Research Report

Building a Validity Argument While Developing and Using an Assessment: A Concurrent Approach for the *Winsight*® Summative Assessment

ETS RR–19-26

Elizabeth Stone
E. Caroline Wylie

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Building a Validity Argument While Developing and Using an Assessment: A Concurrent Approach for the *Winsight*® Summative Assessment

Elizabeth Stone & E. Caroline Wylie

Educational Testing Service, Princeton, NJ

We describe the summative assessment component within a K–12 assessment program and our development of a validity argument to support its claims with respect to intended uses and interpretations. First, we describe the *Winsight*® assessment program theory of action, a logic model elucidating mechanisms for how use of the assessment components can lead to improved student learning in mathematics and English language arts (ELA) and defining the roles of key groups of stakeholders. We explain our process of developing the validity argument that enumerates sources of validity evidence supporting the Winsight summative assessment claims drawn from the theory of action. We examine candidate structures for the argument, describe the steps in which we unpacked stakeholder-level claims and defined subclaims to operationalize components, and discuss types and sources of validity evidence to meet the goals of professional and federal guidelines. Finally, we discuss the research agenda and operational analyses that will provide this validity evidence.

Keywords Test quality; validity; K–12 accountability; USDE peer review; *Winsight*®

doi:10.1002/ets2.12261

Validity is the "cardinal virtue in assessment" (Mislevy, Steinberg, & Almond, 2003, p. 4). This statement reflects, among other things, the fundamental role of validity in test development and evaluation of tests (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). While the conceptualization of validity has evolved (Cureton, 1951; Kane, 2006; Messick, 1989), evidence to support the validity of the intended uses and interpretation of scores from an assessment is required to adhere to the current *Standards for Educational and Psychological Testing* (hereinafter referred to as the *Standards*; AERA, APA, & NCME, 2014). This evidence is also needed to meet federal peer review guidance criteria for K–12 accountability assessments (e.g., U.S. Department of Education [USDE], 2015, 2018).[1] As we note herein, some states also have their own specific standards, and additional evidence may need to be collected to address validity questions that are specific to their contexts.

Traditional K–12 test development begins with a set of grade-level content standards, which are then used to develop a test blueprint that specifies the number of test items per standard (or substandard, depending on the scope of a standard) and, in some cases, the item types or point values (weights) for each item. Examples of K–12 content standards include the Common Core State Standards[2] and the state-specific Florida Sunshine Standards.[3] The standards and blueprint documents (along with performance-level descriptors) drive test development and test form assembly. Items are field tested, and their performance is analyzed and reviewed for possible remediation. Traditional operational statistical analyses are completed at regular intervals with respect to test administration. All of the aforementioned development and analysis activities produce evidence that can be used to support the validity of the assessment uses and interpretations. Post hoc or nonoperational research studies may also be used to collect other kinds of validity evidence. Optimally, the performance-level descriptors and claims about uses of the assessment should be employed to model reporting prototypes at early stages of development to identify what work needs to be planned to provide the desired information to stakeholders (e.g., in prospective score report designs; Zapata-Rivera, Hansen, Shute, Underwood, & Bauer, 2007; Zieky, 2014).

In some cases, particularly when assessments are components of an integrated system, a theory of action or logic model is developed prior to test development; however, even in such cases, assessment systems do not always use the theory of

Corresponding author: E. Stone, E-mail: estone@ets.org

action to examine consequences of test use or as part of an iterative design cycle. The failure to employ a concurrent approach to developing and validating the claims in a theory of action presents several issues. First, the ways in which we anticipate the assessment results being used should drive design, and, therefore, preliminary claims need to be considered early in the process. Second, the resulting information may be obtained too late to have the necessary impact (e.g., on test design and development). Third, the scope of possible analysis types and evidence types that can be collected may be constrained by the test design or delivery system, and it is important to be able to know this to revise approaches to collecting validity evidence. For these reasons, we are fortunate to have the opportunity to consider validity issues from the beginning of the design work on the *Winsight*® assessments in K–12 English language arts (ELA) and mathematics. This more proactive and concurrent approach, which has begun to emerge in the field of K–12 measurement,[4] allows for the validity argument to evolve and be refined along with the assessment throughout the life cycle of the assessment.

Validation efforts have been undertaken by Educational Testing Service (ETS) staff in many areas, including assessment developers, research scientists, and psychometricians conducting operational analyses (Kane & Bridgeman, 2017). These areas work collaboratively to evaluate validity from different perspectives, and there is substantial coordination and collaboration across areas to capitalize on cross-area efficiencies and knowledge and to help ensure that all types and sources of validity evidence that are relevant to the claims in the theory of action are explored.

In this report, we describe a concurrent approach to designing and developing a validity argument for the Winsight summative assessment, a K–12 assessment intended for use by states to meet federal accountability requirements. First, we describe the Winsight assessment program theory of action, a logic model that elucidates claims—paths from assessment components, results, and associated supports to improved student learning on important academic outcomes in mathematics and ELA—for three sets of stakeholders: state education agency (SEA) and local education agency (LEA) decision makers, teachers, and parents/guardians and students. We examine several candidate models or structures for the argument and describe a process of developing a validity argument that enumerates sources of validity evidence that support the Winsight summative assessment claims—drawn directly from the theory of action—with a focus on the interpretations and uses of assessment data that are to be made by the different stakeholders. Then we describe how we unpacked the stakeholder-level claims and defined subclaims to operationalize each component. We next discuss types of validity evidence and examples of their sources before describing the research agenda and operational analyses that will provide this validity evidence.

Winsight Assessment Program Theory of Action

The Winsight summative assessment is one component of the Winsight assessment program. The Winsight summative assessment component is designed to measure ELA and mathematics proficiency for students in Grades 3–8 and high school. While the summative component will eventually have a multistage adaptive format, the initial phases of data collection (the pilot, field test, and initial operational year) will use conventional linear tests with the goal of obtaining item statistics to support the transition to multistage testing. The summative component design includes multiple-choice items, other selected-response items (e.g., multiple-select, inline choice, drag and drop, select in passage, table grid, and zone items), and constructed-response items (e.g., numeric and fraction entry, graph items, fill in the blank, short and extended text).[5] A combination of machine-scoring and human-scoring approaches will be used.

While we focus on the summative assessment component and summarize aspects of the logic model and mechanisms specific to that assessment, it is important to consider the summative assessment within the framework of the full assessment system, which also includes interim assessment (e.g., benchmark and testlet) and formative assessment components. The development of a theory of action for this particular assessment program (Wylie, 2017) is in keeping with the NCME's (2018) position statement on theories of action for testing programs:

> For example, many state K–12 testing program sponsors intend, through test design and score use, to cause change in the behavior of school leaders and teachers. The intended change may occur through the reallocation of resources or the shifting of instructional practices toward improving particular student competencies. In these instances, the testing program takes on the role of an intervention that may, in some ways, be as important as its measurement function. As such, the claims of testing program sponsors are different from the claims associated with more conventional testing programs. (p. 1)

As we describe in this section, the Winsight summative assessment is designed to produce test scores and interpretations; however, it is also designed (as part of a larger program) to improve teaching and learning and to inform resource allocation toward that goal, so it is important to provide a model for the causal mechanisms that we expect to underlie these changes.

The Winsight assessment program was conceptualized according to five undergirding principles (Wylie, 2017):

1. Winsight is intended to improve teaching and learning in the United States. This improvement will result from the availability of assessments that are fair and valid to all groups, that produce reliable scores at the individual and classroom levels, that can provide enriched feedback to stakeholders about student proficiency and level of understanding in key subdomains to inform instruction, and that can guide effective resource allocation at the district and state levels to improve the educational environment. A key aspect of this principle is the focus on accessibility, enacted by the use of universal design principles (Dolan & Hall, 2001). The resulting accessible test design, combined with the availability of test delivery accessibility and accommodations features, will support all students in demonstrating what they know, understand, and can do. There is a specific focus on ensuring accessibility for students in traditionally underserved populations, such as students with disabilities and students who are English learners.
2. Winsight test design will be informed by learning progressions and key practices that articulate the spectrum of understanding of a particular subdomain and the various levels of advancement from novice to master of the knowledge (e.g., Arieli-Attali & Cayton-Hodges, 2014; Confrey, Maloney, Nguyen, Mojica, & Myers, 2009; Deane et al., 2015; Deane, Sabatini, & O'Reilly, 2012; Deane & Song, 2015; Graf & van Rijn, 2016; Heritage, 2008).
3. Winsight reporting and feedback will be targeted to groups of stakeholders who require information at different grain sizes to improve student learning in their own particular spheres of influence.
4. The Winsight formative assessment component in particular recognizes the complementary roles for teachers and students so that the formative assessment provides opportunities for students to develop intra- and interpersonal skills as well as cognitive skills as they engage in self- and peer assessment. Rich tasks and teacher materials are intended to support students and teachers in engaging meaningfully in these interactions (Cohen, Raudenbush, & Ball, 2003).
5. The outcomes displayed in the theory of action are necessarily supported or restricted by policy, economic, educational, and physical environments within which the assessment system is situated.

The logic model for the Winsight assessment program (Figure 1) includes the program components summarized on the left side of the figure. These components consist of the three assessment components along with system supports. The system supports are intended to go beyond supporting test administration, helping to improve stakeholder assessment literacy and use of Winsight reports, model good teaching practices, and support teachers to use system resources such as learning progressions. Use of these components is hypothesized to lead to the intermediate and longer term outcomes that flow from left to right. The figure illustrates the relationships among key intermediate outcomes and how they work together as a system to support improved student learning. The claims in the logic model (numbered arrows) are articulated in the form of relationships between program components and intermediate outcomes or between intermediate and longer term outcomes. Each claim is in the form of an if–then statement. The full theory of action report (Wylie, 2017) provides the claim for each arrow and preliminary logical, theoretical, or empirical evidentiary support for the claim. We note that the theory of action is also labeled as preliminary because it may be modified over time as components are added, removed, or refined and as we better understand how one component or intermediate outcome impacts the system as a whole.

Figure 1 depicts the complete logic model for the full assessment program. A subset of the numbered and color-coded paths shows how each group of stakeholders will use Winsight summative results to improve student learning on important academic outcomes in mathematics and ELA. Following the connections from the summative component, Links 1 through 3, 5, 6, 16, and 18 describe how three stakeholder groups will interact with and use data from this component to ultimately improve student learning outcomes:

- SEA and LEA decision makers use summative results to understand performance in the aggregate and with respect to differences among traditionally underserved students and to effectively allocate resources to schools and classrooms as needed (Links 3 and 5).

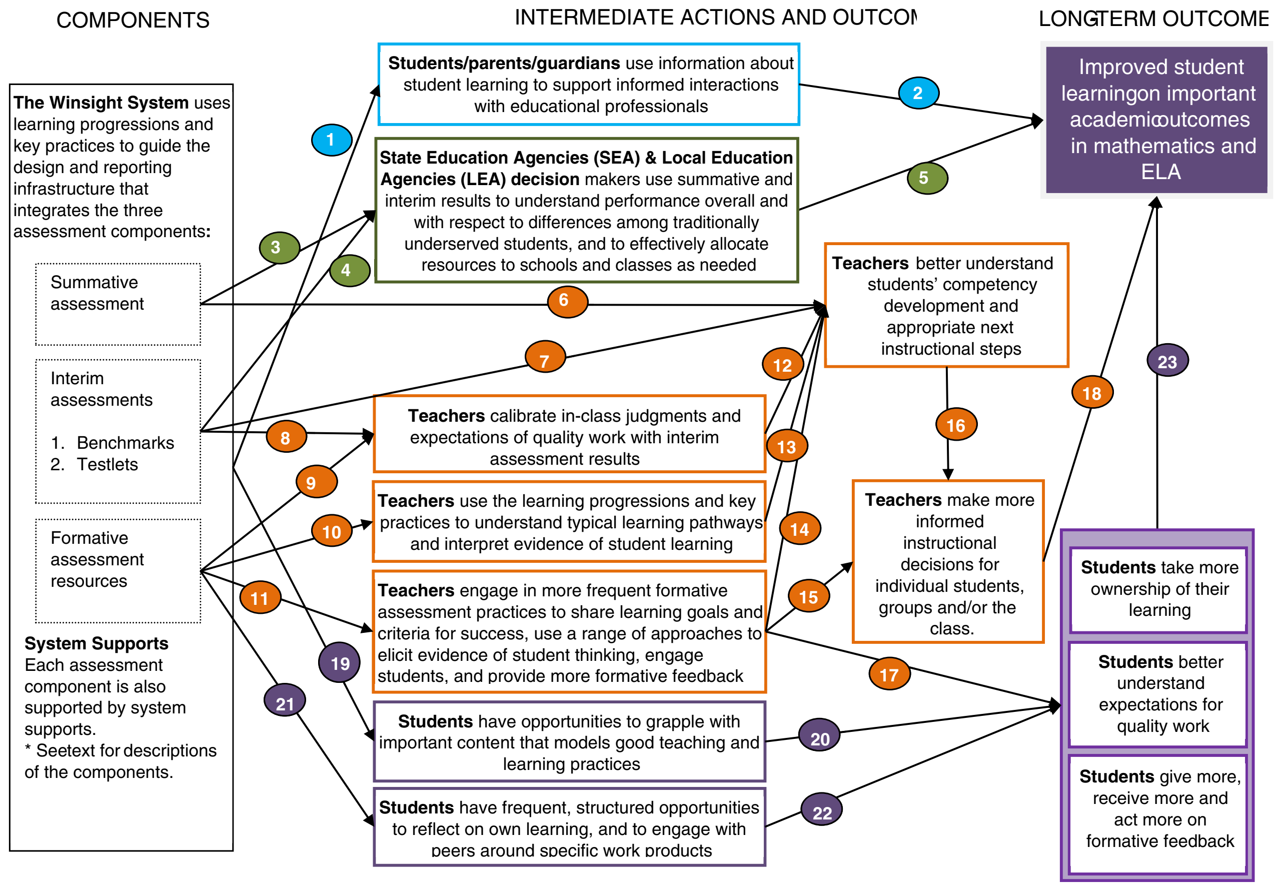

COMPONENTS
INTERMEDIATE ACTIONS AND OUTCOI
LONGTERM OUTCOME
The Winsight System uses learning progressions and key practices to guide the design and reporting infrastructure that integrates the three assessment components:
Summative assessment
Interim assessments
1. Benchmarks
2. Testlets
Formative assessment resources
System Supports
Each assessment component is also supported by system supports.
* Seetext for descriptions of the components.
Students/parents/guardians use information about student learning to support informed interactions with educational professionals
State Education Agencies (SEA) & Local Education Agencies (LEA) decision makers use summative and interim results to understand performance overall and with respect to differences among traditionally underserved students, and to effectively allocate resources to schools and classes as needed
Teachers better understand students' competency development and appropriate next instructional steps
Teachers calibrate in-class judgments and expectations of quality work with interim assessment results
Teachers use the learning progressions and key practices to understand typical learning pathways and interpret evidence of student learning
Teachers engage in more frequent formative assessment practices to share learning goals and criteria for success, use a range of approaches to elicit evidence of student thinking, engage students, and provide more formative feedback
Teachers make more informed instructional decisions for individual students, groups and/or the class.
Students have opportunities to grapple with important content that models good teaching and learning practices
Students have frequent, structured opportunities to reflect on own learning, and to engage with peers around specific work products
Improved student learningon important academicoutcomes in mathematics and ELA
Students take more ownership of their learning
Students better understand expectations for quality work
Students give more, receive more and act more on formative feedback
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Figure 1 Preliminary Winsight assessment program logic model. Reprinted from Winsight™ *Assessment System: Preliminary Theory of Action*, by E. C. Wylie, 2017, Research Report No. RR-17-26, p. 5. Copyright 2017 by Educational Testing Service.

- Teachers better understand students' competency development and appropriate next instructional steps; therefore, teachers make more informed instructional decisions for individual students, groups, and/or the class (Links 6, 16, and 18).
- Students and parents/guardians receive multiple sources of assessment information, including information from the summative assessment, which supports more informed interactions with educational professionals and improved communication between students, parents/guardians, and educational professionals (Links 1 and 2).

These high-level claims from the theory of action are not yet in the form of a validity argument, because these theory of action claims focus on how each stakeholder group will use the Winsight summative assessment data as part of the full assessment program but leave issues of data quality unaddressed. In the sections that follow, we describe the process that we used to build on the claims of the theory of action to articulate the validity argument.

Winsight Summative Assessment Validity Argument Development

Considerations for Choosing a Validity Argument Structure

How did we move from the high-level claims about actions and outcomes in the theory of action's logic model to a formal validity argument? Our next step was to develop a validity argument that would specify assessment claims and associated evidence and that entailed choosing a validity framework that would facilitate explication of this hierarchical structure. Conceptualizations of validity have evolved over time, and validation frameworks have evolved along with them. Validity arguments come in many forms, and we reviewed several approaches before settling on the structure that we ultimately used. The two assessment-related frameworks (Bachman, 2005; Bachman & Palmer, 2010; Kane, 2006, 2013) that most influenced the development of our own validity argument both invoke the Toulmin (1958/2003) argument model, so we begin there. Given that we are developing a K–12 accountability assessment, we were also mindful of the USDE (2015, 2018) peer review guidelines, the standards to which state departments of education are held accountable.

In the Toulmin (1958/2003) model, there are observed data (e.g., a student essay response) that are to be interpreted as demonstrating a claim or a decision made based on the data (e.g., the student has a high level of writing proficiency). To make this inference, there must be several other components of the argument. A warrant must be specified that explains why observing particular data supports the claim (e.g., rubrics are applied in an accurate and consistent manner). Evidence (or backing) must then be presented to support the warrant (e.g., there is evidence of rigorous rater training with the use of benchmark and training examples and a qualification round to certify raters able to score; interrater agreement levels meet or exceed industry standards). There may be alternative (or rebuttal) hypotheses to consider that would explain the data differently (e.g., students may hear or be informed that longer responses tend to result in higher scores independent from proficiency), and these would be evaluated in the face of rebuttal data (e.g., studies to examine the relationship between length and scores to evaluate impact of training on rater bias show that length of essay has no significant impact on resulting scores over and above intended judgments of the essay quality).

One approach to developing a validity argument that we considered was based on the Bachman (see, e.g., Bachman, 2005) model, which applied the Toulmin (1958/2003) argument model to a language assessment. This approach attempted to extend the validity argument, which focused on intended score-based interpretations, to incorporate test use explicitly, which was something Bachman argued had been missing from previous attempts to link validity and consequences of test use. He proposed an assessment use argument (AUA) that connects test scores, interpretations based on those scores, and decisions made on the basis of the interpretations. Bachman proposed four types of warrants as a basic set to which others may be added, depending on the assessment for which the AUA is being created. The first three of these are based on the work of Messick (1989), and the fourth focuses on the assessment in the context of all information being used to make the required decision. The four warrant types are relevance, utility, intended consequences, and sufficiency. Two types of rebuttals were proposed by Bachman: alternate reasons for not making the decision stated in the claim (or for choosing to decide differently) and unintended consequences from making the decision based on the data or from using the assessment. For a more expanded description of an AUA, and an additional example from a language assessment, see Schmidgall (2017).

The other approach that we considered was the Kane (2006) argument-based approach framework, which includes an interpretive argument that "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (p. 23) and also makes use of the Toulmin (1958/2003) argument model. Some of the inference types that may make up an interpretive argument include scoring, generalization, and extrapolation. For each inference, a warrant, assumptions, and evidence that would support the assumptions are supplied. The validity argument involves an evaluation of the interpretive argument to determine its completeness and coherence, the reasonableness of the included inferences, and the plausibility of assumptions supporting the warrants for the inferences (e.g., using validity evidence to do so). Kane subsequently updated his terminology to expand the interpretive argument into an interpretation/use argument to include score usage more explicitly; however, he included equal emphasis on both interpretation and use (see, e.g., Kane, 2013, 2016). Sireci (2013), commenting on Kane's (2013) update, suggested that validation could be made simpler by focusing on three steps: articulating the test's purposes, considering how the test might possibly be misused, and evaluating test purposes and possible misuses according to the five sources of evidence presented in the *Standards* (AERA, APA, & NCME, 2014), discussed subsequently.

The choice of a validation framework may depend to some extent on the purpose and structure of the assessment to be developed and evaluated. As an educational assessment, the Winsight summative assessment component requires validity evidence to support its claims about intended uses and interpretations in accordance with the *Standards* (AERA, APA, & NCME, 2014). Furthermore, to use a K–12 test (such as Winsight) for accountability purposes, states must be able to meet the regulatory and statutory requirements of the Elementary and Secondary Education Act of 1965 and as amended (currently the Every Student Succeeds Act). The USDE (2015) provided nonregulatory guidance with descriptions and evidentiary examples for six sections of peer review critical elements: (a) statewide system of standards and assessments, (b) assessment system operations, (c) technical quality—validity, (d) technical quality—other, (e) inclusion of all students, and (f) academic achievement standards and reporting. The 2018 update (USDE, 2018) includes a seventh critical element section if applicable for the assessment under review: locally selected nationally recognized high school academic assessments. As the name implies, this new section pertains to district-level use of nationally recognized accountability assessments in place of the state assessment. In addition, some elements within sections have changed or been added in the revised version (e.g., to include meaningful consultation in the development of challenging state standards and assessments as part of the "Statewide System of Standards and Assessments" section). Because states must demonstrate that the assessments they use for accountability purposes meet the peer review criteria, states require that potential vendors provide documentation describing how peer review evidence will be collected and available for reporting at the appropriate time in the life cycle of the test. Further, states must demonstrate evidence of a system to continue to monitor and improve their assessments. Therefore, our validity argument was informed not just by frameworks and claims structures but by the specific evidence required to demonstrate that states using the Winsight summative assessment component will be able to meet peer review requirements.

Developing the Framework for the Winsight Validity Argument

We developed validity argument outlines from both the perspectives of Kane (2006) and Bachman (2005), as each included elements that we conceived of as important for supporting the assessment. However, neither structure completely met our needs. The Kane (2006) structure did not allow us to tie in the validity evidence to the peer review guidance as explicitly as we desired, and the language has been viewed as more theoretical and difficult to communicate with stakeholders who are not grounded in validity theory (see, e.g., Sireci, 2013). We also used the Bachman (2005) structure, as illustrated in Wang, Choi, Schmidgall, and Bachman (2012) for the Pearson Test of English Academic (PTEA). As was described previously, this AUA approach involved listing claims and warrants, supporting evidence, and potential rebuttals. Structuring our validity argument this way got us closer to being able to identify the types of evidence that we would also need for peer review.

Our current validity argument is a hybrid of these structures, grounded in validation processes recommended by the *Standards* (AERA, APA, & NCME, 2014) but with a focus on specific peer review evidence requirements. Through the Winsight theory of action, we had articulated for each assessment component how stakeholders should use the data provided. For the validity argument, we added in the more explicit articulation of data quality so that we could define a clear purpose statement for each of the three stakeholder groups as an intended interpretation or use of information from the assessment. Each purpose statement describes who the focus of the statement is, what information or resources will be available to them, what they will be able to do as a result of that information, and finally, what the observable or measurable outcomes will be as a result. In other words, we focused the purpose statements on how particular types of assessment information will be used by identified groups of stakeholders for specific purposes, akin to the approach suggested by Sireci (2013). As we noted previously in the theory of action, we directly associate purpose statements for three groups of stakeholders: (a) SEA and LEA decision makers, (b) teachers, and (c) students and parents/guardians. As we describe in the next section, we broke each purpose down and specified claims and subclaims that would support each part. The use of the term *subclaim* instead of *warrant* is more inclusive and fits our purposes better. The validity argument includes each of the three purpose statements (one per stakeholder group), their specific claims and subclaims, the evidence that we will use to support the subclaims directly, and through those subclaims, the claims and overall purpose statement.

From Framework to Validity Argument

Having arrived at a structure that we believed would support the intended interpretations and uses of the test scores and also allow state clients to meet USDE peer review guidance criteria, we had to articulate the claims and subclaims and determine which types of validity evidence we would need to obtain to demonstrate compliance in each area and to support each claim. Figure 2 displays the validity argument structure and calls out the examples.

Purpose Statements

The theory of action displays the logical paths that show how each group of stakeholders will be able to use the assessment results to improve student learning outcomes. Therefore, the purpose statements for each stakeholder group had to be constructed to reflect that interpretation and use more specifically. For example, the theory of action illustrates

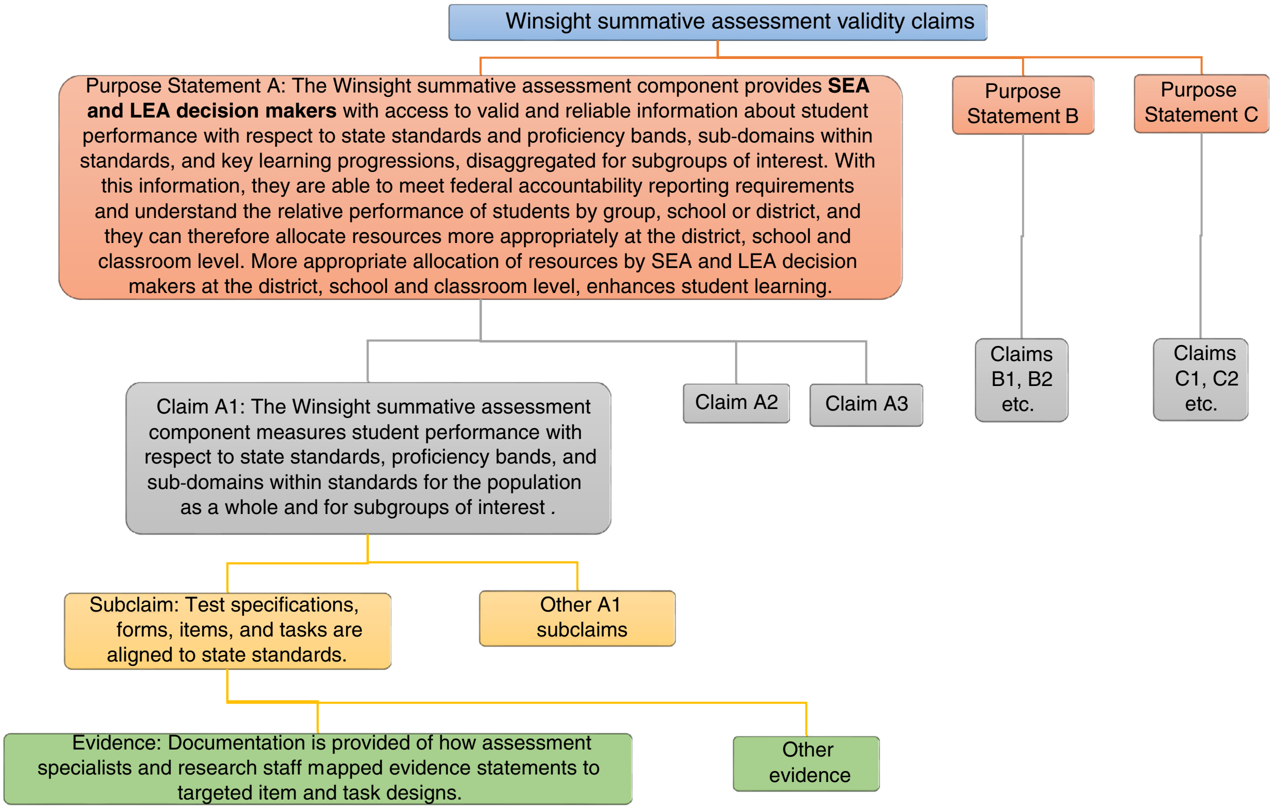

Winsight summative assessment validity claims
Purpose Statement A: The Winsight summative assessment component provides **SEA and LEA decision makers** with access to valid and reliable information about student performance with respect to state standards and proficiency bands, sub-domains within standards, and key learning progressions, disaggregated for subgroups of interest. With this information, they are able to meet federal accountability reporting requirements and understand the relative performance of students by group, school or district, and they can therefore allocate resources more appropriately at the district, school and classroom level. More appropriate allocation of resources by SEA and LEA decision makers at the district, school and classroom level, enhances student learning.
Purpose Statement B
Purpose Statement C
Claim A1: The Winsight summative assessment component measures student performance with respect to state standards, proficiency bands, and sub-domains within standards for the population as a whole and for subgroups of interest .
Claim A2
Claim A3
Claims B1, B2 etc.
Claims C1, C2 etc.
Subclaim: Test specifications, forms, items, and tasks are aligned to state standards.
Other A1 subclaims
Evidence: Documentation is provided of how assessment specialists and research staff mapped evidence statements to targeted item and task designs.
Other evidence

Figure 2 Structure and example components from Winsight summative validity argument. SEA = state education agency; LEA = local education agency.

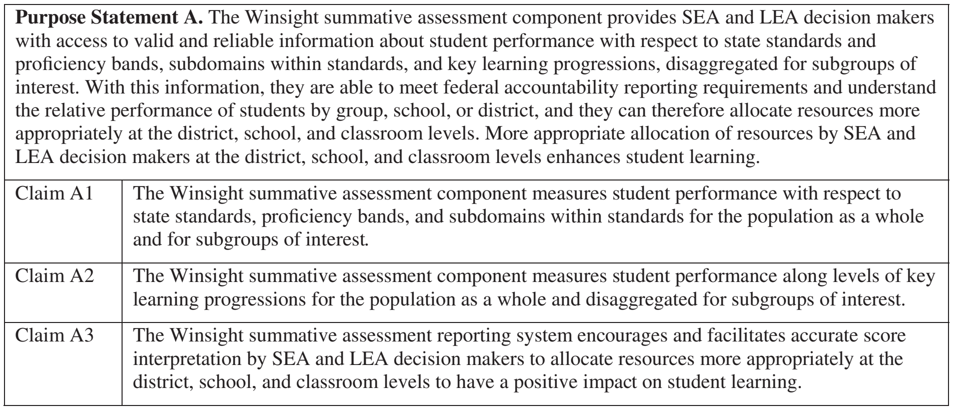
that SEA/LEA decision makers (Figure 2, Purpose Statement A) will be able to use the assessment results to understand aggregate and subgroup-level performance and to use that information to allocate resources appropriately. Unpacking that, we expanded the statement as follows:

> The Winsight summative assessment component provides SEA and LEA decision makers with access to valid and reliable information about student performance with respect to state standards and proficiency bands, subdomains within standards, and key learning progressions, disaggregated for subgroups of interest. With this information, they are able to meet federal accountability reporting requirements and understand the relative performance of students by group, school, or district, and they can therefore allocate resources more appropriately at the district, school, and classroom levels. More appropriate allocation of resources by SEA and LEA decision makers at the district, school, and classroom levels enhances student learning.

This statement describes the stakeholders, which information they will be able to use, and how they will be able to use it. However, the purpose statement itself is not specific enough to operationalize its components to produce evidence statements. Therefore, we parsed the purpose statement into claims. For example, Purpose Statement A (for SEA/LEA decision makers) is composed of three claims, as shown in Figure 3.

Claims

In Figure 3, the first claim, A1, relates to the first part of the purpose statement that SEA/LEA decision makers will have valid and reliable information about student performance, A2 targets the role of learning progressions to provide information, and A3 focuses on the use of the results. We unpacked the other stakeholder-level purpose statements similarly. Purpose Statement B for teachers, Purpose Statement C for parents/guardians and students, and their associated claims (B1, B2, B3, C1, and C2) are provided in the appendix.



Purpose Statement A. The Winsight summative assessment component provides SEA and LEA decision makers with access to valid and reliable information about student performance with respect to state standards and proficiency bands, subdomains within standards, and key learning progressions, disaggregated for subgroups of interest. With this information, they are able to meet federal accountability reporting requirements and understand the relative performance of students by group, school, or district, and they can therefore allocate resources more appropriately at the district, school, and classroom levels. More appropriate allocation of resources by SEA and LEA decision makers at the district, school, and classroom levels enhances student learning.	
Claim A1	The Winsight summative assessment component measures student performance with respect to state standards, proficiency bands, and subdomains within standards for the population as a whole and for subgroups of interest.
Claim A2	The Winsight summative assessment component measures student performance along levels of key learning progressions for the population as a whole and disaggregated for subgroups of interest.
Claim A3	The Winsight summative assessment reporting system encourages and facilitates accurate score interpretation by SEA and LEA decision makers to allocate resources more appropriately at the district, school, and classroom levels to have a positive impact on student learning.

Figure 3 Purpose statement and claims for state education agency/local education agency decision makers. SEA = state education agency; LEA = local education agency.

Subclaims

Having elucidated and unpacked claims, we developed subclaims to describe specific components of each claim. For example, for Claim A1, one subclaim that we specified was “test specifications, forms, items, and tasks are aligned to state standards” (Figure 2) as a building block of being able to make the claim of a valid and reliable assessment that provides information about proficiency relative to state standards. It is specific enough that we could then identify pieces of evidence to support this subclaim and, by extension, Claim A1 and Purpose Statement A.

Evidence

We stopped disaggregating subclaims as soon as they reached a level of specificity where it became fairly straightforward to identify evidence that could be used to support them. For example, in considering the alignment among test specifications, forms, items, tasks, and state standards, one piece of evidence (Figure 2) involves documentation of the decisions made as part of the development process, tracking how assessment specialists translated standards to evidence statements to test blueprints and task models to assessment items following an evidence-centered design process (Mislevy, Steinberg, & Almond, 2002). Additional evidence comes from the documented review process in which external experts engaged to review the evidentiary chain that goes from standards to assessment items. An example of empirical evidence comes from an externally conducted alignment study that is currently under way and due to be completed in 2019. Several methods can evaluate alignment of an assessment (and its items and tasks) with state standards. The Achieve (2006) model, for example, involves collecting evidence from expert rater panel judgments on six criteria: accuracy of the test blueprint, content centrality, performance centrality, challenge, balance, and range (Forte & edCount, 2016). Our alignment study will include a focus on this model, and the resulting findings and recommendations will allow us to identify and remediate issues of coverage in the item pool and test blueprint.

We followed the approach described earlier for each claim, identifying potential evidence to support each subclaim from the perspectives of assessment development, operational psychometric analysis, and research. Some evidence (e.g., blueprint evaluation and standard statistical analyses, such as reliability calculations) is collected in the context of assessment development and routine psychometric analyses of field test or operational data. However, other evidence components may not arise from operational work. In these cases, special research studies can fill in those gaps. As we developed the validity argument and identified sources of evidence, we also developed our research agenda, which would allow us to plan those research studies at the appropriate times in the life cycle of the test (e.g., during development, the pilot, the field test, or from operational administrations).

The concurrent development of validity argument and research agenda also helps with determining the appropriate timing of particular aspects of the work. For example, by making explicit the uses of the assessment information by stakeholders, our attention is focused not only on the prospective approach to developing score reports but also on the importance of communication about appropriate and inappropriate uses of the information. Development of these information tools takes time and resources, along with time needed to then collect evidence from stakeholders to determine whether each stakeholder group understands the information in the way that it was intended. Further revisions and piloting of reports and supports around those reports may then be needed.

In the validity perspective that we worked from (e.g., AERA, APA, & NCME, 2014), validity evidence is categorized as evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of testing. The 2015 USDE peer review guidance did not specifically call out this latter category as a critical element. However, the 2015 guidance did indicate that parents and guardians should be made aware of possible consequences of their children taking an assessment based on alternate achievement standards rather than a general assessment such as the Winsight summative assessment component (e.g., the students must be able to demonstrate their content-area proficiency on the state's general assessment to be eligible for a regular high school diploma). While the emphasis on consequential validity in the peer review guidance is currently not as strong as that in the *Standards* (AERA, APA, & NCME, 2014), federal and state policies may evolve over time, as demonstrated by the aforementioned revisions to the peer review guidance. Because consequential validity evidence in particular typically takes several years of operational testing to acquire, we recommend that assessment developers first articulate consequences of assessment use through the theory of action and then plan to collect evidence of those anticipated consequences.

Identifying Sources of Validity Evidence

As part of the process of moving from the broad statements based on the system-level theory of action and drilling down through increasingly fine-grained levels of claim statements and evidence specifically for the Winsight summative component, we also noted the primary sources of the evidence. Two areas for which it was more challenging to find exemplar approaches and studies were validity evidence based on response processes and on the consequences of testing.

The peer review guidelines explicitly mention validity evidence based on response processes. According to a review by Cizek, Rosenberg, and Koons (2008) of the tests included in the 16th *Mental Measurements Yearbook* (Spies & Plake, 2005), only 1.8% of tests had response-process evidence reported. The lack of this form of research evidence may represent the challenge of obtaining insight into what is occurring in test takers' minds as they work through and respond to test items and tasks. Some techniques that have been used to elicit this evidence include cognitive interviews and think-aloud studies. There are some differences between the techniques with respect to when in the life cycle of the test they are used, how they are implemented, and what degree of evidence is required based on the stakes of the interpretations that will be made from them. For example, cognitive interviews (focusing on comprehension processes) may be used at earlier stages of test development in an exploratory manner, and think-aloud studies may be used in a confirmatory matter (focusing on items and tasks that developers assume are targeting higher level skills, such as problem-solving processes; see, e.g., Bechard, Almond, & Cameto, 2011; Johnstone, Altman, & Moore, 2011; Leighton, 2017; Padilla & Leighton, 2017). Winsight and other online summative assessments have moved toward a greater proportion of items that are not multiple choice—stock-in-trade of No Child Left Behind summative assessments—with claims that the newer, more innovative item types that take greater advantage of the digital environment provide greater coverage of the depth of the standards. These more novel item types may also require students to interact with the items in more complex ways, potentially allowing researchers to understand better whether students are engaging in the level and quality of response processes intended in the item design as the students work through the item and think-aloud. While both cognitive interviews and think-aloud studies provide important pieces of evidence about response processes, they can be time intensive to employ due to the one-on-one nature of the methods; therefore, they tend to be used with smaller sample sizes.

Newer techniques that have been used to obtain evidence of response processes include keystroke logging and eye tracking. These approaches provide data about how test takers construct their responses, navigate the assessment, and interact with assessment content (e.g., how long does the student attend to a reading passage before attempting the comprehension questions), but these and other methods produce large amounts of digital data that need to be sifted to make interpretations (see, e.g., Blascheck et al., 2014, for strategies on visualizing eye-tracking data). Eye tracking can also be time intensive to employ.

The Cizek et al. (2008) review identified that only 2.5% of tests in their study investigated issues of consequential evidence. This finding was further investigated by Cizek, Bowen, and Church (2010), in which additional possible reporting sources for consequential validity evidence were included in the search. Despite widening the search scope, in a review of approximately 2,400 articles focused on validity, none included consequential validity evidence. As the emphasis on test consequences increases in policy and professional guidelines, we might expect to see more studies emerge producing that type of validity evidence. Cizek et al. (2010) discussed in some detail whether notions of validity need to be revised to distinguish between "validation of score inferences" and "justification of test use" and suggested that the version of the then current *Standards* (AERA, APA, & NCME, 1999) be updated to reflect this distinction. This recommendation reflects the Bachman (2005) proposal of the AUA structure including both validity and utilization arguments.

Table 1 shows examples of types of studies or analyses that are planned or are currently being conducted to obtain the types of validity evidence listed in the *Standards* (AERA, APA, & NCME, 2014) to support the high-level claims of the Winsight theory of action and the finer-grained claims of the validity argument. The set of examples is designed to be illustrative but not exhaustive. The first two columns show how the USDE (2015) peer review validity section elements relate to the *Standards*. We separate studies and analyses by group primarily responsible for the work (i.e., assessment development, operational psychometric analysis, and research) to suggest where these evidence collection efforts may occur. Collaboration is necessary between areas for many types of validity evidence to be obtained. In some cases, researchers may take the lead on developing the research questions that direct the validation effort and may coordinate with other areas to execute the work. The types of evidence in Table 1 partially support the validity argument claims.

Note that, in addition to these specific sources of validity evidence, both the USDE peer review guidance (USDE, 2015) and the *Standards* (AERA, APA, & NCME, 2014) include test design and development work, psychometric work (e.g., scoring, reliability), and fairness as stand-alone categories, but these aspects are typically part of any discussion of validity because they contribute to the overall validity argument. Further, documentation of features of score reporting (e.g., the use of simplified language that is free from jargon; supplemental guides that support understanding of student proficiency information, including cautions about how scores should or should not be used or interpreted) provides important peer review evidence. Table 2 shows the overlap between the foundational and operational chapters of the *Standards* with the USDE (2015) peer review guidance critical element categories. While it was outside the scope of this report to map all standards or clusters of standards to the 30 critical peer review elements, it is clear from this table that there is a great deal of crossover between professional standards for all assessments and those specific to U.S. K–12 assessment programs as required by the USDE. Note that this mapping is a work in progress as we develop our research agenda and should not be considered official.

Producing Evidence to Support the Winsight Summative Assessment Validity Argument

The appendix includes a brief version of the validity argument for the Winsight summative assessment component. In this condensed version, we include Purpose Statements A, B, and C and their unpacked components (Claims A1, A2, etc.). Note that there are many areas of evidentiary overlap, especially between Purpose Statements A and B, which focus on SEA and LEA decision makers and teachers, respectively. The information that SEA/LEA decision makers and teachers use is often the same and may be supported by a common set of evidence. However, there are areas in which they are distinct, primarily in how the stakeholder groups use the data, and these different uses require additional and different evidentiary support. For example, Purpose Statement A indicates (as follows from the theory of action) that SEA/LEA decision makers will use the information to allocate resources. Therefore, there are subclaims associated with that particular usage, and evidence will be collected to demonstrate that this usage occurs (e.g., through surveys and focus groups of those stakeholders). Purpose Statement B indicates that teachers will use the information to reflect on the previous year's instruction and to inform their planning for the following year, and this will require different evidence.

As was previously mentioned, support for the validity argument requires evidence from several areas of ETS (assessment developers, psychometricians, and research staff) and will be collected at varying points in the life-span of the test. At the time of writing of this report, we have completed the pilot data collection for the summative component, which took place in 2017 and involved the administration of shortened test forms that, together, covered the content being evaluated for use operationally. During this period, we also conducted preliminary cognitive lab studies. In addition, by the end of 2019, we will have completed a field test, an external item alignment study, and cognitive complexity studies to better understand from a variety of perspectives how the summative items were functioning. Examples of validity evidence that we obtained from those data collections and activities follows.

In the translation from the theory of action to validity claims, one set of claims revolves around the extent to which the summative assessment measures the standards with items that support reliable and valid scores. The primary purposes of the pilot were to conduct a preliminary investigation of ELA and mathematics item types to inform the development of

Table 1 Examples of Validity Evidence to Support U.S. Department of Education Peer Review Guidance and the Standards for Educational and Psychological Testing

Type of validity evidence		Primary group		
USDE peer review	Standards	Assessment development	Operational psychometric analysis	Research
Content	Test content	Documentation and external reviews of development processes Fairness reviews (i.e., accessibility, language load, race/ethnicity, gender, SES, region of country) Alignment study (e.g., including coverage of item pool with respect to depth and breadth of assessed standards)	Psychometric analysis of test forms and items (e.g., item analysis to identify items in need of revision) Mode comparability study	Usability studies for accessibility of new item types Iterative item try-outs
Cognitive processes	Response processes	Expert judgment of item complexity		Cognitive interviews and think-aloud studies Analysis of response time by depth-of-knowledge ratings on items
Internal structure	Internal structure		Factor analysis and differential item functioning analysis overall and on subgroups Item correlations demonstrating relationships between subscores Reliability of scores and subscores; standard errors of measurement Classification consistency and accuracy of cut scores	Factor analysis, DIF, and reliability analyses on low-volume subgroups not part of operational analysis (e.g., students with particular disability subtypes)
Relations to other variables	Relations to other variables		Correlations of state accountability test scores and NAEP scores Relationship of cut score decision to external measures such as course grades and teacher judgments	Correlations of test scores with academic grades or other classroom-based performance measures Evidence of college and career readiness
	Consequences of testing			Stakeholder use of assessment reports Studies investigating intended and unintended consequences among all stakeholders (e.g., washback effects or narrowing of the curriculum)

Note. NAEP = National Assessment of Educational Progress; USDE = U.S. Department of Education. Validity evidence descriptors appear in the *U.S. Department of Education Peer Review of State Assessment Systems Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as Amended* (copyright 2015 by the U.S. Department of Education) and in the *Standards for Educational and Psychological Testing* (copyright 2014 by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education).

Table 2 Mapping of Broad *Standards* Chapters and U.S. Department of Education Peer Review Categories

	Peer review guidance critical element categories					
Standards chapter	Statewide system of standards and assessments	Assessment system operations	Technical quality—validity	Technical quality—other	Inclusion of all students	Academic achievement standards and reporting
Validity	X	X	X	X	X	X
Reliability/precision and errors of measurement			X	X	X	X
Fairness in testing	X	X	X	X	X	
Test design and development	X	X	X	X	X	X
Scores, scales, norms, score linking, and cut scores			X	X	X	X
Test administration, scoring, reporting, and interpretation		X		X		X
Supporting documentation for tests	X	X	X	X	X	X
The rights and responsibilities of test takers	X	X		X	X	X
The rights and responsibilities of test users	X	X	X	X	X	X

Note. The *Standards* chapter appears in the *Standards for Educational and Psychological Testing*. Copyright 2014 by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. The critical element categories appear in the *U.S. Department of Education Peer Review of State Assessment Systems Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as Amended*. Copyright 2015 by the U.S. Department of Education.

the larger item pool; to evaluate item types, timing, and rubrics; and to start to build automated scoring models for short constructed-response items. Psychometricians and assessment developers used statistical methods (e.g., item analysis) to identify items that performed poorly and to review and reject or revise those items. These analyses of a small set of items informed the development of the remainder of the item pool prior to the field test.

Another subclaim from the validity argument is that "summative items and tasks allow students to demonstrate what they know and can do fairly and equitably." One question is whether completing items that focus on argumentation and modeling in a digital environment rather than on paper allows students to demonstrate what they know and can do. To explore this question, during the pilot data collection for a subset of students, we also collected mathematics scratchwork, which students can use while working through problems but which is not included in the scoring. The analysis of the scratch paper in conjunction with the final submitted online responses supports the investigation of whether students were able to transfer their work into the digital platform accurately. For example, if students solved the problem correctly on paper but then entered incorrect information into the platform's calculator or equation editor, this could indicate usability issues or that more instruction or better directions or help is needed. Another source of evidence for this claim comes from usability and cognitive lab studies. One group of studies involved the use of cognitive interview techniques to evaluate technology-enhanced items and their delivery. The aim was to identify issues students have in understanding how to engage with or respond to items. The results will feed back both to assessment developers to inform wording for item types that were confusing for students and to technology developers to inform the assessment interface design. Another set of cognitive interviews investigated whether there is evidence to suggest that the items categorized as having different levels of cognitive complexity do in fact entail those levels of complexity, and an upcoming study with English learners at different levels of English proficiency will examine how those students approach the items.

The field test began in 2018, and data from that stage are being used to obtain item statistics for operational form building; finish building automated scoring models; and identify, evaluate, and remediate items that are not performing well (e.g., using item analysis and differential item functioning, where possible). These efforts further contribute support for claims of valid and reliable measures.

Woven throughout the theory of action and the validity argument is the idea that all students can access the assessment and demonstrate their knowledge, skills, and abilities without the assessment mechanism introducing undue impediments. To that end, a critical component of the work includes accessibility reviews that occur primarily at these early stages of test design and development. As we noted, accessibility for all students, including English learners and students with disabilities, is a key goal for the Winsight assessment program. Universal design principles are used during item and task development with the goal of building a test that is as accessible as possible at every stage. Expert reviews of the items, tasks, and delivery platform (navigability and features) will further ensure that accessibility criteria are being met and will be revisited as there are newly written items and tasks or revised platform characteristics.

As part of the field test, we began collecting process data that will provide additional contextualizing evidence of accessibility and of student progress through the assessment. For example, logfile data from the assessment administration may include whether students use particular accessibility features as well as when and how they use them, which response options they choose and whether they change their answers, and how they navigate through the test (e.g., skipping items and returning or accessing the item review screen). In the future, we intend to expand what is captured to include more fine-grained clickstream data that include every interaction the student has with the platform and can be used to elicit a deeper understanding of test-taking behavior. As an example, keystroke logs can be used to examine how students compose essays (e.g., deleting, copying, and pasting text; time between actions) to identify writing profiles and provide formative feedback (see, e.g., Zhang & Deane, 2015). These data will contribute ongoing evidence that supports some of the foundational subclaims of the validity argument.

The Winsight summative assessment will be available for operational testing during the 2019–2020 school year in a conventional linear testing format. To support subclaims about assessment quality, we will conduct more extensive psychometric analyses (e.g., examining the internal structure of the test, obtaining additional reliability data, creating a vertical scale to allow comparability of scores between grade levels). We will also, however, begin to address subclaims about how assessment results are accessed and used by the various stakeholders to support the “what can they do with the information” portions of the claims. For example, teacher interviews and surveys may provide evidence of whether and how teachers used the results from their students to reflect on the previous year’s instruction and inform the coming year’s instruction. Parent/guardian surveys may provide evidence of whether the reported score information and score interpretation guides helped them to have more informed and productive interactions with their children’s teachers. Log files from the reporting system will also help us triangulate information from the interview and survey data.

Claims A2 and B2 identify the role that learning progressions play in the Winsight assessment data. Operational testing data will provide empirical support for the structures of critical learning progressions for all subgroups of interest. However, it should be noted that this is an example of an area in which evidence will accumulate both across years of operational testing and from other studies and so will not be available immediately. Investigations of score reliability and comparability for subgroups of interest (e.g., students identified by race/ethnicity, English learner status, disability status, socioeconomic status, and migrant status) will depend on sample sizes and may require field test data, operational data, and/or special research study data.

Future operational testing years will include a multistage adaptive format based on psychometric information obtained from the field test and will support additional research such as a longitudinal study of the impact of the uses of the results to provide evidence of student learning outcome improvement.

These are examples of work we have been conducting and work we plan to conduct in support of the validity argument. Necessarily, this is not an exhaustive list, because, as we have suggested, our validity argument will continue to evolve. We have also had the opportunity to subject much of this preliminary evidence to an internal preaudit to try to ensure that Winsight adheres to the *ETS Standards for Quality and Fairness*.[6] These standards are based on the AERA, APA, and NCME (2014) *Standards* and are specific to the types of tests and other materials that ETS produces (Wendler & Kirsch, 2017). Once the Winsight summative assessment component is operational, formal audits will be conducted every 3 years.

Discussion

As should be clear from the description in this report, while various approaches can be taken when developing a validity argument to support interpretations and uses of an assessment, ETS has worked from the start of test design and development to craft an argument for the Winsight program that would feed a research agenda and operational analyses to produce a comprehensive ground-up support for Winsight’s summative assessment claims. The claims, initially outlined in the Winsight theory of action, describe how different groups of stakeholders (SEA/LEA decision makers; teachers; parents/guardians and students) will use the summative assessment results as part of a larger assessment system to contribute to improving student learning outcomes. The evidence to support these claims is identified in the validity argument. When reviewing the PTEA, Wang et al. (2012) noted that there was not sufficient evidence to support all aspects of the AUA, possibly because the test and related materials had not been developed with such an argument in mind. We hope that developing the validity argument as the test is being designed, and keeping those evidentiary requirements in mind, will help ensure that we have plans and are able to collect the appropriate evidence. We are also working to build the validity arguments for the interim and formative assessment components as they are being developed, taking into account their specific intended interpretations and uses.

Validity work for an assessment must continue throughout the life of the assessment, as the test content, intended population, intended uses and interpretations, and relevant legislation evolve. Some validity evidence can be collected from the beginning of test design and development (e.g., documentation of design decisions), whereas other evidence will take several years or administrations of the test to procure, as some evidence is meaningful only when a test is fully operational and when, therefore, test scores actually “count.” However, the collection of such evidence must be planned in advance (e.g., longitudinal data to examine consequences of the test score uses). In addition, new peer review guidelines (e.g., USDE, 2018) take effect periodically and require different validity evidence. New types of data will emerge (e.g., process data to examine accommodation use or to supplement score reporting), and new methods of obtaining validity evidence will be developed. Therefore, validation is not a one-off, stagnant process: It is dynamic and must be both proactive and reactive.

As so many researchers’ and practitioners’ definitions of and frameworks for validity have influenced the development of the Winsight summative assessment validity argument, we hope that the discussion in this report sheds light on the specific actions that we have taken to create that structure. In particular, one of our goals was to encourage deeper understanding of how the theory of action for Winsight directly informed the validity argument for the summative assessment and how we will plan and adjust as needed to collect the evidence that will support the intended uses and interpretations associated with the assessment. By taking a more forward-thinking approach to concurrent design and validation, we can have deeper confidence that we are producing test scores that are comparable across subgroups and represent student proficiency as well as furthering deeper educational goals in instruction and resource management to improve student learning outcomes.

Acknowledgments

The authors would like to acknowledge, in addition to our reviewers, Leslie Nabors Oláh, Cara Laitusis, and Melissa Gholson for their feedback on earlier versions of this report.

Notes

[1] This report was originally written before the 2018 update and focused on the 2015 guidance document. We have updated the text to highlight high-level changes in the revision.

[2] http://www.corestandards.org/

[3] http://www.cpalms.org/Public/search/Standard

[4] See, e.g., the Smarter Balanced assessment theory of action at https://www.education.nh.gov/instruction/assessment/sbac/documents/theory-of-action.pdf and comprehensive research agenda at https://portal.smarterbalanced.org/library/en/comprehensive-research-agenda.pdf

[5] For examples, see the item sampler at https://ws.nextera.questarai.com/tds/#practice

[6] See, e.g., https://www.ets.org/Media/Research/pdf/conf_achgapwomen_kirsh.pdf for a brief overview of that process and its goals. For the *ETS Standards for Quality and Fairness*, see https://www.ets.org/s/about/pdf/standards.pdf

References

Achieve. (2006). *An alignment analysis of Washington State’s college readiness mathematics standards with various local placement tests*. Cambridge, MA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Arieli-Attali, M., & Cayton-Hodges, G. (2014). *Expanding the CBAL™ mathematics assessments to elementary grades: The development of a competency model and a rational number learning progression* (Research Report No. RR-14-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12008

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1 – 34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Bechard, S., Almond, P., & Cameto, R. (2011). Item and test alterations: Designing and developing alternate assessments with modified achievement standards. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 259 – 289). Charlotte, NC: Information Age.

Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2014, June). *State-of-the-art of visualization for eye tracking data*. Paper presented at EuroVis, Swansea, Wales.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*, 732 – 743. https://doi.org/10.1177/0013164410379323

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397 – 412. https://doi.org/10.1177/0013164407310130

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*, 119 – 142. https://doi.org/10.3102/01623737025002119

Confrey, J., Maloney, A., Nguyen, K., Mojica, G., & Myers, M. (2009, July). *Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories*. Paper presented at the 33rd conference of the International Group for the Psychology of Mathematics Education, Thessaloniki, Greece.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621 – 694). Washington, DC: American Council on Education.

Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., & Foley, C. (2015). *Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction* (Research Report No. RR-15-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12063, *2015*, 1, 29

Deane, P., Sabatini, J., & O'Reilly, T. (2012). *English language arts literacy framework*. Princeton, NJ: Educational Testing Service.

Deane, P., & Song, Y. (2015). *The key practice, discuss and debate ideas: Conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12079

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives*, *27*(4), 22 – 25.

Forte, E., & edCount. (2016). *Evaluating alignment in large-scale standards-based assessment systems* [White paper]. Retrieved from https://ccsso.org/sites/default/files/2018-07/TILSA%20Evaluating%20Alignment%20in%20Large-Scale%20Standards-Based%20Assessment%20Systems.pdf

Graf, E. A., & van Rijn, P. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165 – 189). New York, NY: Routledge.

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Washington, DC: Council of Chief State School Officers. Retrieved from http://www.k12.wa.us/assessment/ClassroomAssessmentIntegration/pubdocs/FASTLearningProgressions.pdf

Johnstone, C., Altman, J. R., & Moore, M. (2011). Universal design and the use of cognitive labs. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 425 – 442). Charlotte, NC: Information Age.

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*, 17 – 64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1 – 73. https://doi.org/10.1111/jedm.12000

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In *Handbook of test development* (pp. 64 – 80). New York, NY: Taylor and Francis.

Kane, M., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 489 – 552). https://doi.org/10.1007/978-3-319-58689-2_16

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. https://doi.org/10.1093/acprof:oso/9780199372904.001.0001

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 – 103). New York, NY: Collier Macmillan.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477 – 496. https://doi.org/10.1191/0265532202lt241oa

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3 – 62. https://doi.org/10.1207/S15366359MEA0101_02

National Council on Measurement in Education. (2018). *Position statement on theories of action for testing programs*. Retrieved from https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/NCME_Position_Paper_on_Theories_of_Action_-_Final_July__2018.pdf

Padilla, J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211 – 228). https://doi.org/10.1007/978-3-319-56129-5_12

Schmidgall, J. E. (2017). *Articulating and evaluating validity arguments for the* TOEIC® *tests* (Research Report No. RR-17-51). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12182

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, 99 – 104. https://doi.org/10.1111/jedm.12005

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, England: Cambridge University Press. (Original work published 1958)

U.S. Department of Education. (2015). *U.S. Department of Education peer review of state assessment systems non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended*. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf

U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Retrieved from https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing*, *29*, 603 – 619. https://doi.org/10.1177/0265532212448619

Wendler, C., & Kirsch, B. (2017). Understanding and applying the Standards for Educational and Psychological Testing: A case study of how the standards are applied at Educational Testing Service. *China Exams, 306,* 27 – 35.

Wylie, E. C. (2017). *Winsight assessment system: Preliminary theory of action* (Research Report No. RR-17-26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12155

Zapata-Rivera, D., Hansen, E. G., Shute, V. J., Underwood, J. S., & Bauer, M. I. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education*, *17*, 273 – 303.

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report No. RR-15-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12075

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79 – 87. https://doi.org/10.1016/j.pse.2014.11.003

Appendix

Purpose Statement A. The Winsight summative assessment component provides **SEA and LEA decision makers** with access to valid and reliable information about student performance with respect to state standards and proficiency bands, subdomains within standards, and key learning progressions, disaggregated for subgroups of interest. With this information, they are able to meet federal accountability reporting requirements and understand the relative performance of students by group, school, or district, and they can therefore allocate resources more appropriately at the district, school, and classroom levels. More appropriate allocation of resources by SEA and LEA decision makers at the district, school, and classroom levels enhances student learning.

Claim A1. *The Winsight summative assessment component measures student performance with respect to state standards, proficiency bands, and subdomains within standards for the population as a whole and for subgroups of interest*. This claim has the most subclaims, as it corresponds to the assessment development, psychometric, and research support for the foundational aspects of the assessment. For example, this claim includes subclaims about alignment of test content to state standards, quality assurance, support for scoring methods and outcomes, and issues of fairness. The evidentiary support for these subclaims focuses on documentation of assessment development and expert review along with formal alignment studies; psychometric analysis of the test, tasks, and items; scores and their properties and relationships to external criteria; and research studies investigating how students interact with the testing platform, tasks, and items.

Claim A2. *The Winsight summative assessment component measures student performance along levels of key learning progressions for the population as a whole and disaggregated for subgroups of interest*. This claim relates to the use of theoretical learning progressions to define movement along proficiency spectra, a key element underlying the Winsight summative assessment components. Validity evidence sources include documentation of the theoretical underpinnings and definitions of the learning progressions, documentation of the alignment of items and tasks to the learning progressions, and empirical evidence that the learning progressions and the ordering of levels for subgroups of interest are in evidence through performance on items and tasks aligned with those learning progressions.

Claim A3. *The Winsight summative assessment reporting system encourages and facilitates accurate score interpretation by SEAs and LEAs decision makers to allocate resources more appropriately at the district, school, and classroom levels to have a positive impact on student learning*. This claim requires that SEA and LEA administrators are provided with thorough, timely, and useful information about the test; resulting assessment data and scores; and use and interpretation of scores. Evidence includes documentation of these supporting materials and feedback from the SEA and LEA administrators about the quality and clarity of this information. Additionally, longitudinal data analysis will be used as the test is administered over time to gather evidence to show the impact on educational outcomes of the use of the summative data on state- and district-level decision-making.

Purpose Statement B. The Winsight summative assessment component provides **teachers** with valid and reliable information in terms of individual and aggregated class-level student performance with respect to state standards and proficiency bands, subdomains within standards, and key learning progressions. This information allows teachers to reflect on the previous year's teaching and curriculum to make pedagogical adjustments and to identify students who may be in need of additional resources. It also allows teachers to use the information to modify instructional plans at the start of the school year. When teachers are able to reflect on and adjust their practice, and tailor instruction to individual student needs, there is a positive impact on student learning outcomes.

Claim B1. *The Winsight summative assessment component measures student performance with respect to state standards, proficiency bands, and subdomains within standards for individual students and aggregated at the class level*. This claim's evidence base is the same as that for Claim A1, with one addition. Because the scores will be provided at different levels of aggregation than for the SEA and LEA administrators who are the focus of Claim A1, additional evidence is needed to support the reliability of assessment scores and subscores at the group (e.g., district and classroom) and individual student levels.

Claim B2. *The Winsight summative assessment component measures student performance along levels of key learning progressions for individual students and aggregated at the class level*. Similarly, this claim's evidence is the same as for Claim A2, with one analogous addition: There is also evidence required to show that learning progressions and key practices support meaningful reporting at the individual student level. Such evidence includes decision consistency of the placement of students into learning progression levels, which would offer teachers a more informed starting point with respect to collecting further information to confirm or refute placement.

Claim B3. *Winsight summative assessment reports encourage and facilitate accurate score interpretation by teachers to support their reflection on the previous year's teaching and curriculum to make pedagogical adjustments and to use the data for their current students to modify instructional plans at the start of the school year and identify students who may be in need of additional resources, which will have a positive impact on student learning outcomes*. This claim requires that teachers be provided with thorough, timely, and useful information about the test; resulting assessment data and scores; and use and interpretation of scores. Evidence includes feedback from teachers about the quality and clarity of information about these elements as well as the value the teachers place on using this information to support their reflection and decision-making and their understanding of appropriate score interpretation and use, to avoid negative consequences for students. Additionally, case studies of teachers will provide evidence of how teachers access, process, and apply data for use in supporting instructional planning. Furthermore, teacher surveys and case studies will provide evidence of whether teacher assessment literacy and use of data have improved.

Purpose Statement C. The Winsight summative assessment component provides **students** and **parents/guardians** access to information about a student's learning in terms of total test scores and proficiency levels, subscores, and normative information to compare student performance against district or state averages. This information enables them to have informed discussions with school-level educational professionals about an individual student's progress to be involved with the student's education, to better understand that student's progress and needs, and to garner appropriate supports and resources for that student. Improved communication and involvement between parents/guardians and education professionals have a positive impact on student learning outcomes.

Claim C1. *The Winsight summative assessment component measures individual student performance with respect to state standards proficiency bands and subdomains within standards and provides meaningful comparative data*. As with Claim B, the evidence required to support this claim is the same as for Claim A1 but requires additional evidence of these aspects at the individual student level.

Claim C2. *The Winsight summative assessment reports encourage and facilitate accurate score interpretation by parents/guardians, which supports more informed discussions with school-level educational professionals about individual students' progress, which has a positive impact on student learning*. This claim relates to the information provided to parents regarding their child's individual scores and how they can discuss their child's proficiency and needs. Evidence includes documentation of these materials and feedback from parents about how well these materials support their instructional goals for their child. Case studies or questionnaires will elicit evidence from parents about their broad understanding of score reports to refine score reports as needed. Evidence will also be collected to evaluate whether the frequency and quality of parent–teacher discussions have improved due to the use of these supporting materials.

Suggested citation:

Stone, E., & Wylie, E. C. (2019). *Building a validity argument while developing and using an assessment: A concurrent approach for the* Winsight® *summative assessment* (Research Report No. RR-19-26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12261

Action Editor: Donald Powers

Reviewers: Vince Dean and Richard Tannenbaum