NCSEAA
North Carolina State Education Assistance Authority

Study of Opportunity Scholarship Student Evaluations

North Carolina State Education Assistance Authority

March 1, 2018

NCSEAA
North Carolina State Education Assistance Authority

STUDY OF OPPORTUNITY SCHOLARSHIP STUDENT EVALUATIONS
Section 10A.6 of Session Law 2017-57

INTRODUCTION

In Section 10A.6(a) of Session Law 2017-57, "Current Operations Appropriations Act of 2017," the General Assembly required that the State Education Assistance Authority (Authority) establish a task force to study the evaluation of students receiving scholarship grants through the Opportunity Scholarship Grant Program pursuant to G.S. 115C-562.7(c). The Authority collaborated with the Department of Administration, Division of Nonpublic Education, and the Department of Public Instruction to complete this study.

In addition to representatives from the three collaborating organizations, Task Force members included representatives from nonpublic schools of varying sizes, organizations representing parental school choice (specifically Parents for Educational Freedom in North Carolina), organizations representing nonpublic schools, two independent research organizations, and the public schools, as set out in the legislation. A list of Task Force members appears at the end of this report.

The Task Force met four times (October 23, 2017; November 29, 2017; January 24, 2018; and February 21, 2018) to consider the challenges to and possible methodologies for comparing learning gains or losses between participants in the Opportunity Scholarship Program and selected North Carolina public school students.

TASK FORCE CHARGE

The task force shall study the most effective, valid, and reliable method of evaluating learning gains or losses of students receiving scholarship grants and comparing the learning gains or losses of those students to public school students with similar socioeconomic backgrounds, including the potential for adoption of a nationally normed common test for students participating in the evaluation. In doing so, the task force shall also consider the most reliable manner of establishing causal relationships to student performance outcomes while achieving minimal interference with the operation of the participating nonpublic and public schools, including limited sampling and other suitable research design methods. -- S.L. 2017-57, SECTION 10A.6.(b)

The following terms or phrases from the statutory language are defined at the end of this report: *effective, valid, reliable, comparing learning gains or losses, nationally normed test, causal relationship, sampling,* and *concordance*. The Task Force agreed to these explanations in order to begin discussion of the assignment with a common language. In addition to a definition of the term

causal relationship, relevant challenges to establishing a causal relationship are explained. All terms except for *concordance*, which emerged during Task Force discussions, occur in the language of Section 10A.6 of Session Law 2017-57.

RECOMMENDATION

The Task Force wishes to emphasize to the legislature the commonly held desire among all Task Force members to offer significant positive and measurable educational opportunities for students.

The Task Force has determined that meeting the statutory charge to provide "the most effective, valid, and reliable method" of evaluation and "the most reliable manner of establishing causal relationships to student performance outcomes" requires the following:

1. Evaluators select a random sample of students enrolled in a nonpublic school receiving the Opportunity Scholarship and a random sample of eligible students who applied for the Opportunity Scholarship but were not awarded funds and remained in public school. This second group of students is a comparison group.

2. Both groups of students take the same nationally normed test. Students in the study sample take the test a minimum of two times: once during the Opportunity Scholarship application process to establish baseline performance and again after one year to assess academic growth. To track trends over time in student achievement, additional test administrations (i.e., annually after the initial two test administrations) would be necessary.

3. Once Condition 1 and Condition 2 are met, researchers then will be able to compare the test score performance of students receiving Opportunity Scholarship funds to students in the comparison group and assess the causal impact of the Opportunity Scholarship Program.

This evaluation design is the *most* effective, valid, and reliable design possible, given the context, and would provide the *most* valid and reliable method of measuring a potential causal relationship between students' use of the Opportunity Scholarship and their performance outcomes.

There is one major impediment, as well as other important considerations.

To date, the applications for an Opportunity Scholarship have not yielded sufficient numbers of qualifying non-awarded students to produce a comparison group (i.e., more applicants than

available scholarships). Most of the eligible applicants in grades 2 and higher were offered scholarships.

Other considerations include:

- The public and nonpublic school communities discourage a mandate to administer an additional required test to students that is different from currently administered tests.
- Both the public and the nonpublic schools expressed the need for a different type of evaluation that captures the broader value of the school learning experience that cannot be measured by a nationally normed test alone. A richer approach to an evaluation, including qualitative measures, may be more useful for stakeholders and of interest to policymakers, although potentially more expensive. The Task Force determined that development of these other recommendations was outside the scope of its assignment.
- Costs of tests, third-party test administration, and operationalizing a causal study in an educational environment are significant. An appropriation should be included in any potential legislation to prevent a financial burden on participating schools.
- Administration of tests will incur modest disruption to both public and nonpublic schools for random samples of students.
- The nonpublic schools note that individual assessments used by their schools and submitted to the Authority are carefully selected to be the best measure of the schools' course of study. An imposed assessment may not accurately measure learning gains of students if that test is not aligned with the course of study.

In conclusion, by stipulating that the Task Force shall "consider the most reliable manner of establishing *causal relationships* to student performance outcomes," the legislature has established a high bar for the evaluation. Other less rigorous evaluations may be attainable, albeit with the caveat that those outcomes – though still informative – will be less rigorous as well.

DEFINITION OF TERMS

1. **Effective**: An effective evaluation method is one that fulfills its specified function. Note that *effective* in this case refers to the evaluation method, not to the program or to schools.
2. **Valid**: A valid evaluation method supports the types of inferences the researchers hope to draw. A valid evaluation means that the researchers are collecting the evidence needed to support the inferences or conclusions they wish to make from test scores.
3. **Reliable**: A reliable evaluation method is one that will consistently measure what the method intends to measure.
4. **Comparing learning gains or losses**: A research design must compare test scores from at least two time points to determine the magnitude and direction of any changes in student outcomes.
5. **Nationally normed test**: A nationally normed test (such as IOWA Test of Basic Skills ITBS or Terra Nova) is designed to focus on the comparison of a student's performance to that of a nationwide group of students who completed the same instrument. In contrast, a criterion-referenced test (such as the North Carolina public school End-of-Grade and End-of-Course tests) is designed to measure mastery of specific content knowledge and is not designed to allow a researcher to draw comparisons with other students.
6. **Causal relationship**: To identify a causal relationship, researchers must estimate how people who participated in a program fared compared to how they would have fared if they had not participated in the program. This hypothetical condition is called the counterfactual. Although the counterfactual is never directly observed (because an individual either experienced the program at a particular point in time or they did not), the research design needs to approximate the counterfactual in order to test whether or not a causal relationship exists.
   Two challenges for researchers regarding establishing a causal relationship:
   a. Many factors other than the program can influence outcomes over time. Researchers cannot simply measure outcomes before and after the program. Therefore, researchers must track outcomes for program participants and for a comparison group.
   b. Based on differences in motivation, availability of information, and other factors, people who sign up for a program are systematically different from those who do not sign up for it. As a result, researchers must guard against what is referred often to as "selection bias." Selection bias occurs when the researcher attributes differences in outcomes (positive or negative) to the program being evaluated, when those outcomes are, in fact, attributable to pre-existing differences in people who self-selected to participate in the program and those who did not. Therefore, researchers must identify treatment and comparison groups that, on average, have similar observed and unobserved characteristics.

7. **Sampling**: A sample is a representative subset of a whole population selected for the purposes of drawing inferences about the population as a whole.
8. **Concordance**: The degree of agreement between two measures, such as two different tests, is known as concordance. Reliable test concordance can be established only when the same group of participants take the two tests that are being compared, or when a sufficiently large sample of students takes each test. In the absence of a well-designed concordance study, there may not be sufficient power to detect a statistically significant difference even if one exists.

TASK FORCE MEMBERSHIP AND AFFILIATION

Mr. Don Adams	Berean Baptist Academy
Ms. Diane Allen	Division of Nonpublic Education
Dr. Anna Egalite	North Carolina State University
Ms. Allison Guenther	Parents for Educational Freedom in North Carolina
Dr. Joe Haas	North Carolina Christian School Association
Dr. Rita Haire	Parents for Educational Freedom in North Carolina
Dr. Laura Knapp	RTI International
Ms. Linda Nelson	North Carolina Association of Independent Schools
Dr. Kathryn Marker	State Education Assistance Authority
Ms. Elizabeth McDuffie	State Education Assistance Authority
Dr. Bradley McMillen	Wake County Public School System
Dr. Perry Nicklow	Wesleyan Christian Academy
Dr. John Pendergrass	Catholic Diocese of Raleigh
Mrs. Cheryl Riley	Victory Christian Center School
Dr. Trip Stallings	North Carolina State University
Dr. Tom Tomberlin	Department of Public Instruction