2020

Teachers' Perceptions of the Broad Validity of a High Stakes English Language Proficiency Test

Carolyn N. Waters

TEACHERS' PERCEPTIONS OF THE BROAD VALIDITY OF A HIGH STAKES ENGLISH LANGUAGE PROFICIENCY TEST

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

by

Carolyn N. Waters
Master of Arts in Teaching, School for International Training, 1994
Bachelor of Arts, New York University, 1984

Director: Valerie Robnolt, Ph. D.
Associate Professor, Department of Teaching and Learning, School of Education

Virginia Commonwealth University
Richmond, Virginia
April, 2020

Acknowledgements

This project would not have been possible without the support of a number of important people. First I would like to thank my dissertation committee, beginning with my chair and academic advisor, Dr. Valerie Robnolt, for your support and advice throughout the program and project. I would also like to thank my methodologist, Dr. Jesse Senechal, for helping me discover the power of teacher voice in education research even before I started the program, and for all your help in thinking things through. Many thanks to Dr. Hillary Parkhouse for your encouragement and good ideas, and to Dr. Laura Kuti for engaging with me in conversations about testing over the years but also for your support of this project. In addition, want to thank my teachers and fellow students in the doctoral program. The generosity of professors Dr. Lisa Abrams, Dr. Michael Broda, and Dr. Sergio Chaparro, and of colleagues Dr. Melissa Cuba and Dr. Jorli Swingen, was particularly instrumental in making this project happen.

I am especially grateful to my family, Bert, Paulo, Sue, Rosa, Izzy, Laurie, Mike, Barbara, and Dan for your love, patience, and support. Thank you to my public school friends and colleagues for your unwavering interest, input, and encouragement. A big thank you goes to the 273 teachers who took the time to complete the survey and be interviewed; without your input and insight, this project would not have been possible. Last but not least, I would like to thank my students, especially my ELs and former ELs, for the inspiration you always give me.

Table of Contents

List of Tables

List of Figures

Abstract

TEACHERS' PERCEPTIONS OF THE BROAD VALIDITY OF A HIGH STAKES ENGLISH LANGUAGE PROFICIENCY TEST

Carolyn N. Waters
MAT, School for International Training, 1994
BA, New York University, 1984

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University 2020

Chair: Valerie Robnolt, Ph. D.
Associate Professor, Department of Teaching and Learning
VCU School of Education

The No Child Left Behind Act of 2001 mandated that the English acquisition of all students identified as English learners be assessed annually using high-stakes standardized English language proficiency tests, and the Every Student Succeeds Act of 2015 continues this testing mandate. The WIDA ACCESS for ELLs® has been used for this and other purposes since 2005, and has been adopted by 40 State Education Agencies, including the Virginia Department of Education. However, despite the long-standing and widespread use of this assessment, no comprehensive independent evaluations of the test have been conducted. This mixed-methods study is a conceptual replication of a previous study evaluating the validity of a similar high-stakes standardized English language proficiency test. Using a Broad Validity Framework that considered the test's reliability, criterion validity, and consequential validity, the study surveyed and interviewed Virginia teachers of

English as a second language on their perspectives on the test. Findings suggest that while some teachers believe some sort of test is warranted for accountability and informing decisions, there are many threats to the validity of decisions based on test scores, including potential sources of construct-irrelevant variance, issues with technology associated with the online version of the test, the lapse of time between test administration and the receipt of scores, and questions regarding student motivation and test-taking effort. Furthermore, the study suggests the test has unintended consequences, including negative emotional impacts for teachers and students and a loss of instructional time. Because of questions raised regarding the reliability and validity of the test, study findings suggest the use of multiple measures in high-stakes decision-making for English learners. Furthermore, findings affirm the value of a consideration of teacher input in test evaluations.

Keywords: English learners, English proficiency, Assessment, High-stakes testing, validity

CHAPTER 1. STATEMENT OF THE PROBLEM

The numbers of students identified as English learners (ELs) in U.S. public schools has grown from 3.8 million in 2000 (8.1% of the school population) to 4.9 million in 2016 (9.6% of the school population), making this group the fastest-growing portion of the U.S. K-12 population (National Center for Education Statistics, 2019). In Virginia, 162,374 ELs were enrolled in public schools in 2018-2019, representing 13% of the state preK-12 student population (Virginia Department of Education [VDOE], n.d.-a). For these ELs, standardized English language proficiency (ELP) tests are used for high-stakes decision-making, as well as state and federal accountability reporting (VDOE, 2018; Every Student Succeeds Act [ESSA], 2015; No Child Left Behind Act [NCLB], 2002). While the developers and publishers of ELP assessments conduct technical reviews of the quality of their tests, these tests do not typically undergo independent evaluation. The No Child Left Behind Act of 2001 (NCLB, 2002) mandated the innovation that ELP tests must measure progress and proficiency in "academic English," a construct that still has not been operationalized. In addition, the cut score or level of competence in English necessary for an EL to be reclassified as "proficient," removed from EL status, and thus be considered prepared for academic success has not been agreed upon (Abedi, 2008a; Boals et al., 2015). The Every Student Succeeds Act of 2015 (ESSA), federal policy that went into full implementation in the 2018-2019 school year, not only continues the ELP testing mandate of its predecessor, NCLB, but also raised the stakes in the use of these tests in accountability reporting (ESSA, 2015). Despite the wide use of ELP tests, very little research has been conducted on the quality of these

assessments or their consequences for students, teachers, or schools. Furthermore, no studies have been conducted on the validity of these tests in the context of ESSA policies, which mandate new uses of test scores.

Rationale for Study of the Problem

NCLB required states to create English language proficiency standards linked to state academic content and achievement standards, and to measure proficiency and progress in learning academic English annually using tests aligned with these standards (Office of English Language Acquisition [OELA], 2003). Prior to NCLB, most ELP tests measured general proficiency in social English. The new policy shifted the focus to academic English, and required states to develop a new generation of English language development (ELD) standards and ELP tests, which would cover linguistic competency in four skill areas: reading, writing, listening, and speaking. In addition, these standards and tests covered the four skills across five content language domains: social/instructional, language arts, mathematics, science, social studies (WIDA, 2014). In order to accomplish this complex task quickly for compliance with the new federal law, several states formed partnerships. One such partnership was the WIDA Consortium, which developed ELD standards and ELP assessments aligned with those standards, the *Assessing Comprehension and Communication in English State-to-State* (ACCESS for ELLs® [ACCESS]) tests. To date, the WIDA standards and tests have been adopted by 40 states and territories, including Virginia (WIDA, n.d.-g). With a price tag of $27.75 per eligible K-12 student in Virginia in 2018- 2019 and 2019-2020, and $77.00 for alternate forms for ELs with disabilities (Lane, 2019) the tests cost school divisions in the state over

three million dollars to purchase each year. In Virginia, ACCESS is used not only for accountability reporting required under ESSA (as required previously under NCLB) but also in high-stakes decision-making, including the identification and reclassification of ELs, funding, programming, staffing, student scheduling, and teacher evaluation (WIDA, n.d.-a; T. Jenkins, personal communication, August 29, 2019; B. Russ, personal communication, September 12, 2018). While the developers of the ACCESS test provide technical reports on the test's development, reliability, and annual implementation, remarkably, given the long-standing and widespread use of this assessment instrument, no comprehensive independent evaluations of ACCESS have been conducted, nor has there been any empirical investigation of the consequences of its uses for students or schools.

While NCLB held school districts and states accountable for ELs' progress and proficiency in English, ESSA also assigns this accountability measure to individual schools, and in 2019 for the first time, ELP test scores were reported in School Quality Profiles alongside a school's other achievement data (VDOE, 2018). The theory of action for this new building-level ELD accountability is that by including ELP gains in the evaluation of a school's overall performance, ESSA makes ELs more "visible" at the site of instruction, and signals that serving ELs is an integral part of school quality (West, 2017, p. 75). Virginia's ESSA plan sets a seven-year long-term goal for increasing the number of students achieving progress toward proficiency, with incrementally increasing annual interim goals. Individual students are expected to reach proficiency within five years, as well as to achieve specified rates of progress toward proficiency in the interim. A baseline rate was set

according to 2017 progress results, when 44% of students made sufficient gains to meet progress requirements, and the state's seven-year long-term goal is a progress toward proficiency rate of 58% (Sugarman & Geary, 2018; VDOE, 2018). Another innovation under Virginia's ESSA plan was a new use for ACCESS scores, which are now reported as a proxy for annual grade-level Standards of Learning (SOL) reading test scores for ELs in elementary and middle schools who meet specific criteria (VDOE, 2018). Since ACCESS was designed as an ELP test, not a grade-level reading assessment, the validity of this practice is questionable (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). Another recent change is that when WIDA adopted a new, more rigorous scoring scale in 2016, Virginia responded by changing its progress levels and proficiency cut score to align with WIDA's former scale, designating the highest among a range of scores at different grade-level clusters as the new cut score for all grades, effectively making it harder for some students to attain prescribed progress gains and proficiency rankings. The same year, the state began its transition to WIDA's ACCESS 2.0, adopting the new electronic version of the test (Staples, 2017), the implementation of which was not without problems (Center for Applied Linguistics [CAL], 2017). The validity of decisions based on these ELP test forms and scores has not been studied, and in light of these recent changes, it is particularly timely to do so.

Statement of Purpose

Because accountability policies and methods affect schools, sites where ELP tests are administered and test-based decisions have direct consequences, the

perspectives school personnel such as teachers are important to consider in order to gain a broad understanding of the impact of testing. Thus, the purpose of this study was to investigate teachers' perspectives on the quality and uses of high-stakes ELP testing. Specifically, the study focused on educators' contributions to an understanding of the broad validity of the ACCESS test, particularly in light of recent changes in testing and accountability policy for ELs in Virginia given implementation of the state's ESSA plan (VDOE, 2018) and changes in the test format and its in-school administration (Staples, 2017).

For the past two decades, federal education policy has taken a standards and accountability approach to education equity and opportunity (United States Department of Education [USDOE], n.d.-d). New ESSA policy continues this theory of action with some modifications, the impacts of which remain to be discovered (Orfield, 2016). As NCLB implementation began to take root in public schools, Valenzuela, Preito, and Hamilton (2007) called for the scholarly community to assess the impact of this policy to determine whether and how a standards and accountability approach might fulfill its promise to improve schooling and eliminate the achievement gap between white middle-class students and ethnic minorities, and to reveal the effects of the policy's scrutiny on children, teachers, curriculum and instruction, and on school- and district-level processes. The need for an examination of the impact of testing policy on underserved students, including ELs, has continued relevance given the current evolving policy context. An investigation into the validity of the ACCESS test is warranted given that there are many unexamined questions about the reliability of ELP tests, the validity of high-stakes

decisions made on the basis of scores on these tests, and the impact of changes on people, policy, and procedures, particularly since this test is used in so many states for so many students.

Researcher Positionality

I have been an ESL teacher in a variety of settings in Virginia for over two decades, and have held my current ESL teaching position in a Virginia public school since 2002. I have administered ACCESS every year since Virginia adopted the test in 2008 (VDOE, 2008), and also proctored the previous state ELP assessment, the Stanford English Language Proficiency (SELP) test (Cannaday, 2007), for several years. As a test administrator and professional whose day-to-day work is regulated and confined by the standards and accountability approach to education, and by decisions made on the basis of ACCESS scores in particular, I have direct experience with issues in test administration and the consequences that score-based decisions have for me, my students, and my colleagues. I am well aware of many unexamined questions regarding the validity of the test, and I know that many of my colleagues also have thoughts on this. While I inevitably harbor preconceived notions and opinions regarding the reliability of the test and the validity of the decisions it informs, I am well-positioned to pose questions regarding the test's quality and impacts, because of the insider perspective and trust among peers that my status as a practicing EL teacher affords.

Literature Research Background

NCLB played an important role in bringing the need for English language assessment to the forefront of education accountability (Abedi, 2008b), and ESSA

continues to foreground English learner testing (ESSA, 2015). Because of its explicit prioritizing of adequately assessing English proficiency, NCLB legislation required that schools receiving Title I funding assess EL language development using reliable and valid measures (NCLB, 2002). This mandate still thrives, even though the concept of English proficiency remains highly problematic, and measuring it is even more difficult (Abedi, 2008a; Abedi & Gándara, 2006; Solórzano, 2008). Abedi (2008a) writes, "There is no specific indication of which tests or which cutoff score would indicate an acceptable level of English proficiency" (p. 21). Researchers recommend using multiple measures in conjunction with proficiency scores for reclassifying students (Abedi, 2007; Abedi & Deitel, 2004). While some states include teacher consensus, parent consultation, and/or state academic achievement tests as well as ELP scores for reclassification decisions (California Department of Education, 2017; New York State Education Department, 2015), 29 states including Virginia only require the use of ELP test scores to make reclassification decisions (Linquanti & Cook, 2015). Virginia has designated a score of 4.4 on ACCESS as its sole criterion for EL reclassification (Staples, 2017).

Title III under NCLB introduced the requirement of assessing EL progress and proficiency in academic content language, and required states to align ELD standards with state content standards across content areas, so when NCLB was enacted, a new generation of ELP tests aligned to ELD and content standards had to be created quickly (Fast et al., 2004; USDOE, 2003). These requirements remain in effect under ESSA (VDOE, 2018). Researchers affiliated with the developer and publisher of the ACCESS test have reported on measures taken to enhance the

reliability and validity of the test during development, and publish annual reports on test updates and administration (WIDA, n.d.-d); however, there are no validity studies on ACCESS in the literature, and the WIDA Consortium reports that it is not aware of any third party entities or researchers providing resources or information related to the validity of its tests from their perspective (S. DeWitt, personal communication, June 29, 2018).

Test Construct: Academic English, not Academic Content. The issue of construct-irrelevant interference of English language proficiency on academic content tests administered to ELs in English has been demonstrated in the literature (Abedi & Gándara, 2006; Abedi & Lord, 2001; Kopriva, 2008; Solórzano, 2008). However, the converse, whether and how academic content knowledge interacts with academic English language proficiency in high-stakes testing, has not been studied. For example, to what extent, if any, would a student's proficiency in math mediate their ability to demonstrate proficiency in the Language of Mathematics, one of the content domains addressed in the WIDA standards and ACCESS test (WIDA, 2014). Fast, Carrera, & Conrad (2004) say that ELP assessments are not tests of academic content, "in other words, no external or prior content-related knowledge is required to respond to the test questions. Instead, the academic content language is described as the language that facilitates learning the content, not the mastery of academic concepts" (p.2). What constitutes academic language remains a topic of debate, and content language has not been operationalized for test development (Boals et al., 2015; Rivera, 1984; Valdés, 2004; Wolf et al., 2008). Furthermore, there is no evidence to demonstrate that prior content knowledge is

not confounding measures of language proficiency on ELP tests of content language. Chi, Garcia, Surber and Trautman (2011) found adequate linking between ACCESS and the Common Core State Standards (CCSS) in English Language Arts and Mathematics, and deemed the test in compliance with federal guidance in Common Core states. Lee (2018) found alignment between the CCSS and the WIDA standards to be problematic in terms of both disciplinary practices and cognitive expectations across proficiency levels. Virginia is not a Common Core state, however, and has rejected the adoption of CCSS in favor of its own state standards. Alignment between the Virginia Standards of Learning (SOLs) and ACCESS has not been studied (K. Bach, personal communication, March 2, 2020; J. Costa, March 1, 2020, personal communication; K. Johnson, personal communication, July 26, 2018).

English Language Proficiency and Academic Achievement. Several quantitative studies have specifically examined the correlations between ELP and academic achievement scores. These correlations are of interest since ELP test cut scores are presumed to determine a level of competence in academic language necessary for success in school, and thus predict academic success. Katz, Low, Stack, and Tsang (2004) compared the results of California's ELP measure with the results of English-only standardized achievement tests, and found little relationship between achieving proficiency on the ELP test and students' performance on the academic content tests. Gándara and Rumburger (2009) found that ELs deemed proficient in English based on ELP scores struggled to pass grade-level English language arts tests. In contrast, Ardasheva, Tretter, and Kinny (2012) found that reclassified ELs performed significantly better than their never-EL classmates on

reading and math tests. DuHart (2019) examined the 2017 and 2018 test scores of secondary ELs and former ELs in Virginia, and found a positive correlation between ACCESS scores and Virginia Reading SOL scores. However, data presented in the study show that reclassified ELs far outperformed all other subgroups of students on the SOLs, suggesting the reclassification cut score may be too high, or that ELs may be being held to a higher standard than other students, who are not required to demonstrate proficiency in academic English.

Qualitative Research and Testing Policy. Valenzuela et al. (2007) point to the suitability of qualitative research methods to address problems of practice as well as theoretical questions related to testing and accountability policy, but point out there appear to be very few qualitative studies on these topics in the literature. Sloan (2007) conducted a review of scholarship on the effects of the NCLB-driven high-stakes testing on minority youth, and found that little of the discourse in the academic literature was based on data generated in schools and classrooms, where such policy plays out. He says the existing qualitative research on this topic demonstrates "the power and potential of ethnography to offer clearer, more detailed portraits of the varied ways current accountability policies affect teachers of minority youth, the curriculum and pedagogy that minority youth experience, minority youth in general, and minority youth education" (p. 24).

A handful of qualitative and mixed-methods studies on the academic achievement and English proficiency testing of ELs support Sloan's (2007) claim that qualitative studies reveal unforeseen consequences of testing policy. Pennington (2004) conducted a case study that documented changes that occurred

in an elementary school as a result of its being labeled unacceptable based on student performance on state achievement tests, altering the definitions of literacy and literacy goals of teachers of Latino-Mexican children. McNeil's (2000) qualitative study of teachers detailed the ways that pressures to raise the test scores of minority students led to curricula and pedagogies that focused on tests at the expense of actual learning, and to a systematic teaching to the test in what she terms "defensive teaching" (p. 3). Zacher Pandya's (2011) ethnographic study of day-to-day practices in an elementary classroom found that the overtesting of ELs resulted in time pressures and diminished student self-esteem. Blaise's (2018) ethnography revealed that limited English proficient (LEP) high school students were enormously disadvantaged compared to other students when taking a standardized exam required for graduation, causing some the ELs to drop out and driving some teachers out of the profession.

In a mixed-methods study examining stakeholders' perspectives on an ELP test, Winke (2011) surveyed teachers and other school personnel to examine the validity of the English Language Proficiency Assessment (ELPA), the ELP test formerly used in Michigan for federal accountability compliance. Winke found the testing had negative collateral curricular and psychological effects, and suggested these findings demonstrated that classroom educators can make a valuable contribution to discussions of test validity and highlighted the need for a multidimensional framework for the broad validity of language proficiency tests.

Collectively, the above-mentioned studies reveal unintended consequences of achievement and ELP tests for ELs as administered under NCLB regulations. To

date, little to no independent research has been conducted on the impact or validity of the ACCESS test, and no studies on the consequences of high-stakes testing for ELs have been conducted in the context of ESSA. The current study addressed this gap in the literature by examining the broad validity of the ACCESS test in Virginia in a conceptual replication of Winke's study. [1]

Research Questions

The study employed Winke's two research questions, adapted slightly to suit the current local context; in addition, the study included a third research question to examine the effects of recent policy innovations, including Virginia's adoption of the online form of the test and resetting reclassification criteria, as well as accountability reporting innovations under ESSA:

(1) What are Virginia educators' perspectives on the ACCESS for ELLs® test?

(2) Do educators' perspectives vary according to demographics or teaching environment in which the test was administered?

(3) According to educators, what (if any) are the effects of recent testing policy changes on validity considerations for ACCESS for ELLs® in Virginia?

Methodology

Greene (2007) described mixed-methods research as a way of looking at social phenomena "that actively invites us to participate in dialogue. . . multiple ways of seeing and hearing, multiple ways of making sense of the social world, and multiple standpoints on what is to be valued" (p. 20). Qualitative and mixed-

[1] Paula Winke, the author of the study mentioned above, affirms the need for further research, "because ACCESS is huge" (P. Winke, personal communication, November 17, 2018).

methods research methods are well-suited and necessary for the study of impacts of testing and accountability policy (Valenzuela et al., 2007; Ryan, 2002). Because teachers and other school-based educators work where the tests are administered and where the consequences of testing arise, they are well-positioned to provide information to elaborate on and balance the conclusions of the technical reports presented by the test developers, thus contributing to a broad understanding of a tests' validity (Winke, 2011).

The study replicated Winke's (2011) mixed-methods approach described in *Evaluating the Validity of a High-Stakes ESL Test: Why Teachers' Perceptions Matter.* Like Winke's investigation, the current study used surveys to examine the perspectives of educators on a high-stakes ELP test, in this case the ACCESS test in Virginia. Winke's validated survey instrument was used, with a few adjustments and additions to reflect the current context. It included demographic questions, belief statements about the test that respondents rated on a Likert-type scale, and open-ended questions about educators' opinions and experiences with the test. To extend Winke's survey, the current study also inquired about the impact of innovations in testing policy and procedures. In addition, follow-up telephone interviews were conducted with survey participants chosen to represent a range of perspectives and demographic characteristics, as a member check and to further investigate and clarify emerging themes.

Summary

This study explored the broad validity of ACCESS ELP testing in the context of new ESSA policy implementation in Virginia. While the publishers of this widely

administered high-stakes ELP assessment had provided technical reports on the development of their instrument, the quality and consequential validity of the test had not been independently evaluated. The effectiveness of a new electronic testing format had not been adequately assessed. In addition, impacts of the implementation of the new federal ELP testing and accountability policy under ESSA on practices in schools had not yet been examined. This mixed-methods study investigated the perceptions of school-based educators on high-stakes ELP testing in the context of changing policy in Virginia, in order to examine the broad validity of the test, including the test's effects on students, teachers, curriculum and instruction, and school- and district-level processes.

Key Terms and Definitions

Assessing Comprehension and Communication in English State-to-State (ACCESS for ELLs®, or ACCESS) test. The high-stakes English language proficiency test used in Virginia and 39 other U. S. state education agencies, originally developed by the WIDA Consortium for compliance with the accountability mandates of the No Child Left Behind Act of 2002, currently used to meet the requirements of the Every Student Succeeds Act of 2015 (WIDA, 2018), and also used for other purposes including the reclassification of English learners, funding and programing decision-making, as a proxy for grade-level reading accountability reporting, and the evaluation of teachers. *ACCESS for ELLs® 2.0* refers to the current generation of the test, including the computer adaptive online versions of ACCES launched in 2015 (CAL, 2017),

Broad Validity. Broad validity refers collectively to technical considerations regarding the reliability, concurrent validity, predictive validity, as well as the consequential validity of a test and the test's interpretations and uses (Winke, 2011). *Reliablity,* or the degree to which test scores for a group of test-takers are consistent over repeated administrations of the test, and thus are considered to be dependable and consistent for an individual test-taker, and also the degree to which scores are free from random errors of measurement, is a requirement of test validity. Reliability is a prerequisite of validity. *Validity* refers to the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests, and is the most important consideration in the development and evaluation of tests. If multiple interpretations of a test score for different uses are intended, validity evidence is needed for each interpretation (AERA et al., 2014). Broad validity requires attention to the consequential validity of the test, the test's social, ethical, and practical consequences (Winke, 2011).

The Every Student Succeeds Act of 2015 (ESSA). The most recent reauthorization of the Elementary and Secondary Education Act of 1965, ESSA maintains the standards and accountability approach to educational equity and opportunity of the act's previous reauthorization, the No Child Left Behind Act of 2001, but allows individual states to work out the details their accountability programs.

English Learner (EL). According to the United States Department of Education, an English learner is "an individual who was not born in the United States or whose native language is a language other than English; or who comes

from an environment where a language other than English is dominant; or who is an American Indian or Alaska Native and who comes from an environment where a language other than English has had a significant impact on his or her level of English language proficiency; and who, by reason thereof, has sufficient difficulty speaking, reading, writing, or understanding the English language to deny such individual the opportunity to learn successfully in classrooms where the language of instruction is English or to participate fully in our society" (USDOE, n.d.-c). ELs are also referred to as *English language learners (ELLs),* particularly in less recent documents and publications. NCLB and NCLB-era Virginia policy documents referred to these students as *limited English proficient (LEP),* but this term has been mostly abandoned because of its deficit focus and pejorative nature (García, 2009). García, Kleifgen, and Falchi (2008) proposed the term *emergent bilingual* to highlight the assets of these students and disrupt inequities in their education. Consistent with current federal and Virginia state policy, this paper uses "English learner" (USDOE Office for Civil Rights, 2018; VDOE, 2018). The New York State Education Department also uses the terms *English as a New Language* and *Multilingual Learners* (NYSED, n.d.)

English language proficiency (ELP). This refers to the academic language or language of school that students need to acquire to be successful in the general education classroom (NCLB, 2001; Scarcella, 2003). Much debate still exists on defining the term ELP and its constructs. English language proficiency involves the five language components: phonological, lexical, grammatical, functional, and discourse (Scarcella, 2003). For federal accountability purposes, it also involves the

construct of academic language. The level of English language proficiency at which a student should be classified as *proficient* has not been agreed upon (Abedi, 2004). NCLB required states to develop English language development (ELD) standards to be assessed by ELP tests (NCLB, 2001), and this mandate is affirmed and continued by the Every Student Succeeds Act of 2015 (ESSA, 2015). In recent years, ELP and ELD standards have been amplified to reflect increased rigor in college- and career-readiness standards (Council of State School Officers, 2018; WIDA, 2014).

Long Term English Learner (LTEL). The educational classification given to students who have been enrolled in U. S. schools for more than six years, but who have not yet reached a threshold of proficiency in English and who may be struggling academically due to their English skills (Great Schools Partnership, 2015; WIDA, 2019).

The No Child Left Behind Act of 2001 (NCLB). The reauthorization of the Elementary and Secondary Education Act of 1985 that brought standards and accountability to the forefront of federal education policy, and imposed sanctions on states whose schools were unable to meet increasingly impossible test score requirements.

Standards of Learning (SOLs). Virginia's program of state academic standards and accountability. The SOLs were first adopted in 1995, and SOL testing began in 1998, as students in grades 3, 5, 8, and in high school took assessments in reading, writing, mathematics, history and science (VDOE, 2013). Unlike 45 other states and the District of Columbia (USDOE, n.d.-a), Virginia has not adopted the Common Core State Standards, but continues to use its own SOLs (VDOE, 2010).

WIDA. Originally the Wisconsin, Delaware, and Arkansas Consortium, which formed in 2003 to create English language development standards and an English language proficiency in compliance with the No Child Left Behind Act of 2002 (WIDA, n.d.-e), WIDA adopted the backronym *World-class Instructional Design and Assessment* when Arkansas left to join another consortium. WIDA has dropped this acronym definition and is "just WIDA now" (Mahony, 2017, p. 18). The WIDA standards and its ACCESS English language proficiency tests have been adopted and purchased by adopted by its members, 39 U.S. states and territories, including Virginia (WIDA, n.d.-c).

CHAPTER 2. REVIEW OF LITERATURE

This chapter will review the literature related to questions of the validity of high-stakes ELP tests and the decisions one such test informs in Virginia. It will

begin by presenting a theoretical framework for examining the broad validity of ELP tests, and will review ELP testing literature in terms of this framework. Because little empirical research has been conducted to investigate the consequences of language proficiency for ELs, and, to date, no empirical studies have examined the impact of new ESSA policies on ELP assessment, it will review the literature on the consequences for ELs under the implementation of NCLB, the previous federal high-stakes testing and accountability policy. Because policy changes under ESSA as well as innovations in the administration, scoring, and uses of the test raise new questions about its validity, this chapter will also provide background on the current policy context and review recent modifications in the federal standards and accountability approach to equity and opportunity for ELs, as determined by the Virginia Department of Education's ESSA state plan (VDOE, 2018). Virginia's ESSA plan was fully implemented in 2018-2019, the school year in which this study was conducted. Like earlier federal accountability policy such as NCLB, it had the potential to cause unforeseen, unintended consequences. Additionally, this chapter will examine questions regarding the high-stakes ELP tests used for accountability purposes, specifically the ACCESS test, the ELP assessment adopted by Virginia and 39 other state education agency members of the WIDA Consortium (WIDA, n.d.-f). The changes in uses of ACCESS scores for high-stakes decision-making and accountability reporting under ESSA are relatively new, and their impacts on students, teachers, and schools are still unclear. A first generation of state accountability reports using the new methodologies were issued in the same school year (VDOE, 2018), and the effects of this new reporting have not been examined or

evaluated; thus, questions about the test's reliability, validity, and fairness in the new policy context still remain unaddressed. Finally, this chapter will present a discussion on research methodologies appropriate for examining the consequences of a high-stakes assessment such as the ACCESS test.

Test Validity

The Broad Validity Framework. The *Standards for Educational and Psychological Testing [Standards]* (AERA et al., 2014) define test validity as "the degree to which accumulated evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (p. 11). Winke (2011) developed a framework of *broad validity* (Figure 1) for her study on the validity of the ELPA test. Like ACCESS, the ELPA was an NCLB-compliant consortium-based ELP test used for federal accountability reporting and high-stakes decision-making. Expanding on Norris' (2008) concept of "narrow-vein" validity in language proficiency testing, Winke states that in order to determine whether a test measures and performs well, attention should be given to both its narrow and broad validity. Narrow validity requires that the test be reliable (be internally consistent), have concurrent validity (be consistent with other tests measuring the same construct), and have predictive validity (predict students' future performance). This Broad Validity Framework focuses attention on the consequential validity of the test in addition to attention to its narrow validity. Consequential validity includes the test's social, ethical, and practical consequences: tests should be fair, meaningful, cost-

efficient (Linn, Baker, & Dunbar, 1991; Messick, 1980), developmentally appropriate (Messick, 1994), and able to be administered successfully (Katz et al., 2003).

Winke (2011) used this framework of broad test validity evaluation to examine educators' opinions of the effectiveness of the ELPA "to see if those perceptions could meaningfully contribute to a broad concept of the test's validity" (p. 628), and found that in their roles as test administrators, teachers can indeed offer meaningful perspective on the validity of tests. Because educators' work is framed by test-based decisions, they are uniquely positioned to offer insight in to broad considerations of the quality and validity of a test. While quantitative methods and statistical techniques are useful for evaluating narrow validity considerations of a test, qualitative and mixed-methods approaches are well-suited for examining its consequential and broad validity, which is subjective and relies on evidence-based judgments.

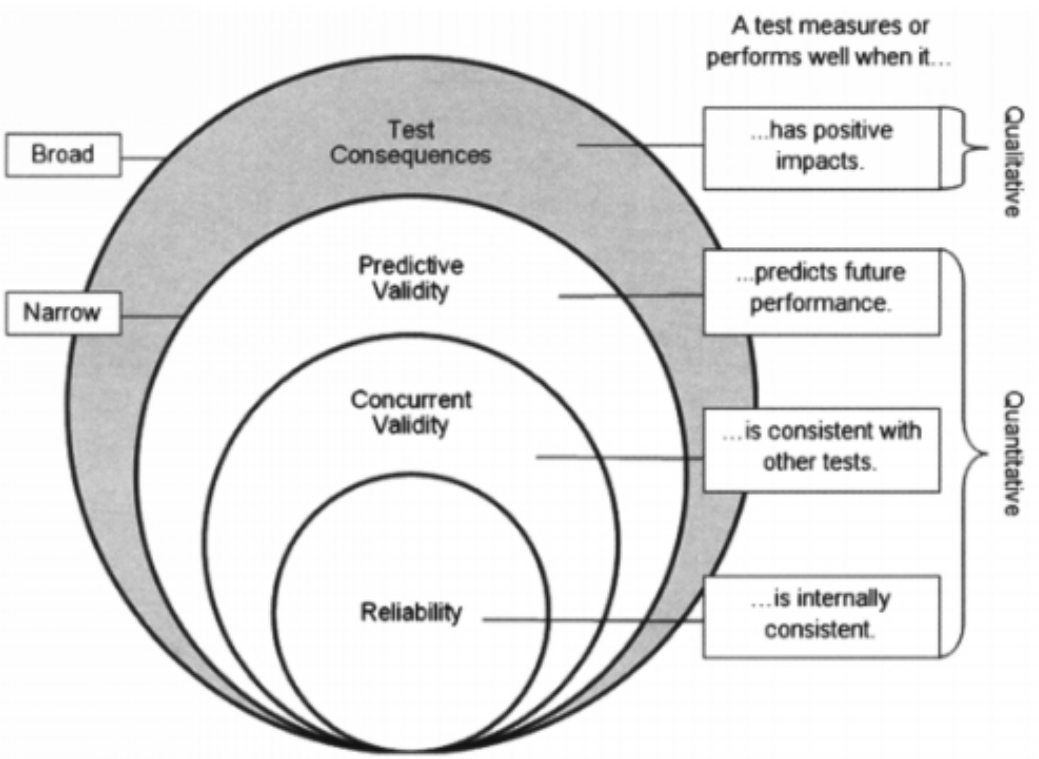

Figure 1: Levels of validity evidence. From "Evaluating the Validity of a High-Stakes ESL Test: Why Teachers' Perceptions Matter," by P. Winke, 2011, *TESOL Quarterly, 45*, p. 632. Copyright 2011 by TESOL Quarterly.

Reliability. It is important to distinguish between the reliability of test scores and the validity of interpretations or uses of the scores. The *Standards* (AERA et al., 2014) clarify:

> Reliability refers to the consistency of scores across multiple administrations of a test. Reliability can be defined as the correlation between scores on two equivalent forms of a test… It is possible for a test to be reliable but for its uses to be invalid or unfair. While the reliability of a measurement is always important, the need for reliability increases as the consequences of interpretations and decisions grow in importance. (p. 33)

WIDA publishes annual technical reports on the ACCESS test (WIDA, n.d.-e), and quantitative evaluations of the techniques used to enhance reliability during the development of the test have been conducted by researchers based at WIDA-affiliated institutions. Porter and Vega (2017) summarize these evaluations, which have included item-level analyses using Rasch methods, examinations of reliability using classical test theory, item response theory, and generalizability theory, test validity review using expert review, concurrent validity tests, and correlations between scale scores across language domains, as well as a report on validity differential item functioning was by gender and ethnicity.

While the developers and publishers of the ACCESS test have provided evidence on the techniques they used to maximize reliability of the test during development, and WIDA publishes technical reports on annual test administration (WIDA, n.d.-h; CAL, 2017), no independent evaluation of the test can be found in the literature, and the WIDA consortium's research office states that it is not aware of any independent confirmation of these evaluation results (S. DeWitt, personal communication, June 29, 2018). Although policy-makers and practitioners are told the test is "research based," they may be unaware that the majority of research on ACCESS is conducted by investigators affiliated with the University of Wisconsin, the Center for Applied Linguistics, and other organizations that benefit from WIDA's proprietary branding and sales of its trade marked products and services. WIDA is affiliated with Wisconsin Center for Education Products and Services, an organization that "works with UW-Madison innovators to commercialize and disseminate copyrightable educational products and services" (Wisconsin Center for

Education Products and Services, n.d.). Fox and Fairbairn (2011) published the only available report on the ACCESS test written by authors not affiliated with WIDA; however, at least 10 of the 14 references they cite were written by WIDA-affiliated researchers, which demonstrates the lack of research by non-affiliated entities.

Abedi (2013) says that despite improvements in ELP tests after the enactment of NCLB, many issues remained to be resolved, including problems with scoring the tests and reporting outcomes, inconsistencies in scoring between the four language domains, and inconsistencies in assessment. Bauman et al. (2007) reported that inconsistencies between states in the weighting of language domains raised concerns about validity. Solórzano's (2008) often-cited meta-analysis of EL testing literature concludes that issues with high-stakes ELP and academic achievement tests (i.e., norming, validity, and technical quality) make them inappropriate for ELs.

In 2015, WIDA launched ACCESS for ELLs® 2.0, a computer adaptive online version of its ELP tests. WIDA's *Annual Technical Report for ACCESS for ELLs ® 2.0 for 2015-2016* (CAL, 2017), states that a "substantial number of interruptions occurred during students' test sessions" due to technical issues with the online testing engine that year. While there were "small but noticeable differences" between the interrupted and non-interrupted students' scores,

> WIDA decided not to correct for interruptions on individual students' score reports; however, WIDA directed the CAL evaluation team not to include students with interrupted sessions in their psychometric analyses. Hence,

psychometric analyses reported on this year's annual report do not included [sic] students with interrupted tests [sic] sessions. (pp. v-vi)

Boals et al. (2015) stress the importance of attention to "mundane" details of assessment delivery in assessment development, such as ensuring visual and textual information in test materials are mistake-free and the distribution of test materials is secure, because these steps play an important role in supporting evidence-based arguments for the validity of the assessment (p. 145). However, WIDA developers do not address the actual administration of the test in schools, where mishaps or systematic issues could potentially have a negative impact on reliability and validity. Questions remain about technical considerations for the online test, which appear to have been a known threat to the reliability of scores, but these scores were nonetheless reported for use in high-stakes decision-making for individual students, schools, and districts.

Test construct: Operationalizing academic language. NCLB played an important role in bringing the need for English language assessment to the forefront of education accountability (Abedi, 2008b). Because the importance is placed on adequately assessing English proficiency, NCLB legislation required that schools receiving Title I funding assess EL students using reliable and valid measures (NCLB, 2002). This proficiency mandate still thrives, even though the concept of proficiency is highly problematic, and measuring it is even more so (Abedi, 2008a; Abedi & Gándara, 2006; Garcia, McKoon, & August, 2006a, 2006b; Solórzano, 2008; Zacher Pandya, 2011). Scholars do not agree on what counts as "proficient." Abedi (2008a) writes, "there is no specific indication of which tests or which cutoff score

would indicate an acceptable level of English proficiency" (p. 21). This lack of agreement on operationalizing the progress and proficiency levels used for test-based decision-making has the potential to negatively impact the reliability and validity of the test.

A primary requirement for standards-based assessment is for the test to represent the construct being measured (Menken et al., 2014; Messick, 1975; 1980). Under NCLB, EL testing and accountability policy has focused academic English as a means to accessing academic content and academic success. Referring to NCLB, the OELA (2003) stated:

> The statute requires English language proficiency standards to be linked to state academic content and achievement standards in reading or language arts and in mathematics beginning in the school year 2002-2003. This is required in order to ensure that LEP students can attain proficiency in both English language and in reading/language arts, math, and science. English language proficiency standards should also be linked to the state academic standards in science beginning in the school year 2005-2006. (p. 10)

ESSA requirements continue the NCLB mandate that ELD standards and ELP tests be aligned with state content standards (VDOE, 2018).

The concept of what counts as academic language is not well defined and has been controversial (Abedi, 2008a; Boals et al, 2015). According to Fast et al. (2004):

> ELP assessments are not tests of academic content, in other words, no external or prior content-related knowledge is required to respond to the test questions. Instead, the academic content language is operationalized as

> the language that facilitates learning the content, not the mastery of academic concepts. The goal is that ELs should be able to demonstrate proficiency in the academic proficiency required to function successfully in school. (p. 2)

This assertion appears to be at odds with the NCLB requirement that ELD standards and ELP tests be aligned with state content standards (NCLB, 2002). Abedi (2008a) states that this question needs the attention of experts in linguistics as well as content and measurement experts in order to operationally define academic language and provide guidelines for test writers, and suggests including teachers in this conversation. The academic content versus academic content language dichotomy is an issue in ELP testing reliability and validity.

Furthermore, when Boals et al. (2015), scholars at WIDA-affiliated institutions, including University of Wisconsin-Madison and the Center for Applied Linguistics[2], report on the theoretical underpinnings for the operationalizing of academic language used to develop the ACCESS test, they offer this caveat: "What we have described here is for illustrative purposes only" (p. 146), emphasizing they do not intend to provide a prescription for ELP test development, since what constitutes the academic language mandated in assessment policy is controversial.

Construct irrelevance in EL testing. Construct validity can be defined as "the degree to which a test measures what it claims, or purports to be measuring" (Brown, 1996, p. 231). The AERA's *Position Statement on High-Stakes Testing* (2000) says that for students who lack mastery of the language in which a test is given, the

[2] Timothy Boals, PhD, the lead author on this report, is the founder and director of WIDA (Wisconsin Center for Education Research, n.d.).

test becomes in part a test of language proficiency. While the literature addresses construct validity and issues of construct irrelevance caused by language proficiency for content tests administered to ELs in English (Abedi & Gándara, 2006; Abedi & Lord, 2001; Kopriva, 2008; Solórzano, 2008), the converse has not been studied; there is no conclusive evidence on whether or how tests of academic English might be confounded by construct-irrelevant academic content knowledge. This possibility is of particular concern in Virginia, where there is little direct evidence that the academic language of the ELP test is aligned with state academic content standards. This is also a concern for students who began their schooling in other countries and who have limited prior exposure to U.S. curricula (Alexander, 2017). AERA et al. (2014) recommend minimizing "confounding of the measurement of a construct with prior knowledge and experience that are likely to advantage or disadvantage test takers from particular subgroups" (p. 54). Abedi (2008a) suggests it would benefit states to review their current ELP tests and evaluate the test items in terms of academic English content in order to ensure construct representation.

Alignment of ELD standards with content standards. In addition, questions remain about the alignment of ELP tests with content standards in Virginia as required by NCLB and ESSA. Chi et al. (2011) found "adequate linking across all grade clusters" between ACCESS and the CCSS in English language arts and in mathematics, and deemed the test in compliance with federal guidance. WIDA's 2012 amplification of the ELD standards were written with explicit intent of alignment with the CCSS (WIDA, 2014). Virginia is not a Common Core state and

does not use these standards (VDOE, 2010). The VDOE provides side-by-side comparisons of the 2010 English Standards of Learning and the Common Core State Standards for English and Literacy (VDOE, 2010), and of the Virginia's 2009 Mathematics Standards of Learning with the Common Core State Standards for Mathematics standards (VDOE, 2011), but has not examined the alignment between SOLs and CCSS in social studies or science, content language standards also tested on ACCESS. While the VDOE states that the comparisons demonstrate the alignment between CCSS and SOLs in English and mathematics, the comparison reports reveal differences in sequencing between the two sets of standards, and similar curriculum is not necessarily taught in the same grade. For example, the VDOE correlates a CCSS Grade 6 Reading standard, "Compare and contrast texts in different forms or genres (e. g., stories and poems; historical novels and fantasy stories) in terms of their approaches to similar themes and topics" to several 9th grade Virginia Reading SOLs (VDOE, 2010). Additionally, a CCSS for Mathematics Grade 5, "Apply and extend previous understandings of division to divide unit fractions by whole numbers and whole numbers by unit fractions" aligns with Grade 6 Virginia Mathematics SOLs (VDOE, 2011). Students at different grade levels take different grade level forms, or "grade level clusters," of the ACCESS test (WIDA, 2019), so it's conceivable that students could be tested on content language before being exposed to the content. Since the publication of these alignment studies, Virginia has revised its English Standards of Learning (Constantino, 2018) and its Mathematics Standards of Learning (Staples, 2018), but the VDOE's comparisons to the CCSS have not been updated to reflect these revisions. A spokesperson for WIDA at the Wisconsin

Center for Education Research stated they were not aware of any studies evaluating the alignment of the WIDA ELD standards or ACCESS test with the Virginia SOLs (S. DeWitt, personal communication, July 23, 2018). The VDOE Assessment Coordinator and ELP Assessment Specialist also reported that their department has not conducted any alignment studies, and that they are unaware of any such evaluations (K. Bach, personal communication, March 2, 2020; K. Johnson, personal communication, March 2, 2020). Therefore, there appears to be no evidence to demonstrate that in Virginia, ELD standards and ELP tests are aligned with state content standards, or that as such, ELs would have been exposed to the language of academic in school before ACCESS tests them on it. Furthermore, there does not appear to be evidence that Virginia is in compliance with the federal policy that requires ELD standards and ELP tests be aligned with state content standards. This lack of alignment evidence raises two important questions: (a) is the ELP test used in Virginia aligned with the state content standards, i.e., is Virginia in compliance with federal law? (b) does the use of the ACCESS test in Virginia raise construct validity issues, because of the possibility that Virginia ELs are actually being tested on content knowledge that is not covered in their state curriculum and not just on academic language proficiency**?**

Validity and ELP tests. The *Standards* (AERA et al., 2014) clarify that typically it is up to the developers and distributors of a test to obtain evidence of test reliability and precision, but in some instances, local users should assume some responsibility for documenting the quality of measurement, particularly when local factors may affect the magnitude of error variance and observed score variance.

While the test developer is responsible for validating that the test scores in fact assess the intended construct, the state is responsible for the validation of the proper interpretation of the test scores and for the evaluation of the policy of using the scores for the decisions that they inform. Abedi (2008b) recommends: "States must always reserve the right to examine the validity and conduct analyses independent of what the test publishers/developers provide, to bring another layer of confidence into their high-stakes assessments" (p. 211). It appears Virginia has not met this responsibility or exercised this right regarding the ACCESS test.

Concurrent validity. The *American Psychological Association Dictionary of Psychology* (APA, n.d., Concurrent validity section) states that concurrent validity is the extent to which "one measurement is backed up by a related measurement obtained at the same point in time. In testing, the validity of results obtained from one test… can often be assessed by comparison with a separate but related measurement." Reviews of pre-NCLB ELP assessments expressed concerns about validity and coverage of academic content (Abedi, 2007; Zehler et al., 1994). Many of these early tests were not based on an operational definition of English language proficiency, had limited academic content coverage, were not aligned with states' content standards, and had psychometric flaws (Del Vecchio & Guerrero, 1995). Asserting that a major strength of the ELP tests developed after NCLB is the number of psychometric and validations studies that were part of the test-development process, Abedi (2008b) found support for concurrent validity in moderate to strong correlations between students' ACCESS scores and their placement in programming based on their states' pre-NCLB ELP tests (i.e., tests which were not tests of

academic language and which Abedi also critiques as inadequate). Because post-NCLB ELP assessments were developed with no interaction between test developers across test projects, and comparison studies have not been conducted, there is not enough evidence to judge the cross-validity of ELP tests (Abedi et al., 2007).

Predictive validity. Predictive validity is defined as "evidence that a test score or other measurement correlates with another variable assessed at some point after the test has been administered or measurement is made" (APA, n.d.). Predictive validity is important in the case of ELP tests, which are presumed to predict student readiness to cope with academic English at a level necessary for success in school (Fast et al., 2004). According to Francis and Rivera (2007) the fundamental validity question regarding ELP tests and ELs is "whether a student who scores in the proficient range of the test can function independently in an English-speaking classroom without specific language supports" (p.20). Because a reclassification decision is a judgment that English learners have adequate linguistic competence to be successful in English-language educational settings, reclassification should predict success on other measures, and failure to do so brings the predictive validity of ELP test scores into question. Several studies illustrate problems with using ELP test scores in reclassification decisions. Slama (2012; 2104) conducted longitudinal research on the ELP growth and reclassification of English learners, and found that foreign-born high school ELs showed faster ELP growth than ELs who were born in the United States, some of whom spent their entire academic careers without managing to test out of EL classification. Once reclassified, students tended to struggle academically. Similarly,

Gándara and Rumburger (2009) investigated the correlation between scores indicating proficiency in English on an ELP test and scores on a test of academic achievement. Referencing 2007 California Department of Education accountability reporting data, they found that while 60% of 10th grade ELs scored "early advanced" or "advanced" on the California English Language Development Test (CELDT), ratings roughly comparable to "proficient" on other ELP tests, only 3% of ELs were able to pass the state English Language Arts (ELA) test. The ELA test assumes that students have a more sophisticated understanding of English and more exposure to English literature compared to the CELDT. Robinson-Cimpian and Thompson (2016) conducted a large-scale longitudinal analysis of data from the Los Angeles Unified School District that spanned a 2007 rescaling of CELDT scores that made it harder for ELs to reclassify. Prior to the reclassification criteria becoming more stringent, negative effects of reclassification on subsequent state ELA achievement test scores and graduation rates were noted. Following the policy change, the authors found no reclassification effect on these measures. They suggest their findings have implications on how reclassification criteria under the new CCSS-aligned standards should be established, and conclude their findings illustrate the interplay between assessment policies for ELs, speak to the importance of rigorously evaluating the alignment between assessment and EL services, and reveal that policymakers can unintentionally create negative reclassification effects by establishing inappropriate exit criteria. In a mixed-methods study of San Francisco ELs, Katz, Low, Stack, and Tsang (2004) compared the results of the CELDT with the results of academic achievement tests, the SAT 9 Reading and Math Tests, and found little relationship

between Proficiency on the CELDT and students' performance on English-only standardized achievement tests. In addition, the qualitative component of the study concludes that testing results do not reveal how ELs are functioning in the classroom; in particular, reclassified ELs demonstrated less oral language participation in class than their English-only counterparts. Katz et al. stress the need for a multidimensional framework of language proficiency.

In contrast, Ardasheva, Tretter, and Kinny (2012) also found evidence of poor alignment of a state ELP test with reclassification decisions, but in the opposite direction. Their non-experimental quantitative study of 18,530 middle school students in Kentucky compared the academic achievement in reading and in math of students who had been reclassified based on ELP test scores with the achievement of both native English speakers and current ELs, and found that reclassified ELs significantly outperformed both native English speakers and ELs. They offer two possible explanations for their findings. First, citing Cummins' Threshold Hypothesis, which states that bilingualism enhances cognitive and linguistic functioning, former ELs might be capitalizing on these bilingual cognitive processing advantages. Second, the reclassification process might actually serve as an academic selection process in which academic ability underlies both success in language acquisition and success in academic achievement, effectively sorting ELs into two groups: those with lower academic ability (the current ELs) and those with higher academic ability (the reclassified ELs). Taken together, these studies indicate that ELP test results do not align well with other indicators of academic performance and

student success. Put differently, ELP tests appear to have questionable predictive validity; they do not do a good job of predicting academic success.[3]

DuHart (2019) examined the test scores of a convenience sample of 324 secondary ELs and former ELs in Virginia in 2017 and 2018, and found a positive correlation between ACCESS test scores and Virginia Reading SOL scores. However, the study neglects to point out that reclassified ELs outperformed all other subgroups of students, with 100% of the reclassified ELs in year two of former EL status in the study (n = 12) passing grade level Reading SOL tests. In comparison, the overall SOL pass rates in Virginia were 80% and 79% during the two years of the study, and the two highest-performing subgroups, Asian students and White students passed the Reading SOL tests at rates of between 86% and 91% (VDOE, 2018). SOL pass rates of former ELs are included in EL subgroup pass rates for four years following reclassification (VDOE, 2018), and the state does not disaggregate former ELs in accountability reporting. While DuHart's (2019) analysis does not include a comparison of pass rates for reclassified ELs and student subgroups, this data seems to suggest that ELs may be held to higher standards than students in all other subgroups, who are not required to take ACCESS or demonstrate proficiency in academic language.

Consequential validity. *Oxford: A Dictionary of Social Research Methods* (Elliot et al., 2016, Consequential validity section) defines consequential validity as

[3] Anecdotally, in 2017-2018 at the Virginia public school where I teach, at least a dozen 8th grade ELs who passed all of their courses and SOL tests (Reading, Mathematics, Writing, Science, and Social Studies) did not achieve the cut score on ACCESS necessary for reclassification. This would appear to suggest poor concurrent and predictive validity of the ELP test, at least for these individuals.

"The social consequences or likely implications of using a particular test/instrument for a particular purpose (and with a particular group). It addresses the 'so what' question and captures the extent that society benefits from that test/instrument." Consequential validity includes the test's social, ethical, and practical consequences: tests should be fair, meaningful, cost-efficient (Linn et al., 1991; Messick, 1980), developmentally appropriate (Messick, 1994), and able to be administered successfully (Katz et al, 2003). Messick (1980) wrote that consequential validity includes:

> …a pragmatic component for the evaluation of actual consequences of test practice... The primary concern for this component is the balancing of the instrumental value of the test in accomplishing its intended purpose with the instrumental value of any negative side-effects and positive by-products of the testing. Most test makers acknowledge responsibility for providing general evidence for the instrumental value of the test. The terminal value of the test in terms of the social ends to be served goes beyond the test maker to include as well the decisionmaker, policymaker, and test user, who are responsible for specific evidence of instrumental value in their particular setting and for the specific interpretations and uses made of the test scores. (p. 1020)

Federal Education Policy

It is useful to provide a testing and accountability policy backdrop for a discussion of the validity of ELP testing, particularly of consequential validity. Because policy contextualizes, drives, and regulates the decisions that are made on

the basis of test scores, it is thus an important consideration in an analysis of test validity.

ESSA is the most recent reauthorization of the Elementary and Secondary Education Act of 1965 (ESEA), civil rights-era federal policy designed to provide educational opportunity for poor and minority students. Prior to ESEA, education policy was the purview of individual states (Orfield, 2016). A key feature of ESEA was Title I, which provided federal financial aid to support schools serving low-income students, particularly in the South, rural areas, and large cities (Kantor & Lowe, 2013). The Bilingual Education Act of 1968 (Title VII of the ESEA) required schools to provide language support services to ELs to ensure access to academic content while simultaneously learning English (Menken, 2010). When several court cases in the 1970s failed to uphold school desegregation plans, educational equity and opportunity approaches began to unravel (American Bar Association, 2013). Subsequently, the influential *A Nation at Risk*: *The Imperative for Education Reform* report (1983) blamed mediocrity in education for the United States falling behind international competitors, and reframed educational policy discourse away from access and equity and toward standards and accountability. The major educational reforms called for by the report took hold nationally when NCLB, the most recent reauthorization of ESEA prior to its current iteration as ESSA, was signed into law in 2002 (USDOE, n.d.-d), representing further expansion of the role of the federal government in U.S. public schools (Orfield, 2016). The theory of action for NCLB was to provide educational access and opportunity to all students, including racial minorities, students with disabilities, economically disadvantaged students, and

what the act termed limited English proficient (LEP) students[4], by drawing attention to the academic underperformance of disadvantaged subgroups, including ELs, and holding states, school districts, and individual schools accountable (Abedi, 2004; Wolf et al., 2008).

NCLB replaced the Bilingual Education Act with Title III, the English Language Acquisition, Language Enhancement, and Academic Achievement Act (Menken, 2010). Under NCLB, Title III held states and school districts accountable for progress toward and attaining English proficiency and for meeting the same academic standards as their English-speaking peers in reading and math (Wright, 2008). The legislation required states to develop clearly-defined ELD standards in each of the four language domains (reading, writing, listening, and speaking), align these ELD standards with state content standards across content areas, and assess ELs annually based on these ELD standards. Thus, a new generation of complex ELP tests had to be created quickly (Fast et al., 2004). These language proficiency assessments would to be used to measure academic English and social language in order to predict ELs' readiness to succeed in English-language school settings; however, the rushed process of developing and launching the new tests led to concerns regarding validity, since they had not been adequately vetted through rigorous research (Wolf et al., 2008). Abedi and Deitel (2004) identified several challenges for EL accountability under NCLB, including the historically low performance and slow improvement of ELs on state tests, measurement challenges, the instability of the EL subgroup caused by high-achieving members exiting the

[4] ESSA replaces the out-of-date pejorative term Limited English Proficient (LEP) students, with English learners (ELs; USDOE, 2015).

group and lower students entering, and the impact on achievement of factors outside of a school's control. To address these challenges, they recommended increased focus on reading instruction, close tracking of EL performance using multiple measures, linguistic modification of content tests, encouraging testing accommodations (while evaluating the validity of these accommodations), including EL reclassification rates in measures of achievement to offset the penalty to schools for removing the highest-achieving students from the subgroup, and meeting an "existence proof" (that the goals required for all schools have been previously attained by at least one school), since no school with a sizeable EL population had actually met requirements of NCLB (p. 785). Virginia's ESSA plan mitigates some of these concerns to a limited extent by including former ELs in the EL subgroup for four years after reclassification instead of the two years they were included under NCLB, extending eligibility for EL testing accommodations to former ELs for two years after reclassification, and using existing proficiency rates achieved in a baseline year to establish incrementally higher accountability targets (VDOE, 2018).

NCLB required 100% of students in all subgroups, including LEP students to demonstrate academic proficiency by passing state tests in reading and math by 2014, and as increasing numbers of schools were sanctioned for failing to meet impossible and unattainable interim targets, public approval of the law plummeted. However, while a reauthorization of ESEA was due in 2007, a divided Congress failed to pass new legislation, and NCLB remained in effect for several years after it had expired. In the meantime, in 2009, without Congressional approval, President Barack Obama ordered the American Recovery and Reinvestment Act. As part of

this economic stimulus program, the Race to the Top Fund incentivized controversial educational programs, including the evaluation of teachers based on student test scores[5], adoption of the CCSS, and the implementation of pro-charter school and school turnaround policies (USDOE, 2009).

Backlash against what was criticized as Executive and federal overreach in education policy eventually led to the reauthorization of ESEA in 2015. After years of Congressional indecision, the complex Every Students Succeeds Act of 2015 (ESSA) was passed within days of its publication, with little time for review or debate (DeBray & Blankenship, 2016; Orfield, 2016). In a newspaper editorial, Boston University School of Education scholars Battenfield and Crawford (2015) characterize ESSA this way:

> The provisions of this 1,061-page bill (about 400 more than NCLB) do not vary radically from the "accountability through testing" mandates that have marked federal education policy through the last 14 years. The main difference is that ESSA hands the accountability ball from the federal government to the states…Testing mandates in ESSA continue the retreat from the anti-poverty focus of the 1965 Elementary and Secondary Education Act. (U.S. News and World Report, 2015, December 4)

Unintended consequences of high-stakes achievement testing for ELs. While to date no research has been published to examine the consequential validity of high-stakes testing for ELs under ESSA, several studies demonstrate the

[5] The evaluation of teachers based on student scores is no longer a requirement under ESSA, but vesitigial test-based teacher evaluation continues as a practice in some Virginia schools, despite WIDA's recommendation against using its scores for this purpose (WIDA, 2015).

unintended impacts of academic achievement and ELP testing on ELs under NCLB, the predecessor of the current federal policy. Sloan (2007) conducted a review of scholarship on the effects of the NCLB-style high-stakes testing on minority youth, and found that little of the discourse in the academic literature had at the time been generated through qualitative research or ethnographic work or based on data generated in schools and classrooms, where such policies are enacted. He says the qualitative research on this topic demonstrates "the power and potential of ethnography to offer clearer, more detailed portraits of the varied ways current accountability policies affect teachers of minority youth, the curriculum and pedagogy that minority youth experience, minority youth in general, and minority youth education" (p. 24). Sloan concludes that the lack of scholarship engaging with teachers and students in classrooms has done much to conceal the complexities involved in improving public education, especially for low-income students of color.

Several qualitative studies have demonstrated that high-stakes academic achievement testing has negative effects on instructional practices for ELs. Pennington (2004) documented changes that occurred at an elementary school in Texas where high-stakes testing and accountability altered the literacy goals and definitions of literacy of teachers of Latino-Mexican children as a result of the school being labeled "unacceptable" based on student performance on state tests. Although the school and teachers still had multiple instruments available to offer a comprehensive view of the literacy abilities of their students, the state test trumped all other measures, and children's literacy abilities were judged solely on the basis of standardized test scores. Similarly, in a qualitative study of teachers, McNeil

(2000) details the ways that pressures to raise the test scores of minority students in social studies lead to curricula and pedagogies focused on tests at the expense of actual learning. This led to a systematic teaching to the test in what she terms "defensive teaching" (p.3). Zacher Pandya (2011) conducted an ethnographic study of the day-to-day practices in an elementary classroom, and found the over-testing of ELs resulted in time pressures and diminished self-esteem. Blaise's (2018) 3-year ethnography conducted in a Boston public high school found that limited English proficient Haitian students were enormously disadvantaged compared to other students when taking the Massachusetts Comprehensive Assessment System, a standardized exam required for graduation. Pressure to pass the graduation test had the effect of limiting the teaching of critical thinking skills and reducing classrooms to test-preparation workshops, causing ELs to drop out, and driving some teachers out of the profession. Alexander (2017) conducted an institutional ethnography of Japanese sojourner students attending a U.S. high school on temporary visas, and found these ELs intentionally and strategically failed the state ELP test, on the advice of their ESL teachers and with the approval of their parents, in order to continue to receive EL accommodations on the very challenging state-mandated high school exit exams, because failure on the exit exams would prevent students from graduating, precluding the possibility of attending college when they return to Japan. These students and teachers questioned the validity of decisions based on the ELP test scores, saying a lowering of the exit criteria pushed students out of ESL services before they were ready. The teachers believed this policy was based on an unwillingness to fund ESL programming for the growing EL population

in the state, rather than on former ELs successfully passing the exit exams, as the state had claimed without providing evidence.

ESSA in the academic literature. Given the transition to the provisions of ESSA is recent, beginning in the 2017-2018 school year and entering full implementation in 2018-2019, no empirical studies on the law's impact on ELP testing have been published. A recent textbook, *The Assessment of Emergent Bilinguals: Supporting English Language Learners* (Mahoney, 2017), which was already obsolete when it was published during the transition to ESSA, instructed practitioners on compliance with out-of-date NCLB-era AYP mandates, pointing to the fact that practitioner literature initially struggled to catch up to the new policy, and there would be some lag time for educators to figure out the requirements and implications of the legislation. A few scholars examined pre-implementation ESSA plans, and addressed the law's potential (or lack thereof) to improve educational equity and opportunity. Hopkins, Malsbary, and Moralez (2016) proposed a framework for examining "components of a state educational system for bi/multilingual students to be considered responsive to federal policy" (p. 32). They applied this framework to a convenience sample of 16 states to demonstrate its utility for understanding needs and programming within a particular state or region, and claimed their framework reveals understandings especially important in "new destination" states where immigrant and EL populations are growing.[6] Fuller, Hollingworth, and Pendola (2017) conducted a document analysis of 50 proposed state ESSA plans, looking for explicit or intentional programs to place effective

[6] Virginia, where EL numbers have been increasing rapidly (Sugarman & Geary, 2017), could be described as a new destination state.

principals in underperforming schools as a means to improve them, and found that only 10% of state plans made any connection between the distribution of effective principals with the recruitment and retention of effective teachers. Education politics, policy, and accountability scholar Arnold F. Shober (2017) suggests that ESSA will allow states to align state assessments to classroom curricula (as opposed to assessing students against their peers or state standards). He says near-universal adoption of the CCSS ensures that teachers will know what their students are supposed to learn, and incorporation of these standards into state assessments could provide feedback to inform the daily practice of teachers. However, because Virginia has not adopted the CCSS (VDOE, 2010, June 24), this observation may not be relevant in the state.

Jennings (2016) called ESSA "NCLB lite" and critiqued it as more-of-the-same standards and accountability approach that has proven to be an indirect and ineffective way to address educational gaps, and as such, predicted that ESSA will suffer the same limitations as NCLB. Anticipating the next round of ESEA reauthorization, he proposed plans for evaluating the impact of the current legislation. Contextualizing ESSA in the context of the history of education and civil rights policy, Orfield (2016) explained that prior to ESEA, the federal government played only a minimal role in education, due in part to resistance by states' rights advocates seeking to avoid federal interference with the segregated schooling in the South. Federal agendas have tended to be more equity-oriented than those of many states, and since ESSA would devolve education policy to the states, new research would be essential in evaluating its ability to promote the equity goals of ESEA.

Federal civil rights laws would continue to be in force, and litigation may be necessary for promoting educational equity if individual states failed to do so. Collectively, these academic perspectives indicated a lack of confidence in ESSA's continuation of a test-driven approach to promote educational equity, and point to the importance of evaluating the impacts of the new accountability policy on practice. No empirical studies have been published addressing these concerns.

Virginia's ESSA plan and changes for EL accountability. When ESSA was signed into law at the end of 2015, state departments of education had to act quickly to develop plans for accountability compliance. After conducting a public review process and submitting several drafts and revisions to the U.S. Department of Education (Botel, 2017; Sodat, 2018; VDOE, 2019a), Virginia's *State Template for the Consolidated State Plan: The Elementary and Secondary Education Act of 1965, As Amended by the Every Student Succeeds Act* received final approval from U.S. Secretary of Education Betsy DeVos in May 2018 (USDOE, 2018). Transition to the provisions of Virginia's proposed plan had already been in effect for the 2017-2018 school year before the plan received final approval (VDOE, 2019b). With full ESSA implementation beginning in 2018-2019, states, school divisions, and schools have only recently interpreted the legislation and begun to devise and implement ways to comply with the changes.

For the most part, under ESSA, NCLB-era ELP accountability requirements remain in effect (VDOE, 2018). NCLB held states and school divisions accountable for meeting Annual Measurable Achievement Objectives (AMAOs) for ELs by demonstrating the language acquisition achievement of LEP students on high-stakes

measures in both the percentages of ELs making progress (by increasing ELP test scores from year to year) and proficiency (by attaining the ELP cut score required for reclassification). ESSA replaced the AMAOs with similar accountability structures, the details of which were determined by the individual states. In addition to states and divisions, individual schools would now be held accountable for a new measure, *progress toward proficiency*, which Virginia operationalized as the percentages of individual ELs increasing their ELP test scores over the previous year by an increment specified on a progress table published by the state. These progress tables were devised using 2017 ACCESS test results as a baseline, and increasing pass rate targets by two percentage points annually. This resulted in much lower pass-rate goals than the previous AMAOs. Under the Virginia plan, ELs are required to achieve the proficiency cut score within no more than five years; however, the progress and proficiency targets do not align, and it is possible for a student to achieve adequate interim progress benchmarks each year and still fall far short of the proficiency score in the required five years (VDOE, 2018).

In another ESSA innovation, ELP accountability reporting has moved from Title III to Title I, so ELP achievement as measured by ACCESS or similar tests began to be reported alongside the other accountability data of individual schools, where English language instruction and ELP test administration take place, with the goal of giving English language acquisition more visibility and consideration in accountability systems. In all academic areas including English language acquisition, Virginia's ESSA accountability plan sets annual interim target pass rates that increase incrementally toward a seven-year long-term goal, aiming to raise the

achievement rates of all schools to above the statewide average levels reported in a baseline school year: 2015-2016 in reading and math and the 2016-2017 levels for English language proficiency (VDOE, 2018). While NCLB's goal of 100% proficiency was impossible to reach, the attainability of this new goal for every individual school to surpass the baseline state average is not yet known, and the changes to practices and programming that could result from efforts to do so remain unexamined.

ESSA eliminated all alternate testing forms except for students with disabilities, capped at 5% of all students (VDOE, 2018), so the Plain English form of the state Math test and the Virginia Grade Level Alternative (VGLA) reading portfolio accommodation for newly-arrived ELs are no longer available. Under the new accountability program, ELs are exempt from one administration of the reading SOL test in their first year of enrollment in U.S. schools (as they were under NCLB). In year two of enrollment, a passing score will be counted in reading if the student passes the SOL or demonstrates adequate growth on the ACCESS test. In year three, a passing score will be counted if the student passes the SOL, demonstrates growth on the SOL progress table without passing the SOL, or demonstrates growth as measured by ACCESS. High school students are not eligible for these options because they must pass SOL tests in order to earn verified credits required for graduation (VDOE, 2018). Thus, starting in school year 2017-2018, an ACCESS score has served a proxy for the grade level Reading SOL in accountability reporting for some elementary and middle school ELs.[7] ACCESS was designed as a test of English

[7]During early implementation of NCLB, Virginia ELs at levels 1 and 2 were permitted to take the statewide ELP test (the SELP test at the time) as a substitute for taking the reading SOL test, but in 2007, the USDOE required that ELs enrolled in U.S.

language proficiency, not grade level achievement in reading, and this use of scores for a purpose for which they were not intended raises serious validity concerns (AERA et al., 2014). The potential exists for these new uses of scores to have unforeseen and unintended consequences (positive or negative) for programming as schools begin to experience the new power of the ACCESS test not only impact English language acquisition reporting but also overall reading pass rates for schools.

The EL subgroup and ESSA in Virginia. An analysis of published California achievement testing data for initially-identified English earners (current and former ELs) and English-only students (Saunders & Marcelletti, 2013) raised the issue of the underestimation of EL subgroup progress in accountability reporting due to a "revolving door" phenomenon in which high-achieving ELs leave the subgroup upon reclassification, while at the same time, lower-proficiency ELs continue to enter. The "Catch 22," as they described it, was that reclassification and removal from the subgroup dooms the subgroup to low scores by design, thus obscuring EL achievement (p. 139). The report recommended that Reclassified Fluent English Proficient students (former ELs) be included in data reporting to provide a broader picture of achievement.

Under NCLB, reclassified students were counted in the LEP subgroup for two years after reclassification. Virginia's ESSA plan increases reclassified ELs'

---

public schools for more than one year take the same grade-level reading tests as native speakers (Menken, 2010) and Virginia introduced the Virginia Grade Level Alternative (VGLA) portfolio assessment, an option in place for students with disabilities, as an alternative to the reading SOL test for ELs at beginning and low-intermediate proficiency levels (Cannaday, 2007).

membership in the EL subgroup to four years. While ELs previously lost eligibility for testing accommodations upon reclassification, reclassified ELs are now permitted to receive testing accommodations in the first two years after exiting EL status. The theory of action for this reporting methodology is that by including proficiency gains in the evaluation of a school's overall performance, the law signals that serving ELs is an integral part of school quality (West, 2017). This increased time in the subgroup has the potential to boost achievement test pass rates for ELs. On the other hand, including reclassified ELs in the reporting category longer could skew perceptions of EL subgroup achievement, giving a false impression of improvement if changes in reporting methodology are not included year-to-year analyses. When changes in reporting methods such as this are implemented but not made transparent, achievement reporting is difficult to understand and interpret, and comparisons of pre- and post-ESSA pass rates could be misleading.

ESSA, like NCLB, requires students in all subgroups, including ELs, to take state tests in reading and mathematics. The American Educational Research Association's *Position Statement on High-Stakes Testing* (2000) says that for students who lack mastery of the language in which a test is given, the test becomes in part a test of language proficiency. Unless the purpose of a test is to evaluate language proficiency, it should not be used with students who cannot understand the language of the test itself. Hopkins, Malsbary, and Moralez (2017) characterize ESSA and previous test-based accountability systems as de facto English-only policy, and say that ESSA offers little guidance and few incentives for English language development and academic growth that builds on the bilingual assets of ELs. In

preparing Virginia's *Revised State Template for the Consolidated State Plan: The Elementary and Secondary Education Act of 1965, as amended by the Every Student Succeeds Act* (VDOE, 2018), the state was required by the USDOE to identify languages other than English spoken by 5% or more of the EL population and to provide a plan for assessments in those languages. Virginia identified Spanish (68.72%) and Arabic (5.44%) as such languages, but declined to provide native-language testing opportunities, stating that "since content instruction is not provided in languages other than English except on a very limited basis, to administer academic assessments in languages in which students are not taught is not considered to be aligned with instruction" (pp. 7-8). The USDOE approved Virginia's plan, which included this failure to comply with its native language assessment requirement, and thus missed an opportunity to potentially improve the reliability and validity of academic achievement testing for ELs in the state.

A study by Abedi and Lord (2001) found that students who are ELs score lower on math tests than proficient English speakers, and that linguistic modification of test items result in higher scores for ELs than non-modified tests. Since ESSA requires all students to take the same test (with the exception of a limited number of students with disabilities), the Plain English forms of the math SOL tests formerly given to beginning- and intermediate-proficiency ELs in Virginia have been eliminated (VDOE, 2017, October 2). Thurlow and Kopriva (2015) suggest that as more states and districts adopt online and computer-based assessments, new technology platforms open the possibility of presenting more testing accommodations for ELs. It is possible to imagine an option to click on a

side-by-side translation of a test item, which would allow for the test to reflect English-language instruction as argued in the Virginia ESSA plan while potentially enhancing the reliability of the test, but no such an accommodation is available (VDOE, 2018). It is important to note that while English is the state's official language, Virginia is not an English-only instruction state, and state law does not prohibit providing instruction or materials in other languages (Code of Virginia, 1996).

Changes in the ACCESS Test and Considerations in Virginia

When NCLB was enacted, the USDOE offered Enhanced Assessment Grants for state education agencies to develop new ELP assessments that would meet the much more stringent NCLB requirements. In 2002, upon receipt of one of these federal grants, Wisconsin, Delaware and Arkansas formed the WIDA Consortium (Boals et al., 2005). With technical support from the Center for Applied Linguistics, the University of Wisconsin system, and the University of Illinois at Urbana-Champaign, the consortium developed a set of English language development standards (Bauman et al., 2007). It took more than three years for assessments aligned with these standards to be fully developed and field-tested (Abedi, 2007), and in 2005, the ACCESS test was operational in three states. Over the years, the consortium steadily gained member states and today, 40 states and territories have adopted the WIDA standards and the ACCESS tests (WIDA, n.d.-c). The Virginia Board of Education adopted the WIDA standards and tests in 2007 and began test administration in 2008 (Cannaday, 2008; Emblidge, 2007; VDOE, 2008). In the 2016-2017 school year, 1,947,902 U. S. students took ACCESS, and 105, 832 took the

test in Virginia (WIDA, 2017a); in 2017-2018, the number of students who took the tests rose to 2,069,398 in total, with 108,293 taking the tests in Virginia. Virginia school districts paid $27.75 per K-12 EL, and $77 for the Alternate ACCESS for each eligible student with disabilities in 2018-2019 (Lane, 2018). This represents $2 per student increase over 2017-2018 base cost (Lane, 2019). The price did not increase for in 2019-2020 (WIDA, n.d.-h). Virginia policy requires ELs to take the test each year until they attain the 4.4 cut score required for identification as "proficient" and reclassification as former ELs (Staples, 2017).

In 2017, WIDA announced that it had changed its score scale to reflect more rigorous Common Core State Standards, and that ELs would have to demonstrate higher levels of language skills on ACCESS in order to achieve the same scores as previous years (WIDA, 2017a). The VDOE responded by reconfiguring its EL progress and proficiency score scales. In a side-by–side comparison of scores on the former and revised scales that year, students who scored 5.0 (the cut score for reclassification) on the former WIDA scale scored between 3.8 and 4.4 on the new scale, depending on grade level cluster. Virginia chose the higher end of this range and reset its cut score, the sole criterion for reclassification, at 4.4 for ELs in all grades, where the previous criteria had been both overall and literacy scores of 5.0 (Staples, 2017). Thus, rescaling made it more difficult for students in some grades to achieve the same score, and this up-scaling and change in decision rule made achieving incrementally higher growth and proficiency rate targets more challenging, as well as making it more difficult for some ELs and easier for others to

exit ESL programming and access other opportunities. These changes complicate the validity of an ACCESS score as the single criterion for reclassification in Virginia.

Validity Questions: Multiple Uses and Multiple Measures

Leading educational testing validity scholar Samuel Messick (1980) urged addressing two questions whenever a test is proposed for a specific use: "First, is the test any good as a measure of the characteristics it is interpreted to assess? Second, should the test be used for the proposed purpose in the proposed way?" (p. 1012). WIDA claims ACCESS scores "have many potential uses," and "work best as a way to aid decision-making," including program entry/exit decisions, establishing when multilingual learners have attained English language proficiency according to state criteria, informing classroom instruction and assessment (including which domains teachers could focus on and how teachers could scaffold instruction), monitoring individual student progress from year to year, and deciding on staffing levels (WIDA, n.d.-a). In addition, ACCESS test scores are the metric for ESL teacher evaluation in some Virginia school districts (T. Jenkins, personal communication, August 29, 2019; B. Russ, personal communication, September 12, 2018), even though WIDA recommends against this practice due to concerns about inadequate sample size and attribution bias (WIDA, 2015). [8]

When a test is interpreted in multiple ways for multiple uses, support is needed for the propositions underlying each interpretation for a specific use. The *Standards* (AERA et al., 2014) state that evidence for interpreting test scores for a

[8] Baker (2013) calls the phenomenon of designing a test for a single purpose but then using it for additional purposes "purpose creep," and says no good solution has been developed to retrofit existing tests to other purposes.

specific purpose does not permit inferring validity for other purposes: "When test scores are interpreted in more than one way… each intended interpretation must be validated" (p. 11). However, it is not clear there is evidence supporting the validity of WIDA's "many potential uses." Menken et al. (2014) say that "to the extent that standards-based assessments are used for multiple and disparate (often unvalidated) decisions, they become increasingly contentious" (pp. 588-589.) Black (2013) refers to the "restricted validity" of the pervasive overreliance on a single annual measure to make high-stakes decisions, noting:

> Systems of summative assessment based only on external testing are unfit for a single purpose, given their inevitably restricted validity, the manner in which they exert pressures that undermine good practices in teaching, and defeat their own stated aim of improving learning. (p. 176)

Consistent with the *Standards* (AERA et al., 2014), ELP assessment scholars recommend using multiple criteria for assessing the level of English proficiency of ELs, particularly for high-stakes decisions such as classification or reclassification of students (Abedi, 2008; Abedi; 2007; Abedi & Deitel, 2004; Boals et al., 2015; Solórzano, 2008). That several studies call into question the alignment between reclassification and academic performance (Ardasheva et al., 2012; DuHart, 2019; Gándara & Rumburger, 2009; Katz et al., 2004; Robinson-Cimpian & Thompson, 2016; Slama, 2012, 2014) seems to affirm this recommendation. In some states, multiple measures are used to inform reclassification decisions. California, for example, requires school divisions to consider teacher evaluation, parent input, and performance on other measures in addition to ELP test scores (California

Department of Education [CDOE], 2017). Multiple criteria are also used for EL reclassification in New York (New York Education Department [NYED], 2015). However, 29 states only require the use of ELP test scores to make reclassification decisions (Linquanti & Cook, 2015). In Virginia, a single overall composite cut score of 4.4 on ACCESS is the sole criterion for reclassification (Staples, 2017; VDOE, 2018). Thus, Virginia's reclassification methodology goes against recommended practice for high-stakes decision-making for ELs.

Also of interest, Abedi (2013) points out an issue with the validity of reclassification decisions based on score calculations: ELs who have been reclassified based on a composite score may not be proficient in one of the language domains. Abedi (2008a) explains that ELP scores can be based on either of two models. A compensatory model relies on the composite score that is the sum of each of the four domain scores (reading, writing, listening, and speaking). Compensatory scoring can result in a higher score in one domain compensating for lower scores in the other domains. For example, a student with a high speaking score but who does not read well could be reclassified, potentially putting that student at risk for academic failure. A conjunctive model, in which students are required to demonstrate proficiency in each of the four domains, regardless of their composite score, is preferable. Under NCLB, Virginia required "proficient" ratings on both an "overall" composite score (based on performance in all four language domains) and a "literacy" score (based on reading and writing subtest scores) for reclassification (Staples, 2017). Beginning in 2017-2018, however, the sole criterion for reclassification of ELs in Virginia schools has been the compensatory "overall" score

(VDOE, 2018).

Several studies have demonstrated poor alignment between reclassification decisions and academic performance, suggesting either that reclassified ELs do not perform as well on other measures as native English-speaking peers (Gándara & Rumburger, 2009; Katz et al,. 2004; Robinson-Cimpian & Thompson, 2016; Slama, 2012, 2014), or that they perform better than other student groups (Ardasheva et al., 2012). DuHart (2019) examined the test scores of a convenience sample of 324 secondary ELs and former ELs in Virginia in 2017 and 2018, and found a positive correlation between ACCESS scores and Virginia Reading SOL scores. However, the study neglects to point out that reclassified ELs (who exit the EL group by achieving the cut score on ACCESS) significantly outperformed all other subgroups of students, and 100% of the reclassified ELs in year two of former EL status in the study (*n=12)* passed grade level Reading SOL tests. In comparison, the overall SOL pass rates in Virginia were 80% and 79% during the two years of the study, and the two highest-performing subgroups, Asian students and White students, passed the Reading SOL tests at rates of between 86% and 91% (VDOE, 2018, August 22). The SOL pass rates of former ELs are included EL subgroup pass rates for four years following reclassification (VDOE, 2018), and state does not disaggregate former ELs from EL subgroup pass rates in published accountability reporting.

Qualitative Research Methods and Test Validity

Qualitative and mixed-methods research are well-suited for examining whether standards and accountability policy fulfills its promise to improve schools and eliminate achievement gaps, and to examine its effects on children, teachers,

curriculum, and instruction (Ryan, 2002; Valenzuela et al., 2007). Test validity scholar Kane (2006) points out, "individual studies in a validity argument may focus on statistical analyses, content analyses, or relationships to criteria, but the validity argument as a whole requires the integration of different kinds of evidence from different sources" (p. 23). This is particularly important when an assessment is used for more than one purpose. Pointing out the need for qualitative research to investigate and evaluate policy implementation under NCLB as it took root in public schools, Valenzuela et al. (2007) highlighted the suitability of qualitative research methods to address problems of practice as well as theoretical questions related to education policy, particularly regarding the effectiveness of standards and accountability reform to meet its stated goal of alleviating the achievement gap between middle-class white students and poor and minority children and their communities. They note there are very few such qualitative and ethnographic studies in the literature.

Winke (2011) examined educator perspectives in a mixed-methods evaluation of the validity of the English Language Proficiency Assessment (ELPA), the high-stakes ELP test used for federal accountability reporting in Michigan prior to that state's adoption of ACCESS in 2013 (WIDA, 2013).[9] Looking beyond the kinds of technical considerations covered in many evaluations of ELP tests, Winke addressed the *broad validity* of the test, a term she used "to refer collectively to reliability, concurrent validity, predictive validity, and consequential validity"

[9] ELPA21, the most recent version of the test, is currently used in seven states, including Arkansas, which was a founding member of the WIDA Consortium (ELPA Consortium, n.d.; WIDA, n.d.-e).

(p.633). A broad validation process not only provides evidence that a test's score interpretations and uses are appropriate, but also investigates the ethics and consequential basis of the use of the test's scores. Employing teacher surveys, Winke found that the administration of the ELPA had negative collateral curricular and psychological consequences, including lost instructional time and feelings of stress, humiliation, and embarrassment. A more narrow, exclusively quantitative evaluation limited to technical considerations could not have revealed these effects of the test. She concludes that while ELP testing may provide beneficial "symbolic recognition" for ELs (they "count"), the testing may also cause some students to be "misrecognized" (represented in inaccurate or harmful ways) or "nonrecognized" (ignored or made invisible; p. 651). The question of different "recognitions" seems particularly relevant to current policy, since ESSA accountability measures aim to make EL achievement more "visible" (West, 2017, p. 75), especially at the individual school level.

AERA et al. (2014) recommend:

> When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those that it is intended to assess or from the test's failure to fully represent the intended construct… A finding of unintended consequences may also lead to reconsideration of the appropriateness of the construct in question. (pp. 30-31)

The judgments of teachers and other stakeholders are important tools for determining the validity of a test (Menken et al., 2014; Ryan, 2002; Winke, 2011).

Ryan (2002) elaborates on the importance of including the perspectives of stakeholder groups, including school administrators, teachers, parents, and students, in the validation of high-stakes assessments. Citing Cronbach, she says one of the tasks of test evaluation involves examining not only the strengths but also the weaknesses of the assessment interpretations and uses. The multiple perspectives can be seen as an antidote to the dilemma of "confirmationist bias," which is the tendency to look for supporting evidence in the validation of test interpretations and uses rather than a more balanced view examining both the strengths and weaknesses of intended interpretations and uses (pp. 8-9). Referring to Messick's construct validation theory, Ryan presents ways in which stakeholders might contribute to the evaluation of six aspects of validation: content, substantive, structural, external, generalizability, and consequential aspects (pp. 9-11). The *Standards* (AERA et al., 2014) also state that it is "useful to consider the perspectives of different interested parties" in the validation process (p. 12). In a historical overview of education policy, Orfield (2016) writes:

> ESEA was developed with active input from two groups of experts appointed by the Administration, drawing on their wide experience and contacts in the education and research worlds. The following major reforms, NCLB and ESSA, were not—they were much more the reflections of ideologies, lobbies, and politics, made with very little consultation with experts on the possible effects of educational policies. (p. 277)

Summary

While NCLB shed light on EL achievement and enhanced ELD standards and accountability, many questions about the reliability and validity of ELP tests remain unexamined. The ACCESS test is widely used for high-stakes decision-making, and under new federal accountability policy, the test scores will be used for new and unexamined purposes. The consequences of these new uses are emerging and remain to be studied. While technical evaluations of the test affirm the test's quality to a limited degree, these findings have not been independently corroborated, and the consequential validity of the test has not been investigated. Validation of a test's uses is the joint responsibility of the test developer and the test user (AERA et al., 2014), but there is no evidence that in the case of the ACCESS test in Virginia, either WIDA or the state has embraced this responsibility. Regarding test validation, the *Standards* (AERA et al., 2014) state, "It should be noted that important contributions to the validity evidence may be made as other researchers report findings of investigations that are related to the meaning of scores on the test" (p. 13).

Teachers, in their roles as test administrators, can offer relevant perspectives on test validity. Furthermore, because their day-to-day work is framed by test-based decisions, they are uniquely positioned to not only shed light on the consequences of a test's uses but to also to raise awareness of some of the questions of reliability and validity that an evaluation of the broad validity of the test entail (Winke, 2011). Qualitative and mixed-methods research are well suited for examining whether standards and accountability policy fulfills its promise to improve schools and eliminate achievement gaps, and to examine its effects on children, teachers,

curriculum, and instruction (Ryan, 2002; Valenzuela et al., 2007). In their literature review on changing assessment policy for ELs under ESSA, Huchinson and Hadjioannou (2017) say the problems for ELs will persist unless significant changes are made, and:

> The time to address assessment concerns for ELs in US schools is now. Without a widespread effort to reevaluate the academic support and corresponding negative consequences for this failure among ELs will have significant consequences for maintaining and sustaining democratic and diversified schools. (p. 121)

CHAPTER 3. METHODOLOGY

Design

This study is a conceptual replication of the investigation Winke (2011) describes in her article *Evaluating the Validity of a High-Stakes ESL Test: Why Teachers' Perceptions Matter.* Winke evaluated the validity of the ELPA, the ELP test used in Michigan to comply with the testing and accountability requirements of NCLB prior to that state's 2013 adoption of the ACCESS test (WIDA, 2013). Gould and Kolb's (1964) *A Dictionary of Social Sciences* defines replication as a scientific method to verify research findings, and "a repetition of a research procedure to check the accuracy or truth of the findings reported" (p. 748). Replication is a key aspect of knowledge building in many areas of research, and has long been recognized as important in education research (Cai et al., 2014). Fisher (1935), a pioneer of modern educational statistics, considered replication to be one of the fundamental building blocks of research design. The importance of replication studies in research has been established in both the physical and social sciences across such varied disciplines as medicine (Begley & Eliss, 2012), linguistics (Chun, 2012; Mackey, 2012; Porte, 2012), psychology (Makel & Plucker, 2014), and advertising (Easley et al., 2000). In contrast to direct replication, which is a narrower, more bounded notion of the exact repetition of an experimental procedure, conceptual replication is a wider notion of replication as a test of more general models, underlying hypotheses, and/or theories (Makel & Plucker, 2014; Schmidt, 2016).

Winke's (2011) study explored the concept of test validity evaluation by examining educators' opinions of the effectiveness of the ELPA "to see if those perceptions could meaningfully contribute to a broad concept of the test's validity" (p. 628). Winke proposed a theory of *broad validity*, which would go beyond providing "narrow" statistical validity evidence--that the test is reliable, has concurrent validity, and has predictive validity --but would also attend to the consequential validity of the test. Consequential validity includes the test's social, ethical, and practical consequences; tests should be fair, meaningful, cost-efficient (Linn et al., 1991), developmentally appropriate (Messick, 1994), and able to be administered successfully (Hughes, 2003). Validity is the overall degree of justification for a test's interpretation and use, "an evaluation, considering all things, of a certain kind of inference about people who obtain a certain score" (Guion, 1978, p. 500).

Research Questions

Winke's (2011) mixed-methods study aimed to understand how educators could shed light on a test's consequential validity. A mixed-methods approach is useful to shed light on the experiences of participants and to identify potential mediating factors (Creswell & Clark, 2018). Thus, Winke investigated two research questions:

(1) What are educators' opinions about the ELPA?

(2) Do educators' opinions vary according to the demographic or teaching environment in which the ELPA was administered?

The hypothesis related to the second question was that educator's opinions would not vary according to demographics or the teaching environment in which the test was administered (pp. 635-636).

The current replication study adapted Winke's research questions to the context of ACCESS testing in Virginia:

(1) What are Virginia educators' perceptions about the ACCESS for ELLs® test?

(2) Do educators' perceptions vary according to demographics or the teaching environment in which the test was administered?

Similar to Winke's study, the null hypothesis related to the second question is that educator's opinions will not vary according to demographics.

In addition, given changes in testing procedures and policy in that potentially impact the test's validity, including Virginia's adoption of the ACCESS 2.0 online test and setting new reclassification criteria, as well as changes in accountability reporting under ESSA, the current study addressed an additional research question:

(3) According to educators, what (if any) are the effects of recent testing policy changes on validity considerations for ACCESS for ELLs® in Virginia?

The project applied the mixed-methods study procedures used by Winke (2011) to study the ELPA in Michigan in order to investigate another high-stakes ELP assessment, the WIDA consortium's widely-used ACCESS test and its administration and uses in a different accountability policy context, Virginia, during the early stages of the state's ESSA plan's implementation. Similar to the ELPA test, which was the focus of Winke's study, the ACCESS test is a standards-based high-stakes assessment developed by a consortium of states to comply with NCLB testing

and accountability mandates (ELPA Consortium, n.d.; WIDA, n.d.-e). The current study surveyed Virginia educators who were involved with test administration and whose work is impacted by test-based decisions. The quantitative portion of the study adopted Winke's Likert-type scale belief-statement items to investigate correlations between demographic factors and perceptions about testing, with a few revisions to reflect the Virginia context and test. The qualitative portion examined data from open-ended comment boxes attached to each of these belief statement items, to provide further insight into the broad validity of the test. In addition to procedures used in Winke's study, follow-up interviews were conducted with survey respondents chosen to represent a range of demographic characteristics and opinions, as a member check and to further explore emerging themes. Surveys were distributed one month after the close of Virginia's ELP testing window, as soon as university Institutional Review Board (IRB) approval was obtained.

Participants and Setting

To recruit participants for the study, the VDOE's September 30, 2018 Fall Membership Reports (VDOE, n.d.-a) were consulted to identify the Virginia school divisions serving the largest numbers of ELs. School division and school websites were then searched to locate English learner educator email addresses. In the interest of efficient recruitment, participants were EL educators contacted in 25 school divisions (among a total of 133 in Virginia) with the largest EL populations that published employee email addresses. Because educator contact information was not available in four of these 25 school divisions, participants were also recruited from the four divisions with the next largest EL numbers. Thus,

recruitment emails were sent to EL educators in 25 of the 29 Virginia school districts with the largest EL populations. This technique of recruiting study participants through publicly-available educator email addresses published on school division websites had been used in previous research conducted at my university, and was approved by my dissertation committee and university IRB. These school divisions ranged in overall enrollment from just over 5,000 students to nearly 190,000 students, with a median district enrollment of about 16,000 students. Table 1 shows representation of six of the eight VDOE Regions (VDOE, n.d.-c), as well as a range of locales (VDOE, n.d.-a) and EL concentrations in these districts (VDOE, n.d.-a). The study recruited EL educators because they directly administer the ACCESS test, oversee test administration, are impacted by decisions based on the test scores, and/or are responsible for making local decisions based on scores, such as deciding on level of linguistic supports and placing students in groups. Survey respondents were given the option of providing contact information if they were willing to participate in follow-up interviews, and participants for interviews were selected based on representation of a range of demographic characteristics and perspectives.

Table 1

School Divisions in Sample

VDOE Region	*n*	Locale	*n*	% ELs	*n*
1 – Central Virginia	3	City	10	< 10.0	12
2 – Tidewater	7	Suburb	9	10.0 – 19.9	6
3 - Northern Neck	2	Rural	6	20.0 – 29.9	4
4 - Northern Virginia	7			30.0 – 39.9	0
5 – Valley	4			40.0 – 49.9	3
6 – Western Virginia	2				
7 – Southwest	0				
8 – Southside	0				

Instrumentation

Winke's (2011) three-part validated survey instrument (see Appendix B) was adapted for the current study (see Appendix A). Winke's survey contained 40 Likert-type scale belief statement items. Her exploratory factor analysis narrowed these 40 statements down to 22 items contributing to the variance in her study. These 22 items were adapted to write the survey instrument for the current study, with a few changes related to differences between the ELPA and ACCESS tests and differences in NCLB and ESSA policy contexts. A few items were added, resulting in a 27 belief-statement survey instrument. The first part of Winke's survey gathered demographic information about participants and contextual information about their schools, including questions about their professional roles, the grade levels of the test with which the participants were involved, who administered the tests, the domains of the test administered, types of schools, and percentage of ELs making up the schools' populations. Items were edited for the current study to reflect the local setting and terminology, and a question about numbers of ELs in the respondent's school(s) was added. At the suggestion of a dissertation committee member, the

current survey also included an item about whether or not the respondent was a native speaker of English. In addition, because educators who have administered the test more times may be more aware of its nuances, because of high teacher turnover rates (Castro et al., 2018), and due to changes in test administration and policy over the past several years (Staples, 2017; VDOE, 2018), the current survey included two additional questions in order to (a) determine participants' years of experience with ACCESS testing, and (b) determine which forms of the test participant administered (the paper version, the more recently implemented electronic version, and/or the Alternate ACCESS for students with disabilities). Furthermore, since consequential validity depends on how test scores are used, a question about the uses of ACCESS scores in local decision-making was included.

The second part of Winke's (2011) survey instrument used forty belief-statements regarding participants' opinions about the test. Winke's participants were asked to indicate on a continuous 10-point Likert-type scale the degree to which they agreed with each statement. Each belief-statement item was followed by a text box in which participants could elaborate on and/or clarify their response if they wished. Winke's exploratory factor analysis resulted in a clear five-factor solution that included salient 22 survey items. The current survey adopted these 22 items, eliminating a few items in Winke's ELPA survey that were not relevant to the ACCESS test (statements about two types of writing questions that ACCESS does not differentiate) and about local scoring (since online ACCESS for ELs 2.0 is centrally scored). The current study added several belief-statement items, regarding recent changes in the ACCESS score scale, which reduced the cut score from 5.0 to 4.4 in

Virginia (Staples, 2017), addressing new uses of ACCESS scores under Virginia's ESSA plan (VDOE, 2018), as well as items regarding students' achievement motivation. Because of the possibility that responses may have skewed to the negative in the original study (Winke, 2011; P. Winke, personal communication, November 17, 2018), care was taken to ensure all belief statement items were written using positive wording.

The third part of the current study used the five open-ended items in Winke's (2011) survey instrument, which invited participants to describe the administration of the test at their schools and to comment on consequences of testing and the uses of test scores. Two additional questions were included to address the recent change in test format from paper to electronic administration and the new uses of scores under Virginia's ESSA plan. As in Winke's instrument, a free-response question provided an opportunity for participants to add anything they wished to say about the test. The final survey item asked participants if they were willing to be contacted by telephone to answer follow-up questions, and if so, to provide contact information. A $15 Amazon gift card was offered as an incentive for participation in these follow-up interviews.

Converse and Presser (1986) recommend using a participatory pilot of a survey instrument before conducting an undeclared pilot. The adapted survey instrument was vetted by conducting a participatory pilot study using cognitive interviews (think-alouds) with five participants, including four current Virginia ESL teachers and doctoral student who had formerly been an ESL teacher in the state. Collectively, the five participants had experience administering the ACCESS test at

all grade level clusters, including the paper and electronic formats as well as the Alternate ACCESS, in five Virginia school divisions, ranging from two to twelve annual administrations. These participants responded to survey items while talking through their thought processes and making suggestions to the researcher. Four of these interviews were conducted face to face, and one was conducted over the telephone. Based on data from the cognitive interviews, Winke's 10-point Likert scale was reduced to the simpler 6-point scale used in the current study, demographic answer choices were added, and survey items were edited for clarity and to better reflect local contexts. The final version of the survey (see Appendix A) was entered into REDCap online survey administration software, and a pilot administered to ten participants confirmed that the application was working well and the survey was ready to launch.

To expand on Winke's (2011) methods, following the survey, semi-structured interviews were conducted with a few willing survey respondents. Semi-structured interviews are a qualitative data collection technique that pairs prewritten interview questions asked to multiple respondents with interviewer-initiated follow-up probes guided by information obtained as the discussion unfolds. They are useful for gathering data on multiple perspectives and viewpoints, particularly when the interviewer is relatively certain that the relevant issues have been identified, but still wishes to provide participants with further opportunity to discuss issues that are important to them (Ahlin, 2019; Wilson, 2013). A semi-structured interview protocol with six open-ended questions about the ACCESS test and ACCESS test administration was developed, as a member check and to explore

emerging themes (see Appendix C). After survey administration was complete, two passes of the data were made to read and hand-code responses, in order to identify preliminary themes and to select potential interview participants representing a range of demographic representation and perspectives. In addition to the six general open-ended questions about ACCESS, the elementary and middle school teachers were asked an additional question based on a theme that emerged from the data related to changes in policy and procedures addressed by Research Question 3, about whether the use of ACCESS scores as a proxy for grade level reading had caused practice change. Middle and high school teachers were asked an additional question about the appropriateness of ACCESS for assessing Long Term English Learners (LTELs), a subgroup of ELs who have not reached proficiency after six or more years of instruction (Great Schools Partnership, 2015; WIDA, 2019), since survey responses had suggested these ELs may be particularly vulnerable to invalid decision-making based on ACCESS scores.

Procedure

Survey administration. Surveys were administered using REDCap software and distributed via email to EL educators in 25 of the 29 Virginia school divisions with the largest EL enrollments on April 29, 2019, one month after the close of the state's English language proficiency testing window, January 21 - March 29 (Lane, 2018), as soon as university IRB approval could be secured. The initial contact email explained the purpose of the study (see Appendix A). A reminder email was sent ten days later, and the survey closed after three weeks. Participation in the study was anonymous, since respondents who believe they can be identified may hesitate to

respond truthfully (Kearney et al., 1984). Explicit statements that personal information would not be disclosed were included in the recruitment emails, surveys, and interviews. Some participants chose to reveal their identities and/or school division affiliations in their responses or when providing optional contact information for follow-up interviews. Of the 2,274 recruitment emails sent with surveys attached sent, 40 bounced back as not delivered, so surveys were distributed to a net of 2,234 potential participants. A total of 273 completed responses were received, yielding a response rate of 12.2%. By way of comparison, the Winke (2011) study received 267 responses. In an attempt to estimate the representativeness of the sample of the total Virginia EL teacher population, three relevant departments at the VDOE were contacted to try to determine the total number of ESL teachers in the state, but they were unable to provide such a statistic. However, by comparing the ratio of teachers to the ELs in the sample districts to the total number of ELs in the state, it was possible to arrive at a rough estimate that the sample represented approximately 10% of the teachers in the state.

Two days after the survey was launched, a complication arose. External research personnel in two large school divisions, where a total of 890 recruitment emails had been sent (to about 39% of the potential participants), notified me that they required prior approval to conduct research in their districts. Because I am employed in another Virginia school division in which the study was conducted, I had sought and been granted research approval from my district, but had not reached out to others, since previous studies at my university had recruited participants using school division emails without prior approval. My school division

had exempted the study from its review process, and the research specialist there had provided a letter regarding the study stating, "… we have determined it does not fall under the purview of the review committee at this time. Because your project is asking for an educator's general perspective rather than information specific to [school division] policy, practice, or procedure, our employees are free to participate at their discretion" (P. Fox, personal communication, February 25, 2019). A copy of this exemption letter, along with university IRB approval documents, a brief description of the study, and the survey instrument to the two divisions in question were provided to the two school divisions. One of the divisions did not contact me again, and I eventually interpreted their lack of further communication as tacit approval of the study. However, the second school division replied that they require all external studies to submit their research screening committee process, regardless of university IRB or other approval, so I submitted an application to their research screening committee for its next review cycle. The division published an August 12 anticipated decision date on its website, but notification of research approval was not received until October 3, more than six weeks after the anticipated decision date and five months after the survey had been launched. Because of pending research approvals, the survey reminder email was not sent to potential participants in these two school districts, and potential interview subjects from this large district were not contacted until approval was granted in October.

Follow-up interviews. After the survey was complete, survey responses were read closely in two iterations and hand-coded to identify preliminary themes.

These readings informed the addition of two questions to the semi-structured follow-up interview protocol (Appendix C), as well as the selection of interview participants. Of the 273 survey respondents, 116 (42.5%) indicated willingness to be interviewed and provided contact information. Nine participants were contacted based on demographic representation and a range of perspectives from positive, negative, or neutral/both. Six of these nine survey participants responded, and six semi-structured interviews were conducted by telephone during the summer of 2019. With the permission of each participant, interviews were recorded and transcribed by the researcher. When the school district that had required prior research approval granted permission to conduct interviews with its employees, four additional survey participants were contacted. One interview was arranged with a participant from that district before the deadline imposed by the research office there, making a total of seven follow-up interviews. Interviews ranged in duration from eight to twenty-three minutes. Following each interview session, $15 Amazon gift card was sent electronically to the participant. Interview participant descriptions are provided in Appendix D.

Data Analysis

The data from the study consisted of four types, (a) demographic and contextual data about participants and their schools, (b) quantitative survey data from the Likert-type scale belief statement items, (c) qualitative survey data from the comment boxes attached to the belief statements, as well as from open-ended questions, and (d) qualitative data gathered in follow-up interviews. Because the goal was to understand how teachers' perspectives could contribute to an

understanding of the ACCESS test's broad validity, an iterative approach was taken to first closely read the 272 survey responses, then conduct follow-up interviews as a member check and to explore emerging themes, then analyze the quantitative data for response patterns in the belief statements regarding the ACCESS test, and finally analyze the qualitative data for clarification and elaboration of the quantitative results. After the data collection was complete, the surveys were entered into SPSS 26 software to analyze the demographic and belief-statement data for teachers' perspectives (Research Question 1). Data were coded on a six-point Likert-type scale, with a score of one indicating the most favorable perspectives and six indicating the least favorable views. Frequencies and descriptive statistics were reported. An exploratory factor analysis was conducted. Teachers' perceptions of factors and survey items were analyzed using one-way ANOVAs and independent samples *t*-tests to investigate whether teachers' perceptions varied according to demographics or the teaching environments in which the test was administered (Research Question 2) and to explore any effects of recent policy changes on validity considerations (Research Question 3). To address the educators' qualitative responses in the text-boxes, data were entered into MAXQDA 2018 qualitative analysis software program, and data segments were coded as either positive or negative in tone. The data were then open-coded by looking for emerging themes while reading responses to each of the questions (Corbin & Strauss, 2008), in two iterations, continually combining and refining the codes. The final codebook (Appendix E) was organized around Winke's (2011) Broad Validity Framework and informed by the research questions, as well as by themes discovered through

quantitative analysis, including achievement motivation, effective administration, and recent changes.

CHAPTER 4. RESULTS

Introduction

The purpose of analysis was to use the quantitative data to identify the characteristics of the participants and the school environments where they administered the test, then to come to a general understanding of their perspectives on the quality and impacts of the test, then to determine whether and how these perspectives might vary according to participant and school characteristics, and finally to examine the qualitative elaboration and open-ended responses to gain deeper understandings of these perspectives. Data analysis was conducted in multiple successive iterations. First, surveys were read closely and hand coded to begin to develop a qualitative codebook and to select interview participants representing a range of characteristics and viewpoints (see Appendix D). Surveys were then entered into SPSS 26 software, and frequencies and descriptive statistics were reported. An exploratory factor analysis was conducted, and one-way ANOVAs and independent samples *t*-tests were conducted to examine variance in survey items and factors by participant and school characteristic. Finally, qualitative data was entered in MAXQDA 2018 software for coding and analysis in multiple iterations, refining the codes each time.

Part One of the Survey: Demographic and Contextual Factors

The first part of the survey included eleven quantitative checkbox items regarding participant demographics and the teaching contexts in which ACCESS was administered. These items were intended to address Research Question 2, "Do educators' perspectives vary according to demographics or teaching environment in

which the test was administered?" The survey received 273 responses, but because participants were instructed to skip any item they did not wish to answer, the response rate for some items was lower than 273. Frequencies and response rates are reported.

Frequencies

Educators' roles, identities, and experience. The first part of the survey asked educators to describe their professional roles. (See Appendix A for the survey instrument.) A majority of participants (*n* = 273) identified as ESL teachers (*n* = 264; 96.7% of the sample). Most identified as ESL teachers only (*n* = 216; 79.12%), but several ESL teachers indicated serving in additional professional roles, including secondary EL content teacher (*n* = 34, 12.45%) and elementary classroom teacher (*n* = 6, 2.2%), and other roles. Participant professional roles are reported in Table 2.

Table 2

Educators' Professional Roles

Please describe yourself. I am a/an…	*n*	%
ESL teacher	264	96.70
ESL teacher (single role)	216	79.12
ESL teacher/secondary EL content teacher	34	12.45
ESL teacher/elementary classroom teacher	6	2.20
ESL teacher/district EL coordinator	2	0.73
ESL teacher/school administrator	1	0.36
ESL teacher/special education teacher	1	0.36
ESL teacher/reading teacher	1	0.36
ESL teacher/instructional coach	1	0.36
ESL teacher/World Languages teacher	1	0.36
ESL teacher (specified itinerant)	1	0.36
Secondary EL content teacher	5	1.82
Principal	1	0.36
ESL Department Chair	1	0.36
ESL instructional assistant	1	0.36
Total responses	272	

Note: Because some participants indicated more than one role, total does not equal 100%.

School contexts in which the test was administered. All 273 participants responded to the survey item asking them to describe the teaching context in which they administered the ACCESS test. Data on school contexts is reported in Table 3. Most teachers described their schools as suburban (*n* = 194; 71.3%), with fewer working in urban (*n* = 67; 24.6%) and rural schools (*n* = 11; 4.0%). School economic demographic information was gathered using two proxy statistics for economic disadvantage (U.S. Department of Agriculture, 2015), free and reduced lunch eligibility and Title I status. More than half of respondents indicated that 40% or more students were eligible for free and reduced price lunch at their schools (*n* = 146; 53.7%), and nearly half reported working in Title I schools (*n* = 130; 47.8%).

The majority administered the test in fully accredited schools (*n* = 173; 84.4% of those responding to this item), while some indicated working schools that were accredited with conditions (*n* = 28; 13.7%), or where accreditation had been denied (*n* = 4; 2.0%).

Table 3

School Contexts

How would you describe your school(s)?	*n*	%
Location		
Suburban	194	71.3
Urban	67	24.6
Rural	11	4.0
Economic Disadvantage		
Title I	130	47.8
40% or more eligible for free/reduced lunch	146	53.7
Accreditation Status		
Accredited	173	84.4
Accredited with conditions	28	13.7
Accreditation denied	4	2.0
Other	5	1.8
At risk for losing accreditation	1	0.4
Alternative school	1	0.4
ESL newcomer hub	1	0.4
School of excellence	1	0.4
Special programs	1	0.4
Itinerant at multiple schools	1	0.4
High immigrant/Latino population	1	0.4

Note: Because some participants indicated more than one contextual factor, total does not equal 100%.

Test /levels, subtests, and test forms administered. Survey participants were asked to indicate the grade-level clusters of the test they administered. The 272 respondents to this item each administered one or more of all grade-level clusters, including the Kindergarten test (*n* = 129; 47.4%), Grade 1 – 2 test (*n* = 152; 55.9%), Grade 3 – 5 test (*n* =257; 57.7%), Grade 6 – 8 test (n = 56; 20.6%), and

Grade 9 - 12 test (*n* = 61; 22.4%). Many participants were involved with more than one grade-level cluster, including multiple elementary-level clusters (*n* = 108, 39.7%), both elementary and middle school clusters (*n* = 12, 4.4%), and both middle and high school clusters (n = 5, 1.8%), as well as all five grade-level clusters (n = 4, 1.5%). Frequencies of grade-level clusters are reported in Table 4.

Table 4

Test Clusters Administered

Grade Level	*n*	%
Kindergarten	129	47.4%
Grades 1-2	152	55.9%
Grades 3-5	157	57.7%
Grades 6-8	56	20.6%
Grades 9-12	61	22.4%
Total responses	272	

Note: Many participants (*n* = 129) administered multiple grade-level clusters

Nearly all of the educators reported administering all four ACCESS skill subtests, with 100% of the 270 who responded to this item having administered the listening and reading tests, which are the two multiple-choice subtests. Nearly all respondents administered the writing and speaking subtests, which require student-constructed responses. The majority of the 271 respondents who indicate the test formats with which they were involved administered the online ACCESS (*n* = 252; 93.4%), but many administered the paper test (*n* = 135; 49.8%) and/or the Alternate ACCESS for students with disabilities (*n* = 76; 28.0%). Many administered more than one form of the test. Table 5 reports frequencies of subtest administration.

Table 5

Tests Administered: Subtests and Forms

	Subtest			Form	
	n	%		*n*	%
Listening	270	100	Online	252	93.4
Reading	270	100	Paper	135	49.8
Writing	268	99.3	Alternate	76	28.0
Speaking	261	96.7			
Total	270		Total	271	

Participants were asked to indicate the number of years of experience they had in administering ACCESS, ranging from one to five or more. Since ACCESS 2.0 was adopted in the state in 2016 (Staples, 2017), teachers with five or more years of experience giving the test would have administered the paper test in the past, even if their schools were currently using the online version. The paper format exposes educators to the content of the test as they handle materials and manually score speaking tests, while the online format, which delivers the test directly to students via individual screens and headphones, affords teachers less exposure to the test content. More than half of the educators surveyed indicated they had administered ACCESS for five or more years (n = 169; 62.1%). Table 6 reports frequencies for number of years administering the test.

Table 6

For How Many Years Have You Administered the ACCESS Test?

Years	*n*	%
One	22	8.1
Two	27	9.9
Three	27	9.9
Four	27	9.9
Five or more	169	62.1
Total responses	272	100.0

Purposes and uses of test scores. The *Standards* (AERA et al., 2014) define test validity as "the degree to which accumulated evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Thus, the survey asked participants about the uses of ACCESS scores at their schools. A majority of educators reported that the test scores were used in decisions about exiting students from ESL programming (*n* = 254; 93.4% of the 272 respondents to this item) as well as placing students in ESL programming (*n* = 210; 77%). Many reported that scores were used for informing classroom instruction (*n* = 210; 77%), as well as decisions about student scheduling (*n* = 201; 73%) and staffing (*n* = 169; 62.1%). Virginia policy establishes an ACCESS score of 4.4 as the sole criterion for reclassification (Staples, 2017), and a large proportion of the educators surveyed affirmed ACCESS scores were used to exit students from the EL subgroup (*n* = 196; 72.1%). More than one fourth of participants reported ACCESS scores are used for teacher evaluation in their schools[10] (*n* = 77; 28.3%). Moreover, qualitative comments revealing school division affiliation indicate that the practice of using ACCESS scores to evaluate teachers occurs in no less than about one third of the 25 Virginia school districts where participants were recruited for this study (*n* = 8; 32%). Frequencies on test purposes and uses are reported in Table 7.

[10] WIDA advises against using ACCESS scores for teacher evaluation due to issues of bias and validity (WIDA, 2015).

Table 7

Score Uses

For what purposes were ACCESS scores used at your school(s)?	*n*	%
Exiting students from ESL programming	254	93.4
Placing students in ESL programming	210	77.2
Exiting students from the EL subgroup	196	72.1
Informing classroom instruction	210	77.2
Student scheduling decisions	201	73.9
Staffing decisions	169	62.1
Teacher evaluation	77	28.3
No response	1	0.4
Other	15	5.5
Showing student growth	5	1.8
School/district accreditation	4	1.5
Department of Justice requirements	3	1.1
Class placement	3	1.1
Counts as "pass" for reading SOL	2	0.7
Deciding ESL service minutes	2	0.7
Assessing language levels	2	0.7
Assigning SOL test accommodations	1	0.4
Small group placement	1	0.4
Student goal setting	1	0.4
Identify student strengths/weakness	1	0.4
Enrichment program design	1	0.4
Total responses	272	

Note: Because most respondents indicated using scores for more than one purpose, the sum of percentages does not equal one hundred.

Participants were asked about the numbers and percentage concentrations of ELs in their schools. Figure 2 shows numbers of ELs enrolled, ranging from schools enrolling fewer than 30 ELs (*n* = 9; 3.3% of the 264 respondents to this item), to schools enrolling 400 or more ELs (*n* = 51, 19.3%). Figure 2 shows percentages of ELs in participants' schools, ranging from less than 5% of the enrollment at their schools (*n* = 25; 9.5% of the 264 respondents) to more than 81% of enrollment (*n* = 7, 2.7%). At most of schools, ELs represented between five and

20% (*n* = 85, 32.2%) or 21 – 40% (*n* = 75, 28.4%) of the student bodies, according to respondents. Percentage concentrations of ELs in schools are reported in Figure 3.

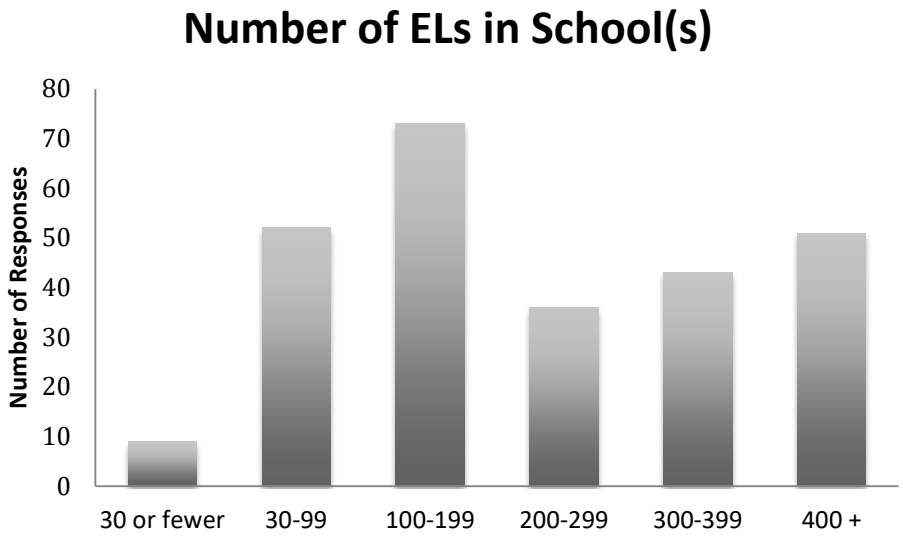

Figure 2: Number of ELs in educators' schools.

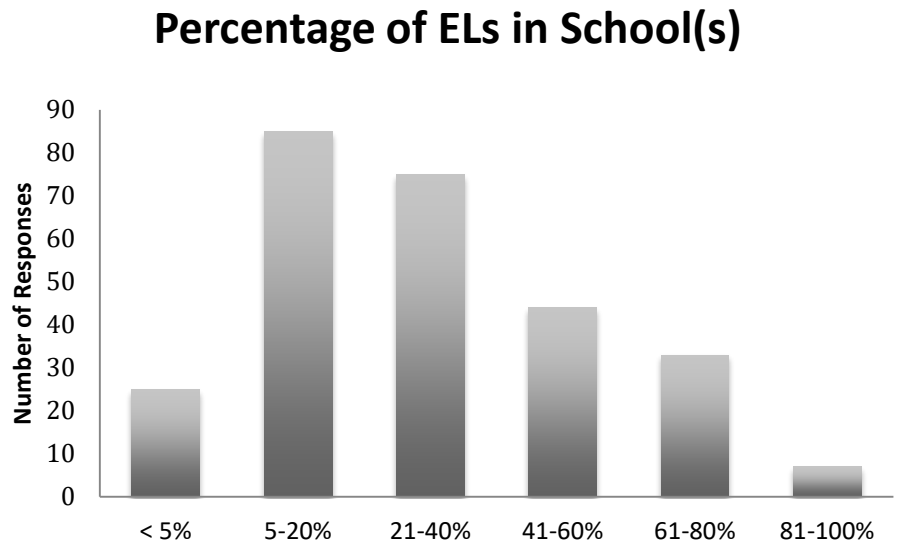

Figure 3. Concentration of ELs in educators' schools.

Parts Two and Three of the Survey: Teachers' Perceptions

Descriptive Statistics for Survey Items

On average, participants' ratings of the positively worded belief statement items ranged from "mostly agree" to "slightly disagree." Table 8 lists a scale for agreement or disagreement with these positively worded Likert-type scale items, ranging from 1-Completely Agree to 6-Completely Disagree.[11] Table 9 lists a scale for agreement or disagreement with these. Thus, since stating that one "mostly agrees" with a positively worded statement would indicate a more favorable perception of that item than "slightly agree," which would indicate a more favorable perception than "slightly disagree," and so on, a score of 1 represents the most favorable rating possible, and a score of 6 represents the least favorable rating possible. Table 9 reports descriptive statistics for teachers' ratings of each of the 27 survey items.

Table 8

Scale for Agreement with Positively Worded Belief Statements

Level of Agreement	Points	Range	Perception
Completely Agree	1	1.0 - 1.4	Completely Favorable
Mostly Agree	2	1.5 - 2.4	Mostly Favorable
Slightly Agree	3	2.5 - 3.4	Slightly Favorable
Slightly Disagree	4	3.5 - 4.4	Slightly Unfavorable
Mostly Disagree	5	4.5 - 5.4	Mostly Unfavorable
Completely Disagree	6	5.5 - 6.0	Completely Unfavorable

[11] A single survey item, "My school(s) received all ACCESS materials on time" was rated in the "completely agree" range; however, constructed comments revealed this item was ambiguous, interpreted by participants to refer to the receipt of test administration items and/or the receipt of test scores, and many participants commented scores were not received in a timely manner.

Table 9

Descriptive Statistics for Survey Items

Item	*n*	*M(SD)*	Agreement
1. My school(s) received all ACCESS materials on time.	270	1.30(0.625)	completely +
2. I feel the training on how to administer ACCESS prepared me well to administer the test.	272	1.76(0.900)	mostly +
3. My school(s) had enough physical space and equipment to administer the ACCESS test smoothly.	271	2.38(1.409)	mostly +
4. My school(s) had enough Internet and server capacity to administer the ACCESS test smoothly.	269	2.00(1.265)	mostly +
5. My school(s) had enough personnel to administer the ACCESS test smoothly.	269	2.16(1.358)	mostly +
6. Teachers had enough support in administering the ACCESS test.	269	2.13(1.278)	mostly +
7. Overall, the administration of the test ran smoothly.	269	1.87(0.980)	mostly +
8. English as a second language (ESL) instruction was positively impacted by the ACCESS test.	266	3.32(1.593)	slightly +
9. The listening test is well designed.	257	2.57(1.201)	slightly +
10. I feel the listening test adequately measures the students' true listening ability.	259	2.75(1.182)	slightly +
11. The reading test is well designed.	258	2.94(1.367)	slightly +
12. I feel the reading test adequately measures the students' true reading ability.	254	3.09(1.368)	slightly +
13. The writing test is well designed.	261	2.82(1.241)	slightly +
14. I feel the writing test adequately measures the students' true writing ability.	260	2.94(1.247)	slightly +
15. The speaking test is well designed.	260	3.21(1.513)	slightly +

Table continues

Table continued

Item	*n*	*M(SD)*	Agreement
16. I feel the speaking test adequately measures the students' true speaking ability.	258	3.51(1.523)	slightly -
17. The students tried to do their best on the ACCESS test.	270	2.39(1.064)	mostly +
18. The students liked the ACCESS test.	265	3.76(1.422)	slightly -
19. The students handled the format of the test well (whether online, paper, or Alternate ACCESS).	266	2.36(1.027)	mostly +
20. The ACCESS test has a positive impact on the students' English language ability.	261	3.59(1.471)	slightly +
21. Overall, the ACCESS test is well designed.	264	2.98(1.228)	slightly +
22. Overall, I feel the ACCESS test is a beneficial test for students.	267	3.27(1.418)	slightly +
23. Overall, I feel the ACCESS test is a reliable measure of English language progress and proficiency.	269	3.09(1.262)	slightly +
24. Overall, I feel the results of the ACCESS test are valid.	268	3.06(1.263)	slightly +
25. The ACCESS scores will accurately reflect the students' actual English language proficiency levels.	266	3.23(1.251)	slightly +
26. Students who do well in their classes do well on the ACCESS test.	263	2.98(1.309)	slightly +
27. A score of 4.4 or proficient on ACCESS is a good indicator that a student will be successful in school.	267	3.55(1.365)	slightly -
Valid N (listwise)	207		

Note: completely + = completely agree, mostly + = mostly agree; slightly + = slightly agree; slightly- = slightly disagree

Exploratory Factor Analysis

Data analysis techniques used in the Winke (2011) study served as a model for much of the analysis in the current study. Before the previous study, no empirical research had been conducted to identify exactly which factors or how many factors contribute to a broad concept of a test's validity. Winke's exploratory factor analysis of her 40-item survey instrument was intended to eliminate any items unrelated to the construct of the consequential validity (see Appendix B). This factor analysis, which retained 22 of the survey items, resulted in a clear five-factor solution. The current study adapted the 22 survey items retained in Winke's factor analysis, and added five more items intended to address the current test and contexts to the survey instrument (see Appendix A). While an explanatory factor analysis was considered for the current study, an exploratory factor analysis was chosen to determine whether other items would be eliminated or different factors would become salient, or whether the five factors identified by Winke would be affirmed.

In the current study, SPSS 26 software employed Chronbach's alpha to conduct an exploratory factor analysis of the quantitative Likert-type scale survey items. A Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy test was run to determine the appropriateness of a factor analysis. KMO analysis of the data yielded a value of 0.897. In a KMO measure, "...values between .8 and .9 are great and values above .9 are superb" (Field, 2013, p. 877.) Bartlett's test of sphericity, which "examines whether a *variance-covariance matrix* is proportional to an *identity*

matrix" (Field, 2013, p. 870) was significant at the < 0.001 level, further affirming the appropriateness of conducting the exploratory factor analysis.

The exploratory factor analysis of the data from the 27 Likert-type scale items resulted in a six-factor solution. All of the survey items loaded onto this six-factor solution; no items were excluded. A principal component analysis extraction method was applied. A Promax rotation method with Kaiser normalization was used. After eliminating all items with communalities less than 0.4, the number of factors to be extracted was determined by the Kaiser criterion, so only factors having an Eigenvalue statistic greater than one were retained. Kaiser's criterion is a rule of thumb that suggests retaining all factors with an Eigenvalue > 1. This method of extraction in a factor analysis appears to be accurate when the number of variables is less than 30, and when the sample size is larger than 250 and the average communality is greater than or equal to 0.6 (Field, 2013; Kaiser, 1960). Data in the current study met these criteria. The six factors explain 68.796% of the variance found in the analysis. Table 10 reports the Eigenvalues and total variance explained by each factor.

Table 10

Eigenvalues and Total Variance Explained by Factor

	Initial Eigenvalues		
Factor	Total	%Variance	Cumulative %
1. General quality	11.143	41.271	41.271
2. Effective admin.	2.645	9.795	51.067
3. Impacts on C. and S.	1.375	5.091	56.157
4. Speaking/listening	1.308	4.845	61.003
5. Achievement motivation	1.146	4.246	65.249
6. Reading/materials	1.006	3.727	68.796

Note: admin. = administration; C. and S. = curriculum and stakeholders

Factor 1 items were related to the general quality of the test. Factor 2 items concerned the effective administration of the test. Factor 3 items related to the test's impacts on curriculum and stakeholders, and or to the test's validity. Factor 4 items concerned the speaking and listening portions of the test. Factor 5 items concerned the students' motivation for achievement. Factor 6 items concerned the reading subtest and on-time receipt of materials. The six factor rotated pattern matrix with values less than 0.4 suppressed is presented in Table 11.

Table 11

Exploratory Factor Analysis: Structure Matrix

	Factor					
Item	1	2	3	4	5	6
22. Overall, I feel the ACCESS test is a beneficial test for all students.	.898		.488			
21. Overall, the ACCESS test is well designed.	.897		.543			
23. Overall, I feel the ACCESS test is a reliable measure of English language progress and proficiency.	.859		.684			
24. Overall, I feel the results of the ACCESS test are valid.	.849		.682			
25. The ACCESS scores will accurately reflect the students' actual English language proficiency levels.	.829		.651			
20. The ACCESS test has a positive impact on the students' English language ability.	.806					
14. I feel the writing test adequately measures the students' true writing ability.	.798		.436			
13. The writing test is well designed.	.778					
9. The listening test is well designed.	.733	.405		.552		
10. I feel the listening test adequately measures the students' true listening ability.	.722			.524		
12. I feel the reading test adequately measures the students' true reading ability.	.695		.676			.645
11. The reading test is well designed.	.695		.538			.684
8. English as a second language (ESL) instruction was positively impacted by the ACCESS test.	.691					
18. The students liked the ACCESS test.	.595				.507	
6. Teachers had enough support in administering the ACCESS test.		.904				
7. Overall, the administration of the test ran smoothly.		.811				

Table continues

Table continued

Item	1	2	3	4	5	6
5. My school(s) had enough personnel to administer the ACCESS test smoothly.		.809				
3. My school(s) had enough physical space and equipment to administer the ACCESS test smoothly.		.758				
4. My school(s) had enough Internet and server capacity to administer the ACCESS test smoothly.		.670	.425			
2. I feel the training on how to administer the ACCESS prepared me well to administer the test.		.549				
26. Students who do well in their classes do well on the ACCESS test.			.768			
27. A score of 4.4 or proficient on the ACCESS test is a good indicator that a student will be successful in school.			.708			
19. The students handled the format of the test well (whether Online, Paper, or Alternate ACCESS).	.492		.537			
16. I feel the speaking test adequately measures the students' true speaking ability.	.525			.892		
15. The speaking test is well designed.	.606		.428	.848		
17. The students tried to do their best on the ACCESS test.					.781	
1. My school(s) received all ACCESS materials on time.						.804

Note: Extraction method: Principal Component Analysis. Rotation method: Promax with Kaiser Normalization.

Results of the exploratory factor analysis in the current study mostly confirm the results of the factor analysis in the Winke (2011) study; not only did both analyses result in similar solutions, but the relative contribution to variance of each factor was also similar. Factor 1 in the current study related to the general quality of the test, and included many survey items, including items relating to the reading and writing subtests. Factor one in Winke's analysis included items related to the

reading and writing tests. Factor 1 accounted for 41.271% of the variance in the current study; similarly, Winke's Factor 1 contributed to 45.7% of the variance in her study. In both studies, Factor 2 items concerned the effective administration of the test, which contributed to 9.795% of the variance in the current study, and 9.73% of the variance in the Winke study. In both studies, Factor 3 items related to the test's impacts on curriculum and learning, contributing to 5.091% of the variance in the current study and 6.08% of the variance in the previous study. In the current study, Factor 4 included items related to the speaking and listening portions of the test, contributing to 4.845% of the variance. Similarly, Winke identified speaking as Factor 4 and listening as Factor 5, and these factors contributed to 5.57% and 4.73% of the variance, respectively. Factor 5 in the current study consisted of items not included in the previous study relating to student motivation for achievement, and contributed to 4.246% of the variance. Table 12 presents a comparison of results of factor analyses in the current and previous studies.

Table 12

Comparison of Factor Analyses

Current Study		Winke Study	
Factor	%Variance	Factor	%Variance
1. General quality	41.271	1. Reading/Writing	45.70
2. Effective admin.	9.795	2. Effective admin.	9.73
3. Impact C. & S.	5.091	3. Impact C. & S.	6.08
4. Speaking/listening	4.845	4. Speaking	5.57
5. Motivation	4.246	5. Listening	4.73
6. Reading/materials	3.727		

Note: The Winke factor analysis data is from "Evaluating the Validity of a High-Stakes ESL Test: Why Teachers' Perceptions Matter," by P. Winke, 2011, *TESOL Quarterly, 45*, p. 640. Copyright 2011 by TESOL Quarterly.

Field (2013) states, "Factor analysis is an exploratory tool and so it should be used to guide the researcher to make various decisions." More specifically, an exploratory factor analysis is useful to "to reduce a data set to a more manageable size while retaining as much of the original information as possible" (p. 698). The exploratory factor analysis for the current study retained all of the items in the data set, perhaps in part because items previously excluded in Winke's factor analysis were not adopted for use in the current survey, and Winke's analysis had already excluded irrelevant items. The first five factors, which related to (a) the overall quality of the test, (b) the ability to effectively administer the test, (c) the test's reliability and validity and thus its impacts on curriculum and stakeholders, (d) the speaking and listening subtests, and (e) students' achievement motivation, all aligned with many of the themes that had emerged from a preliminary reading of the constructed response data in the qualitative portions of the study, and thus proved to be useful for informing further analysis. However, Factor 6, which contributed to only 3.737% of the variance, was not considered further in data analysis for a variety of reasons: reading test items had also loaded on two other factors indicating a weaker contribution of the reading items to variance; the relationship between items concerning the reading test and on-time receipt of material was unclear; the on-time receipt of materials item was revealed to be ambiguous in qualitative comments (with "materials" interpreted as both testing materials and test scores); and the Eigenvalue of 1.006 barely meets the > 1 threshold.

Descriptive Statistics for Factors

Although the exploratory factor analysis did not eliminate any unrelated survey items (Field, 2009), it did provide a six-factor solution. One-way analyses of variance (ANOVA) were conducted to detect differences in five of these factors among demographic subgroups. Post hoc Tukey analyses examined which educator demographic subgroups related to which opinions. Independent samples *t*-tests were also used to compare differences in average responses between demographic and contextual groups.

The average response rates for each factor are listed on Table 13. Of the six factors, on average, respondents had the most favorable perceptions of Factor 2 - the effective administration of the test. The average response score to the six positively worded items that make up Factor 2 was 2.1041, so on average, respondents mostly agreed that test could be administered effectively. On average, study participants expressed slight agreement with the positively worded survey items comprising the other factors, including Factor 3 - the test's impacts on curriculum and stakeholders, Factor 4 - the speaking and listening subtests, Factor 5 - students' motivation for achievement. Factor 1 related to the overall quality of the test and included many survey items regarding testing impacts, design, measures, as well as the reading, writing, and listening subtests.

Table 13

Descriptive Statistics for Factors

Factor	*n*	*M*	Range	Std. Dev.	Agreement
F1. Quality	222	3.1042	1.14 – 6.00	1.01797	Slightly +
F2. Eff. Admin.	247	2.1041	1.00 - 5.43	0.84618	Mostly +
F3. Impacts	225	2.9583	1.00 - 5.69	0.93253	Slightly +
F4. Speaking/listening	251	3.0030	1.00 - 6.00	1.14662	Slightly +
F5. Motivation	264	3.0758	1.00 - 6.00	1.04278	Slightly +
F6. Reading/materials	251	2.4502	1.00 - 5.00	0.95479	Mostly +
Valid *n* (listwise)	207				

In summary, the survey items clustered around six factors. Educators' perceptions of these factors varied to some extent. As a group, educators indicated mostly favorable opinions of the administration of the test. These results affirm the findings of the Winke (2011) study, which found:

> As a group, educators were apprehensive about how effective the exam's administration was. They were, as a whole, troubled about aspects of the different sections of the exam itself. But generally, they were pleased with how the exam impacted certain aspects of the curriculum and the students' English language learning. (p. 641)

Teachers' Perceptions of the Test

Tone. The goal of analysis was to review all the data, then to examine the quantitative data for response patterns in the quantitative portion of the survey, and then to explore the educators' opinions through their qualitative comments. Upon completion of data analysis in SPSS 26, survey data were entered into MAXQDA 2018 qualitative analysis software program. To address educator's responses in the text-boxes that followed the Likert-type scale belief statements and open-ended

items, data segments were coded as either positive or negative in tone. After data segments were coded, twenty-seven survey records were chosen at random and coded by a second researcher (not an EL educator) as a check for reliability and to mitigate potential bias. Approximately 10% of the surveys were reviewed since previous research with large qualitative data sets had established inter-rater reliability on 10% of the data (Chandler, 2003). The level of agreement between the two raters was 93%. Differences of opinion were discussed and resolved. Of the 2,216 total data segments coded, 282 (12.73%) were positive, and 1,934 (87.27%) were negative. It is important to note that while comments elaborating on the belief statements and open-ended survey items were mostly negative in tone on average, quantitative analysis indicated teachers' opinions did not fall into the completely favorable or completely unfavorable range (with the exception of one ambiguous survey item). Results of this coding of tone of qualitative responses are presented in Table 14.

Table 14

Summary of Qualitative Responses

Tone	*n*	%
Positive	282	12.73
Negative	1,934	87.27
Total	2,216	100.00

Quality and impacts of the test. The data were then open-coded by looking for emerging themes while reading responses to each of the questions (Corbin & Strauss, 2008), in two iterations, continually combining and refining the codes. A final coding scheme is presented in Appendix E. Exemplar quotations from the

surveys are presented to illustrate common teacher perceptions, along with the grade level cluster of the test the participant administered. When interview quotations also presented to provide further elaboration of perceptions described on the survey, interview is indicated. Data from the quantitative belief statement survey items suggest that teachers had slightly favorable perceptions of the general quality of the ACCESS test (Factor 1) and slightly favorable perceptions of the test's impacts on curriculum and stakeholders (Factor 3). Qualitative data elaborating on these perceptions reveal that some teachers believe an ELP assessment is needed (*n* = 22): "I feel we do need some information on which to measure their growth, so the ACCESS provides that" (Participant 126; Kindergarten); "Some format of test is necessary to indicate language levels" (Participant 37; 9 - 12); "It's good to have an objective measure like this" (Participant 180; 9 - 12); "I understand the need for a nationally-normed English proficiency test in all four domains" (Participant 223; 1 - 2, 3 - 5).

Favorable perceptions: Using the test to inform classroom instruction. A majority of participants stated that test scores were used to inform classroom instruction in their schools (*n* = 210; 77.2% of 272 responses): "They do help inform instruction and exit students" (Participant 54; Kindergarten, 1–2, 3 – 5). Educators who indicated using test scores to inform classroom instruction had significantly more favorable perceptions of the test. On average, these teachers using scores to inform instruction indicated more favorable perceptions of Factor 1 - the overall quality of the test (*M* = 2.9854, *SD* = 1.00231) than educators who did not use test scores to inform instruction (*M* = 3.5028, *SD* = 0.97687). Educators who indicated

using ACCESS scores to inform instruction (*M* = 2.8655, *SD* = 0.92451) indicated more favorable perceptions of Factor 3 - the test's impact on curriculum and stakeholders, than educators who did not indicate this use of test scores (*M* = 3.2521, *SD* = 0.90443). Educators who indicated using ACCESS scores to inform instruction (*M* = 2.9097, *SD* = 1.12567) also indicated more favorable perceptions of Factor 4, the speaking and listening portions of the test than educators who did not indicate using test scores to inform instruction (*M* = 3.3000, *SD* = 1.17152). Statistically significant differences between the average perspectives of teachers who do and do not use test scores to inform classroom instruction are reported in Table 15.

Table 15

Differences in Means of Factors 1, 3, and 4 by Score Use: Inform Classroom Instruction

Factor	Inform Instruction	*n*	*M(SD)*	*t*	*DF*	*p*
1	Yes	171	2.9854(1.00231)	-3.254	220	0.001
	No	51	3.5028(0.97687)			
3	Yes	171	2.8655(0.92451)	-2.693	223	0.008
	No	54	3.2521(0.90443)			
4	Yes	191	2.9097(1.12567)	-2.320	249	0.021
	No	60	3.3000(1.17152)			

Note: *DF* = degrees of freedom; *p* = 2-tailed significance

Some of the educators who commented about using scores to inform classroom instruction specified using ACCESS scores in conjunction with other information provided by WIDA, such as performance indicators and guidance to teach in the four language domains (*n* = 32): "It informs the teachers about each domain, in my opinion this is very positive" (Participant 179; 6 – 8); "It helps the classroom teachers know what the English linguistic capabilities of a certain ESL student are, and along with the WIDA Con-Do descriptors [sic], it helps that teacher

to have a reasonable expectation of what that student might be able to accomplish" (Participant 130; Kindergarten, 1 – 2, 3 – 5).

Favorable perceptions: Non-native speaker educators. The survey asked participants to describe themselves as native or non-native speakers of English. Most of the educators surveyed identified as native speakers (n = 264; 96.7%), but several were non-native speakers of English (n = 27; 9.9%). Independent samples t-tests determined that educators who identified as non-native speakers of English had statistically significantly more favorable views of several of the survey items than native speakers. Non-native speaker educators indicated stronger agreement with the following belief statement items: "Overall, I feel the ACCESS test is beneficial for students"; "The ACCESS test has a positive impact on the students' English language ability"; "The students liked the ACCESS test"; and "The writing test adequately measures the students' true writing abilities." Furthermore, non-native speaker educators had significantly more favorable views regarding the speaking test. (Among items related to the four subtests, in general, the speaking items were perceived the least favorably). Table 16 reports differences in means between educators who identified as native or non-native speakers of English regarding the quality of the test.

Table 16

Differences in Perceptions: Native and Non-Native Speaker Educators

	Native			Non-Native		
Item	n	*M(SD)*	*Agree*	*n*	*M(SD)*	*Agree*
Overall Beneficial	238	3.35(1.432)	sl.+	27	2.63(1.149)	sl.+
Positive impact Language ability	232	3.68(1.478)	sl.-	27	2.93(1.269)	sl.+
Students liked Speaking Measure	238	3.83(1.411)	sl.-	25	3.16(1.463)	sl.+
Speaking Design	231	3.60(1.506)	sl.-	25	2.84(1.546)	sl.+
Writing Measure	232	3.29(1.506)	sl.+	26	2.56(1.506)	sl.+
	233	3.00(1.247)	sl.+	25	2.44(1.193)	m.+

Note: DF = degrees of freedom; *p* = 2-tailed significance; m.+ = mostly agree; sl.+ = slightly agree; sl.- = slightly disagree

Impacts of the test. Winke's (2011) Broad Validity Framework focuses attention on the impacts of a test. Survey comments were coded according to how they fit in this framework (see Table 17). Constructed response comments shed light on how the unintended consequences of testing result in teachers' perceptions that are not wholeheartedly positive. Some teachers were ambivalent about the impacts of the test and its benefits for students: "It gives us important information but also is extremely time consuming and limits the instructional time ESOL teachers have with students during testing" (Participant 144; Kindergarten, 1 – 2, 3 – 5); "It helps with initial placement and consistency between states/counties, but other academic measures are much more accurate when determining a student's proper placement in courses and need for ESL services" (Participant 196; 9 – 12). Some indicated they thought the test was not necessary because other measures routinely in place were sufficient (*n* = 6), as in "I don't think the test is needed at all,

especially on elementary level. There are plenty of other kinds of testing going on, almost non-stop. In my opinion, once a child is able to pass his reading SOL test, he no longer needs language support" (Participant 191; Kindergarten, 1 – 2, and 3 – 5); "I don't feel we need this test to determine the progress our students are making in the four areas tested. We administer plenty of other tests and have other data points that help determine their progress" (Participant 223; 1 – 2, 3 - 5). Other respondents indicated teachers could more adequately assess their students than the test (n = 6), as in "I don't believe the test is an accurate measure of a student's English ability. I believe the EL teacher should decide if a student is ready to exit, not some test" (Participant 184; 9 – 12).

Table 17

Codes: Broad Validity Framework

Major Category	Subtheme/Code	Frequency
Reliability	Inter-rater – positive	3
	Inter-rater – negative	7
Concurrent	Between forms	15
	Between years	6
Predictive	Align with other tests	2
	Predict classroom performance	2
Consequential	Loss instructional time	
	-Not specified	131
	-Content/classroom	40
	-ESL	37
	-Due to ACCESS test prep	11
	-Prep for other tests	8
	Emotional	
	Students stressed	33
	Students frustrated	21
	Students nervous/anxious	18
	Students singled out/stigmatized	15
	Students tired/fatigued	12
	Teachers stressed	8
	Teachers demoralized	4
	Teachers tired/fatigued	1
	Exit decisions	
	4.4 cut score - Too low	22
	4.4 cut score - Too high	12
	4.4 cut score - Not appropriate	17
	4.4 cut score - Appropriate	2
	Composite scoring not appropriate	7
	Money/Not cost effective	10

Loss of instructional time. Many comments critical of the tests' impacts concerned the loss of instructional time due to preparing for and administering the test. While many educators discussed this loss of time in general terms (*n* = 131), some respondents specified that time taken out for testing had negative impacts on regular content instruction (*n* = 40) and ELD instruction (*n* = 37). "ACCESS test has a highly negative effect on instruction in that students miss hours of classroom

instruction and don't receive support ESL teacher support for an extended (at minimum one month) period of time" (Participant 104; 6 – 8). Additionally, participants indicated time spent on ACCESS test preparation was also a source of lost instructional time (*n* = 11): "We spent several weeks doing the practice tests and talking about personal growth goals prior to testing" (Participant 158; Kindergarten, 1 – 2, and 3 – 5). Furthermore, test administration caused students to miss preparation for grade-level content tests (*n* = 8): "It displaces students from the classroom--i.e. those who are not eligible for an exemption from the Writing SOL--during several critical weeks leading up to the test" (Participant 282; 6 - 8). The teachers' perceptions that the ELP test administration results in a loss of instructional time confirm the findings of the Winke (2011) study, which reported that the administration of the ELPA test significantly reduced ESL services, and caused students to miss out on mainstream classes. While the current survey instrument did not directly address time spent on test administration, many participants (*n* = 66) volunteered that testing interrupted teaching and learning routines for periods of time ranging from several days to over two months. Amounts of instructional time lost to test administration are reported in Table 18.

Table 18

Loss of Instructional Time

Amount	*n*
Several days	4
A week or more	3
Several weeks	6
Two weeks to around one month	34
Around 6 weeks	8
About 2 months	7
More than two months	4
Total responses	66

Timely receipt of test scores. Many participants commented on the lapse in time between test administration and the receipt of test scores (*n* = 60). Several indicated "old" or "out of date" scores negatively impacted the usefulness of these scores: "Our ESL students took the ACCESS from late January to early March. We don't receive their scores until the last week of school in June, so it is difficult for the assessment to drive instruction in April, May, and June. By Sept when the next school year begins, the scores are already 7 months old, and we are expected to plan out instruction based on those 7 month old results" (Participant 130; Kindergarten, 1 – 2, 3 – 5). Several expressed frustration with waiting for scores: "It's widely considered a P.I.T.A. [pain in the ass] whose sole purpose is to disrupt our schedule for 3+ weeks in January every year and then keep us hanging until the last week of school to get our score reports so we can begin to sort out fall planning" (Participant 127; 1 – 2); "I think it sucks that it's only once a year and that the scores take so long to come back. They take the test in February and might score a 3, but by the end of the school year they could be higher, or over the summer they could regress and perform lower coming into the next grade level" (Participant 286; Kindergarten, 1 – 2, 3 – 5, 6 – 8).

Emotional consequences. Educators also discussed negative emotional consequences of testing for students, and in some cases, for teachers: "I dread it every year. It interrupts the momentum we've gained in the first part of the year. I might as well be on leave for a month. The K test is mind-numbing - reading and repeating the same script gets really, really old" (Participant 280; Kindergarten, 1 – 2). Many mentioned stress (*n* = 33, for students; *n* = 8, for teachers): "The students

dislike the test. Taking it is always stressful" (Participant 201; 6 – 8). Several mentioned frustration (*n* =21, students; *n* = 4, teachers): "Sometimes it is so hard they give up and will mark any answer just to finish it. Some students melt down and cry because of the difficulty. They do try until they are too frustrated and upset to complete it" (Participant 89; Kindergarten, 1 – 2, 3 – 5). Nervousness and anxiety were also negative emotional consequences (*n* = 18, students): "The test makes many of my students anxious despite the many opportunities they have to practice with the practice tests" (Participant 161; 6 – 8), as well as fatigue: (*n* = 12, students; *n* = 1, teachers). "It is exhausting and students are tired and unfocused the rest of the day" (Participant 20; 9 – 12). Not only do negative emotional consequences result from taking the test, but they are also seen as a mediator in performance on the test: "There are students who do well in classes that do poorly on the ACCESS test due to stress, test anxiety, or other factors" (Participant 54; Kindergarten, 1 – 2, 3 - 5). Furthermore, teachers indicated students feel stigmatized when they are singled out for ACCESS testing (*n* = 15). "Some students don't have a complete understanding of why they still have to take the test. I often hear, "I speak English." Perhaps they feel it is a punishment or that they feel different than their peers" (Participant 211; 6 – 8); "Long-term English learners are resentful about having to take the test. They identify as Americans and find the label "English learner" to be inaccurate and stigmatizing" (Participant 58; 9 – 12). The Winke (2011) study also found that the ELPA test had negative psychological impacts on students, including stress, frustration, and feeling singled out; however, the previous study did not

report on emotional consequences for teachers. Table 19 reports frequencies of comments related to the emotional consequences of the test.

Table 19

Emotional Consequences of the Test

Consequence	*n*
Students stressed	33
Students frustrated	21
Students nervous/anxious	18
Students singled out/stigmatized	15
Students tired/fatigued	12
Teachers stressed	8
Teachers demoralized	4
Teachers tired/fatigued	1

Note: Some teachers indicated more than type of emotional consequence.

Costs of the test. One consideration in evaluations of consequential validity is that tests should be cost-efficient (Winke, 2011; Linn, Baker, & Dunbar, 1991; Messick, 1980). Although the survey did not explicitly address the financial cost of the test, several participants suggested the test was too costly (*n* = 10). Half of these participants specified they believed the test was a "waste of time and money": "If it wasn't a federally-required test, there's no way we would waste time or money on this joke of an exam" (Participant 181; 9 - 12). Some questioned a profit motive: "More private companies leeching onto the public ed system in an effort to profit from the need to follow uniform standards and have "data" (Participant 188; Kindergarten, 1 – 2, 3 - 5); "I wonder if the changes that WIDA have made in their scoring are based on financial gains. If more students exit the program, then less money is received" (Participant 153; Kindergarten, 1 – 2, 3 - 5).

Effective Test Administration

Enough support. Factor 2, which related to the effective administration of the test, received mostly favorable ratings on average (n =247; M = 2.1041; SD = 0.84618). On average, survey items related to the effective test administration also received mostly favorable ratings (See Table 14). Teachers mentioned support from administrative staff and other colleagues as a factor the successful administration of the test (n = 12): "Strong leadership at this school contributed to the smooth process" (Participant 128; Kindergarten, 1 – 2, 3 – 5); "We have the best testing coordinator ever. Otherwise, administering the ACCESS test would be a nightmare" (Participant 79; 9 – 12). Other comments indicated test administration did not go as smoothly as hoped. Educators cited a lack of support as a factor in problematic test administration (n = 38), for example, "We had no support from administration therefore the students had to be shuffled around, and sometimes in the middle of testing!" (Participant 253; Kindergarten, 1 -2, 3 – 5); "In our school the ESOL leads were expected to do most of the WIDA planning and logistic trouble-shooting even though we were not testing coordinators... we had little admin support, different expectations from each admin, and no interest in sitting down for a debrief after testing to identify areas in need of adjustment for next year's planning" (Participant 136; Kindergarten 1 – 2; 3 – 5). Several teachers indicated they felt ACCESS test administration received less support than other high-stakes tests (n = 19): "Our ACCESS testing does not command the respect (for lack of a better word) that SOL testing gets" (Participant 114; 1 – 2, 3 – 5); "We were on our own - no help with crowd and noise control. Fire drill during testing - nowhere near the same

considerations as during SOL testing" (Participant 173; 6 – 8). Some respondents suggested the new accountability policy had the potential to drive improved support: "I'd like to see more buy-in and support from the school as a whole... considering significant improvement on ACCESS can now replace a failed reading SOL score, I'd like to see school administration promote ACCESS with the same importance as SOLs" (Participant 60; Kindergarten, 1 – 2, 3 – 5, 6 – 8, 9 – 12).

Sufficient physical space and equipment. On average, participants mostly agreed with the item "My school(s) had enough physical space and equipment to administer the ACCESS test smoothly" (n = 271; M = 2.38; SD = 1.409). Very few teachers commented that their schools had sufficient space and equipment, however (n = 3): "My school has the space and equipment, the test runs smoothly. It can be timely since we keep groups manageable and small for them to feel comfortable in the testing environment" (Participant 150; Kindergarten, 1 – 2, 3 – 5). More teachers described issues with space and equipment (n = 50), specifying challenges with finding adequate space (n = 19), insufficient computer labs or available technology (n = 21), and noisy testing environments (n = 10): "Sometimes finding a quiet space to test is an issue since many of us EL specialists share classrooms. Since the ACCESS is given at an odd time of the year compared to end of the year exams, it is also difficult to get computer lab space or get a truly quiet school-wide environment" (Participant 225; Kindergarten, 1 – 2, 3 – 5): "We do not have any classrooms available, so we used the teachers lounge this time- but the vending machines are on, we are near the copy machine, so not the best for quiet" (Participant 129; Kindergarten, 1 – 2, 3 – 5). Several respondents mentioned snow

day cancellations as a hindrance to effective scheduling and test administration (*n* = 15). Table 20 presents numbers of comments about factors in effective test administration.

Table 20

Qualitative Responses: Factor 2 – Effective Administration

Category	Subtheme	*n*
Support	Lack of	38
	Positive	12
	Of administrative staff	12
	Of colleagues	12
	Of testing coordinators	10
	Less than for other tests	19
Training	Learn while doing	10
	Ambiguous	5
Environment	Adequate space and equipment	3
	Issues with space and equipment	50
	Noise	10
	Computer lab/equipment	19
Inclement weather/snow		15

Effective administration in specific teaching contexts. Teachers' perspectives on the effectiveness of test administration varied somewhat according to the teaching contexts in which the test was administered. Specifically, teachers in high-poverty schools and in schools where test scores were used to evaluate teachers had less favorable views, on average, of items related to test administration, and these mean differences were statistically significant.

High-poverty schools. On average, educators in schools where 40% or more of students were eligible for free and reduced lunch had somewhat less favorable views of Factor 2 - effectiveness of test administration than teachers who did not indicate this proxy measure for economic disadvantage. In addition, participants working in schools with high free and reduced lunch eligibility agreed less strongly

with the following belief statements: "Overall, administration of the test went smoothly;" "My school(s) had enough personnel to administer the test"; "Teachers had enough support in administering the ACCESS test"; "My school(s) had enough Internet and server capacity to administer the test smoothly;" and "I feel the training on how to administer ACCESS prepared me well to administer the test." These differences in means were all significant at the $p < 0.05$ level. It is interesting to note that Title I status, which, like free and reduced lunch eligibility, is often used as indicator of socioeconomic disadvantage (National Center for Education Statistics, 2012), was not associated with significant mean differences on any factors or items in this study. Descriptive statistics for factors and items related to effective test administration by 40% or higher free and reduced lunch eligibility are reported in Table 21.

Table 21

Differences in Means of Factor 2 by Free/Reduced Lunch

Item	FRL	*n*	*M(SD)*	*t*	*DF*	*p*	agree
Factor 2	Yes	128	2.2321(0.91148)	2.510	241.447	0.013	slightly +
	No	110	1.19664(0.94943)				slightly +
Smooth	Yes	143	1.98(1.038)	-2.024	267	0.044	mostly +
	No	126	1.74(0.896)				mostly +
Personnel	Yes	143	2.32(1.452)	-2.118	266.480	0.035	mostly +
	No	126	1.74(0.896)				mostly +
Internet	Yes	142	2.16(1.346)	-2.184	267	0.030	mostly +
	No	127	1.83(1.148)				mostly +

Note: FRL = 45%+ students eligible for free and reduced lunch; *DF* = degrees of freedom; *p* = 2-tailed significance

Test administration in schools that use scores to evaluate teachers.

Educators who reported ACCESS scores were used to inform teacher evaluation at their schools (*n* = 77; 28.3% of the 272 responses) indicated less favorable views of

Factor 2 – effective administration, as well as of the following related belief statement survey items: "My school(s) had enough personnel to administer the ACCESS test smoothly" and "My school(s) had enough physical space and equipment to administer the test smoothly." Only one participant comment addressed teacher evaluation: "[The test] might be well designed, but the schools' application of the data is misguided. Teacher evaluations depend on students' ACCESS scores. This is inequitable because other departments offer the chance for remediation after the end of course exams. The ACCESS test is not a content test and so remediation is not offered" (Participant 20; 9 – 12). Table 22 reports differences in the mean scores for teachers where scores are used for teacher evaluation.

Table 22

Differences in Means of Test Items by Score Use: Teacher Evaluation

Item	T. Eval.	*n*	*M(SD)*	*t*	*DF*	*p*	agree
Factor 2	Yes	69	2.3416(1.04669)	2.395	95.236	0.019	mostly+
	No	178	2.0120(0.73738)				mostly+
Personnel	Yes	77	2.57(1.568)	3.302	267	0.002	slightly+
	No	192	1.99(1.230)				mostly+
Space/	Yes	76	2.74(1.569)	2.439	118.525	0.016	slightly+
Equip.	No	195	2.24(1.319)				mostly+

Note: T. Eval. = Teacher evaluation; *DF* = degrees of freedom; *p* = 2-tailed significance

Test content. The content of the test came up as another area of potential threats to construct validity. Teachers commented that the test assessed content knowledge, not just content language (*n* = 29). "The children that had math problems in the reading section were very concerned about getting the math correct. I don't think they should have that on the test in the upper grades. They were afraid to answer and didn't want to go to the next question" (Participant 85; Kindergarten, 1 - 2). Cultural bias was also mentioned (*n* = 5): "The lack of cultural

background often hinders students' reading comprehension" (Participant 191; Kindergarten, 1 – 2, 3 – 5). Participants indicated the test was developmentally inappropriate for younger students (*n* = 25). "It is way too long for the age. A young child who has to read in his second language just does not have the stamina to read and process all that information. The reading sections are developmentally inappropriate. We drive instruction based on DRA levels and the reading levels of the tests are above grade level expectations based on DRA scores. Children who are good readers cannot do well on this test" (Participant 89; Kindergarten, 1 – 2, 3 - 5); "The writing prompts are very wordy in some cases. They cover content that students haven't learned yet. With that said, some students are expected to write an informational, sequenced piece with brand new vocabulary and content that they are unfamiliar with" (Participant 175; Kindergarten, 1 – 2). Some commented that the test did not align with grade level classroom expectations (*n* =24): "Many students can write far better than the test indicates. Most of them are used to taking their time and editing, revising, and discussing their writing with the EL teacher and peers. This allows them to build more complex ideas and vocabulary" (Participant 161; 6 – 8).

The four skill-area subtests. Teachers' perceptions of the content of the test, specifically the four skill-area subtests, the listening, reading, writing and speaking tests, scored in the slightly positive range, on average, with the exception of one item regarding measurement on the speaking subtest, which was rated in the slightly negative range (see Table 13). The qualitative comments related to each of the four subtests were overwhelmingly negative in tone, however, and the speaking

test received the most negative comments. See Table 23 for frequencies of qualitative comments regarding the four subtests.

Table 23

Tone of Comments on Subtests

	Positive	Negative
	n	*n*
Listening	10	104
Reading	15	132
Writing	11	119
Speaking	5	217

Listening subtest. On average, the listening test was perceived the most favorably among the four skill-area subtests. Teachers slightly agreed with positively worded statements about the design of the listening test (n = 257; M = 2.57; SD = 1.201) and the capacity of the listening test to measure true listening ability (n = 259, M = 2.75; SD = 1.182). Furthermore, the listening test received fewer negative comments than other subtests (see Table 23): "I found that most student [sic] were quiet and listened well" (Participant 142, 6 - 8). Another teacher provided this comment to elaborate on the belief statement that the listening test adequately measures students' true listening ability: "If they don't guess, then yes, but since students have to complete questions to finish the test, they tend to guess" (Participant 13; Kindergarten, 1 – 2; 3 - 5). Some participants thought the listening test was too long (n = 9), and tested students' stamina and memory or ability to pay attention (n = 9): "The test was long so some students struggled with stamina. Their scores were not reflective of ability" (Participant 185; 1 – 2, 3 - 5); "The listening test is one of the most difficult sections for US born ELs for as the test becomes more

difficult, it also becomes more boring, and the students find it difficult to maintain their attention level" (Participant 157, 6 – 8; 9 - 12). Another criticism of the listening test was that students could only hear items once (*n* = 17): "The standard test does not allow for information to be repeated. This is unfair for any student. Even as a native English speaking adult - I often ask people to repeat themselves for clarity" (Participant 107; kindergarten, 1 - 2). However, a few participants (*n* = 3) reported their students had discovered a workaround for getting items repeated: "The main problem with the listening test is that when a student [redacted][12], it will re-read the passage/question for them… Being able to [redacted] and re-listen skews results" (Participant 192; 9 -12). Several respondents suggested that listening items requiring students to read answer choices was a potential threat to construct relevant validity (*n* = 10): "There is too much reading. So my students will [sic] high listening skills but low reading skills, don't do well because they can't read the answers. In fact, one of my 1st graders said this was a reading test, not a listening test. I agreed with him" (Participant 129; Kindergarten, 1 – 2, 3 - 5).

Reading subtest. Teachers slightly agreed with positively worded statements about the design of the reading subtest (*n* = 258; *M* = 2.94; *SD* = 1.367) and the capacity of the reading test to adequately measure true reading ability (*n* = 254; *M* = 3.09; *SD* = 1.368). These opinions about the reading test varied by school grade level. One-way ANOVA comparing the means of the two survey items related to the reading test revealed a statistically significant effect of school grade level on agreement with the positively worded belief statement items "The reading test is

[12] Comments with the potential to compromise test security or participant anonymity have been redacted.

well designed" and "I feel the reading test adequately measures the student's true reading ability." Elementary teachers expressed less favorable views than middle school teachers, whose views were less favorable than those of high school teachers. The differences in the means of both reading survey items between elementary and high school test administrators were statistically significant, and are reported in Table 24.

Table 24

One-way Analysis of Variance for the Effects of Grade Level on Reading Subtest Items

	Design			Adequate Measure		
Grade Level	*n*	*M(SD)*	95%CI	*n*	*M(SD)*	95%CI
Elementary	173	3.12 (1.444)	[2.90, 3.34]	171	3.25 (1.410)	[3.04, 3.46]
High	46	2.52 (1.149)	[2.20, 2.88]	44	2.75 (1.203	[2.38, 3.12]

Note: CI = confidence interval

Independent samples *t*-tests were conducted to compare the means of reading item scores of specific grade level cluster tests to those not administering each of those clusters. Teachers administering the grades 1-2 test had slightly less favorable opinions of the reading test design, on average, than those who did not administer the grade 1-2 cluster (*M* = 2.56, *SD* = 1.093, slightly agree). Teachers administering the grades 1-2 test (*M* = 3.35, *SD* = 1.440, slightly agree) had a slightly less favorable opinion of the reading test's capacity to measure students' true reading levels than those who did not administer the grade 1-2 cluster. Teachers administering the grade 3-5 tests also expressed slightly less favorable opinions of the reading test design than those who did not administer the grades 3-5 cluster (*M* = 2.74, *SD* = 1.256, slightly agree). Teachers who administered the grades 9-12 cluster tests indicated slightly more favorable opinions on the two reading test items than those administering the other grade level tests. Teachers who

administered the grades 9-12 test (*M* = 2.51, *SD* = 1.136, slightly agree) had slightly more favorable opinions of the reading test design than those not involved with the grades 9-12 cluster (*M* = 3.06, *SD* = 1.404, slightly agree). Comparisons of means at the other grade levels did not reveal statistically significant results. The significant means differences are reported in Tables 25 and 26.

Table 25

Effects of Grade Level on Means of Reading Test Design

Cluster	*n*	*M(SD)*	*t*	DF	*p*
1-2	145	3.24 (1.483)	4.263	255.271	<0.01
3-5	150	3.09 (1.428)	2.017	256	<0.05
9-12	57	2.51 (1.136)	-3.087	109.286	<0.01

Note: DF = degrees of freedom; *p* = 2-tailed significance

Table 26

Effects of Grade Level on Means of Reading Test Measure

Cluster	*n*	*M(SD)*	*t*	DF	*p*
1-2	143	3.35 (1.440)	3.525	250.761	<0.01
9-12	55	2.75 (1.265)	-2.152	252	<0.05

Note: DF = degrees of freedom; *p* = 2-tailed significance

Qualitative comments regarding reading subtest were mostly negative (see Table 22), and elucidate the more negative viewpoints of the elementary teachers: "The level of reading is too high even for students reading on grade level, especially at 1st and 2nd grades. The passages are too long as well" (Participant 218; Kindergarten; 1 – 2; 3 - 5). Some teachers indicated ACCESS scores would be more useful if the reading scores were aligned with other reading measures and inventories that classroom teachers are familiar with: "I would love to know what the reading level is on the test" (Participant 129; Kindergarten, 1 – 2, 3 - 5);" If we can get reading level correlation-it would be awesome" (Participant 14; 9 - 12).

Writing subtest. On average, teachers were in slight agreement that the writing test was well designed (n = 261; M = 2.82; SD = 1.241) and measured writing ability well (n = 260; M = 2.94; SD = 1.247). There were positive comments about the test (n = 11): "I like how students are given pre-writing steps to scaffold the task" (Participant 164; 9 – 12). Some of these positive comments were accompanied by a caveat: "Although well designed, not all students are comfortable typing on the computer" (Participant 124; 1 – 2, 3 – 5); "I believe it measures some writing skills but a lot of cognition skills" (Participant 190; Kindergarten, 1 – 2, 3 – 5, 6 – 8). Most of the comments regarding the writing subtest were negative (n = 119; see Table 22). As with the listening subtest, teachers pointed out potential threat to construct relevant validity due to the reading, listening, and cognitive skills required to complete the writing test: "The writing prompt is testing them on reading comprehension as well as their writing" (Participant 56; Kindergarten, 1 – 2, 3 – 5); "The writing portion of the test asks students to perform writing skills that seem too advanced. For example, 2nd grade students were asked to write about how [redacted]. The proctor reads text and a series of captions. This test, in my opinion, is testing their listening skills, and their cognitive skills; some students just don't understand what they're supposed to do/write about" (Participant 190; grades Kindergarten, 1 – 2, 3 – 5, 6 – 8). Participants suggested that the test was too lengthy, and that there was not enough time to complete the test: "It is too long. The students tire by the end and do a poor job on the last, most important question" (Participant 103; Kindergarten, 1 - 2; 3 – 5). Some indicated writing test tasks were not developmentally appropriate and not aligned with grade-level classroom

expectations: “The types of prompts don't follow what is taught in school. VA curriculum includes realistic fiction, research reports, memoirs, etc. ACCESS writing prompts test them on Math, Science, and SS. Classroom teachers don't have time to instruct these students in writing formats related to these subject areas... For the prompts and types of writing that are taught in school such as [redacted], students were taught these skills AFTER the ACCESS test” (Participant 153; Kindergarten; 1 – 2, 3 – 5).

The majority of the teachers in this study administered the online version of the test (n = 252; 93.4%). Several of these participants suggested the typing skills required to complete the timed writing portion electronic test is a threat to construct relevant validity (n = 23): “My students in grades 4-5 who are required to type their writing have a lot of difficulty. Typing is a hinderance [sic] because they've never been taught keyboarding and are very slow at it. In addition, their grammar, spelling, capitalization, and punctuation often is lacking when their writing is typed rather than handwritten” (Participant 158; Kindergarten, 1 – 2, 3 – 5).

Speaking subtest. Of the four skill subtests, educators expressed the least favorable opinions of the speaking test, rating the design of the speaking test in the slightly favorable range (n = 260; M = 3.21; SD = 1.513), and capacity of the speaking test to accurately measure speaking ability in the slightly unfavorable range (n = 258; M = 3.51; SD = 1.523). See Table 23 for frequencies of positive and negative comments on each of the four subtests. Moreover, the speaking test received more negative comments than the other subtests (n = 132): “AHHHH, I hate

the speaking test. They don't like talking into the microphone and we can't prompt them to say more if they clam up" (Participant 165; Kindergarten, 1 – 2, 3 – 5, 6 – 8); "I have numerous students who have native-like fluency, yet their speaking scores are Level 2/3. I feel many students do not show their true speaking ability on this test" (Participant 78; 6 – 8). Similar to criticism of the writing subtest, much of the unfavorable comments about the speaking test were related to the online format. Participants suggested having a conversation with a computer and speaking into a microphone was unnatural, and that some of their students were uncomfortable and unsuccessful with recording their own voices: "The students really do not like the online speaking test, even though we practice. It feels awkward talking into a computer. If they make a mistake, they get flustered and hit the stop button (submitting their response before they are ready), instead of completing their answer. They can't talk comfortably and naturally to a computer screen. At our school, we allow them to take it in a very small group and quiet setting to provide some privacy. Even then, kids have said they do not like the speaking part of the assessment" (Participant 116; Kindergarten, 1 – 2, 3 – 5). Several comments suggested the test was too long and/or required too much listening (n = 16), and was a test of attention or stamina (n = 11): "Many children tune out because they have to listen so long and then finally speak" (Participant 237; 6 – 8); "The speaking test is not natural and students have more listening than speaking to perform. The speaking does not adequately measure their speaking ability rather measures their listening comprehension" (Participant 239; Kindergarten, 1 – 2, 3 – 5, 6 – 8).

Several participants acknowledged the potential for standardized administration and centralized scoring of the online speaking test to improve the inter-rater reliability over the locally-scored paper test (*n* = 10), but most of these participants also commented that the benefits of the online testing did not outweigh the problems (*n* = 7): "I understand that having a computer-delivered test does reduce the possibility of teacher influence during test administration, but it also makes the testing experience and environment very unnatural. I do not believe it truly assesses students' speaking ability, and I would say that was reflected in students' scores" (Participant 183; 6 – 8).

Navigating the electronic test. Virginia adopted ACCESS 2.0 online test as an alternative to the paper format beginning in 2016 (Staples, 2017), and the majority of teachers in this study administered the online test (*n* = 252; 93.4%). Qualitative data reveal the electronic format to be the source of a lot of criticism of the test. A few teachers commented that administering the online test was more efficient than the paper test (*n* = 3): "Administering the test online was positive because it allowed the ESL teachers to wrap up testing more quickly" (Participant 106; Kindergarten, 1 – 2, 3 – 5). However, participant comments mentioned a wide range of issues regarding the online test (*n* = 91). Teachers indicated their students were subject to electronic test session interruptions (*n* = 30):"The network often kicks kids out of the test and must be logged back in by test administrator- this might be either the WIDA website test navigator has a limited capacity or the school network" (Participant 170; 9 – 12); "Students were frequently bumped out of the speaking test, had to be moved to a new computer, and subsequently ended up

sitting close to one another when they are supposed to be apart" (Participant 156; Kindergarten, 1 – 2, 3 – 5, 6 – 8).

Participant comments suggest several potential threats to construct relevant validity related to online test administration. Computer literacy was mentioned as a problem (*n* = 30), particularly for younger students and newcomer ELs: "With the way the test is now, we might be getting skewed results because students' dropping and dragging skills come into play, and that's NOT what we're testing... I have never seen anything like it on any other type of assessment, and students are confused by the format as well. A test is NOT the time to throw-in an unusual way of answering a question" (Participant 190; Kindergarten, 1 – 2, 3 – 5, 6 – 8). Teachers brought up that keyboarding skills and typing speed confound scores on the timed writing subtest (*n* = 23): "The online version is difficult because students do not know how to type and lose focus trying to type and remember what they are trying to say" (Participant 89; Kindergarten, 1 – 2, 3 - 5). The ability to operate recording equipment to submit speaking responses was also discussed (*n* = 30): "It is difficult for the students to navigate fairly tricky controls, and some have a hard time overcoming speaking out loud and to a computer" (Participant 138; Kindergarten, 1 – 2, 3 – 5); "I worry at times that true proficiency may not be measured if student's [sic] don't know how to use the technology associated with the speaking test and ACCESS test in general" (Participant 60; Kindergarten, 1 – 2, 3 – 5, 6 – 8, 9 – 12).

The electronic delivery of the speaking test was the topic of many comments (*n* = 30): "Students struggle with the online speaking" (Participant 102; Kindergarten, 1 – 2, 3 – 5). Several pointed out discrepancies between scores on the

paper and online tests: “Ever since the speaking domain has been tested online, instead of in person, regardless of how often a student practices in the lab with the headset with a microphone, every year there is always at least 1 student, sometimes 2, who "freeze up" when the actual testing time happens. They give only minimal responses and do not show their actual capabilities, and therefore receive low speaking scores, which do not accurately reflect their abilities. This never happened when we were administering the speaking test one-on-one, before it was online. Our students were familiar with us, spoke frequently during our ESL classes, and were at ease and tried their best when we tested them in person” (Participant 130; Kindergarten, 1 – 2, 3 – 5); “Many students feel uncomfortable speaking into a microphone or headset so they don't perform as well on the ACCESS as they would in a classroom setting. Their true ability is not really shown” (Participant 285; Kindergarten, 1 – 2, 3 – 5).

Finally, opinions of the computer adaptive features of ACCESS 2.0 were mixed. There were a handful of favorable comments (*n* =3): “The online version of the test is adaptive. I believe this has been a positive change as students are getting questions/tasks specifically targeted at their ability rather than tasks that might cause frustration or boredom” (Participant 164; 9 – 12). Others said their students had been inappropriately sorted by the computer adaptive test (*n* = 5): “Because the reading and speaking tests have to be taken first and the writing test is based on those scores, most of my level 3-4 1st grade students had to take a writing test that was not at their level. They were asked to [redacted], when they are fully capable to take the more advanced writing test” (Participant 80; Kindergarten, 1 – 2); “While I

like the speaking test as well because it grades actual language output, I don't like the fact that the tier is determined by the listening test. I had a student who is a level 2 and has been in our country since birth, test into PreA for speaking. I also had a newcomer who had been here for 2 months test into higher level… I don't like that teachers have no control over placing students into which tier... Computer adapted tests, while useful, are no replacement for actual human teachers with experience and knowledge of students abilities" (Participant 94; 1 – 2, 3 – 5).

Test preparation. Responses to open-ended survey item, "Did students at your school(s) prepare for the ACCESS test? If so, please describe how they prepared" revealed a range of practices in test preparation. About three-fourths of the educators who responded indicated they had taken class time to prepare students for the test (*n* = 164, 74.89%) Several respondents said that since unfamiliarity with computers posed a challenge to responding to test items for some students, they had practiced using technology (*n* = 20). Many used the preparation modules provided by WIDA to practice the online format of the test. Teachers also mentioned familiarizing students with requirements of the test that were different from routine instructional practices and expectations: "We did the practice tests available on the WIDA website. We also practiced speaking to ourselves via Flipgrid.com. Writing to WIDA-like prompts that follow functions for CAN-DOs. Also purposeful practice - use language for discussion, recount, description, or argument as per WIDA" (Participant 197; 9 – 12). Several educators said test preparation activities were provided for targeted groups only (*n* = 25, 11.41%), including younger students, lower proficiency ELs, newcomers, students taking the online

test, and "older strategic students (3-5) who we believed could exit" (Participant 54; Kindergarten, 1 – 2, 3 – 5). On the other hand, some teachers said they did not prepare students for the test except through regular classroom instruction, nor did they familiarize students with test format before administration (*n*= 30, 13.7%). Table 27 provides frequency data for test preparation.

Table 27

Did students at your school(s) prepare for the test?

Test Preparation	*n*	%
Yes	164	74.89
No	30	13.70
Yes, targeted groups only	25	11.41
Total	219	100.00

Factor 5 – Achievement Motivation

Factor 5 – achievement motivation relates to two belief statement items on the survey, "The students tried to do their best on the ACCESS test", with which participants mostly agreed on average (*n* = 265; *M* = 2.39; *SD* = 1.064), and "The students liked the ACCESS test," with which participants slightly disagreed (*n* = 270; *M* = 3.76; *SD* = 1.422). On average, participants indicated slightly favorable views of Factor 5 – Achievement Motivation (*n* = 263; *M* = 3.0742; *SD* =1.0444).

Many qualitative comments, enumerated in Table 24, elaborate on students' motivation to perform. Some participants commented that the test was a positive motivator of achievement (*n* = 22): "It motivates students to raise their level of English proficiency in order to exit the program. Give students a gauge of where they are on the learning continuum" (Participant 103; Kindergarten, 1 – 2, 3 – 5); "Students know that as they become for proficient in English, their opportunities

broaden for electives and for taking classes with the gen ed population. They become more confident" (Participant 142; 6 - 8). Others indicated students lacked motivation to do well on ACCESS: "Sometimes students have not shown what they can really do on this test. Their [sic] is a certain amount of motivation and focus required" (Participant 280; Kindergarten, 1 – 2); "Most of the students hate taking this test. It takes them a very long time and they find it frustrating" (Participant 161; 6 – 8). Several teachers indicated they believed test scores depended on motivation or effort (n = 7): "If students aren't focused on the test or aren't sufficiently motivated to do well on it, then the scores don't reflect students' proficiency in the four domains" (Participant 227; 6 – 8).

A one-way ANOVA conducted to compare the mean ratings of Factor 5 by school level mean revealed statistically significant differences: teachers who administered the elementary tests had more favorable views of achievement motivation than those who administered the middle school tests, who in turn had more favorable views than those who administered the high school tests. Results of the ANOVA of Factor 5 by school level are reported in Table 28. Independent samples t-tests conducted to compare the average ratings of Factor 5 by each grade level cluster of the test to those teachers not indicating that grade level cluster reveal that on average, teachers who administered the Kindergarten tests had the most favorable views of achievement motivation, followed by administrators of the Grades 3 – 5 cluster, followed by Grades 1 – 2, then the Grades 6-8, and teachers who administered the Grades 9 – 12 test expressed the least favorable views. These differences in means were all significant at the < 0.01 level. Results of the

independent samples *t*-tests on Factor 5 by grade level cluster of the test are reported in Table 29.

Table 28

One-way Analysis of Variance: Effects of Grade Level on Factor 5

Grade Level	*n*	*M(SD)*	95%CI	agreement
Elementary	178	2.8174 (0 .96204)	[2.6751, 2.9597]	slightly +
Middle	39	3.5769 (0.97696)	[3.2602, 3.8936]	slightly -
High	46	3.6413 (1.04702)	[3.3304, 3.9522]	slightly -
Total	263	3.0741 (1.04444)	[2.9473, 3.2010]	slightly +

Note: CI = confidence interval; *DF* = degrees of freedom; *p* = 2-tailed significance

Table 29

T-*tests: Effects of Grade Level Cluster on Factor 5*

Cluster	*n*	*M(SD)*	*t*	DF	*p*	difference
K	126	2.7659 (0.98256)	-4.083	262	<0.01	-0.3099
1-2	148	2.8378 (0.97802)	-4.326	262	<0.01	-0.2380
3-5	153	2.8072 (0.93211)	-5.145	262	<0.01	-0.2686
6-8	55	3.4091(0.99112)	2.696	262	<0.01	0.3333
9-12	56	3.7142(1.04819)	5.437	262	<0.01	0.6384
Overall	257	3.0758 (1.04278)				

Note: DF = degrees of freedom; *p* = 2-tailed significance

Some participants commented the students liked or enjoyed the test (*n* = 17), particularly younger students: "K-3 tend to enjoy the test" (Participant 143; Kindergarten, 1 – 2, 3 – 5). Several included a caveat: "Some kids found it fun. Some simply clicked quickly through it. Most simply got through it" (Participant 280; Kindergarten, 1 – 2); "The kindergartners love their test because it is fun and interactive. The older students don't really differentiate between ACCESS and all the other online tests they take - they just tolerate them and move on" (Participant 88; Kindergarten, 1 – 2, 3 – 5).

Others were less motivated, particularly the older students and LTELS: "Students despise the ACCESS test. They take it year after year and most do not take

it seriously" (Participant 16; 9 – 12); "The ones who don't try are the long-term students (identified EL but not in ESL classes because they finished the program long ago). Some of them have attended [redacted] since K" (Participant 248; 9 – 12).

Qualitative comments indicate that not only do younger and older students differ in motivation and effort, but lower and higher proficiency students also differ. "Our lower level students always put forth a great deal of effort but the upper level students who only have one or two ESOL courses and know that ACCESS scores have no impact on graduation are much harder to motivate" (Participant 164; 9 – 12); "Some kids that are higher level English speakers, still blow the test sometimes if they have an attitude" (Participant 36; 1 – 2, 3 – 5).

Sometimes students did not try their best, and guessed or rushed through test items (n = 44): "I feel that the students taking the online test just clicked through everything. They were done in very short amounts of time" (Participant 134; Kindergarten, 1 – 2, 3 – 5). Teachers explained lack of effort on the part of younger and low proficiency students lacked effort for different reasons than older and high proficiency students. Low proficiency and younger students tended to guess on multiple-choice items because the test was long and difficult: "The reading test is not appropriate for the emergent, level 1 and level 2 language learners. The Tier A grade 1-2 test has paragraphs of reading with very little picture support. The test is stressful for students and we often have to just tell them to pick an answer and move on" (Participant 94; 1 – 2, 3 – 5); "I have students who cannot read at all score higher than students who can read. The students who can't read click through the test and end up choosing many of the correct answers" (Participant 153;

Kindergarten, 1 – 2, 3 – 5); "The younger children thought the online test was like a video game and whoever got done the fastest did the best. We spoke to them about it many many times but during testing, they still worked super fast" (Participant 102; grades K, 1 – 2, 3 – 5). Teachers indicated that older students may not find the test relevant or beneficial to them (*n* = 20): "Many higher level high school students don't care anymore about their results because it no longer impacts their ability to graduate" (Participant 196; 9 – 12). Other reasons for lack of motivation and effort were that students didn't take the test seriously or care about the test (*n* = 27): "At the high school level most of the students that have been in the program for several years do not take the test seriously, therefore, their score does not reflect actual ability" (Participant 113; grades 9 – 12).

Participants indicated another reason for low achievement motivation was over-testing (*n* = 21): "The students don't care much. They take so many tests - who can blame them?"(Participant 231; 9 – 12). The over-testing of ELs in particular was mentioned (*n* = 12): "They are tested more than any other students in the school because they have ACCESS on top of SOLs, MAPS, DRA, PALS, benchmarks, classroom tests, simulations, etc." (Participant 55; Kindergarten, 1 – 2, 3 – 5). Some reported students had grown tired of taking ACCESS year after year (*n* = 6); "Many students get WIDA fatigue after a few years and don't make an effort" (Participant 137; 9 - 12). Some suggested the reuse of test items had a negative impact on motivation (*n* = 8): "A student reported that the listening test was the exact same as the one she had taken last year and so she skipped through it quickly because she was already familiar with the answers" (Participant 246; 1 - 2, 3 – 5). In addition,

teachers revealed some students intentionally performed poorly on the test (*n* = 10), "to try to remain in EL services for the testing accommodations on the SOLs" (Participant 161; 6 – 8), or "because they believe ESL classes are easier" (Participant 288; 9 - 12).

Participants suggested that lucky guessing or lack of effort may be inflating the scores of some younger or low proficiency students, and that conversely, guessing or lack of effort may be deflating the scores proficient students and not allowing LTELs to show their true proficiency levels or reclassify: 'When student get higher scores by just guessing an answer it affects the reliability of the test and they exit out of ESOL services that they may still need in the next class" (Participant 264; Kindergarten); "At the high school level most of the students that have been in the program for several years do not take the test seriously, therefore, their score does not reflect actual ability" (Participant 113; 9 – 12). Table 30 reports on comments about achievement motivation.

Table 30

Factor 5 – Achievement Motivation

Tone	Major Category	*n*
Positive	Motivation for achievement	22
	Enjoy/fun	17
Negative	Guess/rush/lack of effort	44
	Don't take seriously/care	27
	Not relevant/beneficial/motivated	20
	Overtesting	21
	Overtesting - ELs	12
	Overtesting - SOLs	7
	Overtesting -"WIDA fatigue"	6
	Overwhelmed/nervous	19
	Do poorly on purpose	10
	Recycled test items	8
	Score reflects effort	7
	Boring	6
	Hate/don't like	5

Multiple Measures

On average, participants slightly disagreed with the survey item "A score of 4.4 or proficient on ACCESS is a good indicator that a student will be successful in school" (n = 267; M = 3.55; SD = 1.365). Comments regarding the cut score, which changed in 2016-2017 (Staples, 2017) were mostly critical (n = 51), and very few participants said they thought the new cut score was appropriate (n = 2): "4.4 is now proficient?!?!" (Participant 214; 1 – 2, 3 – 5); "VA changed what level is now exiting from 6 to 4.4. I feel like that's weird" (Participant 189; 1 – 2; 3 – 5). Some thought the reclassification score was too low, causing ELs to exit programming when they still needed supports (n = 42): "Many students at this level still need services which cannot be provided due to testing out of the program" (Participant 124; 1 – 2, 3 – 5). Others indicated the cut score too high, holding ELs back who

were ready to exit ELD programming (n = 14): "Many students can not pass the ACCESS test with a score of 4.4 but they are fluent and in some cases speak more English than their native language. They are passing their classes. These children stay in the program even though they should be exited because of the rigor of the ACCESS test" (Participant 89; Kindergarten, 1 – 2, 3 - 5). On average, participants slightly agreed with the item "Students who do well in their classes do well on the ACCESS test" (n = 263; M = 2.98; SD = 1.309). Elaborating comments expressed disagreement: "ACCESS scores often do not reflect students' performance on other tests or in the classroom" (Participant 95; Kindergarten, 1 – 2).

A theme that emerged from the qualitative data was discussion around the use of multiple measures in high-stakes decisions, summarized in Table 42. Some participants expressed the opinion that other factors should be taken into consideration when deciding to exit students from EL status and programming (n = 21). Regarding reclassification, one participant wrote, "We no longer have a say, it is strictly based on the results of the test. I feel this is morally wrong for decisions in exiting and staying in. I feel the parents and school should be a part of the decision" (Participant 129; Kindergarten, 1 – 2, 3 - 5). Respondents reported that multiple measures are considered when making high-stakes decisions other than reclassification (n = 31), especially decisions about placement and levels of ELD services: "We use the ACCESS test results as one data point in planning instruction for our ELs" (Participant 163; 6 - 8); "Other factors taken into consideration included student grades, reading levels, performance on local reading and writing inventories, SOL scores, classroom performance, and parent and teacher input"

(Participant 164; 9 - 12). Several respondents said they considered ACCESS scores a "snapshot" and not a comprehensive indicator of a student's proficiency (*n* = 26): "They are a valid representation of what a student could do on that day. What students can do varies from day to day in my experience" (Participant 229; 1-2, 3-5). Table 31 enumerates comments made regarding multiple measures.

Table 31

Use of Multiple Measures in Decision-Making

Major Category		*n*
For exit decisions	Yes	0
	No	183
	Should	21
For other decisions	Yes	31
	No	17
	Should	16
Mentions snapshot/single data point		26
No because DOJ		3

Note: DOJ = Department of Justice mandate

Special Populations: Students with Disabilities and LTELs

Qualitative data revealed particular concern about whether the progress and proficiency of specific groups of students was appropriately measured by the test. One such concern was the appropriateness of ACCESS for students with disabilities (*n* = 32): "I think we should have a bigger discussion about English language learners who are dually identified (have IEPS) and cannot test out of ESOL as a result" (Participant 203; 6 - 8). Of particular concern was whether the 4.4 cut score, which is the only criterion for reclassification, was attainable for these students: "There are students that are in special education that will never be able to exit the program because they cannot pass this test. This is frustrating because of the impact

on the student" (Participant 179; 6 - 8); "If students are low or SPED they are scoring poorly but it is not because language is an issue. We have some students who don't even speak another language besides English but are low, sped, confused by the test, or not good test takers" (Participant 81; 1 – 2). Individualized Education Program (IEP) accommodations routinely afforded students with disabilities not being allowed for ACCESS testing was also mentioned as a confounding factor in the scores of students with disabilities (n = 5): "Dually-identified students (ELs who have learning differences) have a very difficult time with the test as the accommodations allowed by WIDA often do not address their learning disabilities" (Participant 58; 9 - 12). The issue of native speakers of English in EL classification was also raised (n = 5): "I cannot stress enough that the test is not a valid measure for students with intellectual or learning disabilities. I have students who do not speak a language other than English but were put into the ESOL program because a grandparent occasionally speaks a few words of another language around them, but the students are assessed as entering, level 1 English learners" (Participant 187; 1 – 2, 3 - 5); "There is little accommodation for students with IEPs unless they have a severe disability that qualifies them for the alternate access. Most of my SPED [special education] students are simply "stuck" because their scores are not improving" (Participant 158; Kindergarten, 1 – 2, 3 - 5).

The data also raised questions about the ability of LTELS to exit based on a single cut score (n = 14). LTELS are defined as students identified as ELs for six or more years (Sugarman & Geary, 2018), thus the LTEL phenomenon is a concern in secondary schools. Quantitative data suggest teachers view secondary students as

less motivated to achieve on ACCESS than younger students (see Table 14). Long term ELs, by definition, fail to pass the test: "Students who do well in their classroom never see an ESL teacher. They do not think of themselves as an ESL student. Then, once a year, I pull them into a classroom and ask them to try their best on these 4 domains. I feel these are the students who do the most poorly on the test and do not exit as quickly as one would expect" (Participant 229; 1 – 2, 3 - 5); "I feel that [the test] simply pulls US born students into a catch-22 from which they cannot exit… They often do not care as they have other worries" (Participant 157; 6 – 8, 9 - 12).

Because survey data seemed to suggest a potential correlation between the LTEL phenomenon, a major issue in EL equity and accountability, and test-taking effort, secondary follow-up interview participants were asked about the appropriateness of ACCESS for LTELs (see Appendix C for interview protocol). One participant suggested: "For long term… I think, possibly, when they get to higher levels, there should be some other tool used, because I think it's not necessarily a good tool, for possibly other issues that could be going on. I think some students tend to test up to a certain level, and then they don't go any higher" (Participant 98; 9 – 12; interview). When asked about LTELs, another interview participant suggested the test should be normed on the English proficiency of grade-level peers: "I would love somebody to share with me, if… the ACCESS test has ever been normed based upon American English language speakers born and raised in the United States. And I don't believe it has been. And if you fail to do that… measure that have they reached proficiency alongside their peers, at the same age level, same

grade level… It is impossible to determine that they have reached proficiency, because it's some outside entity saying this, but you haven't normed it against what is expected of every other student at this grade level. Additionally, it seems to me that you are putting yet another testing requirement upon students when the state of Virginia for instance, will say you've met your academic requirement for passing your grade-level reading or writing SOL, and I have seen that same student fail to make the test-out on the ACCESS test. So you're telling me that our state says one thing, and yet ACCESS is saying no, you're still not good enough" (Participant 104, 6 – 8, interview). Several survey participants (*n* = 18) also suggested that if native speakers of English took the test, they would not be classified as proficient: "Many fluent speakers who have been in the ESOL program for an extended amount of time tend to plateau on their scores because of both lack of motivation and also because of the increasing rigor of scoring. Honestly, many native English speakers could not get a passing score on this test because it holds ELs to much higher standards than many schools hold their gen ed students"(Participant 174; 9 - 12).

Recent Changes in Testing Policy

Three of the open-ended survey questions addressed Research Question 3, regarding recent changes affecting ACCESS score use asked educators if they were aware of any changes to the ACCESS test, scoring, or uses of test scores in decision-making in recent years. Of the participants who responded to these items, several answered that yes, they were aware of such changes (*n* =14), but more responded that no, they were unaware (*n* = 58), and still others said they were not sure (*n* = 22). Rather than simply stating whether they were aware or not, many participants

responded by explaining their understandings of the changes (*n* = 89): "We transitioned from paper format to online. The cut scores have lowered. Gaining a pass on the SOL Reading test if they show improvement on ACCESS" (Participant 118; Kindergarten, 1 – 2, 3 - 5); "Because the scoring was made more rigorous, the cutoff was adjusted" (Participant 171; 3 - 5). Included in these explanations were mentions that ACCESS scores can now be used as a proxy for Reading SOL scores (*n* = 10) and that ACCESS scores are now a factor in school accreditation (*n* = 7). Among teachers who described recent changes, several misstated details (*n* = 28), such as "We just follow the rules dictated by the country and state. Students are exited at 4.4 but are monitored for three years"[13] (Participant 102; Kindergarten, 1 – 2; 3 - 5); "They lose all their accommodations even if they have an ACCESS of 4.5, are failing reading, are below grade level on the DRA reading benchmark"[14] (Participant 229; 1 – 2; 3 - 5). Table 32 presents coding categories regarding recent changes.

With the goal of gaining a better understanding of the effect of recent changed in order to address Research Question 3, follow-up interview participants were asked about recent changes (see Appendix C). Interview data suggests that some teachers believed that awareness of policy changes had a positive impact, or at least had the potential to make a positive impact. In response to a survey item about whether the ACCESS test affected instruction at your school, one survey respondent wrote: "I think in some ways, it has added credibility to the ESL Program. If they

[13] Virginia policy currently provides for monitoring of ELs for four years; prior to ESSA, former ELs were monitored for two years (VDOE, 2017, October 2).
[14] ELs are still eligible for testing accommodations for two years after reclassification under Virginia's ESSA plan (VDOE, 2017).

show growth on the ACCESS but don't pass the Reading SOL, the school can get a credit for the Reading SOL" (Participant 118; Kindergarten, 1-2; 3-5). When asked to elaborate on this response in a follow-up interview, they clarified: "It has, kind of like I said, given me credibility, because they're ESL, and you know, they're never going to pass the SOL. I get that… So, this way we're showing growth, we're getting a pick up, and I think teachers are investing more time in them than they did before, because they see, "Oh, if we get a point 3, which is a much more reachable goal, then we'll pick up the SOL credit which helps for their accreditation" (Participant 118; Kindergarten, 1-2; 3-5; interview). On the other hand, when another interviewee as asked if their school was more concerned about ACCESS testing since the new policy allows it to count as a pass on the Reading SOL, they responded: "No. (Laughter.) It does not seem to have made any impact on their feelings regarding that… Although I have attempted to do the best PR than I can do, and I am not quiet about it… hasn't made a huge impact… Our multiple different populations have not been making the progress needed, that the school is focusing so much on those groups that need to make the grade to improve and keep the school accredited, and the ELs have been making the grade, and therefore, we're not getting the support that we really need or should have" (Participant 104; 6 – 8, interview). When asked the same questions about the impact of allowing ACCESS growth to count as Reading SOL growth, another participant had not heard of this policy change: "First of all I didn't know that it could be reported as a pass on an SOL. So that's information that I was unaware of" (Participant 132; Kindergarten, 1 – 2, 3 – 5; interview). Follow-up interviews also suggested that when teachers were aware of policy changes, it was

because an individual supervisor had made them aware. The teacher who believed their credibility had improved stated: "It came from DOE to our principal, and the principal told us, and we all passed it out, immediately, as soon as we found out… We were like, that's what we're going to do" (Participant 118; Kindergarten, 1-2; 3-5, interview).

Regarding the awareness of new policy, in response to a survey item about whether teachers had enough support when administering the tests, another respondent wrote: "It was better this year when classroom teachers were educated on how the WIDA helps their student growth scores. They willingly let us take their students when we needed them" (Participant 241; Kindergarten, 1 – 2, 3 – 5). Regarding the impacts of the test, the same participant wrote, "In our district, students want to do well so they can get a pass on the Reading SOL if their score is too low." When asked in a follow-up interview to elaborate on the impacts of this policy change, they said: "I think that helped motivate our students. What happened last year was we got a new director of ESL. And she was wonderful, and she really encouraged us to do professional development with our gen. ed. teachers, and most of 'em really don't know that much. So we actually did two different professional development sessions with them and it, they finally understood why we needed to do this, and how it could benefit them. Um, I think, being allowed to use that pass for the SOLs is good for the students, because sometimes, those SOLs, sometimes, they just don't have the background to be able to answer a lot of those questions on SOLs. But with WIDA, it's more realistic… "When asked if it was an individual from central office who got the word out to everybody, they clarified, "I mean, she really really

encouraged us. Here's the problem. They moved her. She's gone. We have NO director this year. So we're, you know. (Laughter). We're sunk. Somebody needs to watch out for us, so…" (Participant 241; Kindergarten, 1 – 2, 3 – 5, interview).

Table 32

Emergent Themes: Research Question 3

Changes		*n*
Aware	Yes	14
	No	58
	Not sure	22
Articulates changes		118
SOL proxy		10
Accreditation		7
Misstates changes		12
Administrator made others aware		2

Methodological Considerations

Finally, written comments brought into light two concerns regarding the survey design. First, some participants declined to respond to certain survey items, particularly items addressing the four skill area subtests, stating were not able to answer because they couldn't see the content of the test (*n* = 20) or weren't allowed to look at it (*n* = 3); "How can I know if I am not supposed to be looking at the test?"(Participant 59; Kindergarten, 1 – 2, 3 - 5). Most of the participants who explicitly declined to comment on some items still provided their perspectives on the test: "I did not look closely at the test because I am not supposed to. The only one that I have a strong opinion about is the speaking test" (Participant 177; 9 - 12); "Again, I don't look at the test. However, I was surprised by the length of this test. It was quite extensive for such little people (1st and 2nd graders) who are not accustomed to such long tests" (Participant 59; Kindergarten, 1 – 2, 3 - 5).

A second methodological concern revealed by qualitative data was that one of the survey items was ambiguous. Participants interpreted the belief statement "My school(s) received all ACCESS materials on time" to be referring to two different issues: (a) the on-time delivery of test materials prior to administration (n = 14): "We had to email a couple of times regarding tickets due to new arrivals, but everything else was delivered on time" (Participant 18; 9 - 12), or alternately, as (b) the timely receipt of test scores after testing (n = 60): "Test scores are delayed. It takes way too long to get test results" (Participant 132; Kindergarten, 1 – 2, 3 - 5). This ambiguity in part led to the decision not to rely on Factor 6 to inform further analysis, since this was one of two apparently unrelated items that loaded onto Factor 6. Table 33 reports numbers of comments suggesting these methodological issues.

Table 33

Methodological Considerations

Issue	*n*
Decline to comment	
Don't know/can't see	20
Not allowed	3
Timely receipt interpretation	
Materials	14
Results/Scores	60

Summary

Context. This mixed-methods study on teachers' perceptions of the validity of the ACCESS test was conducted in Virginia beginning one month after the close of the state's ELP testing window in the 2018 - 2019 school year, the first year of full ESSA policy implementation. Surveys were sent to 2,234 potential participants in

the 25 Virginia school districts with the largest EL enrollments that published educator email addresses. Responses were received from 273 educators in schools throughout most of the state, nearly all ESL teachers. Follow-up telephone interviews were conducted with seven of these respondents.

Findings. Results of the study provide answers for the three research questions. Research Question 1 asked: What are Virginia educators' perspectives on the ACCESS for ELLs ®? On average, teachers have mostly favorable to slightly unfavorable perceptions of the test. Many believe some sort of ELP test is needed for accountability and decision-making, and ACCESS serves these purposes. Others say the test is unnecessary, that other measures already in place would be sufficient. Educators point out many issues with the test, including problems with test design, administration, and impacts.

Issues with test design include the length, difficulty, and developmental appropriateness of the test. In addition, results suggest many potential threats to the construct relevant validity of the test, including academic content knowledge, poor alignment with classroom and grade-level expectations, cultural bias, and the interference of language domains other than the construct of each subtest (i.e., the reading necessary to complete the listening and writing subtests). The online version of the test in particular is fraught with potential sources of construct-irrelevant variance, including keyboarding fluency, computer literacy, and the ability to operate the technology to record oneself speaking. Problems with the administration of the test relate to support or lack thereof for testing in schools. Testing is seen as more successful where teachers feel supported by administrators,

testing coordinators, and colleagues. The adequacy of space and equipment is another factor in smooth test administration. When there is not sufficient quiet space for testing, or when there is not enough equipment or equipment does not work well, test administration is compromised. The electronic administration of ACCESS 2.0 poses specific challenges, including interrupted test sessions, and the ability of students to type their writing and record their speaking responses. Similar to the Winke (2011) study, data reveal unintended impacts of the test, including the loss of instructional time and negative emotional consequences for students and teachers. The delivery of test scores several months after test administration is a threat to the validity of decisions made based on these "old" scores. Teachers also call into question the validity of high-stakes decisions such as reclassification, programming, support, and teacher evaluation based on a single score from a test that is imperfect, and many teachers call for multiple measures to be used in informing these decisions.

Research Question 2 asked: Do educators' perspectives vary according to demographics or teaching environment in which the test was administered? The null hypothesis is that educators' perspectives do not vary according to demographics or contexts. One-way ANOVAs or independent samples *t*-tests were run for each of the demographic and contextual factors. Statistically significant quantitative results, which are elaborated on by qualitative results, suggest that educator perspectives do vary according to some of these characteristics. Specifically, teachers who are non-native speakers of English view the test more favorably than native speakers; teachers who work in high-poverty schools and

schools with threats to accreditation, as well as teachers who are evaluated based on ACCESS scores, perceive more threats to effective test administration; elementary school teachers view the reading subtest less favorably than high school teachers; and elementary and secondary teachers view their students' motivation to do well on the test as based on different factors, with secondary students seen as less motivated. Thus, the null hypothesis, which stated that educators' perspectives do not vary according to demographics or teaching environments, was rejected.

Research Question 3 asked: According to educators, what (if any) are the effects of recent testing policy changes on validity considerations for the ACCESS for ELLs ® in Virginia? Virginia officially adopted ACCESS 2.0 in 2016 (Staples, 2017), and since that time, schools and divisions have had the option to administer the paper and/or online versions of the assessment. While a few educators believe online test administration is more reliable and efficient, many point out threats to validity related to the online test, threats to concurrent validity between the two test forms, problems with the online platform, and issues of construct irrelevant variance. Furthermore, the study suggests the recent recalibration of the cut score required for reclassification is controversial and not universally understood. A small number of participants indicated that because ELP scores have more impact on school accreditation under new policy, both English proficiency and ACCESS are being taken more seriously. However, there is evidence that in spite of policy changes, ELs and ACCESS testing are not being given more consideration. In addition, there is evidence that teachers and administrators are not fully aware of the changes in policy.

CHAPTER 5. DISCUSSION

Introduction

The study demonstrates that ESL teachers have a lot to say about the ACCESS test and the validity of the decisions it informs, and thus yielded a rich and robust data set. The perceptions of these professionals, who have first-hand contact with the administration of the test, whose professional lives are guided and constrained by decisions based on its scores, and who know their EL students well through daily interactions with them, provide important evidence regarding the validity of the test. Therefore, the findings of the study have important implications for scholarship, policy, and practice.

Implications for Scholarship

Revising the Broad Validity Framework. Findings of the study suggest that two modifications to Winke's (2011) Broad Validity Framework are warranted: (a) that it be expanded to include construct validity, and (b) that both qualitative and quantitative methods be used for obtaining validity evidence at all levels of the framework. Winke's Framework is illustrated in Figure 1, Chapter 2. Proposed revisions to the Broad Validity Framework are illustrated in Figure 4.

Construct validity. Construct validity can be defined as "the degree to which a test measures what it claims, or purports to be measuring" (Brown, 1996, p. 231). Messick (1998) wrote, "All validity is of one kind, namely construct validity. Other so-called separate types of validity—whether labeled content validity, criterion-related validity, consequential validity, or whatever—cannot stand alone in validity arguments" (p. 37).

Construct-irrelevant variance threatens construct validity. The *Standards* (AERA et al., 2014) define construct-irrelevant variance as "variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation" (p. 217). Messick (1989) wrote, "Tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both" (p. 34). He suggested a primary focus of test validation endeavors focus should be identifying sources of construct-irrelevant variance, in addition to determining whether the construct is underrepresented; most threats to the validity of test scores can be classified into one of these two general areas.

Thus, a key component to the validation of any test is evaluating construct relevance. Is each item on the test relevant to the construct tested? Are any mediating factors inhibiting the ability to test the construct? This study revealed a wide variety of potential threats to the construct validity of the ACCESS test. Study results suggest test item responses require content knowledge, cultural background, and cognitive tasks that are above grade level. Completion of test items requires attention, memory, stamina, computer literacy, keyboarding skills, and the ability to operate recording equipment while speaking into a microphone. Furthermore, subtest items require irrelevant language sub-skills, such as the reading skills necessary for understanding answer choices on the listening test.

Messick (1989) wrote that not only is defining what is being measured is the first step to developing a test, but "defining the construct tested is also the first step

in providing validity evidence based on test content because the definition of the construct has a direct impact on score interpretation and use" (p. 65). The test construct of post-NCLB ELP tests, "academic language," is particularly ill defined. What constitutes academic language is a topic of debate, and content language has not been operationalized for test development (Boals et al., 2015; Rivera, 1984; Valdés, 2004; Wolf et al., 2008). Even if we do not have a clear idea of what the test construct is, we can at least discern what the construct is not. Clearly, evaluating typing speed or assessing the ability to produce extemporaneous monologues on unfamiliar topics is not the intended purpose of a test of proficiency in academic English. While construct validity is central to any validity argument, a consideration of construct relevance is particularly warranted when evaluating the validity of post-NCLB ELP tests. Because of the importance of construct validity in test validation, particularly in the evaluation of government-mandated ELP tests with poorly delineated constructs, this study proposed to add construct validity to Winke's (2011) Broad Validity Framework. This addition is displayed in Figure 4.

Mixed-methods to evaluate the broad validity of tests. Cresswell and Plano-Clark (2018) wrote that gaps exist in past research because the exclusive used of quantitative or qualitative approaches only provided a partial view. "There is a need for a more complete understanding through comparing and synthesizing both quantitative and qualitative data" (p. 151). In contrast, Winke (2011) wrote, "Reliability, concurrent validity, and predictive validity can all be measured quantitatively. However, these purely statistical conceptions of validity are rather narrow" (p. 632), and argued that consequential validity, which can be evaluated

using qualitative methods, is an essential component of a broad validity argument. Winke's Broad Validity Framework specifies that while quantitative methods are appropriate for evaluating the reliability, concurrent validity, and predictive validity of tests, qualitative methods are appropriate for exploring tests' consequences, or the consequential validity of tests (see Figure 1, Chapter 2). The findings of this study, however, demonstrate the usefulness of mixed-methods approaches for investigating reliability and validity at all levels of the Broad Validity Framework, in order to achieve a more complete understanding.

For example, reliability refers to the consistency of scores across multiple administrations of a test, regardless of the examiner, time of testing, or setting. While quantitative statistical techniques are an accepted approach to evaluate the reliability of test items, the mixed-methods technique used here, with Likert-type scale survey items followed by open-ended text boxes and interviews, were useful in revealing potential threats to the reliability of the test, such as the range of settings and conditions under which the test was administered (in a quiet or noisy environment, with or without interruptions, etc.). Theoretical analyses, item-level analyses using Rasch methods, and other statistical techniques used by test developers to validate their products (Porter & Vega, 2017) do not have the capacity on their own to reveal such threats to reliability.

Concurrent validity refers to the consistency of scores between tests that are intended to measure the same construct. Elaborative comments on belief-statement items on the survey revealed that scores for students who took the paper and online versions were not comparable; in particular, the constructed-response speaking and

writing subtest scores were lower for students who took the online form of the test, and participants noticed a dip in scores the first year their schools switched from the paper to the online test. While there was some suggestion that these scoring differences may be due to issues of inter-rater reliability in the local scoring of the speaking subtest, more teachers attributed this threat to concurrent validity to the delivery format of the test itself.

The mixed-methods approach of the study also uncovered valuable information about the predictive validity of the test, or evidence that a test score correlates with or predicts future performance. On average, study participants somewhat disagreed with the quantitative survey item: "A score of 4.4 or "proficient" on ACCESS is a good indicator that a student will be successful in school." Qualitative elaborative comments indicated that for some students, attaining the 4.4 cut score does not correlate with academic readiness, and fails to predict continuing linguistic struggles and need for ELD supports; conversely, for other students, falling short of the 4.4 cut score does not always reflect their classroom success and passing scores on grade level standardized exams. Furthermore, some high school students are able to graduate, having successfully passed all of the required coursework and end-of-course exams, but without ever graduating out of EL status because they have failed to score 4.4 or higher on ACCESS.

Thus, while Winke's (2011) Broad Validity Framework (see Figure 1, Chapter 2) proposes the appropriateness of qualitative methods for studying the consequences of tests, the current study demonstrates that not only are qualitative

but also quantitative and mixed-methods approaches useful for studying the consequential validity of the test. For example, scores on belief statement items show the views of educators who identified as non-native speakers of English were more favorable regarding the benefits and impacts on students' English language ability than those of native-speaker teachers; teachers confronting the challenges of high levels of economic disadvantage, conditional accreditation status, and questionable methodology for the evaluation of their professional performance also perceived more challenges to effective test administration in their schools. Furthermore, some quantitative survey items appear to have functioned as prompts for elaborative comments that converged on themes the survey items had not directly addressed. For example, in response to the item "The students liked the ACCESS test," several participants wrote about the experiences of LTELs and the stigma of EL classification. Winke's mixed-method approach in her own study also proved useful for revealing the unintended consequences of the ELPA test.

Because findings of the study demonstrated that both quantitative and qualitative research methods were useful for examining the validity of the test at all levels of the Broad Validity Framework, I propose the framework be modified to include quantitative and qualitative approaches at all levels (see Figure 4).

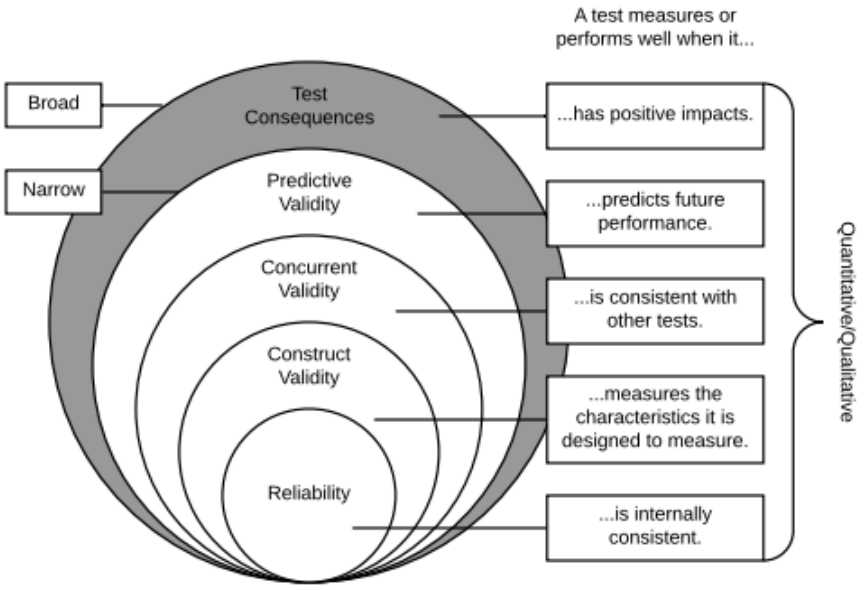

Figure 4. Revised Broad Validity Framework

Practitioner input in validity evaluations. This study affirms Winke's (2011) finding that the perspectives of teachers are indispensable for validating state-mandated tests, and that school-based personnel can identify the unintended consequences of these tests on teaching and learning, information that cannot be examined from just looking at scores. Regarding her survey, Winke (2011) wrote: "The results of this study thus provide evidence that surveying the perspectives of educators is an important way to evaluate the broad validity of a test" (p. 651). The educators who contributed to the current study also provided useful validity data, demonstrating that that asking practitioners about a test that has such a profound impact on their day-to-day professional practice is an important and valuable way to evaluate the broad validity of the test, at all levels of the framework.

Implications for Policy

Validation of the test by states that use it. The *Standards* (AERA et al., 2014) state, "Validation is the joint responsibility of the test developer and the test user... The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used" (p. 13). While validation studies of the ACCESS test have been conducted by researchers in WIDA-affiliated institutions, no independent evaluations have been conducted. Messick (1980) wrote that responsibility for assessment of the consequential value of a test "goes beyond the test maker to include as well the decisionmaker, policymaker, and test user, who are responsible for specific evidence of instrumental value in their particular setting and for the specific interpretations and uses made of the test scores" (p. 1025). Such an evaluation of ACCESS by test users, i.e., state departments of education, is warranted to (a) provide an independent validity assessment not tied to the interests of the test publisher and vendor, and (b) to assess the appropriateness of the test and inferences made based on test scores in individual state contexts.

Furthermore, Winke (2011) suggests that future evaluations of mandated ELP tests include the input of educators who administer them and be undertaken by neutral outside evaluators, since "states… and for-profit agencies often have an incentive to avoid criticizing the tests they manage" (p. 651). Perhaps a group of states could pool resources to commission such an independent evaluation of the ACCESS test, as they did in the early days of NCLB when they formed consortia to develop ELD standards and tests. Winke suggests that such an evaluation study would summarize and present results to the public, and that such a transparent

process "may also increase trust in the states (i.e., the organizations responsible for administration) and any corporate testing agencies they have hired and may encourage discussions about the test and the inferences that can be drawn from it" (p. 651). Findings of the current study suggest that not only user-conducted validation of the test warranted since it has never been independently validated and questions about validity remain, but since teachers lack confidence in the test and the decisions made based on test scores, a transparent process that included teacher input would have the potential to improve assessment and the practice it informs.

Alignment with Virginia curriculum. While there is limited evidence of alignment between an earlier generation of the ACCESS test and a previous version of the CCSS (Chi et. al., 2011; Bailey & Wolf, 2012), assessment personnel at WIDA and at the VDOE confirm alignment between the Virginia SOLs and WIDA ELD standards and assessments has not been studied (K. Bach, personal communication, March 2, 2020; K. Johnson, personal communication, March 2, 2020). This is problematic because if it is possible or true that ELP tests are not assessments of academic content, and prior mastery of academic content-related knowledge is not required to successfully answer test questions, as some ELP testing scholars have suggested (Fast et al., 2004), what constitutes academic content language is poorly operationalized for testing (Boals et al., 2015; Rivera, 1984; Valdés, 2004; Wolf et al., 2008), there is evidence that prior content knowledge may be a threat to the content relevant validity of the test.

Findings of the study suggest that the ACCESS test may contain academic content that is not aligned with the grade level content or the developmental levels

of students as reflected in the Virginia curriculum. Since it is unfair to test the achievement of students based on content to which they have not been exposed, and since this practice would be likely to produce an inaccurate assessment of language proficiency, it is important to assure that the WIDA ELD standards and tests are aligned with the Virginia curriculum, and that students have been afforded the opportunity to learn the content on which they are tested. Thus, an examination of the alignment between ACCESS and the SOLs is warranted, and any mismatches between the ACCESS test and the state curriculum should be addressed.

Alignment of progress and proficiency targets. The study found that, on average, Virginia educators slightly disagreed with the statement, "A score of 4.4 or "proficient" on ACCESS is a good indicator that a student will be successful in school." Some indicated they thought the cut score was too low, causing students to lose linguistic supports before they were ready; others thought the cut score was too high, holding students who were successful on grade-level classwork and tests back in EL classification for too long. Still others believed the appropriateness of the cut score depended on the individual student.

When the VDOE lowered Virginia's reclassification score from 5.0 to 4.4 in 2016, the Superintendent of Public Instruction explained that when WIDA imposed a more rigorous scoring scale, the VDOE conducted a comparison of actual 2015-2016 assessment scores using the old and new scales. Students who had overall and literacy proficiency levels of 5.0 on the old ACCESS scale scored from 3.8 to 4.4 on the new scale, so Virginia decided to use an overall composite proficiency score of

4.4 as the new criterion for proficiency and reclassification[15] (Staples, 2017). In an apparent reconsideration of the NCLB approach of imposing arbitrary numbers as achievement goals, some achievement targets under ESSA are derived from actual scores and pass rates in a base year, using what Abedi and Deitel (2004) referred to as an "existence proof" that the requirements are actually attainable. WIDA's 5.0 cut score, however, originated in the NCLB approach to setting theoretical, incremental score requirements. Under NCLB, ELs were required to achieve one full point of growth or "progress" on ACCESS each year for five years, and to reach a "proficiency" score of 5.0 or higher in five years, despite several research studies showing that a minimum of five to seven years are needed before ELs are sufficiently proficient to benefit from English-only instruction at the same level as their native English-speaking peers (Abedi & Gándara, 2006). Virginia's 4.4 cut score, which was determined based on its equivalency to the old 5.0 score, is a vestige of now-abandoned NCLB accountability methodology. Virginia's ESSA plan, on the other hand, used newer methodology to base growth targets on actual scores in baseline years. As a result, under Virginia's current policy, it is possible for a student to meet annual year-to-year growth target requirements each year for five years, but to still fall far short of the five-year proficiency goal, which requires a 4.4 score. Figure 5, *Composite Proficiency Level Gains*, shows Virginia's annual ELP growth requirements by proficiency range and grade-level cluster. Students are

[15] A Title III document also found on the VDOE website provides a less plausible, conflicting account of this process: "ACCESS for ELLs 2.0 data were used to identify the composite score that was most consistent with passing or failing the state reading assessment…This methodology indicated a composite score of 4.4, which was selected as Virginia's ACCESS for ELLs 2.0 exit criterion." (VDOE, n.d.)

expected to meet or exceed a score of 4.4 in order to reach proficiency in five years. However, based on the targets indicated in the table, it is conceivable that a newcomer starting in Kindergarten making adequate annual gains might only reach a score of 4.0 after five years, a 3rd grade newcomer could score as little as 3.8, and a newcomer starting in 6th grade would only attain a score at 2.8 in the five year timeframe required for reclassification, resulting in the failure of each of these hypothetical students who had met the growth targets to meet proficiency requirements on time.

Composite Proficiency Level Gains

Proficiency Level	Grades K-2	Grades 3-5	Grades 6-12
1.0-2.4	1.0	0.7	0.4
2.5-3.4	0.4	0.4	0.2
3.5-4.4	0.2	0.2	0.1

Figure 5: Composite proficiency level gains. From "Virginia Compliance with Title III Requirements: Purpose of Program and General Uses of Funds of Title III."VDOE, n.d. Retrieved from http://www.doe.virginia.gov/federal_programs/esea/title3/compliance/virginia-compliance-with-title-iii-requirements.dox

Students with disabilities. Study participants pointed out several issues concerning the use of ACCESS in decision-making for dually identified students, or students classified as both ELs and students with disabilities. One such issue is that it is not possible to demonstrate an advanced proficiency level or exit EL classification based on Alternate ACCESS scores, so not only can students with profound disabilities who take this test form never reclassify, but they also must

take the test year after year, regardless of whether their disabilities make them unable to complete some portions of the test. In addition, in some cases, monolingual English-speaking students with disabilities are misidentified as ELs when a checkbox on the home language survey portion of school registration forms flags another language spoken by someone in the home. The student is then misclassified as an EL, and then is unable to test out of classification, due to their disability, not their non-native speaker status. Furthermore, IEP accommodations afforded to students with disabilities in the classroom and on other tests are not permitted for ACCESS testing, making demonstration of progress and proficiency unattainable for dually identified students. It is conceivable that a student with disabilities could be linguistically prepared for classroom success when IEP accommodations in place, but could not demonstrate this readiness based on ACCESS scores when these accommodations are unavailable for testing. Thus, findings of this study suggest that a more holistic evaluation of the English proficiency of students with disabilities is warranted, and that this evaluation should take into consideration the abilities and disabilities of the student. (For example, a student who is not verbal in any language should not be required to demonstrate progress and proficiency in speaking in English.) Furthermore, policymakers should reassess whether and which IEP accommodations might be permitted on the ACCESS test.

The 4.4 composite score for reclassification. Virginia has established a single composite cut score of 4.4 on the ACCESS test as its sole criterion for EL reclassification (Staples, 2017). Findings of this study suggest that this cut score

alone is not always an accurate indicator that a student no longer needs linguistic supports and/or is prepared for academic success in an English-only instructional setting. Study findings suggest that multiple measures should be use for reclassification decisions, and could include teacher and parent input, classroom performance, and other test scores (i.e., SOL, DRA, PALs, and MAP scores) in the decision-making process. Other states, including New York and California, already take into account multiple factors and the perspectives of multiple interested stakeholders in reclassification decisions (CDOE, 2017; NYED, 2015).

In 2016, Virginia switched from requiring both an overall composite ACCESS score and a literacy score that combined reading and writing subtest results to requiring a single overall cut score for EL reclassification. The use of a single compensatory score as a basis for high stakes decision-making raise questions about the validity of these decisions, since scoring higher in one domain can skew a composite score. A conjunctive score, in which students show proficiency in each of the four domains, is preferable (Abedi, 2013). The study found that in some instances, overall scores could be inflated by a high score in a single domain, which may give a false impression of proficiency in other areas. For example, high scores in speaking or listening may obscure low levels of literacy. This sometimes results in linguistic supports and accommodations being removed before an individual EL is ready. Therefore, VDOE should not only reconsider its single 4.4 composite score decision rule which it has set as the only criterion for EL reclassification, but also promote the use of multiple measures and indicators in high-stakes decisions like reclassification.

Local decision-making. Some schools take a holistic approach to local decision-making for ELs, using multiple measures, including ACCESS scores, other reading and writing assessments, SOL scores, grades, classroom performance, and teacher and parent input, to inform decisions such as student scheduling, levels of service, and assignment to small groups. In other schools, including those under Department of Justice guidance, such local decisions are made solely based on ACCESS scores. For reasons mentioned above, as well as because of the untimely receipt of test scores, indications that the test is not always an accurate measure of progress and proficiency for all individual students, and because these scores are just a single measure or "snapshot" as several teachers put it, study findings suggest that schools should consider multiple measures and exercise some flexibility when making local decisions regarding ELs.

Electronic test administration. While most teachers in the study administered the online version of the ACCESS test, many administered the paper test, Alternate ACCESS, or combinations of the three test forms. The study found a range of problems with electronic test: scheduling issues due to insufficient computer lab space, equipment, and server capacity; interrupted test sessions as students get kicked out and have to log back in; difficulty operating computer-based features such as drag and drop items; difficulty typing timed writing test items; difficulty recording oneself speaking; the "unnatural" quality of speaking into a microphone to a computer; and rapid guessing as students "click through" multiple choice items that they are not permitted to skip. Furthermore, findings suggest teachers believe students who take the online test score lower than students taking

the paper test (especially in speaking and writing), that schools that adopted the online test saw a dip in score gains the first year of implementation, and that test preparation activities necessary to familiarize students with the electronic format result in lost instructional time.

When ELPA21, a consortium-based ELD standards and test developer that publishes a test currently used in several states for EL accountability, developed its electronic testing platform several years ago, feedback on the new format was mixed, and concerns were expressed about the level of technological skills needed by ELs to take the test. Specifically, some of the "task types may not be familiar to students, and newly arrived ELLs might not have the necessary keyboarding and mousing skills to access the assessment" (ELPA21, 2015, p. 18). Based on this feedback, ELPA21 revised the test and support platform before launching its electronic test in the 2015-2016 school year. It is not clear that WIDA has made similar adjustments (WIDA, 2017a). The current study suggests that attention to concerns about the online test is warranted. In addition, it suggests that although electronic administration may be slightly more time-efficient, more schools and divisions might consider exercising their right to opt for the paper under Virginia policy, especially for newcomers and for students who don't have strong computer literacy or typing skills.

Policy dissemination. Presumably, the intention of education policy innovation is to improve conditions for students and schools. The stated intent of ESEA was to improve educational opportunity for poor and minority students through Title I funding (Orfield, 2016). The Bilingual Education Act of 1968 was

designed to provide access to academic content and to the English language for ELs (Menken, 2010; Stewner-Manzanares, 1988). NCLB was supposed to provide educational access and opportunity to all students by holding states, school districts, and individual schools accountable for the academic achievement of students in disadvantaged subgroups, including ELs, through high-stakes testing, until 100% of students were deemed proficient (Abedi, 2004; Wolf et al., 2008).

While ESSA continues the opportunity-through-accountability mandate of NCLB, its intention seems to have been to roll back some of the excesses of the previous legislation and to satisfy the political interests of a variety of interest groups (DeBray & Blankenship, 2016). In Virginia's ESSA plan (VDOE, 2018), year-to-year growth on ACCESS can be reported as growth in grade-level reading for some ELs. In 2018-2019, the year of the study, ACCESS test scores were taken into consideration in school-level accreditation decisions, and ELP progress and proficiency rates were reported alongside SOL scores at the school building level for the first time, in accordance with Title I.

The study seems to suggest that when teachers and school administrators are aware of new policy, that policy is more likely to have an impact than when they don't know about it. Some of the participants said that awareness of changes in the impact of ACCESS on a school's reading SOL pass rates and accreditation had resulted in more credibility for their professionalism and work as ESL teachers, focused more attention on their students' language development in all four language domains, and inspired support for better ACCESS testing conditions in their buildings. A few teachers told how a specific individual, such as a principal or

central office ESL specialist, had been instrumental in disseminating the new policy to stakeholders. Others indicated that although they had hoped that since now that ACCESS is now counted more like an SOL test, ACCESS test administration would be supported in the same way as SOLs, but that they had been disappointed. One interview participant described advocating for more support on the basis of the ACCESS test's new impacts on accreditation, but being dismissed, ironically, because of her school's preoccupation with SOL since they were in danger of losing accreditation, a threat that demonstrating growth on ACCESS might have lessened.

On the other hand, many of the ESL teachers in the study said they were not aware of new policy regarding ELs, and several indicated they were aware that policy had changed somehow but didn't know what the changes were. Others who said they were aware of policy changes went on to misstate the details of these changes. Still others said they felt confused, in part because the federal education policy changes came not long after the launch of ACCESS 2.0, followed by WIDA's rescaling of test scores, and Virginia's recalibration of the cut score, and that it was hard to sort out so many changes that happened in such rapid succession. In a few cases, it appeared that lack of awareness of the policy could result in failure to comply with new regulations, such as failure to provide testing accommodations to reclassified ELs in monitor status. This apparent lack of clarity about the new policy, or clarity only because an interested individual had taken it upon themself to "get the word out," seems to point to poor communication and dissemination of the policy. When stakeholders are not aware of the new policy or its intended effects, that policy is less likely to produce its intended effects.

Implications for Practice

Effective test administration. The study found that while test administration proceeded relatively smoothly in some schools, support for ACCESS testing varied from school to school, and ELs in some locations were not provided with adequate testing environments and conditions. Teachers who reported effective testing administration often attributed this success to the support of their colleagues, including principals, specialists, counselors, assessment coordinators, and other teachers. On the other hand, teachers who struggled to administer the test effectively indicated that they needed more support, and that ACCESS testing was not afforded the same consideration as the SOLs or even as lower-stakes tests like MAP. Challenges to effective test administration stemmed from a lack of adequate training for some teachers, who said that despite receiving the training, they had learned "by doing" or from more experienced colleagues. More often, necessary resources such as adequate space, equipment, and Internet/server capacity were limited. The timing of the testing window may have been a factor in effective administration, since ACCESS testing conflicts with lower-stakes assessments and a few mid-year SOLs, but doesn't take place during end-of-year testing, during which schools are more focused on ensuring proper testing conditions. Since ACCESS appears to present particular administration challenges, it is necessary to provide support for test administrators and ensure adequate testing environments and conditions in all schools, to enhance the reliability and validity of the test.

Teacher evaluation based on ACCESS scores. The study found that a considerable proportion of ESL teachers in Virginia are evaluated on the basis of

ACCESS test scores (28.3% of the participants, working in at least 8 Virginia school divisions). WIDA has published guidance recommending that scores of their tests not be used for teacher evaluation (WIDA, 2015), explaining that the caseloads of most teachers constitute small sample sizes that would result in high imprecision, and attribution bias would render spurious the attribution of a student's growth to a single teacher in the presence of potentially confounding factors outside of the teacher's control, such as prior education, home language proficiency, and the influence of other teachers. Thus, "WIDA recommends that educational agencies NOT make high-stakes decisions using growth models based on ACCESS test scores unless the issue of sample size and attribution has been overcome." (WIDA, 2015, p. 2). The *Standards* (AERA et al., 2014) state that for a "use of test scores that differs from those supported by the test developer, the responsibility for providing validity evidence in support of that interpretation for the specified use is the responsibility of the user" (p. 13).

VDOE guidelines for teacher evaluation (VDOE, 2020) say evidence should be provided to demonstrate that work of the teacher results in acceptable, measurable, and appropriate student academic progress. These guidelines also specify "other measures are recommended for use when two valid and direct measures of student academic progress are not available" (p. 43). Until evidence is provided to confirm the validity of interpreting ACCESS scores for teacher evaluation and to refute WIDA's argument against this use of scores, schools and divisions should comply with VDOE policy and WIDA guidance, and stop the practice of using ACCESS scores to evaluate teachers.

Recommendations for Future Research

Could native speakers pass the ACCESS test? Several study participants suggested future research should compare the performance of ELs and grade level native English-speaking peers on the ACCESS test, which some teachers believed native speakers couldn't pass. Carroll and Bailey (2016) write that given the premise that "a proficient level of English-language proficiency can predict success in an English-only instructional setting, it would reasonably follow that non-ELL students who are currently receiving instruction in English-only settings could be used as a "known-to-be-proficient" comparison group" (p. 32). While there has been no recent research comparing the performance of ELs and non-ELs on ELP tests, Stephenson, Jiao, and Wall (2004) conducted such a study of the SELP test (a pre-NCLB ELP test). They used ANOVA to support a claim of the validity of EL proficiency of classifications, based on evidence that ELs score lower on average than non-ELs. However, discriminant analysis by grade-level group membership in the same study indicated SELP test scores classified considerable proportions of non-ELs as "non-proficient" (by grade level cluster: primary, 36%; elementary, 28%; middle, 13%; and high, 17%). Carroll and Bailey (2016) note that it is unfortunate that these findings were not accompanied with recommendations on how ELP classifications should be interpreted when making decisions for individual EL students.

Assessment policy requires students identified as ELs to take ELP tests and demonstrate progress and proficiency in academic language in four language domains. Students not identified as ELs are not subject to this accountability

mandate, and their proficiency in academic English is not scrutinized at this level. While it might be argued (and counter-argued) that all students are indirectly held accountable for academic language proficiency by grade-level reading and writing tests, ELs are the only student subgroup required to demonstrate proficiency in academic language proficiency on speaking and listening tests. Because EL classification can result in relegation to lower academic tracks and limited opportunities for ELs (Callahan, 2005; Solórzano, 2008), such a study would have implications for equity and opportunity for ELs and all students.

Consequences of achievement motivation. Another recommendation for future research is the investigation of the impact on student achievement motivation on ACCESS scores; put differently, what, if any, is the mediating influence of student test-taking effort on ratings of English language progress and proficiency based on ACCESS test scores. A related line of research could investigate possible correlations between trajectories of year-to year growth as measured by ACCESS and models of the expected influence of effort or guessing on test scores.

The study found that students do not always put forth their best effort when taking the ACCESS test: teachers reported that students rush, guess, and click answers at random. One theory that addresses students' motivation to achieve in education is Expectancy Value Theory, which directly links achievement performance to individuals' expectancy-related and task-value beliefs. Expectancy-related beliefs refer to an individual's beliefs on how well they will do on an upcoming task. Task-value beliefs are defined by four components: (a) attainment value, or the personal importance of doing well on a task, (b) intrinsic value, or the

enjoyment the individual derives from performing the task, (c) utility value, or how well the task relates to the individual's goals, and (d) cost, or the negative aspects of engaging in the task (Eccles & Wingfield, 2002). Students' low motivation and the conditions of testing may influence their effort in responding to test items (Zerpa et al., 2011).

The study results suggest that many ACCESS test-takers have low levels of achievement motivation: they don't see the benefits of the ACCESS test; they dislike testing in general and ACCESS in particular; they don't care about the test because it doesn't "count" like an SOL (especially for high school students, who must pass SOL tests to graduate); and they see the test as long, hard, boring and frustrating. Schmitt, Chan, Sacco, McFarland, and Jennings (1999) found that low-effort testing behaviors such as guessing or rushing affect the validity of test results, and can either artificially inflate or deflate estimates of students' proficiency. More specifically, the "lucky guesses" of a student who is not proficient could result in an inflated score that is an overestimation of ability. Conversely, a lack of effort on the test such as rushing and "clicking through" by a student who is actually proficient could result in deflated scores and an underestimation of that student's abilities (Zerpa et al., 2011). In the context of ELP testing, guessing and other low-effort behaviors could result in what Carroll and Baker (2016) refer to as false positive and false negative test results. False positives can result in the promotion and misclassification of non-proficient ELs as proficient, potentially resulting in placement in mainstream settings without appropriate language supports for lower proficiency ELs. False negatives can result in a misclassification as non-proficient for ELs who are

actually proficient, and could result in one or more years with an EL designation in settings below the student's actual linguistic competence. Findings from the current study suggest low-effort testing behaviors resulted the misclassification of ELs; furthermore, lack of achievement motivation and low-effort testing behavior could result in failure of proficient students to attain the 4.4 score necessary to reclassify. If this is true, low achievement motivation could exacerbate the problem of LTELs, defined as students who do not reclassify as proficient after six years (WIDA, 2019). The *Standards* (AERA et al. 2014) state, "Test scores used in psychological assessment ideally are interpreted in light of a number of factors, including … indicators of effort" (p.154). Thus, research on the mediating effects of achievement motivation and test-taker effort on ACCESS scores is warranted. This research might examine possible correlations between the trajectory of growth as measured by ACCESS and models of the influence of guessing or low-effort testing behaviors on multiple- choice tests.

One such data set on ACCESS growth published on the VDOE website is shown in Figure 5, *Composite proficiency level gains,* since these targets reflect the actual progress rates from 2015-2016 to 2016 -2017, the baseline years used for setting accountability requirements by the VDOE (VDOE, n.d.). It is interesting to note that younger and low proficiency ELs showed much higher average growth on ACCESS than older and higher proficiency ELs, i.e., students in grades K – 2 at proficiency levels 1.0 – 2.4 averaged 1.0 point of growth, while students in grades 6 – 12 averaged growth of only 0.1, or only one tenth the number of points on the scale (see Figure 4). Since the moderating effects of low-effort testing behaviors inflate the scores of low proficiency students and suppress the scores of students

with high proficiency, low effort might also manifest as the higher gains at low proficiency levels, and lower rates at higher proficiency levels. Such a pattern can be seen in these average ACCESS growth scores, suggesting a possible correlation.

Effective test administration in schools with high economic disadvantage. While the study found significantly less favorable views of effective test administration among teachers in schools where 40% or more of the students were eligible for free and reduced lunch, surprisingly, the views of test administration among teachers in Title I schools was not significantly different from the average perceptions of teachers in other contexts. Teachers in high free and reduced lunch eligibility schools had significantly less favorable views of the smooth administration of the test, as well as the adequacy of Internet and server capacity, personnel, support, and training. Because both free and reduced lunch eligibility rates and Title I status are commonly used as indicators of socioeconomic disadvantage (National Center for Education Statistics, 2012), further inquiry exploring these differences of perspective might be warranted. Perhaps Title I status affords more staffing and resources, or maybe the difference is related to practices in elementary schools, more of which are in Title I status than secondary schools (VDOE, 2018). A comparison of practices in both types of economically disadvantaged schools could potentially uncover any such differences, and identify successful practices in the Title I schools for replication in all settings.

Considerations for dually identified students. The question of the construct relevant validity of the ACCESS test has particular implications for students with disabilities. The study suggests that intellectual, emotional, and other

disabilities may be confounding the assessment of progress and proficiency of dually identified students, particularly since IEP accommodations provided in the classroom and on other tests are not permitted for ACCESS. The positive impact of linguistic accommodations on the performance of ELs on high-stakes content tests for ELs has been demonstrated (Abedi & Lord, 2001). Future research is needed to study the effects of the withholding of IEP accommodations for ELP testing on the measurement of progress and proficiency for dually identified students. In addition, research is warranted on the predictive validity of ELP tests administered without accommodations and their correlation to success in a classroom with accommodations for students with disabilities.

Perspectives of non-native English-speaker teachers. The study found that teachers who were non-native speakers of English had significantly more favorable views of the overall benefits of the test and its impacts on students' language ability, as well as the quality of the speaking and writing subtests, than native-speaker educators. It might be interesting to further compare these differences in perspectives. Perhaps non-native speaker teachers, as former ELs themselves, could provide insight into an English learner perspective on ELP assessment and accountability.

Limitations

Winke (2011) noted two main limitations of her study that are also limitations of the current study. First, the participants were self-selected, and thus were not a representative random sample of the population. She suggested that educators with strong opinions, particularly those with strong negative opinions,

might have been more likely to respond to the survey, thus skewing results to the negative. However, she points out that the study had a reasonably large sample (267 responses to the Winke survey; 273 participants in the current study), used a mixed-methods design to examine both quantitative and qualitative data. Together these two design features might mitigate this limitation and allow for a meaningful representation of educators' perspectives when random sampling is not possible, for the original and current studies. Because Winke used snowball sampling she did not report a response rate. Because the population of ESL teachers in Virginia was not known, it was not possible to accurately calculate sample representation.

On a related note, in the interest of efficient recruitment of large numbers of participants, my sampling method limited recruitment to school districts with the largest populations of ELs. Winke (2011) found statistically significant differences in the perspectives of educators in schools with small ESL populations, but the current study did not find variance according to population size, perhaps because districts with the smallest numbers of ELs were excluded from the sample.

A second relevant limitation pointed out by Winke was the study sample was limited to educators, nearly all ESL teachers in the current study. The broad validity of a high-stakes ELP test could be more thoroughly studied by triangulating data on the perspectives of other stakeholders, including students, whose lives are most directly impacted by the test, as well as their parents, school administrators, and policy-makers. A broader and more representative sample could potentially provide more insight into the validity of the test and examine whether different stakeholder groups hold different perceptions.

A third limitation is the setting of the study. Virginia is a unique testing context in several ways, so study findings may not be applicable in other states. Virginia's state ESSA plan is unique, as are the accountability plans of each state. The uses of high-stakes ELP test scores differ from state to state. In addition, the alignment of the ELD standards with academic content in Virginia may be different from Common Core states, since there is little or no evidence that the WIDA ELD standards or ACCESS test are aligned with Virginia's content standards, the SOLs (K. Bach, personal communication, March 2, 2020; J. Costa, March 1, 2020, personal communication; K. Johnson, personal communication, July 26, 2018). Furthermore, when WIDA adopted a more rigorous scoring scale in 2017 to align with increased the rigor in the CCSS, Virginia revised its ACCESS progress and proficiency score scales, and adopted a lower proficiency cut score lower than most states (Staples, 2017). Thus, ACCESS scores in Virginia are reported on a scale that is different from the scale used in other states, and high-stakes decisions like reclassification are made based on different criteria. More generally, test validity evaluations may vary across different contexts. As Winke (2011) states, "broad validity is not a fixed property of a test, but can vary depending on the context in which a test is administered" (p. 653). Taken together, factors specific to Virginia may limit the applicability of study findings to other contexts.

A fourth limitation regards barriers to data collection and analysis stemming from protocols and procedures for WIDA ACCESS administration that limit both what test administrators know about the test and what they are permitted say about it. Most of the participants administered the online test, which restricts teachers'

ability to access and assess test content delivered to students through individual screens and headphones. In addition, to safeguard the security of test items, WIDA requires test administrators to sign a non-disclosure agreement, which obliges them to not disclose any test information except for the sole purpose of coordinating or administering the test (WIDA, n.d.-i). Test security is a consideration for the reliability of a test, and the study's structured survey items did not ask participants to reveal test item content. Indeed, several study participants declined to respond to survey items saying they didn't know what was on the test, couldn't see the content of the test, or weren't allowed to look or comment. While survey participation was anonymous, WIDA's nondisclosure requirement had the potential to suppress what participants revealed or even to discourage participation altogether. A few respondents did disclose information about the test, which was redacted or not reported. As a teacher and test administrator, I have signed this non-disclosure agreement, and this precluded my reporting certain details about the test.

Another limitation related to restrictions on teacher speech was the impediment to data collection and analysis caused by requests for retroactive prior research approval by a two large school divisions. Since permission had not been granted when survey reminders were sent out several days later, these reminders were not emailed to potential participants in the two school districts (n = 890). Even though university IRB approval and a letter of research review committee exemption from another school division in the state were provided, lengthy external research application and approval process in one of the school divisions caused a delay in interview recruitment in that large division, and may have reduced the

number of interviews conducted there. It is conceivable that because six months had passed since testing when interview recruitment was finally allowed, respondents had lost interest in being interviewed. It would have been preferable to conduct interviews sooner, when testing was a more recent memory. This experience raised questions of the authority of a public school district to regulate speech, not only for its teachers, but also for an outside researcher using publicly available contact information. More generally, it raises questions about the free speech rights of school district employees, especially given that the participants and school district would not be identified. In the 1965 Tinker vs. Des Moines Supreme Court decision regarding the free speech of students, Justice Abe Fortas wrote, "It can hardly be argued that either students or teachers shed their constitutional rights to freedom of speech or expression at the schoolhouse gate…" (Tinker v. Des Moines Independent Community School District, n.d.). There is a body of literature on what teachers can and can't say inside the school, but less about what they can't say outside of school, about school. There is also a body of inquiry debating whether human subject regulations cross the line from safeguards to censorship, and whether human subjects protections for biomedical research are appropriate for social sciences research (Feeley, M., 2007; Hottenstein, K. N., 2018; Howard, J, 2006; Stark, L., 2007). It is not entirely clear what the free speech rights of educators are in regard to answering surveys related to their professional practice, or to what extent school districts or publishers can inhibit or regulate this speech of teachers; however, it does seem clear that some school division research offices, in an

abundance of caution, are inhibiting the work of researchers who might provide them with needed information.

A final limitation is researcher bias. As a practicing teacher who has administered the ACCESS test every year since it was adopted in Virginia, my students, colleagues, and I have personally encountered many of the problems with the test discussed in the study. I made an intentional effort to control my biases, by recording researcher memos and conducting an inter-rater reliability check with a disinterested researcher. Survey belief statements were positively worded on purpose. Both positive and negative perspectives were intentionally considered and reported. However, all analysis was conducted through the lens of a teacher-researcher whose professional life is constrained by testing and test results, whose work is evaluated based on these test scores, and who admittedly holds pre-conceived opinions of the test.

Final Thoughts

The teachers who contributed to this study offered many valuable insights into the broad validity of high-stakes ELP testing, specifically the validity of the ACCESS as administered in the current testing context in Virginia. Survey respondents offered many practical suggestions for improving the validity of ACCESS testing, including allowing students to skip an unknown item to prevent lucky guesses from skewing scores, shortening the test so that it could be completed in one seating to mitigate negative effects on student affect and loss of instructional time, giving the test earlier in the year so scores could be received earlier or later in the year during end-of-year testing, making the content of the test more relatable to

students, and differentiation for specific populations like newcomers, LTELs, and students with disabilities.

A few teachers suggested creative, out-of-the-box, practical compromise solutions to some of the challenges of ACCESS. One such compromise would be for teachers to administer and record the online speaking test one-on-one to students, and for the resulting recorded speaking sample to be centrally scored by WIDA. This could make the speaking task more natural and comfortable for students and improve the dependability of the audio recording of responses, while still maintaining the enhanced inter-rater reliability that centralized scoring affords. Another idea would be to allow students who score 4.4 on a particular language domain subtest not to retake that subtest in subsequent years. For example, if a student met the cut score in listening one year, the following year they would only take the three remaining subtests, reading writing, and speaking. This could save time and money (if an a la carte provision were made for purchasing tests), and potentially increase the motivation of LTELs and all students to do their best on the test, as their testing burden was lessened and the possibility success seemed more real. Another suggestion is for WIDA to deliver scores in two installments: (a) scores for the online multiple choice subtests, listening and reading, would be delivered first, as soon as those tests were completed in early spring, and (b) results of the human-rater scored constructed-response subtests, writing and speaking, would be delivered later along with overall composite scores. This could partially mitigate the threat to validity of out-of-date scores and give schools at least some up-to-date data for informing instruction in the current year. While such suggestions might improve

the validity of the test, as one interview participant said, “Well, I’m not sure that any test would be… perfect” (Participant 104; 6 - 8), and threats to the validity of the test would likely always persist.

As I write this final paragraph, the governor of Virginia has just announced public schools will be closed for the remainder of the academic year due to COVID-19 (Northam, R. S., 2020). Currently teachers, schools, and school divisions have been forced to consider alternate ways of teaching, and states will likely be granted waivers for high-stakes testing, including ELP assessments. At my school, we finished ACCESS testing the day before the shutdown was announced in our division, but while testing window was still open in Virginia. It’s unclear whether we will be held accountable based on our ACCESS tests, when we will get scores back, or even if we will receive our scores. For us, EL language instruction and content support was compromised for nearly eight weeks while we administered the test in four sittings each to over 250 ELs (S. Teconchuk, personal communication, 2020, March 17). Given the current situation, I am mourning all the time I lost with my students for the administration of a test that was so stressful for them, not to mention and my colleagues and me. If we do eventually get our scores back, it is not clear whether they will lead us to make appropriate decisions for our students, or what effect they will have on public perceptions and official evaluations of our school. This study suggests that not only are there considerable technical and practical issues with the ACCESS test, but, consistent with the Winke (2011) study it seeks to replicate, demonstrates serious concerns about the consequences of the test. Unless a future round of education policy reform eliminates testing and

accountability approach to equity and opportunity which has not proven to be effective at meeting those goals, these threats to the broad validity of the ACCESS test need to be addressed and to be minimized to the greatest extent possible.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14. https://doi.org/10.3102/0013189X033001004

Abedi, J. (2007). *English language assessment and proficiency in the nation: Current status and future practice.* Davis: University of California. https://education.ucdavis.edu/sites/main/files/ELP_Report.pdf

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*(3), 17-31. https://doi.org/10.1111/j/1745-3992.2008.00125.x

Abedi, J. (2008a). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment, 13,* 193-214. https://doi.org/10.1080/10627190802394404

Abedi, J. (2013). Testing of English language learner students, in K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology, Volume 3: Testing and Assessment in School Psychology and Education* (pp. 255-268). Washington, DC: American Psychological Association.

Abedi, J. & Deitel, R. (2004). Challenges in the No Child Left Behind Act for English language learners. *Phi Delta Kappan Magazine, 85*(10), 782-785. https://doi.org/10.1177/003172170408501015

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36-46.

https://doi.org/10.1111/j.1745-3992.006.00066.x

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234. https://doi.org/10.1207/S15324818AME1403_2

Ahlin, E. M. (2019). Semi-structured interviews with expert practitioners: Their validity and significant contribution to translational research. *SAGE Research Methods Cases.* https://doi.org/10.4135/9781526466037

Alexander, M. M. (2017). Transnational English language learners fighting on an unlevel playing field: high school exit exams, accommodations, and ESL status. *Language Policy, 16*(2), 115-133. https://doi.org/10.1007/a10993-015-9390-y

American Bar Association. (2013). Proceedings from 2013 ABA Annual Meeting, August 8-12, 2013, Lessons in Leadership from the Civil Rights Movement: *Timeline of Supreme Court School- Desegregation Cases from Brown to Fisher.* San Francisco, CA. https://www.americanbar.org/content/dam/aba/administrative/litigation/materials/aba-annual 2013/written_materials/20_lessons_in_leadership.authcheckdam.pdf

American Educational Research Association. (2000). *Position statement on high-stakes testing.* Retrieved July 23, 2018 from http://www.aera.net/About-AERA/AERA-Rules-Policies/Association-Policies/Position-Statement-on-High-Stakes-Testing

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

American Psychological Association. (n.-d.). *APA dictionary of psychology.* Retrieved from https://dictionary.apa.org

Ardasheva, Y., Tretter, T.R., & Kinny, M. (2012). English language learners and academic achievement: Revisiting the Threshold Hypothesis. *Language Learning, 62*(3), 769-812. http://doi.org/10.1111/j.1467-9922.2011.00652.x

Bailey, A. L., Wolf, M. K. (2012, January). *The challenge of assessing language proficiency aligned to the Common Core State Standards and some possible solutions*. Paper presented at the Understanding Language Conference, Stanford University, CA. https://ell.stanford.edu/sites/default/files/pdf/academic-papers/08-Bailey%20Wolf%20Challenges%20of%20Assessment%20Language%20Proficiency%20FINAL_0.pdf

Baker, E. L. (2013). *The chimera of validity.* Teachers College Record, *115,* 1-26. Retrieved from https://www.tcrecord.org

Battenfield, M. & Crawford, F. (2015, December 4). Why the every student succeeds act still leaves most vulnerable kids behind. *U.S. News and World Report.* https://www.usnews.com/news/articles/2015-12-14/why-every-student-succeeds-act-still-leaves-most-vulnerable-kids-behind

Bauman, J., Boals, T. Cranley, E., Gottlieb, M., &Kenyon, D. (2007). Assessing Comprehension and Communication in English State to State for English language learners (ACCESS for ELLs©). In J. Abedi (Ed.). *English language proficiency assessment and the nation: Current status and future practice* (pp. 81-93). Davis, CA: UC Davis.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533. http://doi.org/10.1038/483531a.

Blaise, J. G. (2018). The effects of high-stakes accountability measures on students with limited English proficiency. *Urban Education, 53*(9), 1154-1181. http://doi.org/10.1177/0042085915613549

Boals, T., Hakuta, K. & Blair, A. (2005). Literacy development in academic contexts for adolescent English language learners: Policy considerations and future research. In D. Molle, E. Sato, T. Boles, & C. A. Hedgepeth (Eds.), *Multilingual learners and academic literacies: Sociocultural contexts of literacy development in adolescents* (pp.291-308). New York, NY: Routledge.

Boals, T., Kenyon, D.M., Blair, A., Cranley, E., Wilms, C., and Wright, L.J. (2015). Transformation in K-12 English language proficiency assessment: Changing contexts, changing constructs. *Review of Research in Education, 39*(1), 122-164. http://doi.org:10.3102:0091732X14556072

Botel, Jason. (2017, December 18). [Interim feedback letter to Dr. Steven Staples, Virginia Superintendent of Public Instruction]. Retrieved from https:// www2.ed.gov/admins/lead/account/stateplan17/vafeedbackltr122017.pdf

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

California Department of Education. (2017, April 28). *Reclassification guidance for 2017-2018.* https://www.cde.ca.gov/sp/el/rd/reclass1718.asp

Cai, J., Morris, A., Hohensee, C., Hwang, S. Robison, V., & Hiebert, J. (2018). The role of replication studies in educational research. *Journal for Research in Mathematics Education, 49*(1), 2-8. http://doi.org/10.5951/jresematheduc.49.1.0002

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal (42)*2, 305-328. http://doi.org/10.3102/00028312042002305

Cannaday, B.K. Jr. (2007, March 23). *Superintendent's Memo No. 70: Stanford English Language Proficiency Reading Subtest Discontinued as a Substitute the Standards of Learning Reading Test.* http://www.doe.virginia.gov/administrators/superintendents_memos/2007/inf070.html

Cannaday, B.K. Jr. (2008, March 28). *Superintendent's Memo No. 68: Spring Professional Development Training for the World-Class Instructional Design and Assessment (WIDA) English Language Proficiency (ELP) Standards.* http://www.doe.virginia.gov/administrators/superintendents_memos/2008/inf068.html

Carroll, P. E., & Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classificaions for English-language learners and native English speakers in fifth grade. *Language Testing, 33*(1) 23-52. http://doi.org.10.1177/0265532215576380

Castro, Quinn, Fuller, & Barnes (2018). *Addressing the importance and scale of the U.S. teacher shortage*. (Policy brief 2018-1). Retrieved from University Council for Education Administration: http://www.ucea.org/wpcontent/uploads/2018/01/Addressing-the-Importance-and-Scale-of-the-US-Teacher-Shortage.pdf

Center for Applied Linguistics, Language Assessment Division Psychometric and Quantitative Research Team. (2017). *Annual technical report for ACCESS for ELLs® 2.0 online English Language Proficiency Test, Series 400, 2015–2016 administration: Annual technical report no. 12A.* https://www.wida.us/get.aspx?id=2135

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267-296. http://doi.org.10.1016/S1060-3743(03)00038-9

Chi, Y, Garcia, R.B., Surber, C., & Trautman, L. (2011). Alignment study between the Common Core State Standards in English Language Arts and Mathematics and the WIDA English language proficiency standards, 2007 edition, preKindergarten through grade 12. http://achievementstrategies.org/docsTemplates/CommonCore/WIDA_Common_Core_Alignment03_14_11.pdf

Chun, D.M. (2012). Replication studies in CALL research. *CALICO Journal, 29*(4), 591-600. http://doi.org.10.11139/cj.29.4.591-600

Code of Virginia, Chapter 829 (1996) § 1-511. English designated the official language of the Commonwealth. http://www.languagepolicy.net/archives/va.htm

Code of Virginia, Chapter 829 (1996) §22.1-212.1. Obligations of school boards. http://www.languagepolicy.net/archives/va.htm

Constantino, S. M. (2018, January 26). Superintendent's Memo #017-18: Implementation of the 2017 English Standards of Learning and Curriculum Framework. http://www.doe.virginia.gov/administrators/superintendents_memos/2018/017-18.shtml

Converse, J.M. & Presser, S. (1986). Survey questions: Handcrafting the standardized questionnaire. Thousand Oaks, CA: Sage Publications.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research*. (3rd ed.). Los Angeles, CA: Sage Publications.

Creswell, J. W, & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research*, (3rd ed.) Thousand Oaks, CA: Sage Publications.

DeBray, E. & Blankenship, A. E. (2016). Forward: The aftermath of ESSA's devolution of power to states: A federal role in incentivizing equity and building state and local capacity. *Education Law & Policy Review, 3,* ix – xxvii. https://educationlaw.org/images/pdf/2016/elpr vol 3.pdf

Del Vecchio, A. & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque: New Mexico Highlands University, Evaluation Assistance Center—Western Region.

DuHart, J. (2019). *The relationship between English learners' English language proficiency and high stakes assessments in Virginia* (Doctoral dissertation). https://digitalcommons.liberty.edu/doctoral/2149/

Easley, R. W., Madden, C. S., & Dunn, M. G. (2000). Conducting marketing science: The role of replication in the research process. *Journal of Business Research, 48,* 83-92. http://doi.org.10.1016/S0148-2963(98)00079-4

Eccles, J. S. & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Review of Psychology, 53*(1), 109-132. http://doi.org.10.1146/annurev.psych.53.100901.135153

Elliot, M., Fairweather, I., Olsen, W., & Pampaka, M. (2016) *Oxford: A dictionary of social research methods.* New York, NY: Oxford University Press. https://www.oxfordreference.com/view/10.1093/acref/9780191816826.001.0001/acref-9780191816826-e-0072

ELPA21. (2015). *The ELPA21 field test implementation report*. http://www.elpa21.org/sites/default/files/Field Test Implementation Report-September 2015 %281%29.pdf

Emblidge, M. (2007, September 26). Virginia Board of Education Resolution Number 2007-33: Approving World- Class Instructional Design and Assessments (WIDA). http://www.doe.virginia.gov/boe/resolutions/2007/2007-33.shtml

English Language Proficiency Assessment Consortium. (n.d.) *English language proficiency assessment for the century-ELPA21.* http://www.elpa21.org/about/faqs

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Fast, M., Ferrara, S., & Conrad, D. (2004). *Current efforts in developing English language proficiency measures as required by NCLB: Descriptions of an 18-state collaboration.* Washington, DC: American Institute for Research.

Feeley, M. (2007). Legality, social research, and the challenge of institutional review boards. *Law & Society Review, 41*(4), 757-776. http://doi.org.10.1111/j.1540-5893.2007.00322.x

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th edition). London, England: Sage.

Fisher, R. (1935) *The design of experiments.* Edinburgh, United Kingdom: Oliver & Boyd.

Fox, J. & Fairbairn, S. (2011). Test review: ACCESS for ELLs®. *Language Testing 28*(3), 425-431. http://doi.org.10.1177/0265532211404195

Francis, D. J., & Rivera, M. O. (2007). Principals underlying English language proficiency tests and academic accountability for ELLs. In J. Abedi (Ed.), *English language proficiency assessments in the nation: Current status and future practice* (pp. 13-31). Davis, CA: University of California Davis School of Education.

Fuller, E.J., Hollingworth, L. & Pendola, A. (2017). The Every Student Succeeds Act: State efforts to improve access to effective educators, and the importance of school leadership. *Education Administration Quarterly, 53*(5), 727-756. http://doi.org./10.1177/0013161X17711481

Gándara, P. & Rumburger, R.W. (2009). Immigration, language, and education: How does language policy structure opportunity? *Teachers College Record, 111*(3), 750-782. http://alisonlbruno.net/home/EDU_5105_files/Immigration%20language and Education Article.pdf

García, O. (2009). Emergent bilinguals and TESOL: What's in a name? *TESOL Quarterly, 43*(2), 322-326. http://doi.org.10.1002/j.1545-7249-2009.tb00172.x

García, O., Kleifgen, J. A., and Falchi, L. (2008). *From English learners to emergent bilinguals.* Equity matters: Research review number. 1. New York, NY: Teachers College Press.

García, G.E., McKoon, G. & August, D. (2006a). Language and Literacy Assessment of Language-Minority Students. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (p. 597–624). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

García, G.E., McKoon, G. & August, D. (2006b). García, G. E., McKoon, G., & August, D. (2006). Synthesis: Language and Literacy Assessment. In D. August & T. *National Literacy Panel on Language-Minority Children and Youth* (p. 583–596). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Gould, J., & Kolb, W. I. (1964). *A dictionary of the social sciences.* London, England: Tavistock Publications.

Great Schools Partnership. (2015). *The glossary of education reform: Long term English learner.* https://www.edglossary.org/long-term-english-learner/

Greene, J. C. (2017). *Mixed methods in social inquiry.* San Francisco, CA: Jossey-Bass.

Guion, R. M. (1978). Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology, 63*, 499-506. http://doi.org.10.1037/0021-9010.63.4.499

Hopkins, M., Malsbary, C.B., & Moralez, P.Z. (2016). Implementing responsive federal policy for bi/multilingual students. *Education Law & Policy Review, 3,* 31-57. https://educationlaw.org/images/pdf/2016/elpr vol 3.pdf

Hottenstein, K. N. (2018). American institutional review boards: Safeguards or censorship? *Journal of Research Administration, 49*(1), 31-42. https://files.eric.ed.gov/fulltext/EJ1181983.pdf

Howard, J. (2006, November 10). Oral history under review. *The Chronicle of Higher Education, 53*(12), A14-A17. https://www.chronicle.com/article/Oral-History-Under-Review/6566

Huchinson, M. & Hadjioannou, X. (2017). The morphing assessment terrain for English learners in U. S. schools. *English Teaching: Practice & Critique, 16*(1), 110-126. http://doi.org.10.1108/ETPC-02-2016-0037

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.) Cambridge England: Cambridge University Press.

Jennings, J. (2016). Fifty years of federal school law: Back to the future? *Education Law & Policy Review, 3,* 1-30. https://educationlaw.org/images/pdf/2016/elpr vol 3.pdf

Katz, A., Low, P., Stack, J., & Tsang, S.L. (2004, September). *A study of content area assessment for English language learners.* Prepared for the Office of English Language Acquisition and Academic Achievement of English Language Learners, U.S. Department of Education. Oakland, CA: ARC Associates. http://www.arcassociates.org/files/CAELLRpt9-04.pdf

Kearney, K. A., Hopkins, R. H., Mauss, A. L., & Weishiet, R. A. (1984). Self-generated identification codes for anonymous collection of longitudinal questionnaire data. *Public Opinion Quarterly, 48*, 370-378. http://doi.org.10.1093/poq/48.1B.370

Kopriva, R. (2008). Improving testing for English language learners. New York, NY: Routledge.

Lane, J. F. (2018, July 27). Superintendent's Memo #192-18: English language proficiency assessments test administration schedule for the 2018-2019 school year. http://www.doe.virginia.gov/administrators/superintendents_memos/2018/192-18.shtml

Lane, J. F. (2019, August 9). Superintendent's Memo #194-19: English language proficiency assessments test administration schedule for the 2019-2020 school year.

http://www.doe.virginia.gov/administrators/superintendents_memos/2019/index.shtml

Lee, O. (2018). English language proficiency standards aligned with content standards. *Educational Researcher, 47*(5), 317-327. http://doi.org.10.3102/0013189X18763775

Linn, R. L., Baker, E. L., & Dunbar, S. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 5-21. http://doi.org.10.3102/0013189X020008015

Linquanti, R., & Cook, H. G. (2015). *Re-examining reclassification: Insights from a national working session on policies and practices for exiting students from English learner status.* Washington, DC: Council of Chief State School Officers.

Mackey , A. (2012). Why (or why not), when, and how to replicate research. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 21-46). Cambridge, United Kingdom: Cambridge University Press.

Mahoney, K. (2017). *The assessment of emergent bilinguals: Supporting English language learners.* Blue Ridge Summit, PA: Multilingual Matters.

Makel, M. C., & Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts, 8,* 27-29. http://doi.org.10.1037/a0035811

McNeal, L.M. (2000). *Contradictions of school reform: Educational costs of standardized testing.* New York, NY: Routledge.

Menken, K. (2010). NCLB and English language learners: Challenges and consequences. *Theory Into Practice, 49*(2), 121-128.

http://doi.org.10.1080/00405841003626619

Menken, K., Hudson, T., & Leung, C. (2014). Symposium: Language assessment in standards-based education reform. *TESOL Quarterly, 48*(3), 586-5614. http://doi.org.10.1002/tesq..180

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1980). The validity and ethics of assessment. *American Psychologist, 35*(11), 1012-1027. http://doi.org.10.1037/0003-066X35.11.1012

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-100). Phoenix, AZ: Oryx Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23. http://doi.org.10.3102/0013189X023002013

National Center for Education Statistics. (2012) *The condition of education 2012.* Retrieved fromhttps://nces.ed.gov/pubs2012/2012045.pdf

National Center for Education Statistics. (2019, May). *English language learners in public schools.* https://nces.ed.gov/programs/coe/indicator_cgf.asp

New York State Education Department. (2015). *Commissioner's Regulation Part 154: English Language Learners (ELLs) Screening, Identification, Placement, and Exit Criteria.* http://www.nysed.gov/common/nysed/files/bilingual/ellidchartguidance7.1.15.pdf

New York State Education Department. (n.d.) Multilingual learner (MLL)/English language learner (ELL) program quality review and reflective protocol toolkit. http://www.nysed.gov/bilingual-ed/multilingual-learner-mllenglish-language-learner-ell-program-quality-review-and

No Child Left Behind Act of 2001, Pub. L No. 107-110, 115 Stat. 1425 (2002).

Norris, J. M. (2008). *Validity evaluation in language assessment.* Frankfurt, Germany: Peter Lang.

Northam, R. S. (2020, March 23). *Governor Northam orders statewide closure of certain non-essential businesses, K-12 schools* [Press release]. governor.virginia.gov/newsroom/all-releases/2020/march/headline-855292-en.html

Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students. (2003, February). *Draft of Part II, Final Non-Regulatory Guidance on the Title III State Formula Grant-Program Standards, Assessments and Accountability*. Washington, DC: U.S. Department of Education.

Orfield, G. (2016). A great federal retreat: The 2014 Every Student Succeeds Act. *Education Law & Policy Review*, *3*, 273 – 297. https://educationlaw.org/images/pdf/2016/elpr vol 3.pdf

Pennington, J. (2004). *The colonization of literacy education: A story of reading in one elementary school.* New York: Peter Lang.

Porte, G. (2012). *Replication research in applied linguistics.* Cambridge, United Kingdom: Cambridge University Press.

Porter, S.G. & Vega, J. (2007). Overview of existing English language proficiency tests. In J. Abedi (Ed.), *English language assessment and proficiency in the nation: Current status and future practice* (pp. 135-136). Davis: University of California.

Rivera, C. (1984). Language proficiency and academic achievement. Avon, England: Multilingual Matters.

Robinson-Cimpian, J.P., & Thompson, K. (2016). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy and Analysis and Mangagement. 35*(2), 279-305. http://doi.org.10.1002/pam.21882

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 22*(1), 7-15. http://doi.org.10.1111/j.1745-3992.2002.tb00080.x

Saunders, W. M. & Marcelletti, D.J. (2013). The gap that can't go away: The Catch-22 of reclassification in monitoring the progress of English learners. *Educational Evaluation and Policy Analysis, 35*(2), 139-156. http://doi.org.10.3102/016237371246189

Scarcella, R. (2003). *Academic English: A conceptual framework.* (Linguistic Minority Research Institute Technical Report 2003-1). Santa Barbara: University of California.

Schmidt, S. (2016). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. In A. Kazdin (Ed.) *Methodological Issues and Strategies in Clinical Research* (pp. 581-596). Washington, D.C.: American Psychological Association.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*(1), 41–53. https://doi.org/10.1177/01466219922031176

Shober, A.F. (2017). ESSA and state capacity: Can states take accountability seriously? In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act: What it means for schools, systems, and states* (pp. 107-120). Cambridge, MA: Harvard Education Press.

Slama, R.B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educatioal Psychology, 104,* 265-285. http://doi.org10.1037/a0025861

Slama, R.B. (2014). Investigating whether and when English learners are reclassified into mainstream classes in the United States: A discrete-time survival analysis. *American Educational Research Journal, 51.*220-252. http://doi.org.10.3102/0002831214528277

Sloan, Kris. (2007). High-stakes accountability, minority youth, and ethnography: Assessing multiple effects. *Anthropology & Education Quarterly, 38*(1), pp. 24-41. http://doi.org.10.1525/aeq/2007.38.1.24

Sodat, Lynn. (2018, February 1). Accountability updates: Revisions to the Standards of Accreditation and Virginia's federal programs application under ESSA. [PowerPoint presentation; Virginia ESL Supervisor's Annual Conference]. http://www.doe.virginia.gov/federal_programs/esea/tech_assistance.../11-essa-updates.pptx

Solórzano, R.W. (2008). High stakes testing: issues, implications, and remedies for English Language Learners. *Review of Educational Research, 78*(2), 260-329. https://doi.org.10.3102/0034654308317845

Staples, S. (2016, September 23). Superintendent's Memo #244-16: 2016 Mathematics Standards of Learning and Curriculum Framework. http://www.doe.virginia.gov/administrators/superintendents_memos/2016/244-16.shtml

Staples, S. (2017, June 9). Superintendent's Memo #168-17: English Proficiency Criteria for ACCESS for ELLS® 2.0 English Language Proficiency (ELP) Assessment - 2016-2017 Assessment Year. http://www.doe.virginia.gov/administrators/superintendents_memos/2017/168-17.shtml

Stark, L. (2007). Victims in our own minds? IRBs in myth and practice. *Law & Society Review 41*(4), 777-786. http://doi.org.10.1111/j.1540-5893.2007.00323.x

Stephenson, A., Jiao, H., & Wall, N. (2004). *A performance comparison of native and non-native speakers of English on an English language proficiency test.* San Antonio, TX: Harcourt Assessments. http://www.pearsonassessments.com

Stewner-Manzanares, G. (1988). *The Bilingual Education Act: Twenty years later.* Washington, DC: National Clearinghouse for Bilingual Education. https://files.eric.ed.gov/fulltext/ED337031.pdf

Sugerman, J. & Geary, C. (2018). *Fact sheet: Facts about English learners and the NCLB/ESSA transition in Virginia.* Washington, DC: Migration Policy Institute.

TESOL. (2012, March 27). *Convention News: TESOL Announces 2012 Award Winners.* https://www.tesol.org/convention-2018/tesol-2018-convention-news/2012/05/09/tesol-announces-2012-award-winners

Thurlow, M.L. & Kopriva, R.J. (2015). Advancing accessibility and accommodations in content assessment for students with disabilities and English learners. *Review of Research in Education, 39,* 331-369. http://doi.org.10.3102/0091232X14556076

Tinker v. Des Moines Independent Community School District. (n.d.) *Oyez.* https://www.oyez.org.cases/1968/21

United States Department of Agriculture (2015, April). *The community eligibility provision (CEP) fact sheet.* https://fns-prod.azureedge.net/sites/default/files/cn/CEPfactsheet.pdf

United States Department of Education. (n.d.-a).College-and career-ready standards. https://www.ed.gov/k-12reforms/standards

United States Department of Education. (n.d.-b). Every Student Succeeds Act (ESSA). https://www.ed.gov/essa

United States Department of Education. (n.d.-c). *E. D. Data Express. Data about elementary and secondary schools in the US: Definitions.* https://eddataexpress.ed.gov/definitions.cfm

United States Department of Education. (n.d.-d). *Every Student Succeeds Act.* https://www.ed.gov/k-12reforms/standards

United States Department of Education. (n.d.-e.). *Office for Civil Rights. Programs for English Language Learners. Part IV: Glossary*. https://www2.ed.gov/about/offices/list/ocr/ell/edlite-glossary.html

United States Department of Education. (2018, May 3). *Press Release: Secretary DeVos approves South Carolina and Virginia's ESSA state plans.* https://www.ed.gov/news/press-releases/secretary-devos-approves-south-carolina-and-virginias-essa-state-plans

United States Department of Education. (2009). *Race to the Top Program: Executive Summary.* https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

United States Department of Education Office for Civil Rights (2018). *Schools civil rights obligations to English learner students and limited English proficient parents.* https://www2.ed.gov/about/offices/list/ocr/ellresources.html

Valdés, G. (2004). The teaching of academic language to minority second language learners. In A. F. Ball and S.W. Freedman. (Eds.). *Bakhtinian perspectives on language, literacy, and learning* (66-98). Cambridge, England: Cambridge UP.

Valenzuela, A., Preito, L. & Hamilton, M.P. (2007). Introduction to the special issue: No Child Left Behind (NCLB) and minority youth: What the qualitative evidence suggests. *Anthropology & Education Quarterly, 38*(1), 1-8. http://doi.org.10.1525/aeq.2007.38.1.1

Virginia Department of Education. (n.d.-a) *Enrollment and Demographics: Fall Membership Reports.* http://www.doe.virginia.gov/statistics_reports/enrollment/index.shtml

Virginia Department of Education. (n.d.-b). *Virginia Compliance with Title III Requirements: Purpose of Program and General Uses of Funds of Title III.* http://www.doe.virginia.gov/federal_programs/esea/title3/compliance/virginia-compliance-with-title-iii-requirements.dox

Virginia Department of Education. (n.d.-c). *Virginia public school listing: By region.* http://www.doe.virginia.gov/directories/schools/school_info_by_regions.shtml

Virginia Department of Education. (2020, January 10). Guidelines for uniform performance standards and evaluation criteria for teachers. http://www.doe.virginia.gov/teaching/performance_evaluation/teacher/ind

Virginia Department of Education. (2019a). *English language proficiency assessments.* http://www.doe.virginia.gov/testing/english_language_proficiency_assessments/index.shtml

Virginia Department of Education. (2019b). *Every Student Succeeds Act.* Retrieved from http://www.doe.virginia.gov/federal_programs/esea/index.shtml

Virginia Department of Education. (2019c). *Federal Accountability.* http://www.doe.virginia.gov/statistics_reports/accreditation_federal_reports/federal_accountability/index.shtml

Virginia Department of Education. (2019d). *2017 SOA: School Accreditation.* http://www.doe.virginia.gov/boe/accreditation/2017-school-accred.shtml

Virginia Department of Education. (2019e). *Virginia Standards of Learning and Common Core State Standards*. http://www.doe.virginia.gov/testing/common_core/index.shtml

Virginia Department of Education. (2019f). *WIDA ELD standards.* *http://www.doe.virginia.gov/instruction/esl/index.shtml*

Virginia Department of Education. (2018). *Revised State Template for the Consolidated State Plan: The Elementary and Secondary Education Act of 1965, as amended by the Every Student Succeeds Act.* http://www.doe.virginia.gov/federal_programs/esea/essa/essa-state-plan.pdf

Virginia Department of Education. (2018, September 27). *New accreditation standards require continuous achievement for all schools.* http://www.doe.virginia.gov/news/news_releases/2018/09-sep27.shtml

Virginia Department of Education. (2018, August 22). *Press release: Virginia releases 2017-2018 test results.* http://www.doe.virginia.gov/news/news_releases/2018/08-aug22.shtml

Virginia Department of Education. (2017, October 2). *English learners: Guidelines for participation in the Virginia assessment program.* http://www.doe.virginia.gov/testing/participation/el-participation-guidelines.pdf

Virginia Department of Education. (2013). *Appendix A: Historical overview of the Standards of Learning Program.* *http://www.doe.virginia.gov/boe/reports/annual_reports/2013_appendix_a*

sol_history.pdf

Virginia Department of Education (2011, January 13). *Comparison of Virginia's 2009 Mathematics Standards of Learning with the Common Core State Standards for Mathematics.* http://www.doe.virginia.gov/testing/sol/standards_docs/mathematics/sol_ccss_comparison_mathematics.pdf

Virginia Department of Education. (2010, November 18). *Comparison of Virginia's 2010 English Standards of Learning with the Common Core State Standards for English and Literacy.* http://www.doe.virginia.gov/testing/sol/standards_docs/english/sol_ccss_comparison_english.pdf

Virginia Department of Education. (2010, June 24). *Board of Education Reaffirms Support for SOLs; Opposes Imposition of National Standards.* http://www.doe.virginia.gov/news/news_releases/2010/jun24.shtml

Virginia Department of Education (2009, July 31). *Virginia School Division Locale Descriptions.* http://www.doe.virginia.gov/directories/sch_division_locales_schedules/school_division_locale_descriptions.pdf

Virginia Department of Education. *(*2008). *Approving rescission of English Language Proficency (ELP) standards of learning and adopting WIDA ELP standards: Resolution number 2008-15*. http://www.doe.virginia.gov/boe/resolutions/2008/2008-15.shtml

Virginia Department of Education: Virginia Board of Education Committee on School

and Division Accountability. (2017, May 23). *The Every Student Succeeds Act of 2015 (ESSA): Virginia federal programs application part 2.* http://www.doe.virginia.gov/boe/committees_standing/accountability/2017/05-may/essa.pdf

West, M. (2017). The case for ESSA: A proper balance. In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act: What it means for schools, systems, and states* (pp. 75-90). Cambridge, MA: Harvard Education Press.

WIDA. (n.d.-a). *ACCESS for ELLs scores and reports.* https://wida.wisc.edu/assess/access/scores-reports

WIDA. (n.d.-b). *ACCESS for ELLs 2.0 Summative Assessment.* https://www.wida.us/Assessment/access20.aspx

WIDA. (n.d.-c). *Consortium members.* https://www.wida.us/membership/states/index.aspx

WIDA. (n.d.-d). *Download Library.* ACCESS for ELLs: Technical Reports. https://www.wida.us/downloadLibrary.aspx

WIDA. (n.d.-e). *Mission and History*. https://wida.wisc.edu/about/mission-history

WIDA. (n.d.-f). *Participation opportunities for educators.* https://www.wida.us/membership/participation/

WIDA. (n.d.-g). *Virginia State Specific Directions; ACCESS for ELLs Administration (2019-2020).* https://wida.wisc.edu/sites/default/files/state-specific-directions/VA-State-Specific-Directions.pdf

WIDA. (n.d.-h). *WIDA Consortium: Membership and Programs.* https://wida.wisc.edu/memberships/consortium

WIDA. (n.d.-i). WIDA® Non-Disclosure and User Agreement. http://www.k12.wa.us/ELPA21/pubdocs/NDUAforWAAltACCESS.pdf

WIDA. (2019). *ACCESS for ELLs 2.0 Spring 2019 interpretive guide for score reports: Kindergarten-grade 12*. https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf

WIDA. (2019, January). *Research brief: Long-term English learners across 15 WIDA states.* https://wida.wisc.edu/sites/default/files/resource/WIDA-Brief-Long-Term-English-Learners .pdf

WIDA. (2018). *2017 ACCESS for ELLs 2.0 Score Changes*. https://www.wida.us/Assessment/ACCESS 2.0/proficiency.aspx

WIDA. (2017a). *Annual Technical Report for ACCESS for ELLs® 2.0 Online English Language Proficiency Test, Series 400, 2015–2016 Administration Annual Technical Report No. 12A Prepared by: Center for Applied Linguistics.* https://wida.wisc.edu/memberships/consortium

WIDA. (2017b). Consortium report: July 1, 2016 – June 30, 2017. https://www.wida.us/downloadLibrary.aspx

WIDA. (2015, June). *Research brief: Evaluating teacher effectiveness using ACCESS for ELLs*. https://wida.wisc.edu/resources/evaluating-teacher-effectiveness-using-access-ells

WIDA. (2014). *2012 Amplification of the English language development standards, Kindergarten to grade 12.* https://www.wida.us/standards/eld.aspx

WIDA. (2013, January 25). *Blog: Michigan adopts the WIDA standards.* http://widaatwcer.blogspot.com/2013/01/michigan-adopts-eld-standards.html

Wilson, C. (2013). *Interview Techniques for UX Practitioners: A User-Centered Design Method*. Waltham,MA: Elsevier Science & Technology.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly, 45*, 628-660. http://doi.org.10.5054/tq.2011.268063

Wisconsin Center for Educational Products and Services. (n.d.). https://wceps.org

Wisconsin Center for Education Research (n.d.). Timothy Boals: Research scientist. https://www.wcer.wisc.edu/About/Staff/1410

Wolf, M. K., Herman, J. L., Bachman, L. F., Bailey, A. L., & Griffin, N. (2008). Recommendations for assessing English language learners: English language proficiency measures and accommodation uses, recommendations report, part 3 of 3. *National Center for Research on Evaluation, Standards, and Student Testing.* Los Angeles, CA: University of California.

Wright, W. E. (2008). No Child Left Behind Act of 2001, Title III. In J. M. González (Ed.), *Encyclopedia of Bilingual Education.* Los Angeles: Sage. http://sk.sagepub.com/reference/bilingual/n231.xml

Zacher Pandya, J. (2011). *Overtested: How high-stakes accountability fails English language learners.* New York, NY: Teachers College Press.

Zehler, A.M., Hopstock, P.J., Fleischman, H.L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Development Associates, Special Issues Analysis Center

Appendix A: Survey Instrument

Survey on Teachers' Perceptions of the ACCESS Test

Dear Virginia ESL/EL Educator,

I am a Virginia ESL teacher and PhD candidate at Virginia Commonwealth University. I am conducting dissertation research, entitled, "Teachers' Perceptions and the Broad Validity of a High-Stakes English Language Proficiency Test in Virginia," on the 2019 ACCESS for ELLs® (ACCESS) test. I would like to know about educator perceptions of the 2019 administration of the test in Virginia. The questions are concerned with overall perceptions of the test, as well as some of its parts.

Survey items will ask about your experiences with the test, overall perceptions of the test, as well as perceptions of some of its parts; however, you will not be asked to disclose information you received about test, the content of the test, or test items. When the study is complete, I plan to report the results of my study to the VDOE as well as to the ESL professional organizations in the state and beyond. Because teachers and others who work with English learners know about the impacts of ACCESS testing, I feel it is important to include our points of view in conversations about the test.

You are encouraged to participate in the survey to provide the most helpful information concerning the ACCESS test and how it is administered and used in Virginia, though you are not required to participate. No information about whether or not you participate will be known; there are no consequences or risks to not participating. If you are willing to participate in a follow-up interview, please share

your name and contact information and you may be selected for a follow-up interview. Your name will not be shared in any capacity.

I would like you to complete the survey that is accessible from the following link: [link provided] **by May 20, 2019.**

It should take approximately 15-20 minutes to complete.

Please understand the following information regarding consent:

Risks and Discomforts: This study involves no more than minimal risk to you.

Benefits to You and Others: You may not receive any direct benefit from this study, but the information gathered from the survey will help school divisions understand how the ACCESS test is administered and used in Virginia. If you are chosen for a follow-up interview and complete it, you will receive a $15 Amazon gift card.

Costs: There are no costs for participating in this study other than the time you will spend completing the survey.

Confidentiality: Your participation in the survey will be anonymous.

Voluntary Participation and Withdrawal: You do not have to participate in this study. If you begin completing the survey, you may stop at any time without penalty. You may also choose not to answer particular questions on the survey.

Questions: If you have questions, complaints, or concerns at any time, either while you are completing the survey or in the future, please contact:

[Researcher contact information provided]

Your participation: Your participation is voluntary. By filling out the survey, you agree to let the researchers use your data. You may skip any question you do not wish to answer. You may discontinue the survey at any point if you wish.

Please complete the survey below.

Thank you!

Part 1 of 3: Please check all responses that apply.

1. Please describe yourself. I am a/an… (Please check all that apply.)

☐ English as a Second Language (ESL) teacher

☐ Elementary classroom teacher

☐ Secondary English learner (EL) content teacher

☐ School Principal

☐ School Administrator

☐ Other (Please specify: _____)

2. With what level of the ACCESS test were you involved? (Please check all that apply.)

☐ Kindergarten

☐ Grades 1-2

☐ Grades 3-5

☐ Grades 6-8

☐ Grades 9-12

3. With what form of the test were you involved? (Please check all that apply.)

☐ Online

☐ Paper

☐ Alternate ACCESS

4. Who administered the ACCESS test at your school? (Please check all that apply.)

☐ English as a Second Language (ESL) teachers

☐ Elementary classroom teachers

☐Secondary English learner (EL) content teachers

☐ Teachers of other subjects (i.e. biology, physical education, etc.)

☐ School principals

☐ Instructional Assistants

☐ Other (Please specify: _____)

5. What portions of the ACCESS did you administer? (Please check all that apply.)

☐ Listening

☐ Reading

☐ Writing

☐ Speaking

6. How would you describe your school(s)? (Please check all that apply.)

☐ Urban

☐ Rural

☐ Suburban

☐ Magnet

☐ Charter

☐ Title I

☐ 40 percent or more students eligible for free/reduced lunch

☐ Accredited

☐ Accredited with Conditions

☐ Accreditation Denied

☐ Other (Please specify: _____)

7. How many English learners (ELs) are enrolled at your school(s)?

☐ fewer than 30 students

☐ 30 -99

☐ 100 - 199

☐ 200 - 299

☐ 300 - 399

☐ 400 or more

8. Approximately what percentage of the students at your school(s) are classified as English learners (ELs)?

☐ Less than 5 percent

☐ 5 – 20 percent

☐ 21 -40 percent

☐ 41-60 percent

☐ 61 – 80 percent

☐ 81 percent or more

9. For how many school years have you administered the ACCESS test?

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5 or more

10. For what purposes are ACCESS scores used at your school(s)? (Please check all that apply.)

☐ Exiting students from the English learner (EL) subgroup

☐ Placing students in English as a second language (ESL) programming

☐ Exiting students from ESL programming

☐ Staffing decisions

☐ Student scheduling decisions

☐ Informing classroom instruction

☐ Teacher evaluation

☐ Other (Please specify: _____)

11. Which better describes you?

☐ Native speaker of English

☐ Non-native speaker of English

Part 2 of 3: Please respond to each statement below by indicating your level of agreement.

You may skip any statement if it doesn't apply to you or if you do not wish to give your opinion.

If you wish to elaborate concerning a statement, please type your comments in the box.

1. My school(s) received all ACCESS materials on time. Please elaborate if you wish:
2. I feel the training on how to administer ACCESS prepared me well to give the test. Please elaborate if you wish:
3. My school(s) had enough physical space and equipment to administer the ACCESS test smoothly. Please elaborate if you wish:
4. My school(s) had enough Internet and server capacity to administer the ACCESS test smoothly. Please elaborate if you wish:
5. My school(s) had enough personnel to administer the ACCESS test smoothly. Please elaborate if you wish:
6. Teachers had enough support in administering the ACCESS test. Please elaborate if you wish:
7. Overall, the administration of the test ran smoothly. Please elaborate if you wish:
8. English as a Second Language (ESL) instruction was positively impacted by the ACCESS test. Please elaborate if you wish:
9. The listening test is well designed. Please elaborate if you wish:

10. I feel the listening test adequately measures students' true listening ability.
11. The reading test is well designed. Please elaborate if you wish:
12. I feel the reading test adequately measures students' true reading ability. Please elaborate if you wish:
13. The writing test is well designed. Please elaborate if you wish:
14. I feel the writing test adequately measures students' true writing ability. Please elaborate if you wish:
15. The speaking test is well designed. Please elaborate if you wish:
16. I feel the speaking test adequately measures students' true speaking ability. Please elaborate if you wish:
17. The students tried to do their best on the ACCESS test. Please elaborate if you wish:
18. The students liked the ACCESS test. Please elaborate if you wish:
19. The students handled the format of the test well (whether Online, Paper, or Alternate ACCESS). Please elaborate if you wish:
20. The ACCESS test has a positive impact on the students' English language ability. Please elaborate if you wish:
21. Overall, the ACCESS test is well designed. Please elaborate if you wish:
22. Overall, I feel the ACCESS test is a beneficial test for students. Please elaborate if you wish:
23. Overall, I feel the ACCESS test is a reliable measure of English language proficiency. Please elaborate if you wish:

24. Overall, I feel the results of the ACCESS test are valid. Please elaborate if you wish:
25. The ACCESS scores will accurately reflect the students' actual English language proficiency levels. Please elaborate if you wish:
26. Students who well in their classes do well on the ACCESS test. Please elaborate if you wish:
27. A score of 4.4 or "proficient" on access is a good indicator that a student will be successful in school. Please elaborate if you wish:

Part 3 of 3: Please answer the questions below by typing your responses in the boxes.

You may skip any question that you do not wish to answer.

1. Did students at your school(s) prepare for the ACCESS test? If so, please describe how they prepared.
2. Were there any special circumstances at your school(s) that affected the administration of the ACCESS test? If so, please describe.
3. Does the ACCESS test affect instruction at your school, and if so, is it positive, negative, or both? Please describe how ACCESS affects instruction at your school.
4. What effect does the ACCESS test have on the English learners (ELs) at your school? Please describe.
5. Are you aware of any changes to the ACCESS test in recent years? If so, please describe.

6. Are you aware of any changes in the scoring of the ACCESS test in recent years? If so, please describe.
7. Are you aware of any changes in the way ACCESS scores are used to make decisions in recent years? If so, please describe.
8. Please describe the decision-making process for exiting English learners (ELs) at your school. Is the decision to exit based only on ACCESS scores? Are other factors taken into consideration, and if so, what other factors are considered when exiting students?
9. Is there anything else you would like to say about Virginia's ACCESS for ELLs® test?
10. Would you be willing to be interviewed by telephone about your experiences with the ACCESS test? Participants chosen to be interviewed will receive a $15 Amazon gift card.

    ☐ Yes

    ☐ No

If YES, please provide an email address where you can be reached to arrange an interview. Please note that your name, email address, phone number, and personal information will be linked to your survey responses and will no longer be anonymous; however, the researcher will not disclose your identity.

Appendix B: Winke Survey Instrument

ELPA Web Survey

Reflections on the Michigan English Language Proficiency Test (ELPA)

Introduction and explanation of the study: We are faculty members and graduate students within the departments of TESOL, Second Language Studies, and Education at Michigan State University. We would like to learn more about your opinions on the 2007 English Language Proficiency Assessment (ELPA) administered in the state of Michigan. The questions you will respond to are concerned with your overall evaluation of the test as well as your assessment of some of its specific aspects.

PLEASE ONLY TAKE THIS SURVEY **AFTER** THE ELPA HAS BEEN ADMINISTERED AT YOUR SCHOOL IN THE SPRING OF 2007.

The WEB survey will take about **10 or 15 minutes**.

Please note that your responses are anonymous and will only be used for research purposes. That is, no names or identifying information will be collected. When data is presented or published, no names or identifying information will be used. We plan on presenting results at the MI TESOL conference in the fall of 2007. Please feel free to contact us if you have any questions.

[Researcher contact information provided]

Your participation: Your participation is voluntary. By filling out the Web survey, you agree to let the researchers use your data. You may skip any question you do not wish to answer. You may discontinue the survey at any point if you wish.

SIRB's contact information: If you have any questions or concerns regarding your rights as a study participant, or are dissatisfied at any time with any aspect of the study, you may contact Peter Vasilenko, Ph. D., Director to the Human Research Protection Programs at Michigan State University, by phone: (517) 355-2180, fax☹517) 432-4503, email: irb@msu.edu, or regular mail: 202 Olds Hall, East Lansing, MI 48824.

Part 1 of 3: Please check all responses that apply.

1. How are you involved with the school? I am a/an… (Check all that apply.)

☐ English as a Second Language (ESL) teacher

☐ Language Arts Teacher

☐ English literature teacher

☐ Teacher of other subjects (i.e. biology, physical education)

☐ School Principal

☐ School Administrator

☐ Parent of a student in the school who took the ELPA

☐ Student

Other (Please specify: _____)

2. With what level of the test were you involved? (Check all that apply.)

☐ Level 1: Kindergarten

☐ Level 2: Grades 1-2

☐ Level 3: Grades 3-5

☐ Level 4: Grades 6-8

☐ Level 5: Grades 9-12

3. Who administered the ELPA at your school? (Check all that apply.)

☐ English as a Second Language (ESL) teachers

☐ Language Arts Teachers

☐ English Literature Teachers

☐ Teachers of other subjects (i.e. biology, physical education, etc.)

☐ School Principal(s)

☐ School Administrator(s)

☐ Parent(s) of students who took the ELPA

☐ Volunteers from outside the school

☐ Teachers' aides

☐ Other (Please specify: _____)

4. What portions of the ELPA did you administer? (Check all that apply.)

☐ Listening

☐ Speaking

☐ Reading

☐ Writing

5. How would you describe your school? (Check all that apply.)

☐ Urban

☐ Rural

☐ Suburban

☐ Public

☐ Magnet

☐ Charter

☐ Private

☐ Religious-affiliated

6. Approximately what percentage of your school is made up of English Language Learners (ELLs)?

☐ Less than 5 percent

☐ 5 percent

☐ 10 percent

☐ 15 percent

☐ 20 percent

☐ 25 percent

☐ 30 percent

☐ 35 percent

☐ 40 percent

☐ more than 40 percent

Part 2 of 3: Please respond to each statement below by indicating your level of agreement from 1 to 10, 1 being strongly disagree, 10 being strongly agree. You may skip any statement if it doesn't apply to you or if you would like to refrain from giving your opinion. If you have any comments concerning a statement, you may type your comments into the text box at the right.

1. The school received all its test materials on time
2. I feel the orientation on how to administer the well prepared me to give the test.
3. The school had enough physical space and equipment to administer the test smoothly.
4. The school had enough personnel to administer the test smoothly
5. Overall, the administration of the ELPA ran smoothly.

6. The teachers had enough support in administering the ELPA.
7. I feel that those administering the ELPA at the school were qualified to assess the English language ability of the English language learners (ELLs).
8. English as a Second Language (ESL) instruction was positively impacted by the ELPA.
9. Overall, the ELPA is a beneficial test for the English language learners (ELLs).
10. Overall, the ELPA test materials were well designed.
11. Overall, I feel the results of the ELPA are reliable and valid.
12. Students spend too much time preparing for the ELPA.
13. The students did well on the ELPA.
14. The students liked the ELPA.
15. The students tried to do their best on the ELPA.
16. The students were nervous about the ELPA.
17. The ELPA has a positive impact on the students' English language ability.
18. The students' parents wanted their children to perform well on the ELPA.
19. The students' parents were nervous about the ELPA.
20. The listening test is well designed.
21. The administration of the listening test was easy.
22. The listening portion of the listening test was easy for the students to understand.
23. I feel the listening test adequately measured the students' true listening ability.
24. The reading test was well designed.

25. The reading test is too difficult for students.
26. The reading test is too easy for the students.
27. I feel the reading test truly measured the students' true reading ability.
28. The writing test is well designed.
29. The first part of the writing test (about writing conventions) is a positive feature of the test.
30. The second part of the writing test (essay writing) is a positive feature of the test.
31. I feel the writing test adequately measured the students' true writing ability.
32. The speaking test procedures worked well.
33. The rubric for the speaking test is well designed.
34. The rubric for the speaking test was easy to follow.
35. I feel the speaking test adequately measured the students' true speaking ability.
36. My prior knowledge of a student's language ability affected what score I gave him or her o the speaking test.
37. I understand why we have to give the ELPA to our students.
38. I wanted all the students to pass the ELPA.
39. Students who do well in their ESL classes do well on the ELPA.
40. The ELPA scores will accurately reflect the students' actual English language proficiency levels.

Part 3 of 3: Please answer the open-ended questions below by typing your responses in the text boxes.

1. How did students at your school prepare for the ELPA?
2. Were there any special circumstances at your school that affected the administration of the ELPA? If so, please describe.
3. Does the ELPA affect instruction at your school, and if so, is it positive, negative, or both? Please describe how it affects instruction at your school.
4. What effect does the ELPA have on the English language learners (ELLs) at your school?
5. Is there anything else you would like to say about Michigan's ELPA? (P. Winke, personal communication, January 23, 2019).

Appendix C: Telephone Interview Protocol

Telephone Interview Protocol

Once the telephone interview has been set up, the participant will be contacted at the appointed day and time and the following statement will be read:

My name is Carolyn Waters. I'm a doctoral candidate at VCU. Thank you for your willingness to be interviewed. Is this still a good time to ask you a few questions? Your name or any other identifying information will not be recorded or reported. (If this is not a good time, can I call you back? When?) May I record our conversation?

INTERVIEW QUESTIONS:

1. What do you think about the WIDA ACCESS test?

2. Please describe how ACCESS test administration went last year at your school(s). (Possible optional follow-ups: Please describe the test administration at your school. What went well? Were there any issues?)

3. Do you think the WIDA ACCESS test is a good way to assess English learner (EL) progress toward proficiency? (Possible optional follow-ups: Please elaborate. Why or why not?)

Follow-up question for middle and high school teachers: Do you think the ACCESS test is a good way to assess long-term English learners?

4. What do you think about 4.4 as an exit score? (Possible optional follow-ups: Please elaborate. Is it appropriate? Are students ready to be exited at this score? Do you think it's too high or too low?)

5. Have there been any changes for English learners at your school recently? If so, what have they been? (Possible optional follow-ups: Have there been changes in programming or procedures? Do these changes have anything to do with the ACCESS test or ACCESS scores? If yes, how so?)

Follow-up question for elementary and middle school teachers: Is your school more concerned about growth on the ACCESS test now that it can count as a Pass on the Reading SOL?

6. Is there anything else you would like to say regarding the ACCESS test?

APPENDIX D: Interview Participant Characteristics

Interview Participants

Characteristic	Subgroup	*n*
Grade Level	K-5	3
	6-8	2
	9-12	2
Perspective	Positive	3
	Negative	2
	Neutral/Both	2
Professional Role	ESL Teacher	5
	ESL/Secondary	
	EL Content Teacher	1
	ESL/World Languages Teacher	1
Type of School	Urban	1
	Suburban	5
	Rural	1
Number of ELs	Under 30	0
	30-99	1
	100-199	3
	200-299	1
	300-399	0
	400+	2
Title I/40%+ Free or Reduced Lunch	Yes	7
	No	0
Accredited	Yes	5
	With conditions	1
	No	1

APPENDIX E: Qualitative Codes

FIRST PASS
Positive
Negative
BROAD VALIDITY
Reliability
    Inter-rater positive
    Inter-rater negative
Concurrent validity
    Between forms
    Between years
Predictive Validity
    Other measures
    Class Performance Yes
    Class Performance No
Consequential Validity
    4.4 too high
    4.4 too low
    Money/Cost effective
    Loss of instructional time
        ESL
        Content
        Other tests
        WIDA prep
    Emotional
        Ss tired/fatigued
        Ss stressed
        Ss stigmatized
        Ss nervous/anxious
        Ts stressed
        Ts fatigued
        Ts disrespected
    Teacher evaluation
CONSTRUCT RELEVANCE
    Reading on other subtests
    Listening on other subtests
    Developmental level
    No SPED accommodations

- Background knowledge
  - Content
  - Cultural
- Not like class practices
- Shy/timid/'freeze'
- Understand directions
- Difficulty
- Stamina/tired
- Attention/memory
- Opportunity to learn
- Artificial construct
- Online format
  - Keyboarding/typing
  - Stamina/tired
  - Computer literacy
  - Microphone
  - Can't repeat
  - Cheat - can repeat
  - Adaptive positive
  - Adaptive negative

FACTOR 5 - MOTIVATION

- Enjoy/fun
- Motivator/see progress
- Get out of class
- Not motivated
- Hate/don't like
- Boredom
- Frustration
- Stressed
- Lack of effort
- Guess/click through/rush
- Don't take serious/care
- Do poorly on purpose
- Not relevant/no purpose
- Overtesting
  - General
  - SOLs
  - Of ELs
  - "WIDA fatigue"

- Recycled items
- Score reflects effort
- Compliance

FACTOR 2 – EFFECTIVE ADMINISTRATION

Training
- Learn while doing
- Ambiguous time limit Support
- Support
  - Positive
  - From administration
  - From colleagues
  - From CARs
  - Not like SOLs

Online test
- Kicked out/interrupted
- Restrictive scheduling
- Adaptive test positive
- Adaptive test negative
- Difficult despite practice
- Difficult- record self with mic
- Difficult-keyboarding
- Like video game - positive
- Like video game - negative
- No prompt to say more
- No prompt to write more
- Scroll bar didn't work
- Timer distraction

Paper test
- Better than online
- Less efficient than online

Snow days/Inclement weather

SPECIFIC POPULATIONS
- SLIFE
- Newcomers
- Low income
- Low proficiency ELs
- High proficiency ELs
- LTELs
- Disability

- Cognitive
- Emotional/behavior
- Blind
- Deaf
- Preliterate
- Monitor/Former ELs
- Native speakers
  - Classified as ELs
  - Couldn't pass test

RESEARCH QUESTION 3

Aware of recent changes

- Yes
- No
- Not sure
- States/describes
- Misstates/details wrong
- Names individual who told
- SOL proxy
- Accreditation

MULTIPLE MEASURES

- To Exit
  - Yes
  - No
  - Should
- For other decisions
  - Yes
  - No
  - Should
- One data point/snapshot
- Department of Justice

ACCESS AS MEASURE

- Need something
- Other measures adequate
- Teachers know
- Other ELP tests
- Old tests better
- Old tests worse

METHODOLOGY

Materials on time ambiguous

Interpret as ‘results’
Interpret as ‘booklets’
No comment
Don’t know/can’t see test
Not allowed to look
Not allowed to discuss
SUGGESTIONS FOR IMPROVEMENT

Vita

Carolyn N. Waters was born on December 18, 1960, in Richmond, Virginia. She graduated from Open High School in Richmond, Virginia, in 1979. She received her Bachelor of Arts in Romance Languages from New York University, New York, New York, in 1984, and her Master of Arts in Teaching from School for International Training, Brattleboro, Vermont, in 1994. She taught English as a Second Language at Virginia Commonwealth University, Henrico County Public Schools, and J. Sargeant Reynolds Community College. She has been teaching ESL in Chesterfield County Public Schools since 2002.