2014

Test-based Accountability Systems: Concerns for Indiana's Multilingual Learners and Their Teachers

Kathryn Brooks
Butler University, kbrooks@butler.edu

Brooke Kandel-Cisco
Butler University, bkandel@butler.edu

Test-based Accountability Systems: Concerns for Indiana's Multilingual Learners and Their Teachers

KATIE BROOKS,
Butler University

BROOKE KANDEL-CISCO,
Butler University

Indiana's current test-based accountability system grew, in part, out of decades of federal-level educational reform initiatives. This article reviews the history of Indiana's test-based accountability system for schools and details how the system calculates evaluative ratings for Indiana teachers and schools. Additionally, the article analyzes how the Indiana accountability system contradicts what is known about appropriate measurement of English language learners and lists psychometric and validity concerns such as valid assessment, non-random assignment, norming groups, and ceiling/bottom effects. This article calls for a system in which multiple assessments offer rich data for school and teacher evaluations.

Keywords: accountability, teacher evaluation, English learners

The past several years have been marked by rapid change in Indiana education policy. Stakeholders at all levels including children, parents, teachers, and administrators have been affected by changes in standards, testing, evaluation, and public school funding to name a few. In this article, we seek to outline changes in Indiana's education accountability systems and highlight how those changes intersect with what is known about appropriate measurement of English language learners. Additionally, we describe how changes in the accountability system influence ELLs in Indiana.

While the terms English language learner or English learner seem to be the most widely understood and used term to describe a student who is learning English as a second or subsequent language, we purposefully use the term Multilingual Learner to describe these same students. We believe the term Multilingual Learner (ML) better reflects the rich linguistic capacity of students who are learning English. As of 2013, Indiana's K-12 population includes 54, 054 MLs representing over 263 languages. Approximately 63% of Indiana's MLs were born in the U.S., while the other 37% are immigrants to the U.S. (Indiana Department of Education, 2014a). Past trends suggest Indiana will continue to educate increasing numbers of MLs over the next years (U.S. DOE, 2013). When educators of MLs in Indiana understand and can anticipate how current education policy will affect MLs, we are better able to advocate for students, our profession, and as educators.

Indiana's Test-Based Educational Accountability System

History of the system

Indiana's current test based accountability system grew, in part, out of decades of federal-level educational reform initiatives. Educational reform attempts to improve schools through changes in the way they are organized and function day-to-day. Modern educational reform is often traced to the 1983 publication, *A Nation at Risk: The Imperative for Education Reform*. This document suggests that America is "at-risk" of being unable to compete in the world economic marketplace because the current system of education is inadequate. More recently, *No Child Left Behind Act of 2001 (NCLB),* enacted in Indiana in 2002, sets the goal of improving classroom instruction through

- Stronger accountability for results
- Research-based education methods
- More choices for parents (NCLB, 2002)

While NCLB has been criticized for expanding curricula of test preparation and increasing high-stakes testing, NCLB offered some mandates that heightened the profile of MLs in K-12 schools. For example, prior to NCLB schools were able to essentially ignore MLs because data were not available on specific ML education outcomes. Under NCLB, however, schools were required to disaggregate standardized test data for the ML sub-group and to show that the

schools were making progress in providing education (as measured by standardized test) to MLs. No longer could MLs be ignored in distant portable classrooms and whisked away to a special teacher. The NCLB mandated accountability surrounding the education of MLs resulted in increased funding, professional development, and general educational attention that had the potential to benefit the K-12 ELL population (Clewell, Cosentino de Cohen, & Murray, 2007).

A Nation at Risk (1983) and NCLB (2002) have been catalysts for discussions about and changes in education in the U.S. Under NCLB 100% of students needed to attain grade level proficiency in mathematics and reading by 2014; schools failing to attain these goals faced strenuous federal and state sanctions. The 100% proficiency goal was set because to set expectations any lower meant our system was intentionally leaving some children behind. While a goal of grade level proficiency for all students is noble, the 100% target was a drastic departure from historical student proficiency trends on standardized tests (Welner, 2005). Robert Linn (2004), then a researcher at the University of Colorado, examined trend data from the National Assessment of Educational Progress (NAEP) to document the extent to which NCLB 100% proficiency requirements were unlikely to be met. Linn found for eighth grade students, for example, "the rate of improvement in the percentage of students at the proficient level or above in mathematics would need to be 6.5 times as rapid between 2003 and 2014 as it was between 2000 and 2003" (p. 3). Linn and others (e.g., Abedi & Dietel, 2004; Welner, 2004;) predicted the rapid growth in proficiency required by NCLB was unrealistic and the goal was unattainable.

Because the 100% proficiency goal was unattainable, the U.S. Department of Education eventually allowed states to apply for an *Elementary and Secondary Education Act* (ESEA) flexibility waiver from NCLB if they agreed to enact an approved school accountability plan. As of August 2014, 43 states have received waivers and, at the time of this writing, an additional two states are in the process of developing an alternative accountability plan and seeking approval from the U.S. Department of Education. The Indiana Department of Education (IDOE) applied for and received one of these waivers in 2012 (U.S. DOE, 2012). As part of the waiver application, the IDOE proposed replacing the NCLB school evaluation model with a new school evaluation model that interprets standardized test scores in terms of status and growth

(U.S. Department of Education, 2014). In August of 2014, the IDOE's NCLB waiver was renewed by the U.S. Department of Education (U.S. DOE, 2014).

Components of the Indiana Accountability System

Student-level Standardized Tests

Indiana's test-based accountability system includes multiple components. The cornerstone of the system, however, consists of student scores on state standardized tests. Currently, Indiana students are required to take a litany of tests during their K-12 educational career including the Indiana Statewide Testing for Educational Progress Plus (ISTEP+) primarily in language arts and mathematics, but also in science and social studies. Third graders must also take the IREAD-3 and high school students must sit for End of Course Assessments (ECAs) in English 10, Algebra 1, and Biology 1. MLs are further required to be assessed yearly to measure their growth in and attainment of English proficiency using the LAS Links with scores used for schools Annual Measurable Achievement Objectives (AMAOs).

Teacher Evaluation System

The 2012-13 school year was the first year in which teachers were evaluated under the stipulations of legislation passed in 2011. While the legislation did not mandate a particular evaluation system, the law did set certain parameters for teacher evaluation. Under Indiana law, each teacher is rated as ineffective, needing improvement, effective, highly effective (Cole, Murphy, Rogan, & Eakes, 2013). The rating calculation must consider student standardized test scores; only teachers rated in the top two categories are eligible to receive a pay raise (Indiana Department of Education, 2011). Teachers in the lowest two categories are subject to sanctions such as immediate or eventual dismissal. The Indiana Legislature provided no specific guidance on how ESL teachers or other support personnel should be evaluated.

A-F School Ratings

Public Law 221 (P.L. 221) is Indiana's K-12 accountability system. P.L. 221 was passed by the state legislature in 1999, and mandates that public and accredited non-public schools are placed into one of five categories based on results from ISTEP+ and End-of-Course Assessments (IDOE,

nd). Under P.L. 221, Indiana schools have long received accountability scores, but a new iteration of the P.L. 221 accountability system, known as A-F, was approved both by the Indiana State Board of Education and the U.S. Department of Education in February of 2012. This new A-F system allowed Indiana to receive a waiver from the adequate yearly progress requirement of *NCLB Act*. In effect, the U.S. Department of Education's waiver approval gave Indiana flexibility in implementing *NCLB Act* requirements in exchange for an accountability system (A-F) that was focused on increasing student achievement (U.S. DOE, 2012).

While the *NCLB Act* relied on a status model for evaluating school improvement, Indiana's A-F uses a percentile growth model in addition to the status model. Status models measure the percentage of students that pass a state standardized test while the growth models consider how much students grow in performance on standardized tests either in relationship to content knowledge or their peers (Gong, Perie, & Dunn, 2006). In Indiana, public schools, accredited non-public schools, and schools that accept school vouchers are assessed by the A-F percentile growth model grading system. Elementary and middle schools are evaluated on growth and performance while high schools are evaluated on improvement, performance, and graduation rates (Hiller, DiTommaso, & Plucker, 2012).

Under the plan proposed by the IDOE, Indiana schools will be evaluated using a combination of the status and growth models, with the growth model focusing on how students grow in comparison to their peers. Growth modeling has been used in US schools since 1992 when Tennessee started using value added assessment to evaluate school districts. Two forms of growth models are typically used for accountability purposes in U.S. schools: a value added model and a percentile growth model. Value-added models have been used most extensively and for the greatest number of years. The exact variables considered with these models vary across time and state. These models may consider factors such as family income levels, race, ethnicity, language status, gender, and student mobility (Franco & Seidel, 2012). The value added model measures how student test scores change from year to year or over multiple years. These gains in test scores are then used to evaluate teacher and/or school performance.

Percentile growth modeling is the latest iteration of growth modeling used for educational accountability. Betebenner (2009) has

identified two main assumptions underlying this model: a) past student performance serves as a strong predictor of future student performance and b) high quality schools and teachers are better at facilitating growth in standardized test scores than low quality schools and teachers. Percentile growth modeling presents a shift in the conceptualization of student growth. Previous iterations of growth modeling were criterion-referenced. In other words, these models were focused only on how students grew in their achievement in relationship to a certain set of criteria: the state academic standards.

Percentile growth modeling adds a normative component to this growth modeling by comparing how much of an increase a student has on standardized test scores in comparison to students at similar levels of achievement (Betebenner, 2009). For example, if a group of students all have a third grade standardized test scaled score of 350, their growth on a standardized test will be compared with each other. If a particular student from this group scores significantly higher than her peers on the fourth grade test, she will be considered to have high growth in comparison to her scale score peers. Conversely, if she scores significantly lower than her peers on the fourth grade test, she will be considered to have low growth in comparison to her scale score peers. Adding the normative component to the growth modeling addresses concerns expressed by researchers questioning the vertical scaling of content for criterion-referenced standardized assessments in which the standards for grade levels change from year to year (Amrein-Beardsley, 2008). Instead of comparing scores for tests that are often based on different standards, this normative growth model compares students. School scores will show the median growth scores of all the students in the school in comparison to all the students who completed the test.

Indiana's student percentile growth model considers the growth of each student independent of his or her school. The analysis uses quantile regression analysis which will show a relationship between a student's previous test scores and predicted growth in test scores in the subsequent year of testing. Students are grouped (also called blocking) by percentiles or quantiles, of standardized test scaled scores into four different groups:

1. High achieving/high growth
2. High achieving/low growth
3. Low achieving/high growth
4. Low achieving/low growth

Then the student's growth is compared to students in their same quantile group, considered their academic peer group, using growth percentiles. Students are compared to other students in his or her academic peer group for up to three previous years when these data are available for the student. Target growth is set for each academic peer group based on the group's growth trajectory, and students will be rated as high, average, or low growth depending on how well they perform. The target percentile growth will change from year to year depending on the academic peer group performance on standardized tests. A teacher's and a school's growth scores are calculated based on the average growth of students in the class or school.

Concerns with the System: Multilingual Learners, Their Teachers, & Their Schools

Does Test-based Accountability Improve Educational Outcomes?

The primary concern for using test-based accountability system is that *there is no evidence that using student test scores as part of teacher and school evaluation systems results in higher student achievement.* In fact, according to the National Research Council, high-stakes testing and accountability when measured by national measures for more than a decade have produced little to no impact on student achievement, despite great cost and emphasis (Hout & Elliot, 2011). Furthermore, in international comparisons, US 15 year olds maintained their relative standing to other countries in reading and significantly decreased in math from 2000-2009, the years of high stakes testing accountability under NCLB (OECD, 2010).

Validity Issues Related to Indiana's Accountability System

The primary cause of the problems with using standardized test scores to evaluate teachers and schools is the validity of the tests, especially when they are used with ML students. A valid measure assesses what the evaluators believe that it is testing. Without validity, standardized test scores, teacher evaluations, and school A-F grades are meaningless because they are not measuring what the evaluators think they are measuring. In the next section, we present a few reasons that explain why the use of standardized tests for student, teacher, and school level evaluations and high-stakes decisions is invalid for students in general but also for ML students specifically.

Non-random assignment. One of the principles of high quality empirical research is that when comparing different groups, the groups should either be randomly assigned or should have highly similar characteristics. Comparing schools is difficult at best. Students are not randomly assigned to schools, and schools vary greatly in terms of available resources and student demographics and characteristics. This non-random assignment of students to schools and the vast differences in student populations between schools present a significant bias when making cross-school comparisons (Schochet & Chiang, 2010). MLs in Indiana, for example, tend to be clustered in particular schools and school corporations. According to the IDOE, only 27 out of almost 300 Indiana school corporations reported a Limited English Proficient population of at least 10% of the total student body (IDOE, 2014b). Furthermore, even within the ML student subpopulation, the demographic composition of the ML students at different schools can vary widely. For example, one school may have a large number of ML students whose parents are managers and executives for an automotive manufacturer and receive extensive tutoring outside of school, while other schools may have large numbers of ML students who are refugees with significant interrupted formal schooling. While these concentrations of MLs might allow schools to pool instructional resources and language programs, the concentrations are further evidence that comparing schools based on test scores as if all schools are equal is erroneous. In other words, Indiana schools serving MLs are not homogenous and student data from those schools should be interpreted in light of the specific complexities of each school population.

Standardized tests do not measure teacher quality. Multiple factors influence student performance on standardized tests. Betebenner (2009) is one of the developers of Indiana's test-based accountability system. The assumptions that Betebenner (2009) used in developing Indiana's A-F accountability system have serious validity issues and flaws in logic. His first assumption was that high quality schools and teachers are better at facilitating growth in standardized test scores than low quality schools and teachers. By stating this assumption, Betebenner implied that standardized test scores are a valid measure of teacher and school quality. However, this assertion is contrary to almost 50 years of extensive research on the impact of teachers and schools on student achievement. These studies indicate that typically 7-10% of variability

in student performance on standardized tests is attributable to teacher and school level factors (Coleman, 1966, Heubert & Hauser, 1999; Rivkin, Hanushek, & Kain, 1998; Schochet & Chiang, 2010). According to the American Statistical Association, more recent studies focused on basing teacher and school evaluation on student growth shows that only 1-14% of student test score growth can be attributed to teachers (ASA, 2014). Non-school variables such as (1) low birth-weight and non-genetic prenatal influences on children; (2) inadequate medical, dental, and vision care, often a result of inadequate or no medical insurance; (3) food insecurity; (4) environmental pollutants; (5) family relations and family stress; and (6) neighborhood characteristics (Berliner, 2009, p. 1), exert a much greater influence on student achievement than do school-related factors.

The most prominent non-school factors that influence ML student achievement include language differences, parent education level, previous experience with schooling, length of time in U.S. schools, cultural and acculturation issues, and native language literacy development (Abedi, 2002; DeCapua & Marshall, 2010; Garcia & Frede, 2010). Even the developers of the Indiana test-based accountability system acknowledge that teachers who have large numbers of ML students will likely have low growth scores on standardized tests (Diaz-Bilello & Briggs, 2014). Due to weaknesses in connecting student standardized test scores to teacher and school evaluations, the National Research Council of the National Academy of Sciences considers value-added measures of teacher effectiveness "too unstable to be considered fair or reliable" (Heubert & Hauser, 1999) and the American Statistical Association (2014) calls the statistical underpinning of the system "unstable," even under ideal conditions, due to its large error rates.

Characteristics of the multilingual learner. For MLs, the validity of using standardized test scores as a measurement of school effectiveness, or even student learning, is questionable. The test-based accountability system assumes the results of state standardized content tests can be interpreted as valid measures of MLs content knowledge. For example, it is assumed that a standardized test of grade level mathematics content will show the extent to which a student knows and can demonstrate the mathematics content. Yet, this assumption ignores other factors, unrelated to mathematics content, which the test is actually measuring. MLs are, by definition, in the process of learning

English, including academic English. When an ML takes a standardized mathematics test, that test is measuring not only the student's mastery of the mathematics, but is also measuring -and perhaps is mostly measuring- the student's ability to understand the academic English of the test. Certain types of English language that often appear on standardized tests and contribute to construct-irrelevant variance include unfamiliar vocabulary, culturally bound idiomatic language, confusing syntax like double negatives, morphologically complex words, and long sentences with multiple clauses and passive voice (Abedi, 2002; Young, 2008).

Norming group. The norming groups used to make comparisons amongst quantiles present another psychometric issue. Norm referenced interpretation of test results means that one student's scores will be compared against the scores of the "norming group," a group of students' who have already taken the same test. Inappropriate norming groups are known to substantially affect the validity of outcomes on standardized tests (American Educational Research Association [AERA], 1999). This means that a standardized test developed for one group of students is not necessarily valid for a different group of students. Standardized test results for a student who is a ML, for example, should be interpreted with caution if the norming group on which the percentile ranks were constructed did not include English learners. In A-F, the state will not disaggregate sub-groups and will indeed use cross sub-group comparisons to establish a letter grade for schools. Thus, the growth of a ML will be compared to a norming group not necessarily composed of MLs and, thus, the factors that uniquely affect MLs (i.e., language development, cultural differences, prior educational differences, etc.) will not be considered. The AERA's and National Council on Measurement in Education's joint Standards for Educational and Psychological Testing (1999), for example, note that "norms based on native speakers of English either should not be used with individuals whose first language is not English or such individuals' test results should be interpreted as reflecting in part current level of English proficiency rather than ability, potential, aptitude or personality characteristics or symptomatology" (p. 91).

Invalid measures of learning. Indiana's test-based accountability system, including the growth model components of the system, is grounded in student performance on standardized tests, yet standardized tests offer a limited, and often invalid, measure of student

learning. Indiana's academic standards are, in the words of the IDOE "world-class standards" that support students in becoming college and career ready (IDOE, nd). Unfortunately, the high stakes standardized tests purported to measure student mastery of those standards fail to fully assess the rich student learning that occurs in Indiana classrooms. The American Statistical Association (2014) highlighted this issue in a recent report:

> Ideally, tests should fully measure student achievement with respect to the curriculum objectives and content standards adopted by the state, in both breadth and depth. In practice, no test meets this stringent standard, and it needs to be recognized that, at best, most VAMs [value added measures] predict only performance on the test and not necessarily long-range learning outcomes. Other student outcomes are predicted only to the extent that they are correlated with test scores. A teacher's efforts to encourage students' creativity or help colleagues improve their instruction, for example, are not explicitly recognized in VAMs (np).

In other words, standardized tests only measure a small segment of the content and processes students learn in relation to a particular standard and these tests do little to help us understand a student's long-term mastery of the standard.

Ceiling and bottom effects. In addition to norming issues, the growth of the highest and lowest performing students in the proposed A-F system is particularly concerning, due to phenomena called the ceiling effect and the bottom effect. The ceiling effect refers to the tendency for a high performing student's test score growth to be smaller than average because the student's initial score already approaches the highest score possible. In the A-F system, this would be a student whose initial test score falls near the top of the highest quantile. These high performing students have little room to grow based on the standardized test score, and thus, these students and the schools in which they are enrolled could be misconstrued as low performing. A bottoming out effect, in contrast, affects students whose test scores fall near the lowest scores possible, or in the A-F system, near the bottom of the lowest

quantile. These low performing students could show substantial growth based on standardized test scores, yet because they began so low within the quantile, their test performance could still be considered to be insufficient compared to other students whose initial scores were in the upper scores of the quantile. This issue disproportionately affects MLs, especially those MLs just beginning to learn English due to the fact that standardized tests in academic English are often not linguistically accessible for MLs. Thus, the scores of beginning MLs tend to fall within the bottom of the lowest quantile and the language background of MLs adds another source of error in test-based accountability systems (Abedi, 2002). Schools with high numbers of high performing students, low performing students, or high numbers of MLs are likely to receive artificially low grades.

Consequences

An additional group of concerns involves the consequences of the A-F accountability system for schools. Many prominent educational experts have spoken out against the misuse of standardized test scores and their impact on children. These concerns include impacts on student learning and equity issues.

Narrowing of the curriculum

The heavy emphasis on standardized testing over the past decade has led to a narrowing of the curriculum to a focus on low-level basic skills (Hout & Elliot, 2011). In order to keep their jobs when test scores determine teacher evaluations, teachers often choose or are required to focus on test preparation. Furthermore, most schools that are facing sanctions because of high stakes testing have adopted pre-packaged, teacher-proof test preparation programs. This focus on test preparation often greatly limits or eliminates curricula rich in critical and creative thinking skills (Jones, 1999; Jones et al., 2004). MLs in particular need rich and relevant curricula that will support academic language development. MLs are under pressure to simultaneously learn content (mathematics, history, etc.) while also learning academic language. A rich and relevant curriculum allows MLs to make connections between the content and their own life experiences and provides MLs multiple entry points for learning academic language.

Disproportional Impact on High Poverty Schools

Disproportional impact on high poverty schools is an additional concern under the current growth model. According to Franco and Seidel (2012), value added models appear stable for schools that reflect the average demographics for a state. However, for schools that vary significantly in their student characteristics, significant reliability issues arise in using value-added measures for accountability purposes. Scott Elliott (2012), a reporter at the *Indianapolis Star*, examined the impact of the growth model accountability system on Indiana schools. He found that

> For the state's largest high-poverty districts, huge percentages of their schools would see their grades go down — 44 percent in IPS, 53 percent in Gary, 57 percent in Fort Wayne and 65 percent in Hammond. But large, wealthy districts had hardly any schools with grades that fell — zero in Carmel, zero in Zionsville, 12 percent in Center Grove and 20 percent in Hamilton Southeastern.

Franco and Seidel's warnings about the disproportionate impact of growth models on high poverty and schools with diverse student populations are manifested in Indiana schools. This disparity is further highlighted in Elliot's description of what is happening to School 46 in Indianapolis Public Schools:

> Under the new system, School 46 would receive less credit for the good work it does to help students overcome their significant challenges — 91 percent of its students come from families poor enough to qualify for free or reduced-price lunches (annual income of less than $42,000 for a family of four). The school would earn a bonus for raising scores, but only enough to raise its grade to a C.

Despite the fact that School 46 is showing significant growth in student performance on state standardized tests, they would still be labeled as a C school.

The disproportionate impact of the A-F system on high poverty schools also affects MLs. Fry (2008) found that at the national level, the schools in which MLs are enrolled on average have greater proportions of students living in poverty than schools with no MLs. Furthermore,

large, urban school districts in Indiana tend to be accountable for more subgroups and ML students are often identified in multiple subgroups (Burke, DePalma, Ginther, Morita-Mullaney, & Young, 2014). For example, in addition to being part of the limited English proficient (LEP) subgroup, a ML student might also be a part of the Hispanic and free/reduced lunch subgroups. Inclusion in multiple subgroups magnifies the impact that ML students have on teacher, school, and district evaluations.

The growth model system dis-incentivizes high performing teachers from working in low performing schools and working with MLs. As stated in previous sections, since only about 7-10% variability in student performance of on achievement tests can be attributed to teacher and school level factors, the context of where a teacher teaches makes a huge impact on his or her students' standardized test scores. If teachers move from high to low performing schools, they risk lower teacher evaluations, increased criticism, more hostile work environments, lower moral, and possible job loss, not because they are ineffective teachers but that their students have other issues that impact their performance on standardized tests. Under new teacher evaluation systems, teachers' annual performance and salary increases depend, in part, on student standardized test scores.

Shifting Resources Away

Indiana's A-F accountability system is based on flawed science. When NCLB was initiated, it mandated that all educational decisions be based on the US Department of Education's definition of scientifically-based research. When the research did not end up supporting the political agenda of NCLB, policymakers ignored the research. Indiana's accountability system is statistically complicated and complex enough that a layperson, a teacher, or a school administrator would likely be hard-pressed to understand how the system works in practice.

Using quantitative data and statistical models does not good science make. Hoping and believing that the Indiana status/growth models and punitive repercussions for student, teacher, and school evaluation are an effective way to ensure teacher effectiveness does not make the system valid and contradicts what statistical and behavioral science research show as good evaluation and accountability practice. The time, effort, and money spent on the A-F system, which has proven to be an

ineffective lever for school accountability, is a great loss of opportunity for Indiana's children, diverting attention away from research and development of policies that have much greater potential to improve education for all children.

Conclusion

Indiana's children deserve research-based approaches to educational evaluation, not a system based on erroneous assumptions and politics; Indiana tax payers deserve to have their tax dollars spent on effective policies that will have a positive impact on children, schools, and communities. For more than a decade, the reward and punishment policies of standardized-test based accountability have been failed policies for MLs. Continuing to implement the same system of rewards and punishment will not improve educational outcomes, especially for MLs. For teacher and school evaluation systems to be useful tools in informing school improvement efforts, the data gathered and analyzed must be meaningful. The current use of the status and growth models is not measuring teacher and school effectiveness in a statistically significant way because other non-school related factors are influencing test score outcomes to a much greater extent than are school level factors. These factors include, but are not limited to, language difference, cultural difference, and poverty-related factors for teacher and school quality. Punishing or firing educators and closing schools due to the test scores of their students is not going to address these underlying issues. Instead, policymakers need to find ways to provide more support for families and neighborhoods that are facing these challenges. Furthermore, we need to make high stakes decisions about educating our students based on multiple forms of assessments, including a much heavier emphasis on authentic and performance assessments.

ABOUT THE AUTHORS

Dr. Katie Brooks teaches courses in English as a new language and literacy in the College of Education at Butler University. Dr. Brooks is a former English as a Second Language teacher in Indianapolis Public Schools and earned her doctorate from Kansas State University

in Curriculum and Instruction with a specialization in academic and language instruction for second language learners. She has delivered more than 300 presentations at conferences and workshops throughout the United States. Her work on academic instruction for second language learners and school change has appeared in peer-reviewed journals including *Theory into Practice, Multicultural Education,* and *American Secondary Education*.

Inquiries should be directed to kbrooks@butler.edu.

Brooke Kandel-Cisco, Ph.D., is Assistant Professor of ESL and Director of the Master's in Effective Teaching and Leadership Program in the College of Education at Butler University. A former ESL and bilingual teacher, she earned her Ph.D. In Educational Psychology from Texas A&M University. Brooke's research interests include the use of home languages in K-12 education, evaluation of educational programs serving culturally and linguistically diverse students, and teacher research and empowerment. Brooke has published work in Research in Middle Level Education, International Journal for TESOL and Learning, and several edited books.

Inquiries should be directed to bkandel@butler.edu.

REFERENCES

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231-257. doi:10.1207/S15326977EA0803_02

Abedi, J. & Dietel, r. (2004). *Challenges in the No Child Left Behind Act for English Language Learners*. (CRESST Policy Brief No. 7). Los Angeles, CA: National Center for Research in Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/policy/cresst_policy7.pdf

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

American Statistical Association (2014). *Executive Summary of the ASA Statement on Using Value-Added Models for Educational Assessment.* Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A. & Collins, C. (In press). *The SAS education value-added assessment system (EVAAS): Its intended and unintended effects in a major urban school system*. Tempe, AZ: Arizona State University.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. Educational *Measurement: Issues and Practice, 28*(4), 42-51. doi: 10.1111/j.1745-3992.2009.00161.x

Burke, A. M., DePalma, G., Ginther, A., Morita-Mullaney, T., & Young, J. W. (2014). Accountability lessons for Indiana schools serving English learners. Manuscript submitted for publication.

Clewell, B., Cosentino de Cohen, C., & Murray, J. (2007). *Promise or Peril?: NCLB and the Education of ELL Students*. Washington, DC: The Urban Institute.

Cole, S., Murphy, H., Rogan, P., & Eakes, S. (2013). Indiana's teacher evaluation legislation: Implications and challenges for policy, higher education, and professional development. *Education Policy Brief, 11*(3), 1-18.

Coleman, J. (1966). *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.

DeCapua, A., & Marshall, H. W. (2010). Students with limited or interrupted formal education in US classrooms. *The Urban Review, 42*(2), 159-173. Doi: 10.1007/s11256-009-0128-z

Diaz-Bilello & Briggs, (2014). Using student growth percentiles for Educator Evaluations at the Teacher Level. Center for Assessment & CADRE. Retrieved from http://www.nciea.org/publication_PDFs/GrowthPercentileReport%20EDB073114.pdf

Elliott, S. (2012, February 24). Grading system likely to hurt high-poverty schools most. *The Indianapolis Star*. Retrieved from http://icpe2011.com/uploads/Grading_system_likely_to_hurt_high-poverty_schools_most___The_Indianapolis_Star___indystar.pdf

Fry, R. (2008). *The role of schools in the English language learner achievement gap*. Washington, DC: Pew Hispanic Center. Retrieved from http://files.eric.ed.gov/fulltext/ED502050.pdf

García, E. E., & Frede, E. C. (2010). *Young English Language Learners: Current Research and Emerging Directions for Practice and Policy.* Early Childhood Education Series. New York, NY: Teachers College Press.

Gong, B., Perie, M., & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under* No Child Left Behind: *An update to inform design decisions*. Center for Assessment. Retrieved from http://www.nciea.org/publications/GrowthModelUpdate_BGMAPJD07.pdf

Franco, M. S., & Seidel, K. (2012). Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society, 44*(1), 1-29. Doi: 10.1177/0013124511432306

Heubert, J.P., & Hauser, R.M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hiller, S. C., DiTommaso, A., & Plucker, J. A. (2012, Fall). The evolution of Indiana's school accountability system. *Education Policy Brief, 10*(5), 1-16. Bloomington, IN: Center for Evaluation & Education Policy.

Hout, M. & Elliott, S., (Eds.). (2011). *Incentives and Test-Based Accountability in Education*. National Research Council of the National Academies of Science. Washington, DC: The National Academies Press, 2011.

Indiana Department of Education. (nd). P.L. 221. Retrieved from http://www.doe.in.gov/accountability/pl-221

Indiana Department of Education. (nd). Indiana Academic Standards. Retrieved from http://www.doe.in.gov/standards.

Indiana Department of Education. (2014a). *Equity for English Learners: Imagining the Possibilities and #MakingItHappen*.

Indiana Department of Education (2014b). Corporation Enrollment by Special Education and English Language Learners (ELL). Retrieved from http://www.doe.in.gov/accountability/find-school-and-corporation-data-reports

Indiana Department of Education (2011). Model salary schedule. Retrieved from http://www.doe.in.gov/sites/default/files/educator-effectiveness/modelsalaryschedulenarrative.pdf

Jones, B. D., & Egley, R. (2004). Voices from the frontlines: Teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives, 12*(39), 1-29. Retrieved from http://epaa.asu.edu/epaa/v12n39/

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81*(3), 199-203.

Linn, R. (2004). Linn, R. L. (2004). *Rethinking the No Child Left Behind accountability system*. Paper presented at a forum sponsored by the Center on Education Policy, Washington, DC. Retrieved February 4, 2005 from http://www. ctredpol.org/pubs/Forum28July2004/BobLinnPaper.pdf

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

OECD (2010). *PISA 2009 Results: Learning Trends: Changes in student performance since 2000 (Volume V)*, PISA, OECD Publishing. DOI: 10.1787/9789264091580-en

Rivkin. S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, United States Department of Education.

The National Commission on Excellence in Education. (1983, April). A Nation at Risk: The Imperative for Educational Reform. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/title.html

U.S. Department of Education. (2012). ESEA Flexibility. Retrieved from https://www2.ed.gov/policy/eseaflex/approved-requests/in.pdf

U.S. Department of Education. (2013). National Center for Education Statistics, Common Core of Data (CCD), *Local Education Agency Universe Survey, 2002-03 through 2011-12*.

U.S. Department of Education. (2014). ESEA Flexibility Extension Letter. Retrieved from https://www2.ed.gov/policy/eseaflex/secretary-letters/in2extltr82014.pdf

Welner, K. G. (2005). Can irrational become unconstitutional? NCLB's 100% presuppositions. *Equity & Excellence in Education, 38*, 171-179. doi: 10.1080/10665680591002470

Young, J. W. (2008). Content tests for English Language Learners. *R & D Connections, 8,* 1-7.