Are They Trying? Motivation on a State-Mandated Assessment of Career Readiness[1]

Jeffrey T. Steedle, PhD

Abstract

High school students can benefit from demonstrating career readiness on workplace skills assessments, but motivation is a concern in low-stakes test administrations. This study examined four indicators of motivation based on ACT® WorkKeys® data. Compared to adults testing under high-stakes conditions, more high school students were flagged for possible low motivation. Thus, some high school examinees were apparently unmotivated, which would be expected when testing in a relatively low-stakes context. However, the overall levels of motivation among high school and adult examinees were quite similar, which provides support for the interpretation of WorkKeys results from high school administrations.

Introduction

Large-scale assessment of career readiness is overshadowed by increasing usage of the ACT and SAT to measure college readiness and meet federal accountability requirements for high schools. However, several states administer workplace skills assessments, and some allow results to count toward high school performance indices. One such assessment—ACT WorkKeys—provides a certification of foundational workplace skills that is recognized by thousands of employers and job training programs. Yet, some media attention has focused on perceived lack of value for college-bound students (Russell, 2019), and anecdotal reports indicate that some students are unmotivated to perform well.

This study examined whether performing well on a career readiness assessment mattered to student examinees. Analyses addressed the research question, “How does motivation on workplace skills assessments compare between high school students and adult examinees?” Specifically, four indicators of examinee motivation were derived from response patterns, item scores, and response latencies on WorkKeys. Motivation was compared for students testing under different conditions and adult examinees taking WorkKeys for high-stakes purposes. Considering differences in stakes, differences in motivation were expected (Steedle & Grochowalski, 2017). Results indicated small differences in apparent motivation between high school students and adults. This finding is consistent with the notion that most high school students are motivated, which provides support for the validity of inferences drawn from WorkKeys scores of high school students.

Please reference this report as [1]Steedle, J. T. (2020, September). *Are they trying? Motivation on a state-mandated assessment of career readiness*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Method

Data

ACT WorkKeys Assessments measure foundational workplace skills. Most examinees take three assessments: ACT® WorkKeys® Applied Math (34 items), ACT® WorkKeys® Workplace Documents (35 items), and ACT® WorkKeys® Graphic Literacy (38 items). For this study, paper and online testing data were drawn from fall 2017 through spring 2019 for three states. The data were divided into four mutually exclusive groups:

- Census testing state: High school students from a certain state where all public high school students took WorkKeys.
- Course requirement state: High school students from a certain state where students took WorkKeys as a work readiness course requirement.
- Accountability option state: High school students from a certain state where students took WorkKeys to contribute to their school accountability performance indices or to earn a complimentary credential.
- Adult testers: Examinees aged 22 or older who presumably took WorkKeys to qualify for a job or job training program from the same three states (online testing only).

The high school testers were approximately 50% female, 38% Black/African American, 7% Hispanic/Latino, and 47% White. In comparison, the adult testers were approximately 43% female, 41% Black/African American, 4% Hispanic/Latino, and 48% White.

Analysis

Several methods were applied to identify possible low motivation on the three WorkKeys Assessments. First, the percentage of items omitted was calculated, and testers were flagged if it exceeded 20%. In prior research, skipping questions on state standardized tests predicted future educational outcomes even when controlling for test scores (Hernández & Hershaff, 2015).

Next, the C_z index was used to identify unusually long repeating patterns in item responses (e.g., ABABABAB, ABCDABCD, CCCCCCC, etc.). This index was developed to clean data for use in test security analyses (Cui, 2020), but repeating item response patterns can also indicate unmotivated responding. Using cutoffs based on visual inspection of C_z distribution plots, testers were flagged if C_z was greater than 8 for Applied Math, 8 for Workplace Document, and 10 for Graphic Literacy.

Item score patterns deviating substantially from expected were identified using the H^T index (Sijtsma & Meijer, 1992), a non-parametric person-fit statistic that was among the best for detecting simulated aberrant responding (Karabatsos, 2003). Generally, low H^T values are indicative of poor person fit. Testers were flagged according to cutoffs based on a simulated null distribution of H^T conditional on raw scores using simulated scores based on 3PL item statistics and an expected Type-I error (false positive) rate of .05.

Finally, response latencies (in milliseconds) were analyzed for online testers to calculate response time effort (RTE), which is the proportion of items on which an examinee exhibited solution behavior rather than rapid-guessing behavior (Wise & Kong, 2005). Rapid-guessing behavior was inferred from response times less than seven seconds, and testers were flagged if RTE was less than or equal to .80 (i.e., the tester exhibited rapid-guessing behavior on 20% or more of the test).

Results

Table 1 shows descriptive statistics for WorkKeys scores (65–90 scale) and motivation indices. By small degrees, adult testers tended to score highest on WorkKeys, and students in the accountability option state tended to score lowest, in particular those who tested on paper. Some of these students took WorkKeys because they performed poorly on other tests that counted toward school accountability performance indices, and this could partly explain the lower scores. Mean omit rates tended to be lower for students compared to adults. C_z indices were just barely lower on average for adult testers, which indicated shorter strings of repetitive responses. Likewise, mean H^T statistics were quite similar for high school and adult testers. The only notable exception was the group of paper testers in the accountability option state, which exhibited relatively poor mean person fit. As for response latency, adults spent the longest average time responding to items (approximately 70 to 80 seconds per item). Students in the accountability option state spent more time per item on average than students in other states.

Note that differences between high school testing contexts (e.g., census testing vs. course requirement) and differences between testing modes (online vs. paper) were difficult to discern from results because several factors were confounded with context and mode. For example, student ability differed between states and testing modes, and ability was correlated with the motivation indices (see below). Also, examinees testing in different states and different testing modes were likely to be administered different WorkKeys forms, and the motivation indices depended, to some degree, on the items that made up the test forms.

Table 1. WorkKeys and Motivation Index Descriptive Statistics by Assessment, State, and Delivery Mode

Assessment	Group	Mode	N	Mean WorkKeys	SD WorkKeys	Mean Omit %	Mean C_z	Mean H^T	Mean Latency
Applied Math	Census	Online	53,489	77.5	4.8	0.2%	4.8	0.37	60.0
	Census	Paper	45,553	78.5	5.0	0.2%	4.7	0.40	--
	Course	Online	2,354	78.5	4.6	1.1%	4.7	0.40	60.3
	Account.	Online	52,670	77.4	4.2	0.6%	4.8	0.38	65.3
	Account.	Paper	3,167	73.2	3.4	0.5%	5.1	0.27	--
	Adult	Online	8,988	79.0	4.5	1.8%	4.6	0.40	79.9
Workplace Documents	Census	Online	50,168	78.8	4.1	0.2%	4.7	0.34	59.3
	Census	Paper	45,756	79.8	4.2	0.2%	4.5	0.37	--
	Course	Online	1,955	78.7	4.1	0.6%	5.1	0.32	58.2
	Account.	Online	49,736	78.6	3.7	0.5%	4.8	0.34	64.6
	Account.	Paper	3,124	75.1	4.1	0.6%	4.9	0.24	--
	Adult	Online	9,106	80.7	3.7	1.5%	4.7	0.35	77.1
Graphic Literacy	Census	Online	51,186	78.4	4.3	0.2%	5.9	0.29	52.3
	Census	Paper	45,539	78.9	4.4	0.1%	5.8	0.31	--
	Course	Online	2,061	78.5	4.0	0.8%	6.0	0.32	52.1
	Account.	Online	48,152	78.1	3.7	0.5%	5.9	0.31	57.3
	Account.	Paper	3,036	73.8	3.6	0.4%	6.1	0.15	--
	Adult	Online	7,926	79.3	4.0	1.9%	5.8	0.32	71.4

The percentage of examinees flagged by each method is shown in Table 2. Only 0.5% of examinees were flagged for omitting 20% or more items, but the percentages for adults were consistently higher than for students. Among students, those who took WorkKeys as a course requirement were more likely to be flagged for omitting items. For each of the three assessments, adult examinees were less likely to be flagged for repetitive responding based on the C_z index. Flagging for low H^T person-fit statistics was slightly more common for paper testers in the census testing state for Applied Math and Workplace Documents. The H^T flagging rate for adults was tied for lowest on Graphic Literacy, but some groups of high school testers had lower average H^T than adults on Applied Math and Workplace Documents. Consistent with adults' longer average response times, the response time effort flagging rate was consistently lowest for adults. Among students, it was highest in the accountability option state, especially on Applied Math.

Despite differences, the percentages of high school and adult examinees flagged for unmotivated responding were quite similar. For Applied Math, 17.5% of high school examinees were flagged in at least one of the analyses, compared to 14.8% of adults. The corresponding percentages were 16.0% and 13.5% for Workplace Documents, and they were 15.1% and 13.3% for Graphic Literacy. Note that the flagging rates depended on decisions about the cutoff values, so it was difficult to infer anything about the "true" level of unmotivated responding from the results. However, the more important results for this study were the comparisons between high school and adult examinees.

Table 2. Flagging Rates by Assessment, State, and Delivery Mode

Assessment	Group	Mode	N	Omit	C_z	H^T	RTE
Applied Math	Census	Online	53,489	0.3%	2.5%	3.5%	13.5%
	Census	Paper	45,553	0.2%	2.5%	4.5%	--
	Course	Online	2,354	1.5%	3.0%	3.9%	16.0%
	Account.	Online	52,670	0.7%	3.0%	3.9%	21.9%
	Account.	Paper	3,167	0.4%	4.6%	2.8%	--
	Adult	Online	8,988	3.0%	1.8%	3.4%	7.6%
Workplace Documents	Census	Online	50,168	0.3%	2.2%	4.8%	10.5%
	Census	Paper	45,756	0.3%	1.8%	5.8%	--
	Course	Online	1,955	1.1%	2.8%	4.7%	16.9%
	Account.	Online	49,736	0.7%	2.4%	5.0%	17.7%
	Account.	Paper	3,124	0.8%	2.8%	3.1%	--
	Adult	Online	9,106	2.5%	1.6%	5.2%	4.7%
Graphic Literacy	Census	Online	51,186	0.2%	1.8%	4.8%	10.3%
	Census	Paper	45,539	0.1%	2.5%	4.7%	--
	Course	Online	2,061	1.3%	2.3%	5.2%	17.0%
	Account.	Online	48,152	0.6%	2.1%	4.7%	17.1%
	Account.	Paper	3,036	0.2%	3.3%	3.9%	--
	Adult	Online	7,926	3.2%	1.7%	3.9%	5.2%

Note: **Omit**: Flagged if the omit rate was 20% or higher. **C_z**: Flagged if $C_z > 8$ (Applied Math and Workplace Documents) or $C_z > 10$ (Graphic Literacy). **H^T**: Flagged if H^T was less than the α = .05 cutoff value for a given raw score. **RTE**: Flagged if rapid guessing was detected on 20% or more of items.

Table 3 shows the agreement rate percentages for the motivation flags. For example, of the examinees flagged for a high omit rate, 1.8% were also flagged for high C_z, 0.1% were also flagged for low H^T, and 9.3% were also flagged for low RTE. C_z and RTE had the highest levels of agreement (21.6% to 27.8%). That is, about 25% of examinees flagged for high C_z were also flagged for low RTE. This result is consistent with the notion that rapid responding is likely to exhibit repetitive response patterns.

Table 3. Agreement between Motivation Flags

Assessment		% Omit Flag	C_z Flag	H^T Flag	RTE Flag
Applied Math	% Omit Flag	--	1.8%	0.1%	9.3%
	C_z Flag	0.4%	--	6.3%	27.8%
	HT Flag	0.0%	4.3%	--	19.5%
	RTE Flag	0.4%	6.2%	6.3%	--
Workplace Documents	% Omit Flag	--	0.9%	0.0%	5.3%
	C_z Flag	0.2%	--	6.4%	25.4%
	HT Flag	0.0%	2.6%	--	11.7%
	RTE Flag	0.3%	5.9%	6.5%	--
Graphic Literacy	% Omit Flag	--	1.7%	0.1%	8.2%
	C_z Flag	0.4%	--	5.6%	21.6%
	HT Flag	0.0%	2.5%	--	12.9%
	RTE Flag	0.4%	5.0%	6.7%	--

Correlations between WorkKeys scores and motivation indices are shown in Table 4. Overall, the directions of the correlations indicated that examinees with higher WorkKeys scores were less likely to exhibit unmotivated responding behavior. Specifically, examinees with higher WorkKeys scores were less likely to omit items, less likely to exhibit repetitive responding, more likely to have good person fit, and more likely to exhibit solution behavior (i.e., not rapidly guess). Subsequently, students with higher WorkKeys scores were less likely to be flagged for unmotivated responding. However, the correlations between WorkKeys scores and H^T flags were approximately zero, which was expected because the H^T flagging cutoffs were conditional on raw scores.

Table 4. Correlations between WorkKeys Scores and Motivation Indicators

Test	% Omit	C_z	H^T	RTE	% Omit Flag	C_z Flag	H^T Flag	RTE Flag
Applied Math	-.09	-.13	.36	.23	-.07	-.09	.00^{ns}	-.18
Workplace Documents	-.13	-.17	.32	.37	-.09	-.10	-.01	-.28
Graphic Literacy	-.11	-.12	.33	.32	-.08	-.07	.01	-.23

Note: All correlations were statistically significant at the $p < .05$ level unless otherwise noted (ns = non-significant).

The differences in flagging rates shown in Table 2 might not apply equally to examinees across the range of WorkKeys scores. For example, considering the negative association between WorkKeys scores and flagging rate, differences in flagging rates might have been smaller for higher scoring examinees. Perhaps such examinees have a low likelihood of exhibiting unmotivated responding regardless of the testing context. Figures 1, 2, and 3 illustrate differences in flagging rates by WorkKeys decile for Applied Math, Workplace Documents, and Graphic Literacy, respectively. Overall, flagging rates for C_z and H^T were very low on all three assessments. Only in the first to second deciles was there a discernable difference, with high school testers being flagged slightly more often than adult testers. In general, adult testers were more likely to be flagged for omitting more than 20% of items. Differences in flagging rates for omitting items were largest in the low score range, and they progressively decreased as WorkKeys scores increased. In contrast, high school testers were more likely than adult testers to be flagged for low RTE. As with other flagging rates, differences were greatest for low-scoring examinees, and the differences decreased with increasing WorkKeys scores. Thus, the observed differences in flagging rates between high school and adult WorkKeys testers (Table 2) were mainly driven by the relatively large differences between lower scoring examinees.

Figure 1. Flagging Rate Conditional on Applied Math Decile

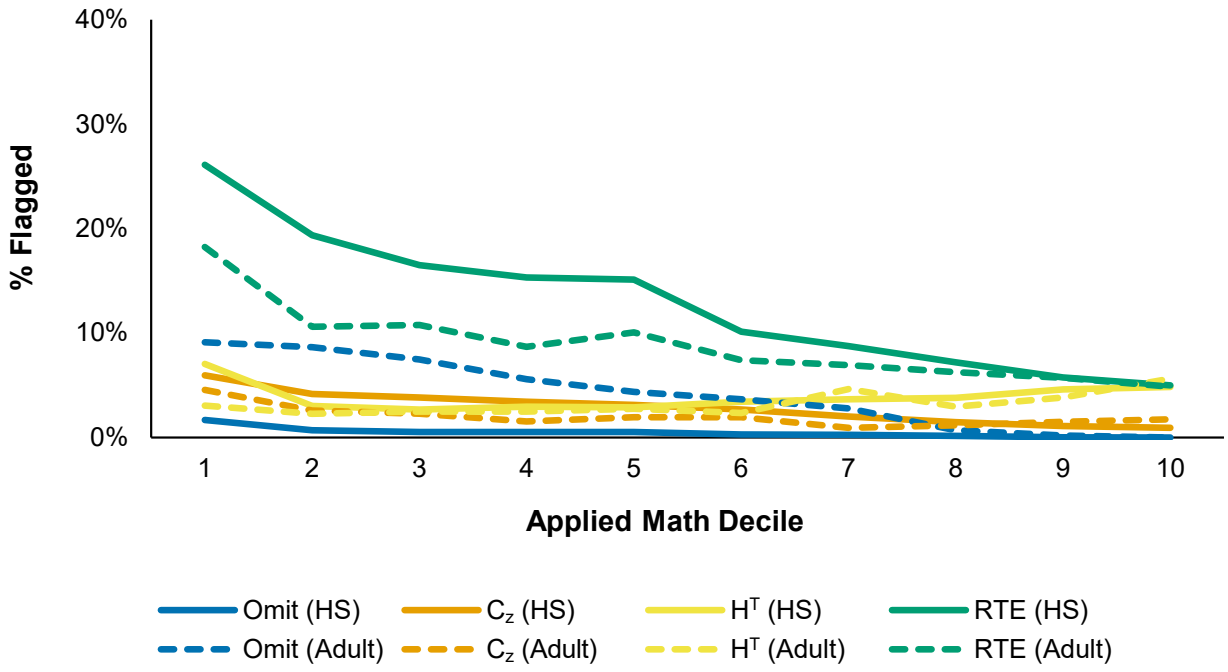


Figure 2. Flagging Rate Conditional on Workplace Documents Decile

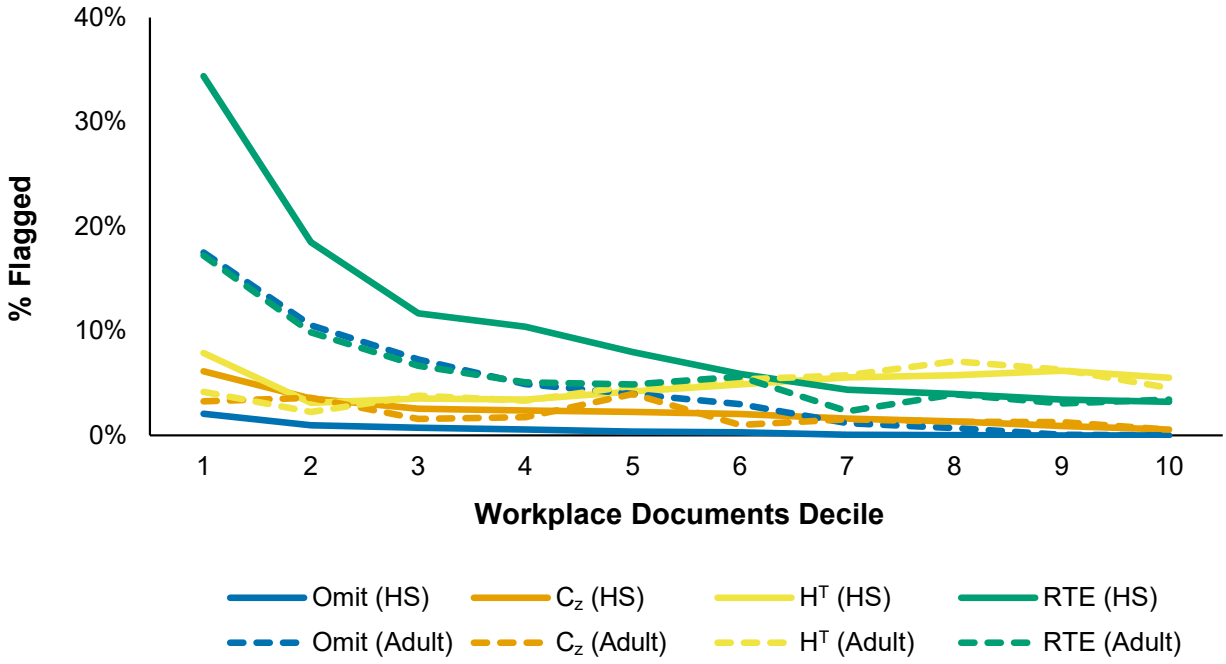
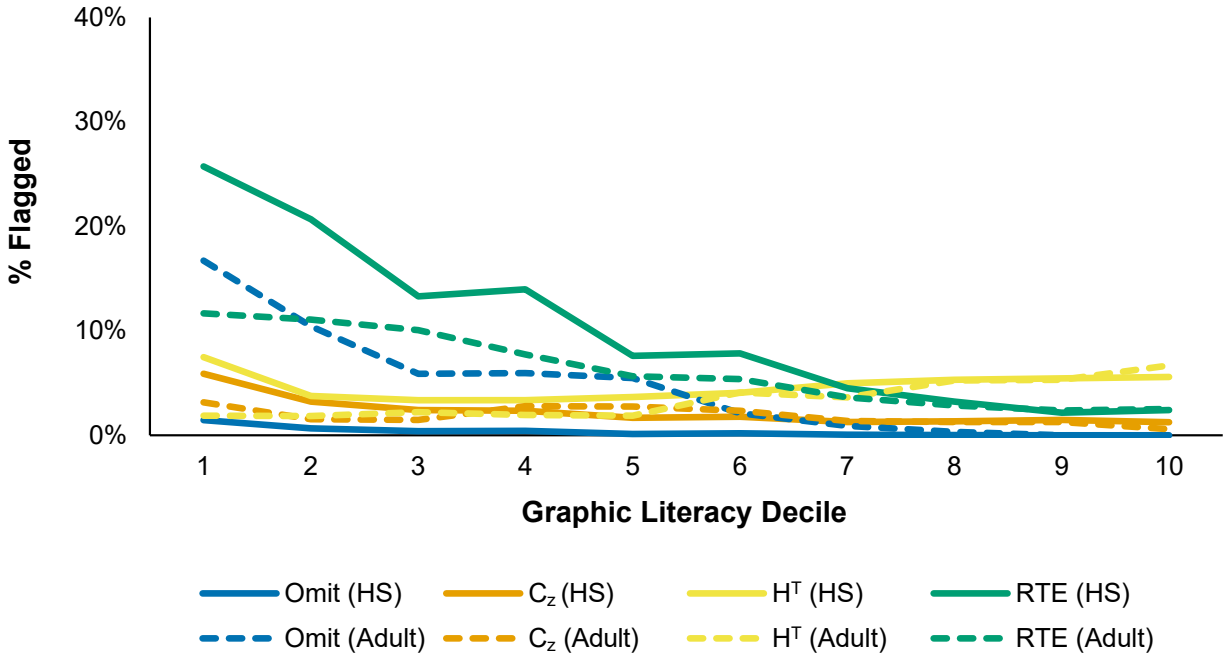


Figure 3. Flagging Rate Conditional on Graphic Literacy Decile



Discussion

This study investigated possible differences in motivation between high school students and adults taking a workplace skills assessment. The students generally took WorkKeys under low-stakes conditions, though some students were aware of the possible benefits of earning a National Career Readiness Certificate® based on WorkKeys performance. In contrast, adults took WorkKeys to qualify for a job or job training. Considering differences in test stakes, adult examinees were expected to exhibit higher motivation than high school examinees. The adult examinee group had slightly higher average test scores, and correlations indicated that higher scoring examinees were less likely to exhibit unmotivated responding. Note, however, that the difference in average achievement was likely, in part, a reflection of differences in motivation. Consistent with expectations, the adult examinees less frequently exhibited repetitive responding, poor person-fit, and especially rapid guessing behavior, though they were more likely to omit items. These differences between adult and high school examinees primarily manifested among lower scoring examinees.

Lower apparent motivation among high school students might call into question the interpretation of test results for the high school students. However, the differences between high school and adult examinees were small in magnitude. Specifically, 86% of adults were not flagged for unmotivated responding by any method, compared to 84% for high school students. Thus, the overall levels of motivation among high school students appeared reasonable if one considers motivation among adults testing under high-stakes conditions as a useful point of reference. In short, some high school students were apparently unmotivated—as expected for a low-stakes assessment—but many students were motivated, and this lends support to the use and interpretation of workplace skills assessments in high schools. To address possible concerns about motivation, states might improve communication to schools and students about the potential value of demonstrating career readiness.

References

Cui, Z. (2020). On a new algorithm for removing repeating patterns in similarity analysis. *Educational and Psychological Measurement, 80*(3), 446–460. doi:10.1177/0013164419882980

Hernández, M., & Hershaff, J. (2015, October). Skipping questions in school exams: The role of non-cognitive skills on educational outcomes. Paper presented at the Latin American and Caribbean Economic Association Annual Meeting, Medellín, Colombia.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298. doi:10.1207/S15324818AME1604_2

Russell, K. (2019, April 9). 'This test is worthless': Lawmakers, kids who take WorkKeys test say it is so simple it's insulting. *WXYZ Detroit.* Retrieved from https://www.wxyz.com/news/this-test-is-worthless-lawmakers-kids-who-take-workkeys-test-say-it-is-so-simple-its-insulting

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement, 16*(2), 149–157. doi:10.1177/014662169201600204

Steedle, J. T., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment, 22*(2), 111–123. doi:10.1080/10627197.2017.1309276

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Jeffrey Steedle, PhD

Jeffrey Steedle is a lead psychometrician in Assessment Transformation directing the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. Jeff holds advanced degrees in education, statistics, and educational psychology, and his research interests include assessment validation and motivation on achievement tests.