Measuring Durable Skills: Does Participant Engagement Vary Across Item Types?

Kate E. Walton and Jeff Schiel

High participant engagement (i.e., being interested and focused) is key to ensuring high-quality data in research and applied settings. Lack of engagement may bring about fatigue and careless responding, which, in turn, could weaken reliability and validity estimates in research and lead to inaccurate estimates of a participant's tendencies or skill levels in both low- and high-stakes practical applications. To combat careless responding when assessing durable skills (e.g., collaboration and resilience), some researchers and practitioners rely on item types that are deemed more stimulating than typical self-report Likert items, for example. Some (e.g., ACT, 2024) use situational judgment tests (SJTs), forced-choice measures, or performance-based measures in addition to, or in place of, Likert items.

The real-world feel and face validity of SJTs, for example, leave assessment developers and users with the assumption that these items are relatively stimulating for examinees (Lievens et al., 2008). Indeed, researchers have anecdotally reported that participants experience higher engagement with SJTs (e.g., Olaru et al., 2019); however, to our knowledge, no data that speak directly to this issue have been reported in the literature. In the current study, we directly compare indicators of engagement across four item types.

Current Study

Participants

Participants were 2,861 students in Grades 10–12 who registered to take the April 2026 National ACT® test. Demographic information collected at the time of ACT registration indicates that 63.6% of these students were female and 34.4% were male (1.9% preferred not to respond, and the information was missing for one individual). Of the 2,858 students for whom race/ethnicity information was available, 1,843 were White (64.4%), and the next three groups with the largest percentages of participants were Hispanic/Latino (10.8%), Asian (7.1%), and Black/African American (6.4%).

Procedure

After taking the ACT test, a random sample of 60,000 students in Grades 10–12 received an email inviting them to participate in unincentivized research by completing a brief (i.e., less than 2 minutes) survey. The email simply said that the research was on skills, tendencies, and preferences, and that ACT was interested in learning more about these characteristics and how to measure them. Students were randomly assigned to one of four conditions in the survey, each of which presented a different set of item types. Note that condition assignment served as the independent variable in this study, and students' actual responses to the items described below were not of interest.

© 2026 by ACT Education Corp. All rights reserved. | R2618

- In one condition (*n* = 715), students were administered Likert items measuring Conscientiousness (3 items), Agreeableness (3 items), and Emotional Stability (3 items) on a 6-point scale, including Strongly Disagree, Disagree, Somewhat Disagree, Somewhat Agree, Agree, and Strongly Agree. An example of a Likert item is "I manage my time well."

- In another condition (*n* = 715), students were administered unidimensional forced-choice dyads measuring Conscientiousness (3 dyads), Agreeableness (3 dyads), and Emotional Stability (3 dyads). Participants were instructed to select the statement that was most like them (e.g., "I usually handle stress well" versus "I rarely handle stress well").

- In the third condition (*n* = 715), students were administered three SJTs (one measuring Conscientiousness, one measuring Agreeableness, and one measuring Emotional Stability) that had three behavioral responses each. Students were instructed to indicate their likelihood of exhibiting each of the three behaviors on a 5-point scale including Very Unlikely, Somewhat Unlikely, Neither Unlikely Nor Likely, Somewhat Likely, and Very Likely. An example SJT scenario is "Your two best friends have gotten into an argument. How likely are you to do each of the following?" Three behavioral responses follow:

    - Work with them together to resolve the conflict.

    - Talk to each of them separately to find out what happened.

    - Ignore them both until they hopefully work things out.

- In the final condition (*n* = 716), students were presented with the start of nine sentences (e.g., I dislike . . .) and asked to complete each one. No particular constructs were deliberately assessed with the sentences.

At the end of the survey, students in all conditions were asked to complete three 6-point Likert items (Cronbach's alpha = .84) to gauge their engagement with the survey (the dependent variable). This scale was newly created for this research. These items read as follows:

- I found this survey to be engaging.

- I would like to complete this type of survey again in the future.

- This survey was more fun than most other surveys I have completed.

Median completion times in seconds are as follows (with the average time per item in parentheses):

- Condition 1/Likert: 68 (7.6)

- Condition 2/forced-choice: 93 (10.3)

© 2026 by ACT Education Corp. All rights reserved. | R2618

- Condition 3/SJT: 116 (12.9)
- Condition 4/sentence-completion: 257 (28.6)

Analyses and Results

To assess engagement, we examined completion rates across conditions. We first calculated the percentage of students who completed the first item administered in their condition out of all students assigned to that condition (see Table 1). While first-item completion rates were similar among the Likert, forced-choice, and SJT conditions, with rates surpassing 80%, the first-item completion rate for the sentence-completion condition was much lower at just over 41% (second column of Table 1). Likewise, far fewer participants in the sentence-completion condition (30.3%) completed the entire survey, which included the engagement items, than did participants in the other three conditions (e.g., the entire-survey completion rate for the Likert condition was 79.9%; third column of Table 1). After completing the first item in their condition, the students were more likely to complete the entire survey, but rates varied across conditions (fourth column of Table 1). Once again, the completion rate was lowest among those in the sentence-completion condition (73.6%), and this was considerably lower than the rate for the Likert condition, for example, which was nearly 98%.

Table 1. Completion Rates Across Conditions

Condition	Percentage of students who completed the first item	Percentage of students who completed the entire survey	Percentage of students who completed the entire survey out of those who completed the first item in their condition
Likert	81.7%	79.9%	97.6%
Forced-choice	86.0%	74.0%	86.2%
Situational judgment test	81.0%	69.8%	86.2%
Sentence-completion	41.2%	30.3%	73.6%

Next, we evaluated participant engagement across the four conditions by calculating group means (see Table 2) for the three-item engagement measure and comparing those means using a one-way analysis of variance (ANOVA). There was a statistically significant overall effect ($F_{3, 1,812} = 2.77$, $p = .040$). We carried out a post hoc Tukey's test to determine which groups statistically differed from one another. The Likert condition had significantly lower engagement than the sentence-completion condition (mean difference = −.24; 95% confidence interval = −.47, −.01; $p = .043$; $d = -.20$).

© 2026 by ACT Education Corp. All rights reserved. | R2618

Table 2. Descriptive Statistics for the Engagement Measure for Each Condition

Condition	*n*	*M*	*SD*
Likert	571	3.64	1.11
Forced-choice	529	3.74	1.12
Situational judgment test	499	3.79	1.05
Sentence-completion	217	3.88[a]	1.40

[a] This condition's level of engagement was significantly higher than the Likert condition's.

Ad Hoc Cognitive Ability Analyses

We surmised that cognitive ability might play a role in the effect of condition on level of engagement, so we carried out an analysis of covariance (ANCOVA) using ACT Composite scores as a covariate. Given research documenting a strong correlation between ACT scores and cognitive ability estimates (Koenig et al., 2008), ACT scores served as a proxy for cognitive ability. There was a statistically significant overall effect ($F_{3,\ 1,731} = 3.00$, $p = .030$). We carried out post hoc pairwise comparisons to determine which groups statistically differed from one another. The Likert condition had significantly lower engagement than the sentence-completion condition (adjusted mean difference = −.25; 95% confidence interval = −.49, −.00; $p < .048$).

Discussion

To our knowledge, little to no empirical evidence supports the idea that some item types are more engaging than others, so we carried out a study to determine whether this is the case. We found that participants were less likely to engage with and complete a survey when presented with a sentence-completion task than when presented with Likert items, forced-choice dyads, or SJTs. Likewise, after completing a sentence-completion task, individuals were less likely to complete a following three-item measure of self-reported engagement than if they had completed a series of Likert items, forced-choice items, or SJTs. This could be due to the fact that individuals in the sentence-completion condition took substantially longer to complete the entire survey than individuals in the other three conditions. However, as evidenced by the mean scores on the three-item engagement measure, participants self-reported higher engagement with the sentence-completion task than with the Likert items, and this effect held when we controlled for cognitive ability; in other words, if you can get buy-in and entice individuals to engage with a sentence-completion task, they will actually find it relatively stimulating. In this low-stakes study, participants had the option to withdraw at any time with no repercussions. Our findings, however, suggest that sentence-completion tasks might be preferred in a high-stakes scenario when people are less likely to withdraw. It would be prudent to take measures to optimize test taker engagement; yet, it is worth noting that the effect was small.

Also worth noting are the absolute values of the mean engagement scores. Across all conditions, they fell between 3 and 4 (on a 1–6 scale), which correspond to response options of Somewhat Disagree and Somewhat Agree, respectively. This indicates that participants found the durable skills assessments mildly engaging at best. One should also bear in mind that this sample of students self-selected into the study and therefore likely felt more engaged than the average student. Like many online surveys, this one had a considerable rate of nonresponse.

© 2026 by ACT Education Corp. All rights reserved. | R2618

We do not know whether nonresponding students' perspectives on engagement would have been similar to those of the responding students, and it is possible that significant bias occurred in the engagement measures. Moreover, although we controlled for cognitive ability in our ad hoc analyses, we should call out that the students were relatively high achieving, with a median ACT score of 25, which corresponds to the 83rd percentile.

Given the small effects, there is not overwhelming evidence to push a researcher toward using any one item type over another. Along with indicators of engagement, other factors ought to be considered, like the higher cognitive load associated with SJTs and the fact that individuals may not initially want to participate in a more time-consuming sentence-completion task.

© 2026 by ACT Education Corp. All rights reserved. | R2618

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

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, *36*, 153–160. https://doi.org/10.1016/j.intell.2007.03.005

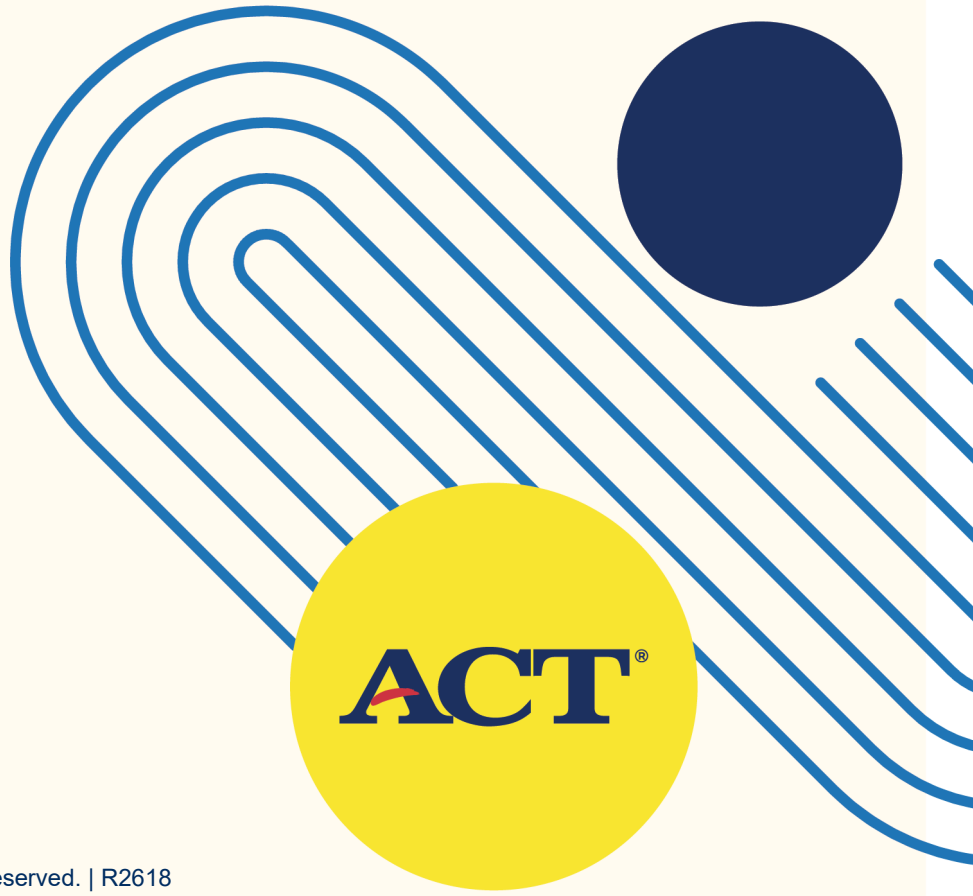
Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Olaru, G., Burrus, J., MacCann, C., Zaromb, F. M., Wilhelm, O., & Roberts, R. D. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS ONE*, *14*(2), 1–19. https://doi.org/10.1371/journal.pone.0211884

© 2026 by ACT Education Corp. All rights reserved. | R2618

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.



© 2026 by ACT Education Corp. All rights reserved. | R2618