Brief Conscientiousness Scales: How Low Can You Go?

Kate E. Walton

When an assessment designer develops a scale or a practitioner selects a scale for use, length is often an important consideration. Administration time is clearly a key concern, and shorter assessments are often more desirable. For example, brief assessments are more suitable for research with populations lacking the cognitive or emotional resources needed to withstand lengthy tests (Allen et al., 2022). Moreover, shorter assessments may enhance data fidelity. That is, participants are generally more willing to complete a brief assessment than a lengthy one, and they are generally more inclined to spend more time and effort on their responses. Some researchers and practitioners even promote the use of single-item measures (e.g., Allen et al., 2022).

There can, however, be drawbacks to extremely brief measures. A primary argument is that single-item or brief measures lack sufficient evidence of reliability (generally defined as the amount of error in a measurement; American Educational Research Association [AERA] et al., 2014). A second argument is that brief measures may also have insufficient evidence of validity (defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"; AERA et al., 2014, p. 11) because they may not capture all the complexities of multidimensional psychological constructs (Allen et al., 2022). As an example, the personality trait of conscientiousness has been shown to be a hierarchical factor with six facets (Roberts et al., 2005). A brief three- or four-item scale may not have the breadth necessary to capture all important aspects of conscientiousness and therefore may not have robust validity evidence.

There is certainly a tradeoff between scale length and psychometric concerns. The two are, in fact, directly linked. Generally, when scales are shortened, reliability is reduced, and when scales are lengthened, reliability is improved, provided the items added to the scale are comparable psychometrically (AERA et al., 2014). Scale reliability, in turn, affects validity. Typically, to the extent that scores are unreliable (i.e., reflect measurement error), their ability to accurately predict criteria is compromised.

In the present collection of studies, we examine whether shortened conscientiousness scales maintain acceptable levels of reliability and validity. We also examine whether scale shortening results in unintentional exaggerated subgroup differences. We focus on conscientiousness (defined as "socially prescribed impulse control that facilitates task- and goal-directed behavior, such as thinking before acting, delaying gratification, following norms and rules, and planning, organizing, and prioritizing tasks"; John & Srivastava, 1999, p. 121) because this trait has the strongest associations with academic and job performance (Zell & Lesick, 2021).

Study 1

Participants and Measures

Participants were 24,400 students in Grades 6–8. Twelve thousand two hundred seventy-three (50.3%) students identified as female, and the rest identified as male. Information on race/ethnicity was not gathered for nearly 95% of the sample; therefore, we were unable to examine differences across racial groups.

Students took the Mosaic™ by ACT®: Social Emotional Learning Assessment (for more information, see ACT, 2021). The assessment measures five skills, one of which is sustaining effort, which aligns with conscientiousness (ACT, 2021). The assessment is multi-method, containing three distinct item types, but for the purposes of the current study, we focus on the Likert items. There are eight Likert items measuring conscientiousness.

Analyses and Results

Reliability

We computed Cronbach's alpha for the full eight-item scale and proceeded to remove one item at a time until only three remained. The item removed at each step was the one that would either have the least detrimental effect on reliability or improve the reliability estimate the most. Note that estimates of .70 or higher are generally considered acceptable. Alpha values, found in Table 1, ranged from .68 for the three-item scale to .76 for the six-item scale.

Table 1. Study 1 Reliability Estimates

Number of items	Cronbach's alpha
8	.71
7	.72
6	.76
5	.74
4	.71
3	.68

Validity

We correlated the full-length and shortened scales with GPA (which students self-reported on a 12-point scale) to evaluate the effect of shortening the scale on test-criterion validity (i.e., whether the scale predicts a given criterion). Table 2 provides the correlations as well as the squared correlations, which equate to the amount of variance in GPA accounted for by conscientiousness. The amount of variance decreased from 17% for the seven- and eight-item scales to 12% for the three-, four-, and five-item scales.

Table 2. Study 1 Test-Criterion Validity Estimates

Number of items	*r*	r^2
8	.41	.17
7	.41	.17
6	.36	.13
5	.35	.12
4	.35	.12
3	.35	.12

We next examined evidence for content validity (i.e., the extent to which the item content covers the construct it is intended to measure). Seven subject matter experts mapped the items to six conscientiousness facets—industriousness, order, self-control, responsibility, traditionalism, and virtue (Roberts et al., 2005). With an eight-item scale, four facets were covered, and with scales of between three and seven items, three facets were covered (see Table 3).

Table 3. Study 1 Content Validity Counts

Number of items	Number of facets covered
8	4
7	3
6	3
5	3
4	3
3	3

Group Differences

We examined whether the magnitude of the difference between female and male scores remained constant across scales of varying length. Standardized mean differences (*d*), found in Table 4, decreased in magnitude as scale length decreased.

Table 4. Study 1 Gender Differences

Number of items	*d*
8	0.22
7	0.23
6	0.19
5	0.18
4	0.17
3	0.15

Note. Positive values indicate that female students scored higher than male students.

Study 2

Participants and Measures

We sought participation from a random sample of 30,000 students who took the ACT® test on the September 2023 national test date. They were not incentivized to participate, and they were assured that their involvement and responses would not impact their ACT scores.

We have complete data for 1,707 participants. One thousand one hundred ninety-eight (70.2%) participants identified as female, 474 (27.8%) identified as male, seven (0.4%) identified as another gender, 27 (1.6%) preferred to not respond, and the information was missing for one participant. One thousand one hundred thirty (66.2%) participants identified as White, 182 (10.7%) identified as Asian, 137 (8.0%) identified as Hispanic/Latino, 94 (5.5%) identified as Black/African American, 79 (4.6%) identified as two or more races, one (0.1%) identified as American Indian/Alaska Native, 79 (4.6%) preferred to not respond, and the information was missing for one participant.

Participants completed a six-item conscientiousness scale developed by two subject matter experts (Walton & Anguiano-Carrasco, 2024). They also self-reported their GPA and responded to two items designed to evaluate the test-criterion validity evidence of the conscientiousness scale. One item asked about the tendency to challenge oneself to work harder, and the second asked about the tendency to check homework for errors before turning it in.

Analyses and Results

Reliability

We computed Cronbach's alpha for the full six-item scale and removed one item at a time until only three remained (Table 5). Cronbach's alpha decreased from .68 for the full-length scale to .61 for the three-item scale.

Table 5. Study 2 Reliability Estimates

Number of items	Cronbach's alpha
6	.68
5	.66
4	.63
3	.61

Validity

We correlated the full-length and shortened scales with the three outcomes to evaluate the effect of shortening the scale on test-criterion validity (Table 6). For each outcome, the amount of variance accounted for was higher for the three-item scale than for the full-length scale.

Table 6. Study 2 Test-Criterion Validity Estimates

Number of items	GPA *r*	GPA r^2	Challenging oneself *r*	Challenging oneself r^2	Checking work *r*	Checking work r^2
6	.26	.07	.33	.11	.35	.12
5	.27	.07	.33	.11	.35	.12
4	.28	.08	.36	.13	.36	.13
3	.29	.08	.34	.12	.38	.14

We next examined content validity evidence using the same six-facet solution as in Study 1 (Roberts et al., 2005; also used in Studies 3 and 4). Four facets were covered with the six-item scale, three facets were covered with the four- and five-item scales, and two facets were covered with the three-item scale (see Table 7).

Table 7. Study 2 Content Validity Counts

Number of items	Number of facets covered
6	4
5	3
4	3
3	2

Group Differences

We examined whether the magnitude of the difference between female and male scores remained constant across scales of varying length. We also examined the magnitude of the differences between students identifying as White and those identifying as Asian, Black, or Hispanic. Results can be found in Table 8. Female–male and White–Black differences were greater for the three-item scales than for the six-item scales, but White–Asian and White–Hispanic differences were smaller for the former than for the latter.

Table 8. Study 2 Subgroup Differences

Number of items	Female–male *d*	White–Asian *d*	White–Black *d*	White–Hispanic *d*
6	0.21	0.29	0.26	0.20
5	0.23	0.25	0.27	0.19
4	0.30	0.29	0.36	0.29
3	0.26	0.19	0.32	0.18

Note. Positive values indicate that female students scored higher than male students and White students scored higher than students from the other groups.

Study 3

Participants and Measures

The participants of Study 3 were from an online Amazon Mechanical Turk (MTurk) sample of 1,768 adults with a mean age of 36 years (*SD* = 11.1). One thousand one hundred four (57.4%) identified as male, 744 (42.1%) identified as female, and 10 (0.1%) identified as other or declined to respond. One thousand one hundred forty-six (64.8%) identified as White, 359 (20.3%) identified as Asian, 134 (7.6%) identified as Black/African American, 83 (4.7%) identified as Hispanic/Latino/of Spanish origin, 27 (1.5%) identified as American Indian/Alaska Native, and 19 (1.1%) identified as other.

Participants completed a multi-trait multi-method assessment, ACT® WorkKeys® Essential Skills(for more information, see ACT, 2024). For the purposes of this study, we focused on eight Likert items that measure the construct of work ethic, which aligns to conscientiousness. Participants also completed a measure of the Big Five, the 10-item Big Five Inventory (BFI-10; Rammstedt & John, 2007). The Big Five factors are conscientiousness, agreeableness, emotional stability, openness to experience, and extraversion. Because participants took both assessments, we were able to evaluate the full-length and shortened scales' convergent and discriminant validity evidence. The work ethic scale should correlate most strongly with the BFI-10 conscientiousness scale (convergent validity) and should correlate to a much lesser degree with the other, unrelated scales (discriminant validity).

Analyses and Results

Reliability

We computed Cronbach's alpha for the full eight-item scale and removed one item at a time until only three remained (Table 9). The reliability of the three-item scale (.82) was slightly lower than that of the full-length scale (.85).

Table 9. Study 3 Reliability Estimates

Number of items	Cronbach's alpha
8	.85
7	.86
6	.85
5	.84
4	.84
3	.82

Validity

We correlated the full-length and shortened scales with the BFI-10 to evaluate the effect of shortening the scale on convergent and discriminant validity (Table 10). For conscientiousness, convergent validity estimates ranged from .53 for the three-item scale to .58 for the six- and

seven-item scales. Discriminant validity estimates generally improved as the scale length decreased.

Table 10. Study 3 Convergent and Discriminant Validity Estimates

Number of items	C	A	ES	O	E
8	.57	.34	.30	.27	.16
7	.58	.35	.32	.27	.13
6	.58	.34	.33	.26	.13
5	.57	.33	.33	.24	.15
4	.57	.32	.34	.23	.18
3	.53	.32	.31	.22	.14

Note. C = conscientiousness. A = agreeableness. ES = emotional stability. O = openness to experience. E = extraversion.

We next examined content validity evidence. Scales containing six to eight items covered four facets, and scales containing three to five items covered three facets (see Table 11).

Table 11. Study 3 Content Validity Counts

Number of items	Number of facets covered
8	4
7	4
6	4
5	3
4	3
3	3

Group Differences

As we did in Study 2, we examined gender and racial/ethnic group differences across scale lengths (see Table 12). The magnitude of difference between female and male participants decreased with shorter scales, as did the magnitude of difference between White and Black participants. The differences were slightly higher for White–Asian and White–Hispanic comparisons with the three- vs. eight-item scale.

Table 12. Study 3 Subgroup Differences

Number of items	Female–male *d*	White–Asian *d*	White–Black *d*	White–Hispanic *d*
8	0.27	0.12	−0.10	−0.01
7	0.22	0.19	−0.08	−0.00
6	0.21	0.17	−0.07	0.01
5	0.20	0.17	−0.04	−0.01
4	0.17	0.13	−0.02	0.00
3	0.21	0.17	−0.01	0.06

Note. Positive values indicate that female participants scored higher than male participants and White participants scored higher than participants from the other groups.

Study 4

Participants and Measures

Study 4 included an online sample of 173 employees with a mean age of 43 years (*SD* = 13.3). One hundred eighteen (68.2%) identified as female, 54 (31.2%) identified as male, and one (0.6%) declined to respond. One hundred eleven (64.2%) identified as White, 35 (20.2%) identified as Black/African American, 16 (9.2%) identified as Hispanic/Latino/of Spanish origin, four (2.3%) identified as Asian, one (0.6%) identified as American Indian/Alaska Native, and six (3.5%) identified as other.

As in Study 3, employed participants completed the WorkKeys Essential Skills assessment. Again, we focused on eight Likert items that measure work ethic, which aligns to conscientiousness. The employees' supervisors rated their job performance on a 24-item performance measure.

Analyses and Results

Reliability

We computed Cronbach's alpha for the full eight-item scale and removed one item at a time until only three remained (Table 13). This resulted in the same order of item elimination as in Study 3 (therefore, content validity counts remain the same). Estimates ranged from .82 for the three-item scale to .87 for the eight-item scale.

Table 13. Study 4 Reliability Estimates

Number of items	Cronbach's alpha
8	.87
7	.88
6	.88
5	.87
4	.85
3	.82

Validity

We correlated the full-length and shortened scales with the supervisor ratings to evaluate the effect of shortening the scale on test-criterion validity (Table 14). The amount of variance in supervisor ratings decreased from 7% for the full-length scale to 4% for the three-item scale.

Table 14. Study 4 Test-Criterion Validity Estimates

Number of items	*r*	r^2
8	.27	.07
7	.28	.08
6	.29	.08
5	.26	.07
4	.21	.04
3	.20	.04

Group Differences

We again examined gender and racial/ethnic group differences across scale lengths (see Table 15). Given the small sample size, we did not examine differences between Asian and White participants or between Hispanic and White participants. Female–male differences ranged from .02 for the six-item scale to .08 for the three-item scale. The White–Black difference was in the opposite direction and was more pronounced for the three-item scale than the eight-item scale.

Table 15. Study 4 Subgroup Differences

Number of items	Female–male *d*	White–Black *d*
8	0.06	−0.03
7	0.03	−0.03
6	0.02	0.01
5	0.04	0.07
4	0.04	0.14
3	0.08	0.22

Note. Positive values indicate that female employees scored higher than male employees and White employees scored higher than Black employees.

Conclusions

As expected, the reliability estimates were worse for the shortened scales; however, the differences were negligible, and in most cases, the shortened scales still reached an acceptable level of reliability. In some instances (Study 2), test-criterion validity estimates actually improved for the shortened scales. In the instances when they deteriorated as scale length decreased, the effects were small. Convergent validity estimates decreased slightly for the shortened scales, but discriminant correlations did as well, which is a benefit. As expected, scales with fewer items had poorer content validity. Finally, with one exception (Study 4, White–Black differences), subgroup differences were smaller (or nearly the same) for brief scales than full-length scales.

In general, shortening conscientiousness scales had little to no negative impact on reliability and several validity estimates, and we therefore argue that brief conscientiousness scales can be used in place of lengthier ones.

References

ACT. (2024). *ACT® WorkKeys® Essential Skills Technical Manual*. https://www.act.org/content/dam/act/unsecured/documents/pdfs/ACT-WorkKeys-Essential-Skills-Technical-Manual.pdf

ACT. (2021). *Mosaic™ by ACT®: Social Emotional Learning Assessment*. https://www.act.org/content/dam/act/unsecured/documents/mosaic-sel-tech-manual.pdf

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldbert, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, *58*(1), 103–139. https://doi.org/10.1111/j.1744-6570.2005.00301.x

Walton, K. E., & Anguiano-Carrasco, C. (2024). *Generating social and emotional skill items: Humans vs. ChatGPT*. ACT.

Zell, E., & Lesick, T. L. (2021). Big Five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, *90*(4), 559–573. https://doi.org/10.1111/jopy.12683

ABOUT ACT

ACT is transforming college and career readiness pathways so that everyone can discover and fulfill their potential. Grounded in more than 65 years of research, ACT's learning resources, assessments, research, and work-ready credentials are trusted by students, job seekers, educators, schools, government agencies, and employers in the U.S. and around the world to help people achieve their education and career goals at every stage of life. Visit us at www.act.org.