Mosaic™ by ACT® Social Emotional Learning Assessment: Evaluating Student Group Differences Across Item Types

Kate E. Walton & Jeremy Burrus

Mosaic™ by ACT® Social Emotional Learning Assessment (hereafter referred to as Mosaic) is a multi-method assessment of five key social and emotional (SE) skills. The SE skills align with the Big Five personality traits and are called Sustaining Effort, Getting Along with Others, Maintaining Composure, Keeping an Open Mind, and Social Connection. Details about the assessment and evidence of its reliability and validity are reported by ACT (2021).

Multi-Method Assessment Approach

Most SE skill assessments rely solely on single stimulus Likert items where students indicate on a rating scale how much they agree that a particular statement reflects their skills (e.g., *How much do you agree with the following statement?: I get my schoolwork completed on time.*). Validity evidence for this item type is ample; however, like any item type, there are some associated disadvantages. For example, reference effects may influence students' responses because they have to ask themselves "*Compared to whom?*" For example, in very high achieving schools, it may be the case that most students complete schoolwork on time, and therefore might rate themselves lower on this item than if they were in a lower achieving school. In addition, if they felt the need to do so, students can easily engage in impression management on this item type. That is, it is plain to see that completing schoolwork on time is an attractive quality, so students might be inclined to inflate their responses to this item.

Given these issues, Mosaic includes Likert items in addition to two other item types, situational judgment tests (SJT) and forced choice (FC) items. The Mosaic SJTs entail providing a student with an imagined scenario, followed by five possible responses to that scenario.

Students are asked to indicate the likelihood of having each of those responses. SJTs are not immune to impression management, but research suggests that it is reduced with SJTs in comparison to single stimulus items (Hooper et al., 2006), and in light of this, some have argued for the use of SJTs even in high-stakes settings (Whetzel & McDaniel, 2016).

Another item type proposed to reduce the possibly of impression management is FC (Stark et al., 2011). On Mosaic, respondents are presented with three statements in blocks and asked to indicate the statement that is most and least like them. The blocks contain multidimensional items. For example, one item may capture intellectual interest (e.g., *I enjoy solving complex problems*), and one may capture the tendency to work hard (e.g., *I am a hard worker*). Impression management is difficult given that both items are desirable, yet respondents cannot select both items as being most like them. Meta-analytic data (Cao & Drasgow, 2019) suggest that single stimulus measures are subject to far more participant manipulation than FC measures in high-stakes scenarios; though certain characteristics of FC measures may make them more or less resistant to impression management (Cao & Drasgow, 2019; Walton et al., 2021). Moreover, reference effects are eliminated because the respondent is comparing self with self rather than self with other.

In operational use, the three item types are aggregated to produce a single score per SE skill. For the purposes of the current study, however, we examine them separately.

Group Differences

Previous research has examined group differences on SE skill-related constructs. Foldes and colleagues (2008) provide a meta-analytic overview of differences across racial groups and concluded that there were negligible differences at the broad skill level but some notable differences at the narrower facet level. Others have reported on large-scale studies of gender differences. Costa et al. (2001) examined data from over 23,000 participants from 26 cultures and concluded that gender differences are small relative to variation within genders, but

some effect sizes reached .44. This was based on a Likert-type measure, however, and likewise, the studies included in Foldes and colleagues' meta-analysis mainly included Likert items. Moreover, they did not examine item type as a moderator of the effects. Whetzel et al. (2008) performed a systematic review of mean race and gender differences on SJTs and concluded that White respondents outperformed Asian, Black, and Hispanic respondents, and females outperformed males. Drasgow et al. (2012) reported group (gender and ethnic) differences on their FC measure and concluded that differences were minimal, and some favored the minority or protected group.

ACT (2021) reported SE skill high school student group differences at the aggregate level, and the biggest difference was on Keeping an Open Mind (d = .27) with underrepresented students (all students who did not identify as White or Asian) scoring higher than White students (note that Asian students were not included in this group either because of a small sample size). Larger differences were observed when examining differences between female and male students with the highest reaching .46 (female students scored significantly higher on Getting Along with Others). It is important to determine whether certain item types lead to greater student group differences than others as this can help us understand processes at play when students respond to SE skill items and can inform future scale development and refinement. We carried out the current studies to examine the extent to which student group differences by race and gender emerge on SJT and FC SE skill measures compared to traditional Likert items.

Method

Participants

Participants were high school students in grades 9-12 who completed Mosaic between August 2020 and January 2021 (N = 3,720). Students' ages ranged from 13 to 18 (M = 15.31 years, SD = 1.21). The gender breakdown was as follows: female = 1,894, male = 1,711, another gender = 45, and 70 did not provide this information. The race/ethnicity breakdown was

as follows: American Indian or Alaska Native = 55, Asian = 123, Black or African American = 324, Hispanic or Latino/a = 503, Native Hawaiian or other Pacific Islander = 15, White = 2,308, Bi/multiracial = 257, and 53 selected "other" and 82 did not provide this information. White and Asian students[1] were combined and compared with underrepresented minority (URM) students, which included four groups routinely considered URM by the National Science Foundation and National Institutes of Health (e.g., National Institutes of Health, 2020; National Science Foundation, 2017) – American Indian, Black, Hispanic, and Pacific Islander. Students not identifying as any of these six groups were not included in the analyses below.

Measure

Mosaic is administered online and is taken in school, typically during a single class period. In the current sample, there were eight six-point Likert items per skill with Cronbach's alpha values ranging from .75 to .84 (M = .78). Scores were calculated by averaging the eight items per scale (after reverse scoring negatively worded items). There were two SJTs per skill with five behavioral responses per SJT, and students indicated the likelihood of having the described response on a five-point scale. For each skill, SJT scale scores were obtained by averaging eight or nine of the ten items (some items were excluded to increase reliability; items were reverse scored in the event of negatively keyed items), and alpha values of these scales ranged from .56 to .76 (M = .69). The FC items were arranged in multidimensional blocks of three yielding 30 items total and six per skill. Students were asked to indicate the item in a block that was most like them and least like them. Items that were most like them received a score of 3, items that were not selected received a score of 2, and items that were least like them received a score of 1. FC scale scores were created by averaging responses to four, five, or six items (some items were excluded to increase reliability), and alpha values ranged from .46 to .59 (M = .51). Lower internal consistency estimates are typical of SJTs, which are often multidimensional (Whetzel & McDaniel, 2009). Likewise, reliability estimates of ipsative FC

scores are questionable at best (Meade, 2004). In addition, students provided demographic information and GPA, which is reported on a 12-point scale from E/F (below 65%) to A+ (97-100%).

Results

Independent samples *t*-tests were carried out and standardized effect sizes were computed to compare the groups' SE skill scores by item type. To help contextualize the differences, the same comparison was done for GPA. Confidence intervals (95%) for the effect sizes were calculated to determine whether effects differed across item types. If two effect sizes' confidence intervals did not overlap, those effect sizes are called out as being notably different from one another.

Racial Groups

Asian and White students (M = 9.74 on the 12-point scale; note that 10 = A- / 90-92%), which is just below an A-) had significantly higher GPAs than URM students (M = 8.85; note that 9 = B+ / 87-89%) with a standardized mean difference of .43. In contrast, the highest mean difference on a SE skill (Keeping an Open Mind, Likert scale) was .22, and URM students scored higher. Only three effect sizes exceeded an absolute value of .20, and the average effect size across all 15 SE skill scale scores was .04. The average was close to zero because the highest scoring student group was about evenly split across the 15 scale scores. See Table 1 for all statistics.

There were several significant student group differences, and in some instances, Asian and White students scored higher while in others, URM students scored higher. When making specific comparisons across item types (e.g., Likert Sustaining Effort vs. SJT Sustaining Effort), we found some confidence intervals that did not overlap. The magnitude of the student group differences on two SJT scales and one FC scale differed significantly from their respective Likert scales' group differences. Specifically, the difference in Maintaining Composure scores was

larger on SJT items (*d* = .21), with Asian and White students scoring higher, than on Likert items (*d* = -.03). The difference in Keeping an Open Mind scores was larger on Likert items (*d* = -.22) than on SJT items (*d* = -.01). Likewise, the difference in Keeping an Open Mind scores was larger on Likert items (*d* = -.22) than on FC items (*d* = -.02). One FC scale's group difference effect size was significantly different from its respective SJT scale's. Specifically, the difference in Maintaining Composure scores was larger on SJT items (*d* = .21) than FC items (*d* = .05). Finally, across the three item types, group differences by skill were largely consistent. For example, Asian and White students scored significantly higher on all measures of Sustaining Effort.

Table 1. Descriptive Statistics and Mean-Level Differences Across Racial Groups and Item Types

		Asian/White *M*(*SD*)	URM *M*(*SD*)	*t*	*p*	*d*	CI
GPA		9.74 (2.01)	8.85 (2.30)	10.26	< .01	.43	(.34, .51)
Likert	Sustaining Effort	4.45 (.80)	4.36 (.79)	2.96	< .01	.12	(.04, .19)
	Getting Along with Others	4.86 (.62)	4.79 (.72)	3.05	< .01	.12	(.04, .20)
	Maintaining Composure	3.93 (.70)	3.95 (.73)	-.74	.46	-.03	(-.11, .05)
	Keeping an Open Mind	4.32 (.68)	4.47 (.73)	-5.63	< .01	-.22	(-.30, -.14)
	Social Connection	4.18 (.75)	4.22 (.77)	-1.43	.15	-.06	(-.13, .02)
SJT	Sustaining Effort	3.85 (.60)	3.75 (.63)	5.48	< .01	.15	(.08, .23)
	Getting Along with Others	3.84 (.55)	3.79 (.60)	2.32	.02	.09	(.01, .16)
	Maintaining Composure	3.83 (.56)	3.71 (.62)	5.29	< .01	.21[a]	(.13, .28)
	Keeping an Open Mind	3.45 (.56)	3.45 (.55)	-.73	.47	-.01[a]	(-.08, .07)
	Social Connection	3.31 (.52)	3.39 (.54)	-3.80	< .01	-.15	(-.23, -.07)
FC	Sustaining Effort	2.39 (.39)	2.31 (.37)	5.46	< .01	.21	(.14, .29)
	Getting Along with Others	2.62 (.33)	2.57 (.35)	3.68	< .01	.14	(.07, .22)
	Maintaining Composure	2.19 (.44)	2.17 (.43)	1.19	.24	.05[b]	(-.03, .12)
	Keeping an Open Mind	2.15 (.45)	2.15 (.45)	-.46	.65	-.02[a]	(-.09, .06)
	Social Connection	2.18 (.43)	2.18 (.41)	-.21	.83	-.01	(-.09, .07)

Note. Asian/White GPA n = 2,226, SE skills n = 2,431. URM GPA n = 783, SE skills n = 897. [a]Differs significantly from Likert d. [b]Differs significantly from SJT d.

Gender Groups

Female students (M = 9.82) had significantly higher GPAs than male students (M = 9.14) with a standardized mean difference of -.32. SE skill effects exceeded this value in many instances. They ranged from a low (in absolute magnitude) of .03 to a high of -.64. There were many significant student group differences, and in most cases, female students outscored male students. See Table 2 for all statistics.

Female students scored higher than male students on three of the five Likert scales, all SJT scales, and one FC scale. When making specific comparisons across item types, we found some confidence intervals that did not overlap. The magnitude of the student group differences on three SJT scales and four FC scales differed significantly from their respective Likert scales' group differences. In terms of Likert vs. SJT differences, female students scored higher than male students on Getting Along with Others and Social Connection, but the effect sizes were larger on SJT items (Getting Along with Others d = -.64; Social Connection d = -.49) than on Likert items (Getting Along with Others d = -.48; Social Connection d = -.08). Male students scored higher on the Likert measure of Maintaining Composure (d = .08), but female students scored higher on the SJT measure of Maintaining Composure (d = -.28). In terms of Likert vs. FC differences, male students scored higher on Likert and FC measures of Maintaining Composure, but the effect was greater on FC (d = .35) than Likert items (d = .08). Female students scored higher than male students on Likert items but not FC items for Sustaining Effort (Likert d = -.46 vs. FC d = .03), Keeping an Open Mind (Likert d = -.36 vs. FC d = .18), and Social Connection (Likert d = -.08 vs. FC d = .07). Finally, we compared the effect sizes across SJT and FC items. Although female students scored higher on both SJT and FC Getting Along with Others items, the effect was greater for SJT (d = -.64) than FC items (d = -.43). Female students scored higher on the remaining four SJT scales as well, whereas male students scored higher on the four corresponding FC scales (Sustaining Effort SJT d = -.55 vs. FC d = .03;

Maintaining Composure SJT *d* = -.28 vs. FC *d* = .35); Keeping an Open Mind SJT *d* = -.39 vs. FC *d* = .18); Social Connection SJT *d* = -.49 vs. FC *d* = .07).

Table 2. Descriptive Statistics and Mean-Level Differences Across Gender Groups and Item Types

		Asian/White *M*(*SD*)	URM *M*(*SD*)	*t*	*p*	*d*	CI
GPA		9.14 (2.26)	9.82 (1.97)	-9.14	<.01	-.32	(.25, .39)
Likert	Sustaining Effort	4.25 (.80)	4.61 (.75)	-14.02	<.01	-.46	(-.54, -.39)
	Getting Along with Others	4.69 (.67)	4.99 (.58)	-14.39	<.01	-.48	(-.56, -.41)
	Maintaining Composure	3.97 (.72)	3.91 (.71)	2.64	<.01	.08	(.01, .16)
	Keeping an Open Mind	4.24 (.73)	4.49 (.66)	-10.96	<.01	-.36	(-.44, -.29)
	Social Connection	4.17 (.77)	4.23 (.74)	-2.63	<.01	-.08	(-.15, -.01)
SJT	Sustaining Effort	3.47 (.52)	3.75 (.50)	-16.22	<.01	-.55	(-.63, -.48)
	Getting Along with Others	3.53 (.48)	3.82 (.43)	-19.25	<.01	-.64[a]	(-.72, -.57)
	Maintaining Composure	3.71 (.60)	3.87 (.56)	-8.27	<.01	-.28[a]	(-.35, -.20)
	Keeping an Open Mind	3.34 (.50)	3.54 (.52)	-12.03	<.01	-.39	(-.46, -.32)
	Social Connection	3.20 (.51)	3.45 (.51)	-14.90	<.01	-.49[a]	(-.56, -.42)
FC	Sustaining Effort	2.38 (.40)	2.37 (.38)	1.31	.26	.03[ab]	(-.05, .10)
	Getting Along with Others	2.53 (.35)	2.67 (.30)	-11.97	<.01	-.43[b]	(-.51, -.36)
	Maintaining Composure	2.26 (.43)	2.11 (.43)	10.37	<.01	.35[ab]	(.28, .42)
	Keeping an Open Mind	2.19 (.47)	2.11 (.43)	5.53	<.01	.18[ab]	(.11, .25)
	Social Connection	2.20 (.43)	2.17 (.43)	2.33	.02	.07[ab]	(.00, .14)

Note. Male GPA *n* = 1,503, SE skills *n* = 1,711. Female GPA *n* = 1,743, SE skills *n* = 1,894. [a]Differs significantly from Likert *d*. [b]Differs significantly from SJT *d*.

Discussion

SE skill assessments typically use only Likert item types, and although they have advantages such as having a large body of validity evidence, they have shortcomings such as being relatively susceptible to impression management. Mosaic includes two additional item types, SJTs and FC, to mitigate shortcomings associated with any single item type. There is ample evidence evaluating racial and gender group differences on SE skill-related constructs, but to our knowledge, there is none comparing these three item types in a single sample, and this is the first time doing so with the Mosaic scales.

The results pertaining to racial group differences suggest that Likert, SJT, and FC item types are similar. Although there were some instances of the group of Asian and White students being more strongly favored on an SJT or FC scale vs. a Likert scale (e.g., d = .21 for FC Sustaining Effort vs. d = .12 for Likert Sustaining Effort), the effects were small overall, and the differences in the magnitude of the effects were small as well.

The difference in effects was much more pronounced when comparing male and female students. For example, female students scored higher than male students on each of the five SJT scales but only one of the FC scales. Even when effects were in the same direction, sometimes they were much larger for one item type than another (e.g., d = -.49 for SJT Social Connection vs. d = -.08 for Likert Social Connection). Gender differences do seem to be moderated by item type with FC items seeming to function differently than Likert or SJT items. It is possible that this is the result of student group differences in susceptibility to reference biases or tendency to engage in impression management; however, our data do not speak to this, so this would be an avenue for future research.

Limitations and Future Directions

There are several additional areas for future research with Mosaic as well as SE skill assessments in general. First, the specific content across the various item types was not constant, so responses are subject to more than method variance. Second, students took

Mosaic as part of a formative assessment system, so we cannot make inferences to high-stakes settings. Third, future research can contribute to the literature by attempting to include samples that are more representative of the population of high school students in terms of SE skills and demographics, and with larger sample sizes, URM minority groups would not have to be aggregated, which may disguise some effects. Moreover, additional student groups can be examined such as higher vs. lower income students. Finally, measurement invariance was not considered prior to examining student group differences. This work is in preparation. As a follow up, we will continue to examine group differences to determine whether our findings for Mosaic are in line with prior research, to determine what process may be at play to explain observed differences (e.g., increased cognitive load on SJT and FC items relative to Likert items), and to inform any future revisions made to Mosaic.

Acknowledgements

Thank you to Jeff Allen and Cristina Anguiano-Carrasco for their review of this paper.

References

ACT. (2021). *Mosaic™ by ACT®: Social emotional learning assessment*. Author.

Anglim, J., Bozic, S., Little, J., & Lievens, F. (2017). Response distortion on personality tests in applicants: comparing high-stakes and low-stakes medical settings. *Advances in Health Sciences Education*, *23*, 311-321.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*, 1347-1368.

Costa, Jr., P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322-331.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions*. Drasgow Consulting Group.

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology*, *61*, 579-616.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of SJTs: faking, coaching, and retesting issues. In J. A. Weekley, & R. E. Ployhart (Eds.), *Situational Judgment Tests: Theory, Measurement and Application* (pp. 205-232). Lawrence Erlbaum Associates.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*, 531-552.

National Institutes of Health. (2020, February 7). *Populations underrepresented in the extramural scientific workforce*. https://diversity.nih.gov/about-us/population-underrepresented

National Science Foundation. (2017, January). *Women, minorities, and persons with disabilities in science and engineering.* https://www.nsf.gov/statistics/2017/nsf17310/digest/introduction/

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2011). Constructing fake-resistant personality tests using item response theory: High-stakes personality testing with multidimensional pairwise preferences. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 214-239). Oxford University Press.

Walton, K. E., Radunzel, J., Moore, R., Burrus, J., Anguiano-Carrasco, C., & Murano, D. (2021). Adjectives vs. statements in forced choice and Likert item types: Which is more resistant to impression management in personality assessment? *Journal of Personality Assessment*, *3*, 1-35.

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment*, *20*, 23-45.

Westrick, P. A., Marini, J. P., Young, L., Ng, H., Shmueli, D., & Shaw, E. J. (2019). *Validity of the SAT® for predicting first-year grades and retention to the second year.* College Board. *https://collegereadiness.collegeboard.org/pdf/national-sat-validity-study.pdf*

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*, 188-202.

Whetzel, D. L., & McDaniel, M. A. (2016). Are situational judgment tests better assessments of personality than traditional personality tests in high-stakes testing? In U. Kumar (Ed.), *The Wiley Handbook of Personality Assessment* (pp. 205-214). John Wiley & Sons.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291-309.

Notes

1. Differences in estimated statistics were minimal when Asian students were removed and only White students were compared with URM students, and overall conclusions remained the same.

ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Grounded in 60 years of research, ACT is a trusted leader in college and career readiness solutions. Each year, ACT serves millions of students, job seekers, schools, government agencies, and employers in the U.S. and around the world with learning resources, assessments, research, and credentials designed to help them succeed from elementary school through career. To learn more, visit http://www.act.org/.