to better understand how these traits and skills are understood in ECA contexts; and (2) test for measurement invariance between the non-cognitive items administered in this study and in other studies.

References

Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. (2008). The economics and psychology of personality traits. Journal of Human Resources, 43, 972-1059.

CASEL. (2013). What is socioemotional learning? Retrieved from: http://www.casel.org/social-and-emotional-learning

Dodge, K. A. (2003). Do social information processing patterns mediate aggressive behavior? In B. Lahey, T. Moffitt, and A. Caspi (Eds.), Causes of conduct disorder and juvenile delinquency (pp. 254-274). New York: Guilford Press.

DiStefano, C., Zhu, M., and Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research and Evaluation, 14(20), 1-11.

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

Dweck, C. S. (2008). Can personality be changed? The role of beliefs in personality and change. Current Directions in Psychological Science, 17(6), 391-394.

Goldberg, L. R. (1990). An alternative description of personality: The Big-Five factor structure. Journal of Personality and Social Psychology, 59(6), 1216.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality, 40(1), 84-96.

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

Jones, S. M., Brown, J. L., and Lawrence Aber, J. (2011). Two- Year Impacts of a Universal School-Based Social- Emotional and Literacy Intervention: An Experiment in Translational Developmental Research. Child Development, 82(2), 533-554.

Kline, R. B. (2011). Principles and practice of structural equation modeling. New York: Guilford Press.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., and Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. Behavior Research Methods, 43(2), 548-567.

Mann, L., Burnett, P., Radford, M., and Ford, S. (1997). The Melbourne Decision Making Questionnaire: An instrument for measuring patterns for coping with decisional conflict. Journal of Behavioral Decision Making, 10(1), 1-19.

McAdams, D. P. (1995). What do we know when we know a person?. Journal of personality, 63(3), 365-396.

Muthén, B., Kaplan, D., and Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika, 42, 431-462.

Muthén, L.K., and Muthén, B.O. (1998-2012). Mplus user's guide. Seventh Edition. Los Angeles, CA: Muthén and Muthén.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. Journal of Personality and Social Psychology, 84, 411-422.

Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. Journal of Personality and Social Psychology, 94, 718-737.

Torrente, C., Alimchandani, A., and Aber, J. (in press). International perspectives on social-emotional learning. Handbook on social and emotional learning: research and practice.

Torrente, C., Johnston, B., Seidman, E., and Gross, A. (2014). Improving learning environments and children's social-emotional wellbeing in the Democratic Republic of the Congo: Preliminary results from a cluster randomized trial. Paper presented at Society for Research on Educational Effectiveness (SREE), Washington, DC.

Yeager, D. S., and Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. Educational Psychologist, 47(4), 302-314.