Investigation of 2018 ACT Score Declines Final Report

Leslie Keng, Ph.D. and Michelle Boyer, Ph.D.

National Center for the Improvement of Educational Assessment

January 20, 2020

Table of Contents

Acknowledgement

The Center wishes to acknowledge the ACT team's engagement and responsiveness during the investigation. We respect ACT's willingness to open its procedures and processes for scrutiny by an independent organization. We also understand the need for discretion in certain areas where information is proprietary or privileged. We specifically would like to thank Drs. Wayne Camara, Melinda Taylor, Samuel Haring, Benjamin Andrews, and Dongmei Li for their support and efforts throughout the process.

EXECUTIVE SUMMARY

ACT requested assistance from the National Center for the Improvement of Educational Assessment (Center for Assessment) to investigate declines of scores for states administering the ACT to its 11th grade students in 2018. This request emerged from conversations among state leaders, the Center for Assessment, and ACT in trying to understand the 2018 score declines, particularly in census testing states.

The primary goals of the Center's investigation were to (a) identify possible explanations for the observed score declines based on information provided by ACT, (b) make recommendations to ACT and states for additional analyses to confirm or uncover explanations for the observed trends, and (c) suggest changes to ACT's existing processes or procedures that can help expedite the detection of such trends or mitigate potential sources of error in the future.

The Center pursued three general lines of inquiry to understand potential contributions to the ACT score declines in 2018. We examined:

1. ACT's established **procedures**,
2. **implementation** of its established procedures, and
3. select **contextual factors** related to the test-taking populations.

Using documentation, data, and interview responses provided by ACT, the Center sought to answer specific questions for each line of inquiry.

Primary Findings

This investigation focused on exploring possible explanations for identified score declines across multiple states using the same operational form in either, or both, 2017 and 2018. Although we have not been able to conclusively determine that scores on this form were erroneous, we found at least one potentially serious condition threatening the comparability of ACT scores across forms: the changing national population participating in the ACT over the years. Specifically, this report raises concerns about the possible impact of observed changes in ability distributions of the equating samples between 2014 and 2015. The primary ACT form administered in most census-testing states in 2018 was equated in 2015. If the observed changes in ability distribution in the national population led to an equating solution substantively different from a solution based on little or no year-to-year variation in population ability, the adjustments to form difficulty would contain some level of error that could affect all forms equated that year, including the primary form administered in 2018 to students in Louisiana, Missouri, North Dakota, and Tennessee.

The Center for Assessment has not seen evidence to confirm that the ACT scaling and equating procedures are robust to observed changes in the national test-taking population, a fundamental tenet of ACT's equating design. The Center also identified additional sources of risk in existing ACT procedures and documentation that may threaten the validity of ACT score interpretations. We offer several recommendations for ACT to defend the validity of its scores more completely as its test-taking populations continue to change and expand. Our recommendations fall into two areas: analyses and procedures.

Recommendations for Analysis of Scale Stability and Invariance of ACT Scores

The main technical concern we noted throughout this investigation regards the stability of the ACT scale and the robustness of the ACT scaling and equating procedures to the changing test-taking populations. ACT should engage in more direct testing of the population invariance assumptions associated with the ACT scaling and equating procedures and regularly document the findings. We further recommend that ACT provide more comprehensive documentation about the fidelity of implementation to support claims that ACT scores are valid for all examinees, taking all forms, in each year since the current scale was established in 1989.

Although ACT conducted a study of the invariance of ACT equating solutions across different ability groups, that study was limited to examining equating results based on data from a single, unspecified test form. In this study, approximately 10000 examinees taking a single test form were divided into 3 different samples based on ability groups. ACT's scaling and equating procedures were then applied based on each group, and each of the three outcomes was compared to scaling and equating results based on the full sample. We question whether the study design and interpretation of results apply sufficiently to the context under examination. We recommend that ACT share the details of their study with its technical advisory committee and make the results available to measurement experts and other stakeholders for independent evaluation.

We also recommend several areas for additional study to more fully interrogate the robustness of ACT's scaling and equating procedures to observed population changes, as well as to uses beyond college readiness for different populations. As the test provider, ACT is responsible for providing evidence in support of three key assumptions:

1. the ACT scale, established 30 years ago and in a different testing context, still produces valid and comparable scores for all examinees across forms and years;
2. ACT's scaling and equating procedures produce examinee scores that are invariant to the observed changes in the ACT national test-taking population in recent years, and
3. evidence to support claims that ACT's scaling and equating procedures produce examinee scores that are valid for intended uses in statewide testing programs.

The Center for Assessment did not find such evidence in our investigation.

Two ideas for studies that might support further examination of invariance assumptions under ACT's testing conditions are provided in this report. ACT may also consider partnering with states and other measurement experts to fully investigate and provide evidence for the three key assumptions.

Recommendations for Procedures and Processes

Definition of 'national' population

The main technical concern we note throughout our investigation is the possible impact of changing ACT test-taking populations. To fully investigate the impact of population changes on equating over time, we urge ACT to adopt a precise definition of what constitutes the target "national" population and how well that target is met each year that new forms are put on the ACT scale. This definition is critical for supporting accurate score inferences because new forms are equated to the ACT scale each year based on samples drawn from the national population.

ACT equating samples come from a single national testing date (October). The Center for Assessment could not find the rationale for this choice in any of the ACT materials, or how well this sample generalizes to all ACT test-takers. The Center concluded that the lack of clear definition in the sampling frame for equating may have implications for the year-to-year comparability of examinee scores and scale stability. ACT's sampling specifications must include more precise definition of "national" and a clearer description of the sampling frame. Accordingly, the Center recommends ACT produce detailed sampling specifications with precise definitions of the target ACT linking population in order to facilitate the ability to monitor changes overtime that might threaten scale stability and the comparability of scores.

Specifications and technical documentation

The Center identified important gaps in the specifications and technical documentation for several key steps in ACT's standard operating procedures. These gaps included a lack detailed guidelines and specifications for test construction, sampling, data processing, scaling and equating analyses, and quality control.

The Center found the current technical reporting process to be largely ad hoc. ACT produces technical reports periodically. The reports do not include form level information about the technical quality of the tests, which misses important sources of internal validity evidence. At a minimum, we recommend including the types of invariance analyses discussed previously and annual studies to support evidence of score validity, particularly in the context of state census testing. We also recommend including form-specific information in technical documentation such as reliabilities, classical item analysis results, decision accuracy and consistency, and evidence of content alignment with intended frameworks or standards. The Center did not have

the opportunity to review external content alignment studies. We urge ACT to make these studies publicly available and incorporate the results into appropriate technical documentation that can be used by participating states to support the validity of ACT use in each specific context. In short, full and regular technical reporting will better serve state clients who must provide evidence of technical quality to their stakeholders and as part of the federal peer review process.

Internal communication

Our interviews with ACT staff indicated that the test development workflow across functional groups involved limited psychometric review of test forms. We recommend ACT improve its standard operating procedures to formalize communication between psychometrics and test development teams to better support coherent quality assurance of forms development. Formal psychometric guidelines for test development staff use in test construction, and to guide formal psychometric QA is one example.

Implementation outcomes

The Center for Assessment did not observe official records of outcomes from key steps of the annual test development, data processing, and psychometric processes. In particular, no documentation showed that ACT met various criteria or targets in the standard operating procedures each year. Many testing programs capture such evidence by completing quality control checklists, with annotations to convey salient observations or issues. A systematic and planful archive of implementation outcomes would engender confidence, both internally at ACT and externally, that standard operating procedures were implemented with fidelity—particularly in circumstances such as the score decline in 2018.

Context effects

Due to the limited availability of contextual data, we focused our investigation on ACT procedures and implementation, and whether either could have contributed to the score declines in 2018. The investigation into contextual factors was limited to those that are related to ACT's procedures and implementation (e.g., impact of population shifts on the equating sample.) Although there are many contextual factors that could reasonably explain the score declines, they cannot be characterized with confidence before examining ACT procedures and their implementation.

Nonetheless, there are some possible analyses that are within ACT's purview to pursue concurrently with recommended analyses and improvements. For example, we recommend ACT take a closer look at the item-level (p-value) comparisons between the equating sample and the state-specific examinees to identify whether score declines can be attributed to decreased performance in specific content standards. We also recommend that ACT conduct additional

analyses on the 2018 primary-form samples and the impact that differences in ability distribution and demographic characteristics might have on the robustness of equating outcomes.

INVESTIGATION OF 2018 ACT SCORE DECLINES FINAL REPORT

Several census testing states, in 2018, expressed concern regarding the unexpected declines in their ACT 11th grade results compared with previous years. The table below summarizes the changes in the statewide average ACT composite scores from 2014 to 2018.

Changes in statewide average ACT composite scores, 11th grade (2014 to 2018)

	Change			
State	**2014 -2015**	**2015-2016**	**2016-2017**	**2017-2018**
Oklahoma	.	.	0.0	-0.7
Louisiana	0.0	0.2	0.0	-0.6
Kentucky	0.0	0.2	0.2	-0.5
North Dakota	-0.2	-0.1	0.1	-0.5
Ohio	.	.	.	-0.5
Tennessee	0.1	0.5	-0.1	-0.4
Wisconsin	.	0.1	0.0	-0.3
Alabama	0.0	0.4	-0.3	-0.3
Mississippi	.	0.6	-0.3	-0.2
Utah	0.0	0.1	0.1	-0.2
Wyoming	-0.1	0.2	-0.3	-0.2
Arkansas	-0.3	0.1	0.0	-0.1
North Carolina	0.1	0.0	0.0	-0.1
South Carolina	.	0.3	-0.5	-0.1
Montana	-0.1	0.1	-0.3	-0.1
Hawaii	0.3	0.1	0.1	0.0
Nebraska	.	.	.	0.0
Nevada	.	0.0	-0.1	0.1
Missouri[1]	.	0.4	-0.5	.

[1] Missouri does not have a score change value for "2017-2018" because it was no longer an ACT census testing state in 2018. However, Missouri was concerned about the score decline in observed in 2017 and agreed to make its test data available via ACT for this investigation. In 2017, the primary ACT form administered in Missouri was the same primary form given to most of the ACT census testing states in 2018.

While some states observed score declines in previous years, the 2018 declines are noteworthy in both their prevalence and their magnitude. Further, the four ACT subject areas evinced similar declines:

Change in average ACT subject area scores by state from 2017 to 2018

State	English	Math	Reading	Science	Composite
Oklahoma	-0.5	-0.8	-0.9	-0.5	-0.7
Louisiana	-0.4	-0.5	-0.7	-0.9	-0.6
Kentucky	-0.4	-0.6	-0.5	-0.7	-0.5
North Dakota	-0.5	-0.5	-0.4	-0.4	-0.5
Ohio	-0.3	-0.4	-0.5	-0.4	-0.5
Tennessee	-0.2	-0.3	-0.3	-0.5	-0.4
Wisconsin	-0.5	-0.1	-0.2	-0.3	-0.3
Alabama	-0.2	-0.2	-0.3	-0.2	-0.3
Mississippi	-0.3	-0.3	-0.1	-0.3	-0.2
Utah	-0.2	0.1	-0.2	-0.3	-0.2
Wyoming	-0.2	-0.3	-0.3	-0.2	-0.2
Arkansas	-0.2	-0.1	-0.1	-0.1	-0.1
North Carolina	-0.2	0.0	-0.1	-0.2	-0.1
South Carolina	-0.3	0.0	0.0	-0.2	-0.1
Montana	0.0	-0.2	-0.1	-0.2	-0.1
Hawaii	-0.3	0.2	0.0	0.2	0.0
Nebraska	0.1	-0.1	0.1	0.2	0.0
Nevada	0.1	0.1	0.2	0.0	0.1

ACT investigated these declines, asserting "the score decline in states from 2017 to 2018 cannot be attributed to factors (such as a scoring error or test form issues) other than changes to examinee performances in the states. However, it was unclear what specific cause might be contributing to score declines." In view of these inconclusive findings, ACT approached the Center for Assessment, as an independent third-party organization, to conduct an investigation of the 2018 ACT score declines in the 11th grade test results.

The Goals and Approach of the Center's Investigation

The primary goals of the Center's investigation were:

1. Identify possible explanations for the observed score declines based on information provided by ACT;
2. Make recommendations to ACT and states regarding additional analyses to confirm, or explore possible explanations for, the observed trends; and
3. Suggest changes to ACT's existing processes or procedures that can expedite the detection of such trends and mitigate potential sources of error.

To support the investigation, the Center received permission from four states – Louisiana, Missouri, North Dakota, and Tennessee – to examine their ACT test results.

We organized our investigation into three general lines of inquiry, representing three factors that may have contributed to the ACT score declines in 2018. These factors, and the respective questions framing our investigation, are as follows:

1. Established ACT procedures: Are there any standard operating procedures at ACT that may have contributed, directly or indirectly, to the 2018 score declines?
2. Implementation: Were there any irregularities in the implementation of ACT's standard operating procedures in 2018 that may have contributed, directly or indirectly, to the 2018 score declines?
3. Contextual factors related to the test-taking populations[2]: Are there any conditions or trends in test-taking populations, overall or in each state, that may have contributed, directly or indirectly, to the 2018 score declines?

The Center began by identifying categories of possible threats to score comparability, generating a comprehensive list of corresponding questions (see Appendix A), and then narrowing this list to a smaller set of questions relevant to ACT's designs and methods. We used these questions guide our request for pertinent materials from ACT (see Appendix B). The Center gathered additional information during several interview-style webinars with ACT staff. The Center discussed the scope and rationale for specific materials requests with ACT, asked for clarification on the contents and intended uses of materials provided, and inquired about information that was not available in the provided documents or files. ACT staff provided the Center with most of the requested information.

[2] Note that because the data and information for this investigation were primarily provided by ACT, our inquiry into the contextual factors is limited to those that are related to ACT's procedures and implementation. However, Louisiana has conducted detailed analyses related to these questions.

Investigation Results

This section summarizes the Center's investigation results, framed by the three stated lines of inquiry announced above. Under each line of inquiry, we address specific questions related to these broad categories in which possible threats to score comparability may occur.

Established ACT Procedures

This general question guided our investigation of ACT procedures: ***Are there any standard operating procedures at ACT that may have contributed, directly or indirectly, to the 2018 score declines?*** The Center posed, and sought answers to, specific questions about standard operating procedures for (a) test development, (b) scoring and data processing, and (c) psychometrics.

Test development

The Center primarily referenced the following information from ACT to address questions related to ACT's test development procedures: The ACT technical report, test specifications, statistical target summaries, prediction of test difficulty process/equations, test content change documents, and webinar notes.

Are the test construction procedures appropriate for the test design and purpose/use of scores?
The Center's findings here were inconclusive because ACT did not provide documentation of standard operating procedures and psychometric guidance for test construction. That said, we do raise several concerns.

From interviews with test development staff, the Center saw that item development and selection procedures focus appropriately on the alignment of item content with established test specifications and targeted statistical properties. ACT test development personnel described the following sequence of events when constructing new ACT test forms each year:

1. Develop item content.
2. Develop parallel test forms by matching test specifications.
3. Review items/forms for content, bias, fairness, and domain representativeness.
4. Evaluate forms for content cluing within the same form and across the form battery (i.e., in other subject areas).
5. Verify answer keys and conduct differential item functioning (DIF) analyses on each field test as well as on operational items after test administration.
6. Implement exposure control protocols governing the frequency and intervals that items are potentially seen by examinees (e.g., by examines who take the ACT test multiple times).

ACT examines item statistics such as p-values, point biserial correlations, and differential item functioning (DIF) values after forms are administered operationally; further, rigorous key checks ensure that all items are properly machine scored. We learned in the interviews with ACT staff that this information is not used in routine checks for possible changes in population performance on specific items and forms across administrations. It would be useful to have procedures in place to evaluate item statistics across years on the anchor form—not just for comparing p-values to understand population ability differences, but also to evaluate changes in the item-total correlations and reliabilities as a means to inform judgments about whether the form is performing similarly across years. This is important evidence for supporting invariance assumptions, as changes in model fit can threaten equating solutions from one population to another. The Center evaluated the item statistics for one of the primary ACT forms administered in 2018. A summary of the evaluation and key takeaways are provided in the *Contextual Factors* section.

The absence of detailed and well-documented test development procedures poses a risk to the consistency of their application. Such consistency is evidence that content and statistical considerations in test design and development are appropriately applied within and across years during forms construction. The Center recommends that ACT implement formal standard operating procedures for psychometric guidance to ensure that test forms consistently match blueprint and statistical requirements.

Does the design for field testing new items provide an appropriate basis for decisions about test construction? No, the Center for Assessment questions the efficacy of the field test design and procedures for supporting consistently high-quality form construction.

ACT conducts field tests of items when constructing operational test forms. Our interviews disclosed that field test items are not administered to necessarily representative or highly motivated samples of examinees—which contradicts claims in the technical manual. That said, ACT has developed statistical models to predict item difficulty with reasonable accuracy, as confirmed by a general match between target and actual item difficulty across the forms examined. As a result, ACT test forms appear to generally meet form-difficulty targets, understandably with some variation. However, since ACT's scaling and equating procedures do not separate examinee ability and item difficulty, the influence of item difficulty cannot be disentangled from that of the abilities of the population or sample of students who take a given ACT form.

Is alignment among ACT test content, the ACT learning frameworks, and state standards well documented? Not determined at this time. Although ACT staff stated that alignment studies of ACT vis-à-vis various states' standards have been done, these studies are not publicly available

(we recommend they be made so), and the Center did not request or review these studies due to scope and time limitations of this investigation.

We addressed this question primarily through interviews with ACT staff. Although we learned of routine internal review of content alignment with respect to ACT frameworks during test construction, we know of no results of an external review of alignment with these frameworks.

Documents detailing the transition of reporting categories to reflect subscore performance relative to Common Core standards address how this transition supported ACT alignment in this regard. This process did not involve changes to ACT test specifications—only its secondary alignment of some content to Common Core standards.

Since alignment of test content with what students are expected to learn is fundamental to valid interpretations of scores, states must require independent evaluations of the alignment between ACT content and state-specific standards and/or curricula. Such efforts would allow state stakeholders to better understand ACT results in view of what students are expected to know and be able to do.

Are the test specifications sufficiently detailed to support alignment consistently across forms? Yes, the test specifications appear to contain appropriate detail to support adequate form-to-form content alignment.

Is appropriate guidance and psychometric review employed to ensure that targets in the specifications are met? No. Through interviews with ACT staff, the Center learned that the statistical targets provided in the test specifications constitute the primary psychometric guidance for test developers during forms construction. Once forms are produced, the only stated psychometric review occurs during the key check process after field test and operational data are analyzed.

Are statistical targets for test construction appropriate? Yes, the overall targets seem reasonable.

Are quality assurance procedures reasonable and clear enough to be followed with fidelity? Because ACT did not provide the Center its quality assurance procedures for test construction, we cannot comment on this.

Scoring and Data Processing

The Center relied on the following materials and information from ACT to address questions related to ACT's scoring and data processing procedures: interview notes about equating sample and sampling frame, key check procedures, equating sample cumulative score distributions and demographics (i.e., gender and ethnicity), and notes on testing schedules and sampling timelines.

Are processes in place to ensure scoring keys are correct and applied consistently and correctly? Yes, ACT routinely conducts key checks following both field test and operational administrations.

Is the sampling design for equating well documented and appropriate for use in an equating design that targets randomly equivalent groups? No, ACT did not provide complete sampling specifications to the Center during this investigation.

ACT did document an occasion in 2018 in which the October testing date yielded an insufficient sample size for equating and consequently required sample augmentation. This document explains the detailed and careful selection process used to add examinees to the 2018 equating sample. But while providing insight into the 2018 sampling frame and equating, the document does not detail how similar it was to the sampling frame in previous years or to its sampling specifications moving forward.

One specific challenge in understanding ACT's sampling frame for equating is the lack of clarity around a definition of "national" population. There are a variety of definitions for the national population, depending on the context. National data have been characterized as:

- all data from national test dates (individually and collectively),
- all data for graduating students,
- all national test data for juniors, and
- all data from both national- and state-level administrations.

There are significant implications for test scores depending on which definition is employed. Further, these data may or may not include retakes, best scores from two or more administrations, or special make-up test results. It appears that national data sets are also produced for ad hoc analyses that depend largely on when the analyses are conducted.

This presents a fundamental challenge for the Center to fully understand the characteristics of, and possible changes in, the sampling frame for equating over time. As ACT staff disclosed in the interviews, the sampling frame appears to be data collected exclusively from the October national test date. This lack of clarity is a significant barrier to determining (a) whether the equating sample is representative of the national test-taking population, (b) whether the national population has been changing in recent years, or (c) whether this national sample includes sufficient representation of examinees in the state and districts in which the ACT is administered. We recommend ACT provide a detailed rationale for selecting the October test date sample as representative of the total ACT test-taking population. What is it about that sample that satisfies representation requirements? Is that representation constant over time? Are there changes in the

October samples over time that matter for equating? Can equating solutions generated based on this sampling frame be accurately applied to presumably different examinee populations?

Also, equating specifications are not sufficiently clear to independently replicate the process of drawing a sample for equating. Although a third party, HumRRO, replicated ACT's scaling and equating procedures this replication did not include an evaluation of sampling. Rather, the samples used in the ACT operational equating were provided to HumRRO. This lack of information about ACT's sampling procedures causes uncertainty about the nature of the population from which equating data are sampled and how this population has changed. *National* appears to have many definitions, from what ACT staff members disclosed, and this makes it difficult to discern whether the equating sample reflects a constant sampling frame or a changing one. We assume the sampling frame changes with any changes in the characteristics of the October national test date examinees, which could threaten year-to-year scale comparability and stability (depending on the nature and size of change in the national population).

Are there documented rules for the exclusion of invalid student records? None that were provided so documentation that clarifies any inclusion or exclusion rules would be helpful for supporting data quality verification efforts.

Are quality assurance procedures reasonable and clear enough to be followed with fidelity? Beyond sampling specifications, the Center did not request full detail regarding data management and processing outside of quality assurance conducted by psychometric staff at ACT. Given more time, the Center might have asked for process documentation from ACT staff responsible for scoring and the technologies supporting scoring. Data handoffs typically have strict QA procedures in place as it is transferred within a testing organization from scoring through analysis and reporting.

Psychometrics

The Center referenced the following information from ACT to answer the questions related to ACT's psychometric procedures: ACT Technical Manual; ACT 2018 equating analysis specifications; quality assurance procedures; multiple reports on score decline investigation results (summary of score change investigation and invariance studies report); anchor selection procedures; 2015 equating data; HumRRO equating replication report; and national, equating-sample, and state-level reports.

Are the scaling and equating procedures appropriate for the test design and purpose/use of scores?
ACT uses a random groups equipercentile linking procedure for its annual scaling and equating process (Kolen & Hanson, 1989). When all assumptions for this well-established method are met, it is expected to produce appropriately equated scores across forms and administrations.

However, we did not find evidence that some of the most critical assumptions were met (e.g., there are no substantive changes in test administration conditions or characteristics of the test-taking population).

Are the specifications about the operational psychometric procedures present and appropriately detailed to ensure consistency? Not entirely; the specifications are detailed with respect to the propriety equating tools and steps used at ACT. Specifications include instructions for using both publicly available (RAGE-EQUATE; Kolen & Brennan, 2004) and proprietary codes and tools for equating. Because of incomplete documentation of the methodological details, however, a fully independent external replication would not be possible without training on the equating tools and processes employed. (For example, the aforementioned HumRRO replication required specific onsite training by ACT psychometric staff.)

Are the scaling and equating procedures robust to changes in population ability distributions? The Center could not reach firm conclusions regarding the influence of population changes on the 2018 ACT results. However, we note several patterns related to the ACT test-taking population that warrant further investigation.

ACT uses classical test theory (CTT) methods for scaling and equating new test forms each year, which carry both benefits and limitations (Kolen & Hanson, 1989; Kolen & Brennan, 2004, 2014). When the target population is stable over time, a distinct benefit of these methods is the ability to control conditional standard errors of measurement along the full scale so that they are roughly equivalent. However, because CTT methods cannot separate examinee ability from item and test difficulty, their accuracy is heavily reliant on stable examinee populations, both within and across years. Within years, the random equivalence of groups taking each form is essential to the accuracy of equating, but well-designed spiraling designs can manage this requirement reasonably well. ACT appears to have robust spiraling procedures in place and to be following them rigorously each year.

ACT should test the invariance of equating solutions across different examinee abilities and demographic groups over time. The ACT scale was developed in 1989 to measure college readiness for the college-bound—having 1989 ability and demographic characteristics. ACT should provide evidence that appropriate adjustments of form difficulty are made for every form produced under conditions of a changing national population, as well as for the various state populations for which the test is used.

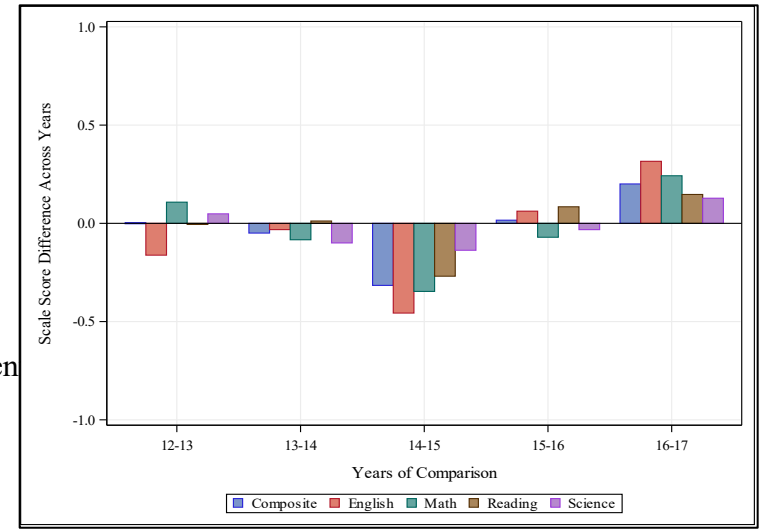
ACT statistically adjusts scores to control for the impact of test-difficulty variation. We accept that the assumption of population invariance has limitations insofar as exact random equivalence within a constant population is never achieved. Nor do we have perfect model fit, either overall or for the various student groups. Even methods related to item response theory (IRT), where

item parameters have theoretically robust invariance properties, require that we routinely monitor item and scale drift to support validity claims regarding scores. Population changes can impact equating results; the practical question concerns the point at which such changes are large enough to meaningfully affect examinee scores, individually or for specific groups.

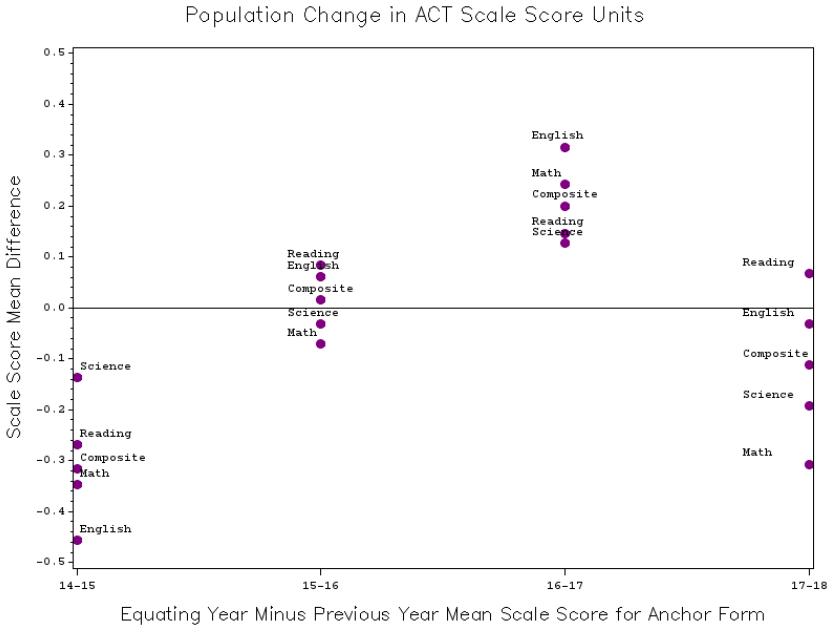
The purpose of the *Standards for Educational and Psychological Testing* is "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended uses" (AERA, 2014, p. 1). Particularly relevant to the present context is Standard 5.6, which requires test publishers to demonstrate the stability of their scales over time. Also relevant are Standards 5.13, 5.14, and 5.15, which call for evidence of form-to-form and year-to-year comparability of scores. These professional standards, coupled with the limitations of classical test theory (whereby we cannot separate item difficulty from examinee ability), demand skepticism and scrutiny regarding the claim of invariance over changing populations (Hambleton, Swaminathan, & Rogers, 1991; Kolen & Brennan, 2004, 2014).

When new ACT forms are administered operationally, they are equated to the ACT 1989 base scale through an anchor form. The anchor form is included with the new forms as part of the spiraling sequence during operational administration. Whereas anchor designs, in an IRT context, can leverage the invariance properties of item parameters in scaling and equating, ACT's classical equating procedures rely on equivalent groups taking each form. One way to routinely check assumptions of invariance, in an IRT paradigm, is to examine the stability of IRT item parameters over time. Operational implementation of equating based on randomly equivalent groups requires strong assumptions regarding the year-to-year population invariance. While finding evidence of effective practices for producing within-year randomly equivalent groups across ACT test forms, we did not see evidence that ACT routinely checked the invariance assumption across years.

From interviews, the Center understands that ACT chooses the anchor form in a given year from the pool of forms equated in the previous year. Consequently, population changes in ability can be viewed by displaying the differences in the mean performance of the anchor form in each pair of years. ACT performed this analysis for 2013 through 2017, the results of which are shown in this figure:

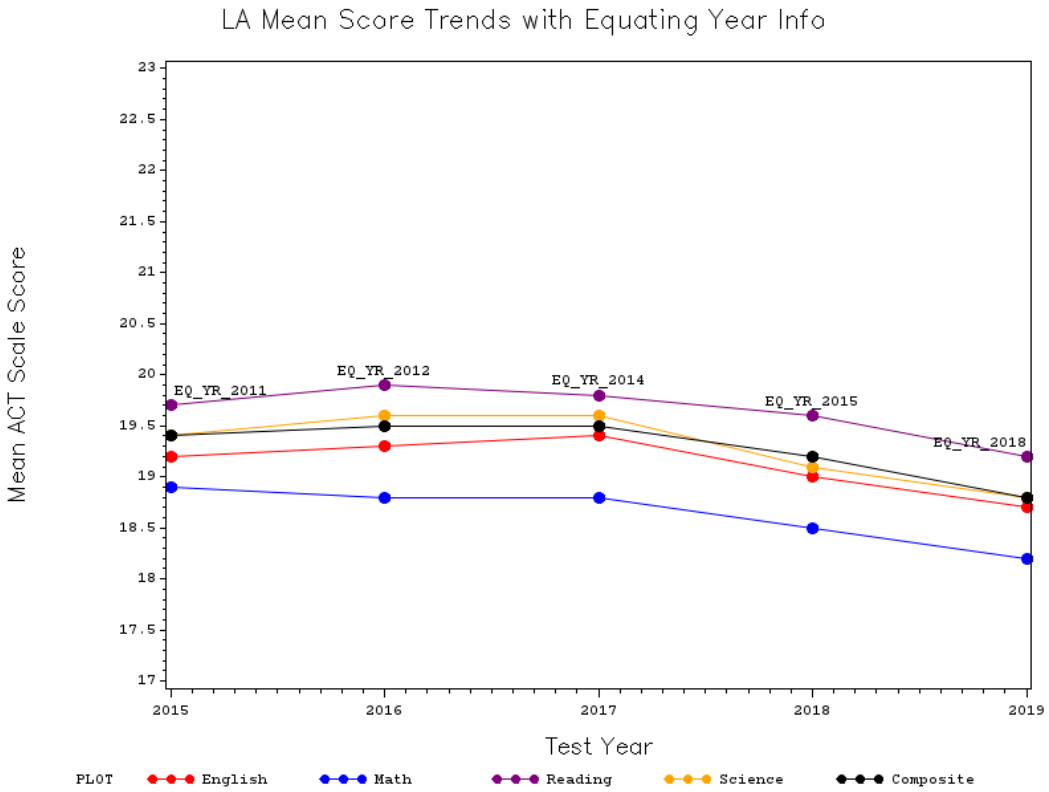


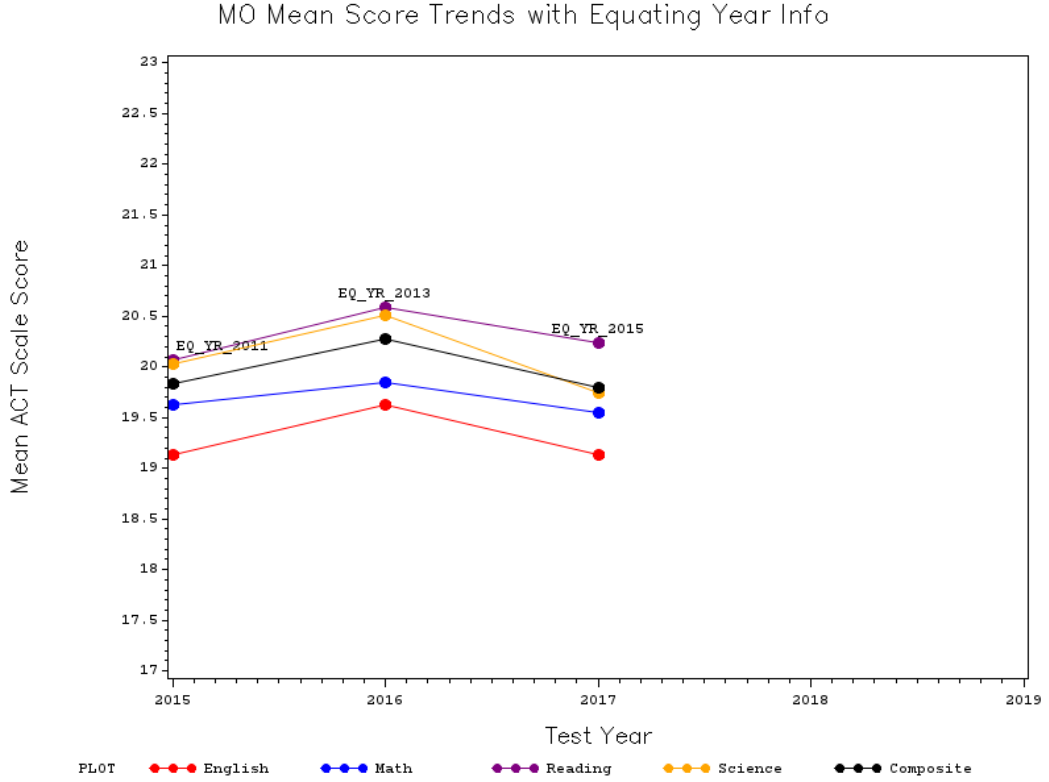
ACT also provided the Center with anchor form data from 2015 to 2018, so the view is extended (in a different format) in the following figure:

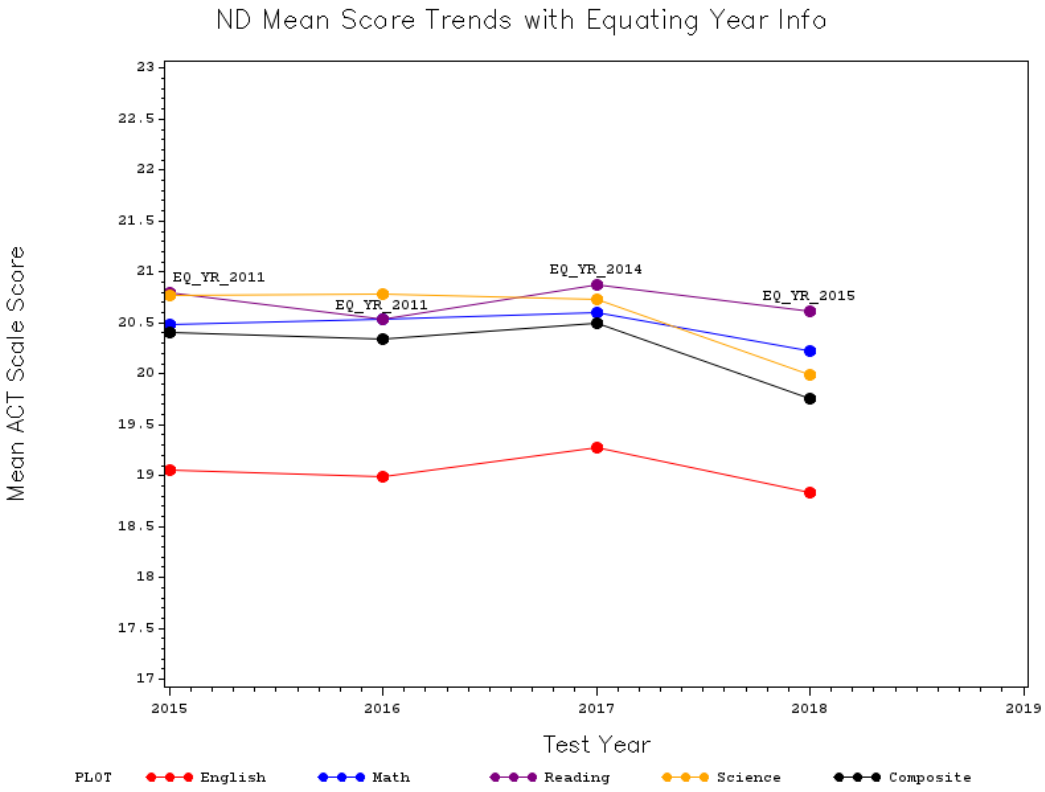


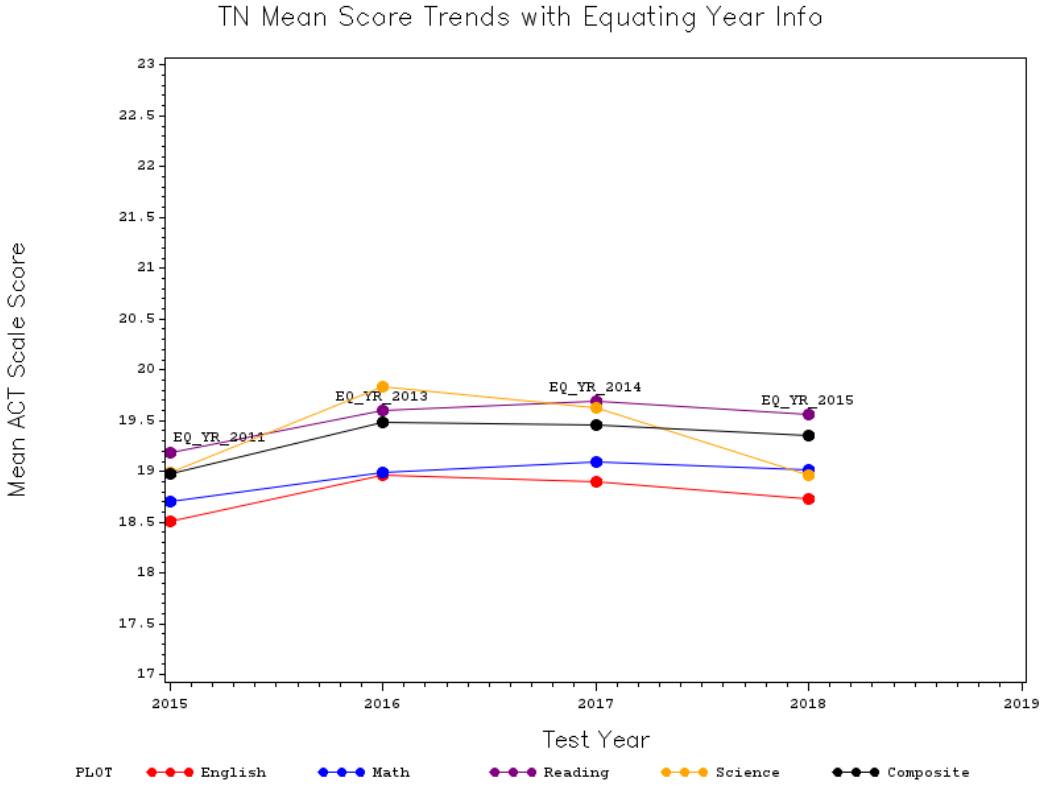
Changes from 2012 to 2013, 2013 to 2014, and 2015 to 2016 show small mean differences that vary in direction across subjects, suggesting random year-to-year population variation. In contrast, changes from 2014 to 2015, 2016 to 2017, and 2017 to 2018 show larger, systematic differences in performance.

We observed these same trends in performance when we tracked form use forward from the year in which the form was equated, which caused us to wonder whether (and, if so, how) the ACT equating results are influenced by the observed year-to-year population changes. In fact, we observed these trends generally in the four states we examined. The following four figures portray the mean scale score by subject area, within equating year.


LA Mean Score Trends with Equating Year Info
Mean ACT Scale Score
Test Year
EQ_YR_2011
EQ_YR_2012
EQ_YR_2014
EQ_YR_2015
EQ_YR_2018
PLOT English Math Reading Science Composite


MO Mean Score Trends with Equating Year Info
Mean ACT Scale Score
Test Year
EQ_YR_2011
EQ_YR_2013
EQ_YR_2015
PLOT English Math Reading Science Composite


ND Mean Score Trends with Equating Year Info
Mean ACT Scale Score
Test Year
EQ_YR_2011
EQ_YR_2011
EQ_YR_2014
EQ_YR_2015
PLOT
English
Math
Reading
Science
Composite


TN Mean Score Trends with Equating Year Info
Mean ACT Scale Score
Test Year
EQ_YR_2011
EQ_YR_2013
EQ_YR_2014
EQ_YR_2015
PLOT
English
Math
Reading
Science
Composite

ACT should examine the patterns between equating year and state trends for other operational forms to determine if the same pattern surfaces across forms, populations, and years. For example, ACT might consider:

1. Additional state trend lines (by operational form used)
2. National trend lines (by operational form used)
3. Trends for the other forms that were scaled in 2015, 2016 and 2018 (as these years showed larger systematic changes in population ability)

An important question is whether population changes have introduced bias into ACT's equating solutions. This is a difficult question to answer using ACT results, unfortunately, because ACT does not produce analyses that disentangle form difficulty from population ability. This question forces us to think about how the ACT scale is carried forward through administration of an anchor form in the ACT equating context. For example, if a form was newly scaled in 2014, where the population mean score in English was approximately half a scale score point higher than when it was subsequently administered as the anchor in 2015, would we get the same conversion table for the new 2015 forms if we equated based on the 2014 data for the anchor form? The logic in the ACT equating design (and all other anchor form-based equating designs) is that the anchor form provides the basis for the link between new forms and the existing scale, and that this link should account for population differences. But again, such methods are based on strong assumptions—and strong assumptions require strong evidence.

Foundational to item response theory (IRT), the assumption of population invariance means that item parameters estimated from one random sample of the announced population would be the same if these parameters had been obtained from a different random sample of the same population. In contrast, classical test theory (CTT) does *not* assume population invariance and, instead, relies on equivalent group designs for test equating. The equating links among forms—as seen through a scoring table—are treated in CTT as *de facto* invariant. Consequently, any error associated with this year-to-year link would be brought into the equipercentile mapping of the forms scaled in subsequent years. IRT is not free from these challenges, to be sure. But there are mathematical procedures for evaluating threats to invariance and, in turn, proceeding accordingly. In the present context, however, the Center was unable to conclude, based on the information provided, whether there was drift at the item, form, or scale level. ACT should be able to provide evidence of population invariance to support their validity claims.

ACT shared one population invariance study, which divided a large dataset into thirds by ability group at the school level and then performed equating for each group. When observing mean differences for the total group, we see some evidence that the ACT equating approach is robust to changes in the ability levels of the students in the equating sample. However, the report is a high-level summary of a cross-sectional look at a sample of examinees taking a specific form in a single (unspecified) year. Further, the mean differences vary across the scale. That the

differences were large relative to the 2014-2015 change is important but, for three reasons, does not tell the full story.

First, the middle ability group should result in the smallest difference from the total group, but this is not the case. Second, where the differences occur along the scale is important. When differences that matter fall at the readiness indicator, for example, the conclusion regarding a student's college readiness is affected. Also, if differences that matter appear at points in the scale where states have more students than the national sample, any systematic effect due to a changing national population would be more dramatic. Third, using the total group result as the criterion with which to measure differences in score scale conversion across the three equating scenarios will contaminate these comparisons. The equating solutions are not independent of the criterion if one third of the examinees are the same in each comparison. Therefore, the Center does not believe these results are sufficient for concluding that ACT's equating procedures are robust to all possible population changes across years. It does not directly address the specific context in which the 2018 score declines occurred. We recommend that ACT share the details of their study with its technical advisory committee and, further, make the results publicly available so that measurement experts and other stakeholders can independently consider the results.

The Center is unable to draw formal conclusions about the influence of population changes on the ACT 2018 results. Although the patterns noted here may in the end be spurious, we believe they warrant further investigation. The Center proposes a two-step approach. First, we suggest examining whether the general pattern (i.e., pattern of anchor form differences) bears out across forms equated in different years based on anchors that show smaller and larger year-to-year shifts (up or down) in population ability. Second, we recommend conducting additional studies to evaluate the influence of changing populations on equated results using the methods employed by ACT, particularly under conditions where there are larger population changes in the mean ability compared with conditions where the changes are smaller or simulated as zero. The design could leverage actual population parameters to create data sets with the characteristics of those noted empirically in the ACT equating data. One way to study the impact of changing populations is through equating where response data are simulated based on known population differences (e.g. distributions on an anchor form in current year versus previous year). The equating results from changing distributions can be contrasted with unchanging distributions across forms of similar and different difficulty. If population changes produce similar raw-to-scale score conversions and impact for different examinee populations (i.e. ability distributions), this would be evidence that the ACT procedures are robust to changes in the national population of ACT examinees.

The Center found that at least one state, Louisiana, administered the primary 2018 form in question with the original 2015 anchor form on which it was equated. This presents the possibility of an second study. Louisiana also administered the 2015 anchor form as a makeup

form during the 2018 administration and, therefore, was not taken by a representative population. However, the sample size for the makeup form appears to be reasonable. By using a sample matching method, such as propensity score matching, ACT could generate randomly equivalent groups between the primary and makeup forms and apply its standard equating procedures. While we would not endorse the equating outcomes from such an exploratory study to score students, it nonetheless could serve as a "stress test" for the ACT equating procedures.

Are quality assurance procedures for equating reasonable and clear enough to be followed with fidelity?

Yes, the processes implemented by ACT appear to be appropriate for ensuring accuracy and consistency of implementation. The checks and replications are based on best practice for high-stakes large-scale testing programs that conduct annual equating. The ACT procedures are as follows:

1. Check all equating programs (i.e., one person rerunning last year's equating to make sure all programs work as expected).
2. Set up all equating folders and input files with computer programs (i.e., one person doing the set up, and another person checking and providing signoff).
3. Run statistical analysis prior to equating (i.e., two people running the analysis).
4. Run equating program to produce output for all smoothing values (i.e., two people checking all the input files, and one person clicks the button to run).
5. Review all the output, and run equating program interactively to choose smoothing values. (At least three psychometricians are involved. The equating output files and smoothing plots are reviewed, and multiple criteria as listed in the procedure are considered to choose a smoothing value. Final smoothing values are determined with consensus of all psychometricians.)
6. Review summary statistics to evaluate equating results (moments and distributions of the equated forms are reviewed by all psychometricians involved).
7. Produce conversion tables, and provide them for scoring (all conversions are double- or triple-checked before they are provided for scoring).

Implementation

For each administration, appropriate standard operating procedures must be implemented with fidelity to the specifications. Here, our guiding question was: ***Were there any irregularities in the implementation of ACT's standard operating procedures in 2018 that may have contributed, directly or indirectly, to the 2018 score declines?***

The Center focused on the 2018 implementation of ACT's operating procedures for the primary test forms administered in Louisiana, Missouri, North Dakota, and Tennessee. Material provided by ACT and our webinar interviews with ACT staff constituted the primary source of information regarding ACT's implementation processes. ACT provided the Center with test form specifications by subject area from 2014 to 2018, and test form match-to-specifications

summary by subject area for three of the primary forms given to census testing states during the time period (including the primary form in 2018). We also obtained several documents—*The Evolution of English Language Arts on the ACT*, *The Evolution of the ACT Mathematics Test*, *The Evolution of the ACT Science test (1989 to present)*, and *ACT Item and Unit passing criteria*—as well as the flowchart *ACT National Forms Development (EchoAdapt ATA)*.

As in our examination of ACT procedures, we organized our findings around specific questions: test development, data processing, and psychometrics.

Test development

Were content specifications and test development procedures followed? Yes, with the few exceptions noted below.

The provided ACT test form specifications include content and cognitive-level targets. These targets vary by subject area and comprise:

- English—passage word count, depth of knowledge, and diversity (i.e., units coded for gender, ethnicity, region, and urban/rural).
- Mathematics—LIG,[3] depth of knowledge, and diversity (i.e., items coded for gender), and total word count.
- Reading—item word count, depth of knowledge, passage word count, and diversity (i.e., units coded for gender, ethnicity, region and urban/rural).
- Science—item cognitive level, depth of knowledge, content domain (i.e., biology, chemistry, earth/space science, and physics), and form word counts.

The primary forms in the match-to-specifications summary generally met the content and cognitive-level targets for each ACT subject area. There were, however, several exceptions. For example, the number of items in two of the cognitive level categories for science fell outside the target range in 2018, and the depth-of-knowledge targets for mathematics were not met in 2018. While these violations appear minor, we recommend ACT document in the spreadsheet *why* the content or cognitive-level targets were not met in each case.

Is alignment with blueprint/test specifications well documented? Yes, the forms generally met the blueprint targets. The blueprint targets for each subject area include:

- English—total items (overall and by reporting category), total units and number of items per unit.

[3] Note: "LIG" was not spelled out anywhere in the documentation reviewed by the Center.

- Mathematics—total items (overall and by reporting categories).
- Reading—total items (overall and by reporting categories).
- Science—total items (overall and by item format), units and reporting categories.

Are there any reasons to believe that content alignment with either the standards or the enacted curriculum may have drifted? No, we see no evidence of content drift.

ACT documentation revealed no substantive changes in the test specifications at the overall content domain levels of English, mathematics, reading, and science between 2014 and 2019. That said, ACT reorganized the scheme for classifying the content in 2014 to 2015 (in response to states shifting to Common Core State Standards), which resulted in a new score reporting framework for each of the tests. Unlike previous subscores, however, the new reporting category scores were not equated year-to-year. These changes would not be expected to contribute to any scale drift at the overall domain levels for English, mathematics, reading, and science.

Are statistical targets for test construction reasonably met for each form? No, ACT did not always meet the statistical targets, nor did they clearly document these shortfalls.

For each ACT subject area, the match-to-specifications spreadsheet included technical/statistical targets for item difficulty (p-value mean, standard deviation, and distribution) and item discrimination (biserial correlation mean, standard deviation, and above/below .30). The primary forms administered in Louisiana, Missouri, North Dakota, and Tennessee in 2017 to 2019 did not always meet the statistical targets. For example, the 2017 primary form did not meet the biserial-correlation threshold for items (i.e., > .30), and all three primary forms had overall mean p-values substantially below the target. ACT should document in its match-to-specifications spreadsheet why statistical targets were not met in each case. Further, we recommend ACT provide clearer explanations for the statistical targets, possibly in a single document concerning all aspects of test construction. These explanations should include the rationale for each statistical target, any tolerance for not meeting a particular target, and whether meeting one statistical target is more critical than meeting another (i.e., in case trade-offs needs to be made).

Is item difficulty reasonably well aligned with the examinee population? Because none of the ACT test construction materials included criteria or considerations for the examinee populations, the Center for Assessment was unable to answer this question conclusively. That said, in reviewing item statistics for the primary ACT forms given in Louisiana from 2016 to 2018, we saw a widening gap in the p-values between the equating samples and the examinees.[4]

[4] We shared these item-level comparisons with ACT staff in a webinar.

Are quality assurance procedures reasonable and clear? The Center for Assessment was unable to answer this question conclusively because of unclear documentation.

Are the scoring keys produced from the test construction process accurate? Yes, but additional documentation is necessary.

The flowchart *ACT National Forms Development (EchoAdapt ATA)* shows the general workflow for assembling test forms using an automated test assembly algorithm. One of the final steps is "Keysheets and M2S." We assume this represents the generation of answer keys (keysheets) and match-to-specifications (M2S) summaries for verification and approval at the end of the test construction process. While this workflow seems reasonable, ACT did not provide evidence that the keysheets from the 2014 to 2018 test construction process were indeed verified for accuracy. We can only assume, therefore, that this step was implemented. For the integrity of its test construction process, we recommend that ACT officially record and archive this evidence annually.

Data Processing

ACT provided little information about data processing and, consequently, the Center could not meaningfully address any of the four question above. Our interviews with the ACT psychometric staff provided the only confirmation that ACT implemented the data processing steps with fidelity. The Center recommends that ACT include tracking components in its operational equating process, such as an issues log and/or a data processing checklist, to officially document the annual implementation process.

Psychometrics

The Center referenced the following information from ACT to address the five questions above: ACT equating analysis specifications, HumRRO's *ACT Equating Replication Technical Report*, detailed equating output from 2015, score conversation tables from 2015 equating, and the document *ACT Anchor Form Recommendation for the 2016 Equating Study*. We gathered additional information about the equating implementation process in several webinar interviews with the ACT psychometric staff.

The Center's investigation focused on the 2015 equating because this was the year in which the primary ACT form given in 2018 in Louisiana, North Dakota, and Tennessee was equated. This also was the primary ACT form given in Missouri in 2017.

Were scaling and equating analyses performed with fidelity to specifications? Yes, the specifications for the scaling and equating analyses were followed completely and correctly.

The ACT psychometric staff walked us through the scaling and equating process and, further, provided detailed explanations of the resulting statistics and plots. As noted earlier, ACT commissioned HumRRO to replicate the standard procedures for equipercentile equating and application of equating results for four of the forms equated in 2015. The results of the independent replication can be found in HumRRO's *ACT Equating Replication Technical Report*. HumRRO was able to reproduce the results from ACT's original equating. However, it is important to note that neither the data cleaning procedures nor the equating sample selection procedures were part of HumRRO's replication.

Was sampling performed with fidelity to specifications?
The Center for Assessment could not answer this question conclusively because we received no evidence that sampling for equating meets the "nationally representative" criterion.

ACT did not provide sampling specifications. Although we were told in interviews with ACT staff that sampling was representative of the national population, we cannot say, in the absence of evidence, whether ACT conducted the sampling with fidelity to the corresponding specifications. This is a critical issue that ACT needs to clarify, to be sure. We strongly recommend that ACT details its sample selection process and evaluation criteria as either part of its equating analysis specifications or in a separate sampling specifications document. ACT also should regularly archive evidence that the equating samples are representative of the ever-changing ACT test-taking population.

Are all sampling and analysis results well documented? Yes, ACT provided detailed records of its scaling and equating analysis results through the equating output and score conversation tables they provided for the 2015 equating process. However, information about the corresponding sampling results was sporadic and difficult to evaluate.

In the report *ACT Anchor Form Recommendation for the 2016 Equating Study*, ACT summarized the statistical characteristics of the anchor form and new forms that were equated in 2015. Compared with the anchor forms, the new forms generally were easier in English and more difficult in math, reading, and science. The form that later would become the primary form for Louisiana, North Dakota, and Tennessee in 2018 (and Missouri in 2017) has different characteristics than its "sibling" forms in 2015. Specifically, the English test was more difficult than the anchor form while the science test was similar in difficulty. ACT argued that there is no evidence that such disparities adversely affected the equating of this form. As documented throughout this report, we do not agree with this conclusion.

Were operational scoring keys adequately checked? Yes, while ACT did not provide the Center with specific key check results, we have no reason to doubt that ACT implemented the key check

process with fidelity after the 2018 operational administration based on the key check procedural descriptions and conversations with ACT staff regarding this routine analysis.

Were quality assurance procedures implemented with fidelity? Yes, in general, the standard quality assurance procedures were followed.

The Center collected evidence regarding the implementation of quality assurance procedures for the 2015 equating process primarily through interviews with the ACT psychometric staff. If the ACT psychometric team does not already do so, we recommended documenting key decision points (e.g., selection of the anchor form, choice of smoothing values, etc.) Evidence of matching results and any reasonableness checks from the implementation of ACT's quality assurance procedures should also be archived.

Contextual Factors Related to the Test-taking Populations

Contextual factors, such as changes in the ACT test-taking population, enacted curriculum, and relevant policies or programs, also may explain the observed score declines in 2018. This third and final line of investigation seeks to answer the question: ***are there any conditions or trends in test-taking populations, overall or in each state, that may have contributed, directly or indirectly, to the 2018 score declines***?

We posed a set of more specific questions related to contextual factors, many of which are unique to each state's context. ACT staff was unable to provide state-specific information not directly tied to ACT test results. Therefore, the Center's investigation focused on contextual factors associated with ACT test development, scoring, and psychometric processes. We list all questions for completeness, but only address those questions for which ACT provided relevant information.

The Center relied on the following information from ACT: *ACT Technical Manual* (Fall 2019, version 3), *ACT Profile Report: Louisiana State Testing 2018-2019 (Grade 11 Tested Students)*, *ACT Profile Report: Missouri State Testing 2016-2017 (Grade 11 Tested Students)*, *ACT Profile Report: North Dakota State Testing 2018-2019 (Grade 11 Tested Students)*, State-level ACT raw and scale score descriptive statistics for the primary ACT forms from 2015 to 2018, State-level ACT scale score frequency distributions for the primary ACT forms from 2015 to 2018, State-level item statistics (p-values and biserial correlations) for the primary ACT forms from 2015 to 2018, *ACT's 2018 State Testing Score Change Investigation* report, and ACT's presentation to the Louisiana Technical Advisory Committee (November 2018).

Have there been any notable changes in test reliability/measurement error? Because ACT does not annually update reliability-related information, the Center for Assessment was not able to answer this question.

The document *ACT Technical Manual* provides overall test reliability information, such as standard error of measurement and classification consistency. In our interview with ACT staff, however, we learned that ACT does not annually obtain and report reliability information. For example, reliability statistics in the provided *ACT Technical Manual* are for the 2015-2016 school year. Further, this information is not reported at the state level. As such, we are unable to determine whether any notable changes have taken place in this regard. We recommended ACT compute, evaluate, and report this information regularly and at the state level.

Have there been any notable changes in student demographics?
No, but with one minor (if unclear) exception.

ACT provided profile reports for grade 11 students who participated in state testing in Louisiana, Missouri and North Dakota; the Missouri report was for 2016-2017, whereas the Louisiana and North Dakota reports were for 2018-2019. (Tennessee was not a census testing state and, therefore, did not have a profile report.) Each profile report included five-year trends for race/ethnicity and the percentage of students taking what ACT defined as a core curriculum.

The race/ethnicity composition of these grade 11 test-takers generally was stable across the five years reported.

Across these three states, there was a drop in the percentage of ACT examinees taking a core curriculum. In Louisiana, this percentage was 60% in 2014-2015, 50% in 2017-2018, and 40% in 2018-2019. Similarly, in North Dakota the percentages were 48%, 44%, and 33%, respectively; while in Missouri, the percentage went from 39% in 2014-2015 to 33% in 2016-2017. It is unclear whether these changes reflect differences in course-taking patterns or differences in how states and ACT define *core* curriculum? We recommend each state compare the performance of this group of students with those not taking a core curriculum on the ACT and other related academic measures.

Have there been any notable changes in alignment to state standards? This needs to be investigated by each state.

Have there been any notable changes in identifiable instructional practices or programs in the states? This needs to be investigated by each state.

Have there been any notable changes in difference in performance trends between the state's high school (grade-level or end-of-course) assessments and the ACT? This needs to be investigated by each state.

Have differences among items on item-performance statistics changed over years in specific content areas? The answer to this question depends on the analyses that ACT's content teams may choose to conduct.

ACT provided the Center with state-level item statistics (p-values and biserial correlations) for the primary ACT forms from 2015 to 2018. ACT also provided these item statistics for the equating sample for the year in which the primary forms were equated. As a proof of concept, the Center compared each state's item statistics for science with those for the equating sample. The table below shows the comparisons of statistics for ACT science items taken by students in Louisiana in 2018 versus by students in the *national* equating sample in 2015:

Item	LA - 2018		Equating - 2015		Difference	
	p-value	PBIS	p-value	PBIS	p-value	PBIS
1	0.80	0.41	0.91	0.29	-0.11	0.12
2	0.79	0.45	0.91	0.32	-0.11	0.13
3	0.76	0.43	0.87	0.31	-0.11	0.12
4	**0.64**	**0.52**	**0.85**	**0.41**	**-0.22**	**0.10**
5	0.46	0.34	0.56	0.36	-0.10	-0.02
6	0.28	0.28	0.37	0.36	-0.09	-0.08
7	0.69	0.46	0.84	0.35	-0.14	0.11
8	0.71	0.31	0.76	0.19	-0.05	0.12
9	0.52	0.42	0.67	0.33	-0.15	0.09
10	0.55	0.43	0.68	0.39	-0.13	0.04
11	0.54	0.37	0.66	0.35	-0.13	0.03
12	0.48	0.31	0.62	0.29	-0.14	0.02
13	0.38	0.25	0.43	0.33	-0.05	-0.08
14	**0.59**	**0.45**	**0.75**	**0.42**	**-0.17**	**0.03**
15	0.54	0.37	0.70	0.29	-0.16	0.09
16	**0.56**	**0.59**	**0.76**	**0.53**	**-0.20**	**0.06**
17	0.55	0.43	0.71	0.41	-0.16	0.02
18	0.70	0.46	0.83	0.38	-0.12	0.08
19	**0.43**	**0.39**	**0.65**	**0.46**	**-0.21**	**-0.07**
20	0.19	0.19	0.29	0.39	-0.10	-0.21
21	0.52	0.44	0.65	0.39	-0.12	0.05
22	0.55	0.37	0.66	0.39	-0.11	-0.02
23	0.43	0.41	0.57	0.32	-0.15	0.09
24	0.46	0.44	0.62	0.42	-0.16	0.02
25	0.46	0.31	0.53	0.34	-0.08	-0.03
26	**0.39**	**0.45**	**0.57**	**0.45**	**-0.18**	**-0.01**

Item	LA - 2018		Equating - 2015		Difference	
27	0.44	0.30	0.53	0.39	-0.10	-0.09
28	**0.50**	**0.50**	**0.68**	**0.49**	**-0.18**	**0.02**
29	0.59	0.46	0.74	0.43	-0.15	0.03
30	**0.41**	**0.52**	**0.57**	**0.50**	**-0.16**	**0.02**
31	0.43	0.40	0.58	0.44	-0.15	-0.04
32	**0.33**	**0.39**	**0.51**	**0.48**	**-0.18**	**-0.08**
33	**0.27**	**0.46**	**0.45**	**0.55**	**-0.18**	**-0.10**
34	0.27	0.27	0.39	0.37	-0.12	-0.10
35	0.42	0.26	0.48	0.33	-0.06	-0.07
36	0.31	0.43	0.44	0.47	-0.13	-0.04
37	0.48	0.34	0.62	0.35	-0.14	-0.01
38	0.23	0.35	0.34	0.41	-0.11	-0.06
39	0.35	0.24	0.38	0.30	-0.03	-0.06
40	0.22	0.21	0.29	0.25	-0.07	-0.04
Total	19.22		24.42		-5.20	

As the p-value differences show, these items were harder for Louisiana examinees than for those in the equating sample. We observed a similar pattern for ACT science items on the 2016 and 2017 primary forms given to census testing states, with the magnitude of the differences increasing over the three years. This is not an unexpected finding given what we see at the test level. However, the item-level comparison revealed items demonstrating substantially larger differences. We flagged items with a difference of .15 or greater (shown in **bold**), which should be shared with content experts to see if there are patterns concerning the content standards to which the flagged items are aligned (or other item characteristics). We shared the science results with ACT psychometric staff to, in turn, share with the ACT content teams for evaluation and feedback. If the ACT team found the item-level analysis results helpful, the Center offered to conduct the same analysis for the other content areas. We did not hear back from the ACT team.

Is there evidence of item drift on the primary forms administered in the four states? The Center for Assessment was not able to conduct this analysis, although our item-level comparisons provide the basis for evaluating potential item drift on the primary forms. But this evaluation would require additional student-level demographic information and prior achievement data so that matched samples could be generated to permit valid comparisons between the equating and state samples. We recommend ACT (or the states) conduct such an analysis and, in turn, consider the results within the context of any notable findings from the content standards evaluation suggested under the previous question.

Is there evidence from related scores or measures to corroborate the observed ACT score declines for the four states in 2018?

No, the 2018 ACT results are not corroborated by other data.

In its report *2018 State Testing Score Change Investigation*, ACT compared the 2017 and 2018 mean ACT scores of "national" examinees in the census testing states as well as the corresponding mean high school grade point average. Neither comparison showed a decline similar to the 2018 ACT score decline in the four states. Each state should have access to additional high school academic measures, and, if not already done, we recommend the conduct of a similar comparison on such measures. In doing these comparisons, each state should take steps to maximize the comparability of the two groups of students with respect to their demographic characteristics and prior achievement.

Have score declines such as the ones observed for the four states in 2018 occurred on the ACT before? No, the 2018 score declines observed were noticeably larger than in previous years.

Using the state-level scale score descriptive statistics provided by ACT, the Center compared the effect size of the ACT scores between consecutive years on the primary forms administered from 2015 to 2018 in Louisiana, North Dakota, and Tennessee, and from 2015 to 2017 in Missouri. We show these comparisons for ACT composite score in the following tables.

Louisiana – ACT Composite from 2015 to 2018

Year	# Test Takers	Mean Score	Standard Deviation	Mean Score Difference (Current - Previous)	Pooled Standard Deviation	Effect Size (Cohen's d)
2015	50,833	18.61	4.38			
2016	49,166	18.87	4.41	0.26	4.39	0.06
2017	47,947	18.91	4.48	0.04	4.44	0.01
2018	**47,557**	**18.33**	**4.57**	**-0.58**	**4.53**	**-0.13**

Missouri – ACT Composite from 2015 to 2017

Year	# Test Takers	Mean Score	Standard Deviation	Mean Score Difference (Current - Previous)	Pooled Standard Deviation	Effect Size (Cohen's d)
2015	54,663	19.84	4.96			
2016	52,641	20.28	4.94	0.44	4.95	0.09
2017	**50,361**	**19.79**	**5.13**	**-0.49**	**5.03**	**-0.10**

North Dakota – ACT Composite from 2015 to 2018

Year	# Test Takers	Mean Score	Standard Deviation	Mean Score Difference	Pooled Standard	Effect Size (Cohen's d)

				(Current - Previous)	Deviation	
2015	6,332	20.41	4.51			
2016	6,306	20.34	4.59	-0.07	4.55	-0.02
2017	6,117	20.50	4.60	0.16	4.59	0.03
2018	**6,121**	**20.16**	**4.60**	**-0.34**	**4.60**	**-0.07**

Tennessee – ACT Composite from 2015 to 2018

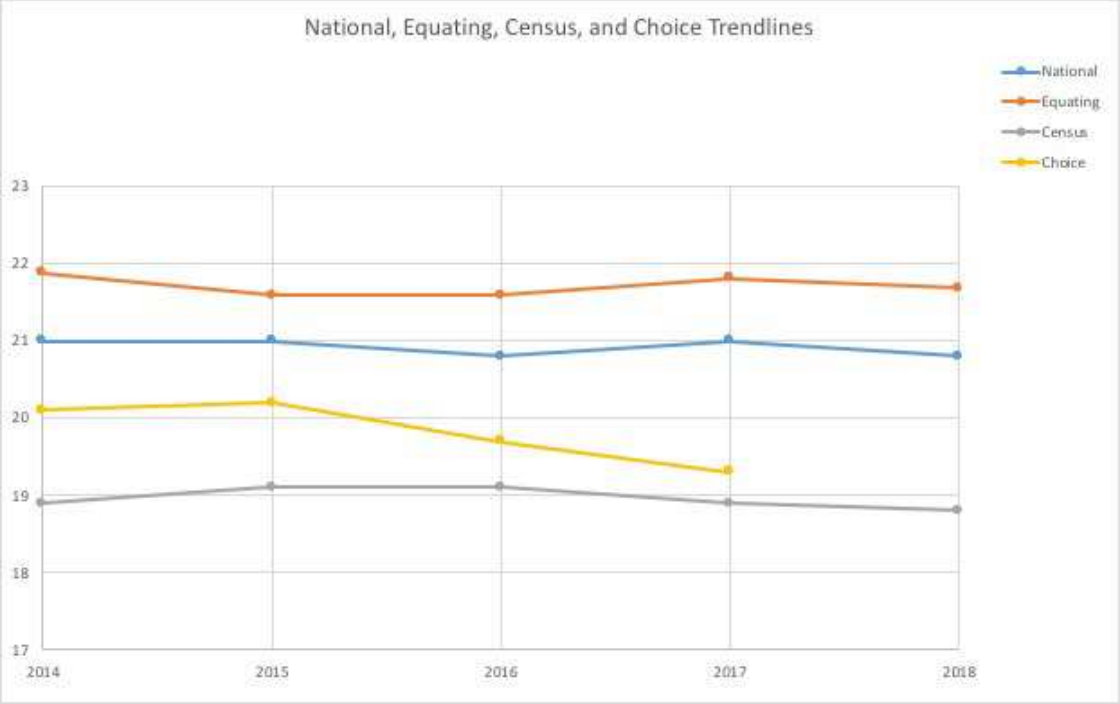
Year	# Test Takers	Mean Score	Standard Deviation	Mean Score Difference (Current - Previous)	Pooled Standard Deviation	Effect Size (Cohen's d)
2015	41,178	18.98	4.99			
2016	51,520	19.48	4.86	0.50	4.92	0.10
2017	39,542	19.46	4.90	-0.02	4.88	0.00
2018	**51,070**	**19.07**	**5.10**	**-0.39**	**5.01**	**-0.08**

We offer two general observations regarding these effect sizes. First, all values are less than .20, the popular criterion for claiming an effect size to be "small." Our second observation pertains to the **bolded** rows in each table. Here, states gave the same primary form, equated using the "nationally representative" sample in 2015. Notice that overall performance in each state decreased after it, in the previous year, had either increased or stayed the same. This trend likely points to a form effect for that specific primary form, a possibility we raised with the ACT psychometric staff. In response, they referred us to ACT's November 2018 presentation to the Louisiana Technical Advisory Committee. This presentation showed two states (Hawaii and Nebraska) whose mean ACT composite score on this primary form was approximately the same as the year before, and one state (Nevada) whose ACT composite score increased (by a tenth of a point). These counterfactuals notwithstanding, we cannot ignore the other 15 states whose mean ACT composite scale score dropped with the administration of the primary form in question.

This final point brings us back to our recurring concerns about what constitutes the "nationally representative" sample, how the equating sample is chosen each year, and, in particular, how it was chosen in 2015. To reiterate the most salient point, the scaling and equating procedure used by ACT is *not* population invariant (Kolen & Brennan, 2004). ACT established the original base scale in 1989. The primary test-taking population in 1989—and for at least the two decades that followed—was college-bound juniors, who took the ACT on the national administration dates. The ACT test-taking population and number of administrations each year both have changed considerably since 1989, particularly in recent years with more students taking the ACT as part of state or district testing. We therefore recommend that ACT address the following questions:

1. Is the ACT scale, established 30 years ago and in a different testing context, still valid?
2. Is the equipercentile-based relationship, which is the core of ACT's year-to-year equating procedure, invariant to the change in test-taking population observed in recent years?
3. How robust are the ACT scaling and equating procedures if the equating sample is not representative of the achievement and demographic characteristics of *all* ACT test-takers, whether they take the test as part of the national administration or within the context of state and district testing?

The figure below, provided by ACT, compares the mean scale scores of various samples over the past five years. These trends, which are relevant to our third question above, show that equating-sample performance is consistently higher than that of the other groups of examinees nationally, and in states and districts.



Because of the timeline and scope of our work, we cannot answer these three questions conclusively. However, the burden of proof for addressing them falls on ACT as the test provider. Toward that end, we encourage ACT to conduct additional analyses to show that its equating procedures and implementation processes account for these and other contextual factors.

Conclusions and Recommendations

This investigation focused on exploring possible explanations for identified score declines across multiple states using the same operational form in either, or both, 2017 and 2018. Although we have not been able to conclusively determine that scores on this form were erroneous, we found at least one potentially serious condition threatening the comparability of ACT scores across forms: the changing national population participating in the ACT over the years. Specifically, this report raises concerns about the possible impact of observed changes in ability distributions of the equating samples between 2014 and 2015. The primary ACT form administered in most census-testing states in 2018 was equated in 2015. If the observed changes in ability distribution in the national population led to an equating solution substantively different from a solution based on little or no year-to-year variation in population ability, the adjustments to form difficulty would contain some level of error that could affect all forms equated that year, including the primary form administered in 2018 to students in Louisiana, Missouri, North Dakota, and Tennessee.

The Center for Assessment has not seen evidence to confirm that the ACT scaling and equating procedures are robust to observed changes in the national test-taking population, a fundamental tenet of ACT's equating design. The Center also identified additional sources of risk in existing ACT procedures and documentation that may threaten the validity of ACT score interpretations. We offer several recommendations for ACT to defend the validity of its scores more completely as its test-taking populations continue to change and expand. Our recommendations fall into two areas: analyses and procedures.

Recommendations for Analysis of Scale Stability and Invariance of ACT Scores

The main technical concern we noted throughout this investigation regards the stability of the ACT scale and the robustness of the ACT scaling and equating procedures to the changing test-taking populations. ACT should engage in more direct testing of the population invariance assumptions associated with the ACT scaling and equating procedures and regularly document the findings. We further recommend that ACT provide more comprehensive documentation about the fidelity of implementation to support claims that ACT scores are valid for all examinees, taking all forms, in each year since the current scale was established in 1989.

Although ACT conducted a study of the invariance of ACT equating solutions across different ability groups, that study was limited to examining equating results based on data from a single, unspecified test form. In this study, approximately 10000 examinees taking a single test form were divided into 3 different samples based on ability groups. ACT's scaling and equating procedures were then applied based on each group, and each of the three outcomes was compared to scaling and equating results based on the full sample. We question whether the study design and interpretation of results apply sufficiently to the context under examination. We

recommend that ACT share the details of their study with its technical advisory committee and make the results available to measurement experts and other stakeholders for independent evaluation.

We also recommend several areas for additional study to more fully interrogate the robustness of ACT's scaling and equating procedures to observed population changes, as well as to uses beyond college readiness for different populations. As the test provider, ACT is responsible for providing evidence in support of three key assumptions:

1. the ACT scale, established 30 years ago and in a different testing context, still produces valid and comparable scores for all examinees across forms and years;
2. ACT's scaling and equating procedures produce examinee scores that are invariant to the observed changes in the ACT national test-taking population in recent years, and
3. evidence to support claims that ACT's scaling and equating procedures produce examinee scores that are valid for intended uses in statewide testing programs.

The Center for Assessment did not find such evidence in our investigation.

Two ideas for studies that might support further examination of invariance assumptions under ACT's testing conditions are provided in this report. ACT may also consider partnering with states and other measurement experts to fully investigate and provide evidence for the three key assumptions.

Recommendations for Procedures and Processes

Definition of 'national' population

The main technical concern we note throughout our investigation is the possible impact of changing ACT test-taking populations. To fully investigate the impact of population changes on equating over time, we urge ACT to adopt a precise definition of what constitutes the target "national" population and how well that target is met each year that new forms are put on the ACT scale. This definition is critical for supporting accurate score inferences because new forms are equated to the ACT scale each year based on samples drawn from the national population.

ACT equating samples come from a single national testing date (October). The Center for Assessment could not find the rationale for this choice in any of the ACT materials, or how well this sample generalizes to all ACT test-takers. The Center concluded that the lack of clear definition in the sampling frame for equating may have implications for the year-to-year comparability of examinee scores and scale stability. ACT's sampling specifications must include more precise definition of "national" and a clearer description of the sampling frame. Accordingly, the Center recommends ACT produce detailed sampling specifications with precise

definitions of the target ACT linking population in order to facilitate the ability to monitor changes overtime that might threaten scale stability and the comparability of scores.

Specifications and technical documentation

The Center identified important gaps in the specifications and technical documentation for several key steps in ACT's standard operating procedures. These gaps included a lack detailed guidelines and specifications for test construction, sampling, data processing, scaling and equating analyses, and quality control.

The Center found the current technical reporting process to be largely ad hoc. ACT produces technical reports periodically. The reports do not include form level information about the technical quality of the tests, which misses important sources of internal validity evidence. At a minimum, we recommend including the types of invariance analyses discussed previously and annual studies to support evidence of score validity, particularly in the context of state census testing. We also recommend including form-specific information in technical documentation such as reliabilities, classical item analysis results, decision accuracy and consistency, and evidence of content alignment with intended frameworks or standards. The Center did not have the opportunity to review external content alignment studies. We urge ACT to make these studies publicly available and incorporate the results into appropriate technical documentation that can be used by participating states to support the validity of ACT use in each specific context. In short, full and regular technical reporting will better serve state clients who must provide evidence of technical quality to their stakeholders and as part of the federal peer review process.

Internal communication

Our interviews with ACT staff indicated that the test development workflow across functional groups involved limited psychometric review of test forms. We recommend ACT improve its standard operating procedures to formalize communication between psychometrics and test development teams to better support coherent quality assurance of forms development. Formal psychometric guidelines for test development staff use in test construction, and to guide formal psychometric QA is one example.

Implementation outcomes

The Center for Assessment did not observe official records of outcomes from key steps of the annual test development, data processing, and psychometric processes. In particular, no documentation showed that ACT met various criteria or targets in the standard operating procedures each year. Many testing programs capture such evidence by completing quality control checklists, with annotations to convey salient observations or issues. A systematic and planful archive of implementation outcomes would engender confidence, both internally at ACT

and externally, that standard operating procedures were implemented with fidelity—particularly in circumstances such as the score decline in 2018.

Context effects

Due to the limited availability of contextual data, we focused our investigation on ACT procedures and implementation, and whether either could have contributed to the score declines in 2018. The investigation into contextual factors was limited to those that are related to ACT's procedures and implementation (e.g., impact of population shifts on the equating sample.) Although there are many contextual factors that could reasonably explain the score declines, they cannot be characterized with confidence before examining ACT procedures and their implementation.

Nonetheless, there are some possible analyses that are within ACT's purview to pursue concurrently with recommended analyses and improvements. For example, we recommend ACT take a closer look at the item-level (p-value) comparisons between the equating sample and the state-specific examinees to identify whether score declines can be attributed to decreased performance in specific content standards. We also recommend that ACT conduct additional analyses on the 2018 primary-form samples and the impact that differences in ability distribution and demographic characteristics might have on the robustness of equating outcomes.

Limitations

The Center did not replicate the equating of the ACT primary form(s) in question as this work has been completed by an independent third party, HumRRO, and the supporting documentation for the replication work appeared complete (with the caveats noted earlier in this report). Per discussions with ACT, the Center only received student-level ACT data from the 2015 equating process. These were the same data provided to HumRRO for its equating replication study and, therefore, were readily available for sharing within the timeline of our investigation. This limited the types of empirical analyses the Center could perform. Our findings and recommendation are therefore based primarily on reviews of ACT materials, interviews with ACT staff, evaluation and simple analyses of the data summaries provided by ACT, and the outcomes of empirical analysis that ACT had conducted.

References

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

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

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer-Verlag

Kolen, M . J ., & Hanson, B . A . (1989) . Scaling the ACT Assessment . In R . L . Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp . 35–55) . Iowa City, IA: American College Testing Program

APPENDIX A: UNDERLYING FRAMEWORK FOR THE INVESTIGATION

Kolen (2007) listed three conditions of measurement that influence linking/equating:

1. Tests' content
2. Administration conditions
3. Characteristics of the population

In practice, possible threats to score comparability in these areas might be elaborated in a general framework of conditions that influence score comparability. This framework considers four areas in which we might group the elements that influence the precision of score linking: 1) design/procedures, 2) implementation, 3) administration, and 4) examinee population/contextual factors. Not all elements in this framework apply to the current ACT investigation. Only those that support investigation of ACT-specific procedures, implementations, and contextual factors are included in the Center's Investigation Plan.

Design/procedural elements that influence measurement

- Blueprint/test specification alignment with framework and/or standards
    - Is alignment well documented?
    - Are the test specifications sufficiently detailed to support alignment consistently across forms?
    - Are test construction procedures sufficient to ensure consistency in meeting targets across forms?
    - Are statistical targets for test construction appropriate and reasonably met for each form?
    - Are there any reasons to believe that content alignment with either standards or the enacted curriculum may have drifted?
- Range of item difficulty
    - Is overall target item difficulty well defined?
    - Is item difficulty reasonably well aligned with targets? With the examinee population?
- Sampling design
    - Does the sampling design for equating support the intended inferences to be made about test scores?
- Field test design (and use of item statistics)
    - Does the design for field testing new items provide an appropriate basis for decisions about test construction, use of item parameters in scoring and/or equating?
- Test length
    - Is the test length sufficient to support reliable scores across equated forms?

- Item statistical quality
    - Is item quality sufficient to contribute acceptable levels of information about examinee performance?
        - Is overall target item discrimination well defined?
        - Are procedures to examine differential item functioning well defined
        - Is target test score reliability well defined?
- Data cleaning procedures
    - Is the rationale appropriate and clear for how examinees are included or excluded from scaling and equating analyses?
- Appropriateness of scaling model/parameter estimation procedures
    - Is the model sufficiently robust to changes in the population?
- Appropriateness of equating method for intended score inferences
    - Is the randomly equivalent groups method robust to changes in the target population (national population sampling frame)?
    - Is there evidence of threats to the equivalency assumptions due to the sampling approach?
    - Are assumptions met?
    - Is invariance across forms and groups met?

<u>Implementation conditions that influence measurement</u>

- Any number of construct irrelevant features of item content, their rendering, or their scoring rules (keys/rubrics)
    - Are the correct keys used?
    - Have items been appropriately reviewed for alignment, bias, appropriateness of reading load and text complexity, unnecessary features of the prompt or item content?
    - Are higher levels of differential item functioning (DIF) present on any operational and/or linking items?
- Item statistical quality
    - Is item quality sufficient to contribute acceptable levels of information about examinee performance?
        - Is overall target item discrimination met?
        - Are items and tests sufficiently free of DIF
        - Is test score reliability sufficiently high and consistent across administrations?
- Field test (FT) item positions
    - Are examinees sufficiently motivated for standalone FT administrations?
    - Is the use of item statistics generated from FT data appropriate?
- Adherence to data cleaning procedures
    - Were the data cleaning procedures followed accurately?

    - Were there unexpected characteristics of the data that were not covered by the rules? If so, how was that handled? Is there reason to believe that the new handling procedure affected the outcome of equating?
- Adherence to scaling method
    - Were scaling procedures followed accurately?
- Adherence to equating method
    - Were equating procedures followed accurately?

Administration conditions that influence measurement

- Accommodations
    - Have there been notable changes in the numbers of examinees receiving accommodations, or the nature of accommodations that might be expected to influence the equating function/results?
- Timing (calendar date and testing time allotted)
    - Are there planned or unplanned differences in testing windows?
    - Is enough testing time consistently allowed to avoid speededness?
- Item exposure/security
    - Are there known or suspected item exposures (aberrant or by design)?
- Sampling implementation
    - Was the equating sample drawn correctly according to the sampling frame and specifications?
    - Was there significant sample attrition (at any level, e.g. state, district, school, classroom, student) from design? What was the nature of attrition—missing at random, not at random?

Examinee/population conditions that influence measurement

- Opportunity to learn (OTL)
    - Are there identifiable gaps or changes in OTL?
- Motivation
    - Are examinees motivated? Is there any reason or condition for why they might not be equally motivated over the administrations of equated test forms?
- Population change
    - Have any significant shifts in the characteristics of the examinee population occurred across administrations?
- Curricular change
    - Have any identifiable changes to the enacted curriculum taken place that may affect the examinee performance overall or performance of specific types of items?

APPENDIX B: LIST OF MATERIALS REQUESTED BY THE CENTER

The Center requested data and documents from ACT that we deemed necessary for answering the questions posed in this investigation. This complete list of materials that the Center requested included:

1. General development/administration/equating analysis schedule
2. Copy of detailed specifications for psychometric analyses
3. Description of the QA procedures used for data cleaning and equating analyses
4. Technical reports and memos from 2016 to 2019
5. 2018 and 2019 equating samples and census data with complete state and individual and school demographic indicators (minimally: gender, race/ethnicity, IEP/504 status, EL status, and if available disadvantage status and any school level indicators, e.g. region, size, public/private/charter, proportion FRL/disadvantage...)
6. Copy of any other analyses conducted to date to investigate score drops, including descriptive summaries by student groups.
7. Description of test design changes in 2017 or 2018 such as:
    - Content/blueprint
    - Item exposure rates
    - Field test item positions
    - Scaling/equating procedures (including exclusions rules, data processing/cleaning rules)
8. Any ACT results from the spring 2019 administrations for the four states
9. Test development specifications, such as detailed design info blueprint, plus finer grain information on depth of knowledge, passage difficulties/types, etc.
10. Any documented psychometric guidance for test development and psychometric review of test form procedures
11. List of data/materials provided to HumRRO (in 2018) for its equating replication
12. ACT's 22-point investigation documentation
13. ACT survey questions given to test takers (and summary of responses, if possible)
14. Spreadsheet of historical form assignments (across states and year equated – life cycle and usage views)
15. A plot of trendlines of mean scale scores from 2015-2019 that includes lines for:
    - National sample
    - Equating sample
    - Census States (combined)
    - District/school choice states (combined)
    - Plus, clarification about the differences in these samples and their respectively labels. For example, does the national sample represent all states? Or only the

non-census states? Does the national sample represent the sampling frame for the equating sample?

16. Item statistics for the primary form(s) administered to the four states from 2015 to 2019, the anchor form(s) that was used to equate the primary form(s), and any "sibling" forms; that is, forms with which the primary form(s) was equated.
    - Include any statistics or criteria that are used for anchor form selection; at the minimum, we would like to have p-values and item-total correlations for all operational/scored items on the forms.
17. The national sample raw score distributions for each of these forms in each of these years (2015-2019).

Additional materials requested and received after review of items 1 to 17:

18. 2014 statistical summaries for the anchor form used in 2015
19. Form designations for national testing, and the following items:
20. Scale score mean and standard deviation stats for equating samples
21. Elaboration of the equating specs to including methodological details
22. Sampling design for equating, to include the quantifiable ways that the equating sample representative of the national sample, i.e. how is representation assured—cell design and/or ability/demographic matching routine?
23. Details for predicting operational item difficulty based on field test item difficulty
24. The 2015 raw to scale score tables for all forms equated in 2015, and for the anchor form