THE EVALUATION OF "THE WRONG ANSWER PROJECT" AS VALIDITY EVIDENCE FOR THE SOCIAL CONSEQUENCES OF TESTING

Darius D. Taylor
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Educational Assessment, Evaluation, and Research Commons, Other Theatre and Performance Studies Commons, and the Social Justice Commons

Recommended Citation
https://doi.org/10.7275/34925415 https://scholarworks.umass.edu/dissertations_2/2863

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
THE EVALUATION OF “THE WRONG ANSWER PROJECT” AS VALIDITY EVIDENCE FOR THE SOCIAL CONSEQUENCES OF TESTING

A Dissertation Presented
by
DARIUS DION-DWAYNE TAYLOR

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY
May 2023
College of Education
Research, Educational Measurement, and Psychometrics
THE EVALUATION OF "THE WRONG ANSWER PROJECT" AS VALIDITY EVIDENCE FOR THE SOCIAL CONSEQUENCES OF TESTING

A Dissertation Presented

by

DARIUS DION-DWAYNE TAYLOR

Approved as to style and content by:

_______________________________________
Lisa A. Keller, Chair

_______________________________________
Judyie Al-Bilali, Member

_______________________________________
Sharon Rallis, Member

_______________________________________
Jamila Lyiscott, Member

_______________________________________
Shane Hammond
Associate Dean for Student Success
College of Education
DEDICATION

To my people.
ACKNOWLEDGEMENTS

Thank you, God, for your grace and FAVOR! MOM! We started from the bottom now we here! Hope you proud of this fruit of your loins. It’s you and Paw Paw that showed me what hard work and hustle looks like. Pops, thanks for movin us up to the east side, so we could finally get a piece of this pie. Dwayne, thank you for your love, joy, and creativity. Troyce, thank you for your “Y”, brainz, and genetics. Uncle Garry, thanks for forcing me to take practice ACT tests with the white folk and showing me there is life outside of Lake County. Gram, Godmommy-TT-Ant Terra, and Ms. Rosie thank you for your prayers! Bro. Marlon, Uncle Ken, Auntie Ife, and Auntie Tamar I appreciate just watching you to see #BlackExcellence. Danyelle, Chandrea, and DayChan thank you for seeing me and encouraging me. Godmommy Auntie Cheryl and Goddaddy Uncle Carlus thank you for your great exhortations speaking into my potential. To my academic/professional mentors: Roberta Donahue, Raquel Qualls-Hampton, Emily Spence, Osayi Akinbosoye, Zhuo Jiang, Anthony Petrosino, Lisa Keller, Jennifer Randall, Iya Judyie, Steve Sireci, Sharon Rallis, Jamila Lyiscott, and Erica Moore, so much gratitude for nurturing, opportunities, pushing, inspiring, rectifying, teaching and protecting me! To my peer mentors: Trevor Fronius, Joshua Marland, Molly Faulkner-Bond, and Raifu Durodoye, I appreciate your influence/observational learnings. To the homies: Garry Scott, Bestie, Lily, Chrissy, DreamGyal, Funmi, Mikey, Shawny, Mari, Sabi, WW, CheChe, V, CC, & Sungy, ya’ll the REALEST. To my spiritual parents: Pastors Aaron & Mercy Myles-Jenkins and Apostle James Duncan thank you for imparting in me. And most of all my honeybuttercup and visionary MoMo, I love you!
ABSTRACT

THE EVALUATION OF “THE WRONG ANSWER PROJECT” AS EVIDENCE FOR THE SOCIAL CONSEQUENCES OF TESTING

MAY 2023

DARIUS DION-DWAYNE TAYLOR, B.S., TRUMAN STATE UNIVERSITY

MPH, UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa A. Keller

The use of educational tests for making high stakes decisions has had societal consequences for decades. Parents, teachers, and administrators have been willing to pay off, lie, cheat, and steal so that their children, students, and they themselves would not fall prey to the negative consequences of subpar performance on educational assessments. Respected psychometric scholars have supported Samuel Messick’s claim over the years, but their advocacy has caught minimal traction. I founded an initiative in 2019 – The Wrong Answer Project – that shows promise as a vehicle for collecting validity evidence based on the social consequences of testing and raising the voices of perpetually marginalized groups. My hope is to ignite grassroots solutions for achieving more equity centered designs in educational testing. This study is an investigation of the usefulness of my initiative by exploring the pilot data from the 2019 implementation using an evaluability assessment approach. The findings from the assessment supports strategy that evolves my project to a more manageable product in service of social justice for all who have been oppressed by high stakes tests.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>V</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>VI</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>VIII</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>IX</td>
</tr>
</tbody>
</table>

1. **INTRODUCTION** | 1 |
| BACKGROUND | 4 |
| HISTORY | 16 |
| My Response | 17 |
| The “Intervention” | 21 |
| The Wrong Answer Project as Program/Intervention | 22 |
| Phase 1: Development and Planning | 23 |
| Phase Two: Implementation | 25 |
| Phase 3: Dissemination | 29 |
| Analysis | 29 |
| Resulting Data | 30 |
| A Vision Forward | 33 |

2. **METHODS** | 34 |
| Research Questions | 34 |
| Evaluability Assessment | 34 |
| Validity Evidence | 39 |

3. **RESULTS** | 46 |
| Insight Team: Key Player & Expert Roles | 46 |
| Clarify the Program Design | 51 |
| Explore the Program Reality | 55 |
| Assess the Plausibility of the Program | 68 |
| Reach Agreement on any Needed Changes in Program Design or Implementation | 76 |
| Reach Agreement on the Focus and Intended Use of any Further Evaluation | 83 |

4. **DISCUSSION & NEXT STEPS** | 95 |

**APPENDICES**

A. **Wrong Answer Intervention Protocol** | 105 |
B. **Testing Perceptions Survey** | 127 |
C. **Facilitation Protocol** | 128 |

**BIBLIOGRAPHY** | 130 |
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Graffiti Writing Themes</td>
<td></td>
</tr>
<tr>
<td>2. Group Dialogue Themes</td>
<td></td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Feedback representation of unified test validity.</td>
<td>6</td>
</tr>
<tr>
<td>2. ACT Proficiencies, AY 19-20</td>
<td>13</td>
</tr>
<tr>
<td>3. ACT Proficiencies, AY 18-19</td>
<td>14</td>
</tr>
<tr>
<td>4. SAT Proficiencies, AY 19-20</td>
<td>15</td>
</tr>
<tr>
<td>5. SAT Proficiencies, AY 18-19</td>
<td>15</td>
</tr>
<tr>
<td>6. W.A.P. Insights Team - Key Players</td>
<td>47</td>
</tr>
<tr>
<td>7. W.A.P. Water Cycle Evaluation</td>
<td>87</td>
</tr>
<tr>
<td>8. Ms. Engel Prepares Class for Test Day</td>
<td>97</td>
</tr>
<tr>
<td>9. W.A.P. 2019 Pilot Playbill &amp; Test</td>
<td>97</td>
</tr>
<tr>
<td>10. Audience Graffiti Reflection</td>
<td>98</td>
</tr>
<tr>
<td>11. Audience Graffiti Response to Prompts</td>
<td>99</td>
</tr>
<tr>
<td>13. Test Time Spoken Word Highlight</td>
<td>100</td>
</tr>
<tr>
<td>14. Audience Solutions from W.A.P. 2019 Pilot</td>
<td>100</td>
</tr>
<tr>
<td>15. Actor Stuck on Test Spotlight</td>
<td>101</td>
</tr>
<tr>
<td>16. Audience Graffiti Reflection</td>
<td>101</td>
</tr>
<tr>
<td>17. Movement Expression On Test Day</td>
<td>102</td>
</tr>
<tr>
<td>18. Decent into Enlightenment</td>
<td>103</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

The consequences of high stakes testing have been a prevalent topic amongst societal issues for decades. In March of 2019, over 50 people (including well known celebrities) were convicted of a white-collar crime – racketeering. This involved them privately conspiring to commit fraud and manipulate testing administration or scores just so their children could gain admittance to elite colleges and universities. The parents paid very large sums of money to weigh the odds of a high stakes testing outcome in their child’s favor. In the greater scheme of things, these were mostly white affluent families attempting to retain their elite status by insuring the future of their offspring. A decade prior in 2009, 178 teachers and principals from 44 Atlanta Public Schools were identified for manipulating test scores on the Criterion Referenced Competency Test (CRCT). All to avoid the consequences that the 2001 No Child Left Behind accountability system placed on poor student performance. School staff used their positions of power to influence high stakes testing outcomes in their favor, regardless of the downstream effects it would have on their students. Both are examples of issues in society that resulted from the use of a high stakes test for a particular purpose.

According to the Glossary of Education Reform,

“A high-stakes test is any test used to make important decisions about students, educators, schools, or districts, most commonly for the purpose of accountability—i.e., the attempt by federal, state, or local government agencies and school administrators to ensure that students are enrolled in effective schools and being taught by effective teachers. In general, “high stakes” means that test scores are used to determine punishments (such as sanctions, penalties, funding reductions, negative publicity), accolades (awards, public celebration, positive publicity), advancement (grade promotion or graduation for students), or compensation (salary increases or bonuses for administrators and teachers).” (High Stakes Testing, 2014)
From both examples above, one may inquire that if parents, administrators, and teachers were willing to go to such criminal extremes to obtain the favorable outcome of successful test performance, was the actual flaw in the use of these tests for such a high stakes purpose? The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), hereafter referred to as the Standards, describe five sources of validity evidence that should be used to validate a test score for a particular purpose. These are: validity evidence based on test content, response processes, internal structure, relations to other variables, and the consequences of testing. The inclusion of the fifth source was influenced by Samuel Messick (1989) who, in defining validity as a unitary concept, detailed a component of validity he called “the consequential basis of test use.” (Often times referenced in the literature as consequential validity). Messick pointed out that in validating a test score for a particular purpose, it is important to consider the social consequences (whether intended or unintended) of its use. He claimed that “the appropriateness of the intended testing purpose and the possible occurrence of unintended outcomes and side effects are major issues” (p. 85).

Whether one is on the job, in the classroom, or at home, every system of human life has a testing, assessment, or evaluation component. As a diverse and inclusive society in America, to remain unbiased in our assessment practices, we must continue to evaluate the social impact of testing and the values laden within test score interpretation. But how

---

1 Neither the AERA et al. Standards nor Messick support the use of the term “consequential validity, and I concur.
do we evaluate these values and their impact? Messick (1989) presented a feedback loop (presented in Figure 1) for test validation that urges developers to consider social implications at multiple steps in the process. He presents a dynamic and interactive process for validating the use of a test for a particular purpose. This study focuses on the component of the feedback loop that encourages test constructors to evaluate social consequences as implications for test use.
Background

Messick’s feedback representation of unified test validity (Figure 1) begins with an iterative process of construct validity evidence, which leads to efficient score interpretation. If score interpretation aligns with values (and assumptions) associated with the defined construct, then the score (and test) can be considered for a proposed use. This is phase one of the feedback loops. It is essential for scores to meaningfully and appropriately quantify the defined construct. If score interpretation does not align with values (and assumptions), further construct evidence is required.

Phase 1: Test interpretation & Value Implications

During test construction the desire is to create a test that produces a score that is aligned with the values of the construct to be measured. According to Messick (1989) the values of the construct should be represented in: the labels placed on certain scores (or score intervals); the theory used to claim that the score has certain meanings with respect to the construct; and the ideology used to claim that the construct can be measured by a test and the score derived will be meaningful in the valuation of that construct. Messick (1989) claimed that values laden in test score interpretations can be explored via a dialectical approach between two or more scientists with opposing views, or through a single scientist presenting rivaling perspectives. In both cases rhetoric is produced that contributes to the conversation of the values laden in test scores. As a part of the validation feedback loop, Messick highlights the need to use this evidence as a sounding board for making the decision to proceed forward to use the test for its proposed purpose, or to stop and re-evaluate the testing purpose and/or the approach to how the construct is being measured.
In the event that further construct validity evidence is required, it may be necessary to reimagine the theory derived for the construct from its respective nomological network. Cronbach and Meehl (1955) stated that “when observations will not fit into the network as it stands, the scientist has a certain freedom in selecting where to modify the network. That is, there may be alternative constructs or ways of organizing the net which for the time being are equally defensible.” (p. 290) For example, for a test that is constructed for the purpose of measuring an 8th grader’s proficiency in geometric functions, the test constructor aims to produce a score from the items that can be interpreted as varying levels of proficiency in geometric functions. However, in evaluating what Kane (1992) identifies as the “interpretive argument” test scores may not map back to geometric proficiency, but instead the ability of students to follow rote memorization instruction from their teacher.
Figure 1. Feedback representation of unified test validity.

From Messick (1989)
While it may not be the desire of the test developer, to utilize a test for a meaningful purpose, it may behoove the scientist to repurpose the test or rewrite items more conducive to measuring proficiency in geometric functions that also taps into components of the nomological network more proximal to the construct of interest. Such choices in defining the construct are likely to have direct consequences for instruction, and thus also for student learning.

*Phase 2: Relevance & Utility*

Once a test is considered for a proposed use, a similar feedback loop is employed to assess whether test scores are defensible with respect to relevance and utility for an applied purpose. This is the second phase in the feedback loop. The investigation of relevance and utility has the potential to span a wide array of techniques given the proposed hypothesis or interpretive argument of the test score. These techniques are employed to capture convergent, discriminant, and predictive validity evidence that informs the decision as to whether a test or test score is defensible for a proposed use.

For example, if a 7th grade English teacher claims that her classroom assessments foster learning that increases Lexile scores for her students due to the “positive testing effect,” validity evidence should point to the test score as the greatest indicator of Lexile score achievement. However, if relevance and utility are not defensible, for example, validity evidence may reveal that the number of books read above the student’s grade level was the greatest indicator of Lexile achievement, then the proposed use of the test or methods of relevance and utility inquiry are re-assessed. If these effects are confirmed (proving classroom assessments defensible), social consequences are evaluated.
**Phase 3: Social Consequences**

The third phase of Messick’s feedback representation of unified test validity investigates the societal implications of the test (or testing system) being used for a particular purpose. Are there differential attitudes, knowledge, and behaviors across subgroups in society as a result of this test being used for the particular purpose? If no negative social consequences exist, it is safe to assume the test can be used for a particular purpose.

Although Messick’s unified validity approach is well recognized and respected (e.g., Kane, 2006; Lane, Parke, & Stone, 1998; Poplum, 1997; Shephard, 1993), the extent to which social implications and intrinsic values are actually considered during test validation efforts is unclear. The discovery of negative social consequences should require researchers to loop back to understanding the proposed use and defensibility with respect to relevance and utility. However, as Engelhard (2014) and Haertel (2012) pointed out, many testing programs do not even consider social consequences of testing in their test validation efforts.

It is critically important to study the consequences of testing because it helps test developers and users evaluate whether test score interpretations have deviated from their intended uses (Engelhard, 2014; Haertel, 2012; Lane, 2014; Lane & Stone, 2002; Lane, Parke, & Stone, 1998; Messick, 1980 & 1989; Poggio, Ramler, & Lyons, 2018). Such an investigation leads to an understanding of the societal values placed on specific tests, which may also explain biases associated with the use of certain constructs, scores and score interpretation. As Messick (1989) described, “Values are important to [consider] in score interpretation not only because they can directly bias score-based inferences and
action, but because they could also indirectly influence in more subtle and insidious ways the meanings and implications attributed to test scores” (p.59).

Thus, considerations of potential consequences of testing begin with deciding that a test is needed and defining the construct to be measured. Whomever defines the construct determines the meaning. What happens when construct definitions are created without considering the demographics or intrinsic values of the definer (gender, race, socioeconomic status, conservative, libertarian, etc.)? A potential negative consequence of such a process is an overly narrowed definition of the construct that is centered on a particular, dominant culture, and marginalizes other worldviews. Similar to what Randall (2021) posits in respect to the White majority’s historic hegemonic influence on construct definition. While the issue of values associated with test score interpretation is not the direct focus of this paper, in test development it is important to understand the downstream effects that values have on test score interpretation and use. Test developers assess value by identifying: the implications of labeling a construct (claiming that it represents a particular idea), the use of a particular theory as a source for construct and item development (in the realms of competing theories), and the extent of societal belief in the measurement theories (just because it’s been theorized does not mean society agrees with this particular epistemology).

Messick (1989) stated “even if adverse testing consequences derive from valid test interpretation and use, the appraisal of the functional worth of the testing in pursuit of the intended ends should [consider] all of the ends, both intended and unintended, that are advanced by the testing application, including not only individual and institutional effects but societal and systemic effects as well” (p. 85). Lane (1998) mentioned “the evaluation...
of the plausible negative effects can begin with, but is not limited to, the examination of potential misuses of the assessment and the assessment results” (p. 25). Thus, before a test user or testing program decides to place a test into production it is important to explore how test scores and the test itself could potentially be misused to avoid consequences a priori.

Calls for evaluation of the social consequences of testing were also emphasized by Cronbach (1989), who encouraged test developers to aim to capture all potential negative consequences from various stakeholders to ensure one has obtained a well-rounded perspective. Thus, it is important to consider ALL related parties in this early exploration of potential misuses to obtain diverse points of view that could contribute to this conversation. As Lane (2002) pointed out “state assessment programs designed for school accountability affect administrators, educators, students, parents, future employers, and the community. Which individuals are most affected depends on the purpose of the assessment” (p. 24). There is a clear attention to the need for multiple diverse voices in the assessment of social consequences.

Further, the assessment of social consequences lends itself to both quantitative and qualitative methodologies. In reference to Koretz, Barron, Mitchell, and Stecher (1996); Pomplun (1997); and Chudwosky and Behuniak, (1997); Lane (1998) mentioned that “researchers are now beginning to examine the consequences of assessment programs by using various procedures such as surveys of principals and teachers and focus groups” (p. 25). She continued, “there is a need for more direct evidence, such as evidence obtained through the analyses of classroom instructional and assessment activities, in addition to information obtained through surveys and interviews” (p. 25).
According to the NCME 2019 Conference Proceeding by Taylor and Sireci, not much has been fulfilled to meet this need since Lane made this call (in 1998!)

The joint usage of quantitative and qualitative methodologies allows for the triangulation of results. However, in a review of literature related to “consequential validity,” Poggio, Ramler, and Lyons (2018) suggested the jury is still out as to whether the testing community has much interest in providing sound validity evidence related to social consequences. In an exploration of over 100 publications, Poggio et al. (2018) concluded, “at this time, we recommend that strong bodies of work are [needed] to be discussed and shared among investigators to consider and enhance such validation efforts to include more formally concrete empirical consequential validation evidence whenever it can be considered, expected and justified” (p. 31).

In our NCME Poster Paper, Steve Sireci and I, took a similar approach to exploring published work focused on validity evidence for the social consequences of testing. We were interested in empirical research that employed the third phase in Messick’s feedback loop, which focuses on the appraisal of effects to society. We expanded on Poggio et al.’s methodology by identifying research published at two pivotal times related to test consequences: 1) the decade between the published unified validity theory from Messick (1989) and the release of the Standards (1999) and 2) the research published after the Standards (1999) until December 30th, 2018. Additionally, we expanded our literature search to include a review of technical manuals and publications from the major testing companies in the nation (ACT, SAT, NWEA, PARCC, and NAEP). The results of our study showed very little work has been done in this area, and the few studies that have been done provide mixed results.
There is a need for a continued call to action for test developers and measurement specialists to consider potential testing consequences from the earliest stages of test development, as Messick stipulated more than 30 years ago. The AERA et al. (1999, 2014) Standards, and many prominent validity theorists, have advocated for the evaluation of social consequences as an important part of test validation. However, the degree to which these standards and recommendations have been followed by testing practitioners is unclear, buried, disguised, and ignored. This negligence has a considerable societal impact that perpetuates marginalized educational outcomes for minority groups.

*Current State of High Stakes Testing*

There were roughly 3.6 million 12th grade students enrolled in public school in 2019 and 2020 (USDE, 2020). About 1.7 million graduating high school students take the ACT. About 2.2 million take the SAT. Both are considered high stakes assessments because the exams are often used to determine high school graduation and/or matriculation into post-secondary educational opportunities. As such, the ACT and SAT have defined benchmarks for college and career readiness based on the performance of their examinees. It is no secret that there are historic discrepancies between racial/ethnic groups in respect to educational performance (Howard, 2010).

Prior and current, achievement data tells us that Black and Brown students do not measure up to their White counterparts. Recent ACT and SAT college and career readiness benchmarks show continued marginalization between Black and Brown minorities and the White majority. Figures 2 through 5 (below) tell the story. Of the 2019 and 2020 high school graduating students who took the SAT and ACT, 45% and 26%
met college and career benchmarks according to their respective composite exam scores.

Which makes a statement that according to the SAT less than half of their graduating examinees are ready for college and career. This reduces to about a quarter of students for ACT examinees.

**Figure 2.** ACT Proficiencies, AY 19-20

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>English</th>
<th>Math</th>
<th>Science</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>54%</td>
<td>69%</td>
<td>46%</td>
<td>45%</td>
<td>33%</td>
</tr>
<tr>
<td>Black</td>
<td>19%</td>
<td>30%</td>
<td>12%</td>
<td>12%</td>
<td>6%</td>
</tr>
<tr>
<td>AIAN</td>
<td>20%</td>
<td>28%</td>
<td>13%</td>
<td>13%</td>
<td>7%</td>
</tr>
<tr>
<td>Latinx</td>
<td>31%</td>
<td>43%</td>
<td>24%</td>
<td>22%</td>
<td>14%</td>
</tr>
<tr>
<td>All</td>
<td>45%</td>
<td>58%</td>
<td>37%</td>
<td>36%</td>
<td>26%</td>
</tr>
</tbody>
</table>

*Note: AIAN – American Indian/Alaskan Native*
Figure 3. ACT Proficiencies, AY 18-19

For the 2020 graduating class who took the SAT, 20% of Black examinees and 17% of Native (AIAN) examinees met the college and career benchmarks for all subject areas compared to 56% of White examinees. Additionally, over half of the Black and Native examinees did not meet college and career readiness benchmarks in any of the SAT subject areas. The ACT reported 6% of Black and 7% of Native examinees meeting benchmarks for all the subject areas amongst the 2020 graduating class. While 33% of their White counterparts met all benchmarks. Latinx students win the award for most achieved subgroup. Most recent ACT and SAT scores report 14% and 28% of Latinx students meeting the college and career readiness benchmarks, respectively.
**Figure 4.** SAT Proficiencies, AY 19-20

SAT Proficiencies, 2019-20 Graduates

<table>
<thead>
<tr>
<th></th>
<th>Reading/Writing</th>
<th>Math</th>
<th>Composite</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>76%</td>
<td>59%</td>
<td>56%</td>
<td>19%</td>
</tr>
<tr>
<td>Black</td>
<td>44%</td>
<td>21%</td>
<td>20%</td>
<td>54%</td>
</tr>
<tr>
<td>AIAN</td>
<td>38%</td>
<td>20%</td>
<td>17%</td>
<td>60%</td>
</tr>
<tr>
<td>Latinx</td>
<td>53%</td>
<td>30%</td>
<td>28%</td>
<td>45%</td>
</tr>
<tr>
<td>All</td>
<td>66%</td>
<td>48%</td>
<td>45%</td>
<td>27%</td>
</tr>
</tbody>
</table>

*Note: AIAN – American Indian/Alaskan Native*

**Figure 5.** SAT Proficiencies, AY 18-19

SAT Proficiencies, 2018-19 Graduates

<table>
<thead>
<tr>
<th></th>
<th>Reading/Writing</th>
<th>Math</th>
<th>Composite</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>80%</td>
<td>59%</td>
<td>57%</td>
<td>18%</td>
</tr>
<tr>
<td>Black</td>
<td>46%</td>
<td>22%</td>
<td>20%</td>
<td>53%</td>
</tr>
<tr>
<td>AIAN</td>
<td>39%</td>
<td>21%</td>
<td>18%</td>
<td>58%</td>
</tr>
<tr>
<td>Latinx</td>
<td>55%</td>
<td>31%</td>
<td>29%</td>
<td>43%</td>
</tr>
<tr>
<td>All</td>
<td>68%</td>
<td>48%</td>
<td>45%</td>
<td>30%</td>
</tr>
</tbody>
</table>
History
As society reflects on the historic social impact of high stakes testing, it would be remiss to ignore the origins of current testing practices in America. The format of standardized multiple choice testing dates back to the 1800s. The United States passed the Pendleton Civil Service Act in 1883 which required government employees to be selected based on merit and sit for a comprehensive exam. It is the earliest record of high stakes decisions made from a standardized assessment in the states. The civil service exam was modeled after the Chinese system. It truly is interesting to first think about the origins of this system and who currently performs the best on standardized high stakes tests. Additionally, though the Pendleton Act occurred after the 1863 Emancipation Proclamation, my people had no real rights until the Civil Rights Act of 1963. One hundred years in the dark, in respect to opportunities limited by high stakes examinations.

During these proverbial dark ages, by 1905 the field of assessment entered the “modern era” of mental measurement with the release of the Binet-Simon scales of mental ability. According to Gallagher (1998), this scale was the “first formalized assessment based on a measure of tasks of graded complexity. The Binet-Simon scale was developed to identify students that did not have the mental capacity to benefit from standard public education. It was originally developed in French then converted to English. Later Lewis Termin, a prominent eugenicist revised the scale to what it is today. A eugenicist! I will stop there.

Like most origin stories in the U.S., the history of testing started with old White men who knew what was best for everyone. There was no seat at the table for women, Black people, Native people or anyone else who was not a White man.
My Response

I am Darius Dion-Dwayne Taylor third born to my mother youngest born of my bio-dad. My fathers are an interesting story. I have three: one biological, one that signed my birth certificate, and my stepdad (affectionately called pops or padre). I am a Black man who grew up in the hood of North Chicago, IL though at 12 years old we moved up to the proverbial “east side” middle class life when pops stepped on the scene. I’m a first-generation college student and have gotten the farthest scholastically in my family thus far. My identity is mostly defined by my faith and unconditional love I have for people. I have limited ties to my family heritage as my mom was adopted and my bio-dad is only now beginning to scratch the surface of my family history with me. There is a special place in my heart for serving Black, Brown, and Native peoples. My peoples. My service is to provide healing, love, and hope for my peoples to have opportunities to not just survive the systems of society, but thrive and succeed across all life outcomes. A big part of that success follows from education.

From the start of my doctoral studies at UMass, I had been informally capturing data from words that family and friends would talk at me when I told them I was studying educational testing. The words and narratives expressed were full of indignation urging me to put my best foot forward to change the system!

Their words were seeds planted in me.

Me – an eclectic soil nourished with love for my people and heart for change.

Partial writings of spoken word and theatrical dialogue sprung forth.

With the buddings of these writings, I turned to my art family born from my performance experiences at UMass. My art mother and Associate Professor in Theater for
Social Change, Iya Judyie Al-Bilali, and my ace, Jen Onopa, an established producer/director in devised theater and MFA-Directing Alumni. We brainstormed together how I could potentially blossom the buds that I began with the seeds from my people.

Amidst this brainstorming process, the heavens opened in the form of a grant funding opportunity. The National Council on Measurement in Education (NCME) Mission Fund Grant Program released a call for proposals during the 2018-19 academic year. The Mission Fund was established “to advance the science and practice of measurement in education, and to provide individuals and organizations with financial support for projects, research, and travel that address this mission directly.” In 2018-2019, the Mission Fund sought to fund initiatives that were “designed to promote a broader understanding of high-quality assessment practices and appropriate test use among diverse groups of assessment stakeholders.”

I distinctly recall releasing a loud vocalization of gratitude, hopeful that I could attain access to a stream of funds that would allow for me to really give back to my community (and maybe get my dissertation done in the process). It was so timely. And even more timely that the current NCME President at the time was one of the faculty in my program, Steve Sireci. He advised me to pitch my project as a scalable initiative rather than a mere dissertation project. Steve’s advice encouraged me to operationalize my envisioned societal impact with confidence that if funded my project would be the proof of concept for a greater work. My proposed project was awarded $10,000 to create and implement an arts-based intervention that interrogated the perceptions of stakeholders within educational assessment and encouraged dialogue about democratic
solutions that could influence educational policy and ultimately create better experiences in testing for all.

My NCME Mission Fund project titled, “Wrong Answer: Examining Stakeholder Voices in High Stakes Testing” was produced with great honor and gratitude. My implementation team and I operationalized the values of the Mission Fund Grant Program guided by Samuel Messick’s 1989 validity theory as a unitary concept. Specifically, we employed Messick’s proposition that validity evidence for the social consequences of testing should be evaluated in an iterative feedback loop when constructing and determining test value and use.

The 2014 Standards for Educational and Psychological Testing *Standard 1.25* states:

“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 it is intended to assess or from the test’s failure to fully represent the intended construct.”

The decades of perpetual racial and ethnic academic achievement gaps in the U.S. are signals of unintended consequences of using test scores for high stakes decisions. *Wrong Answer* interrogates whether racial gaps are construct irrelevant or construct relevant? Test sensitive or test insensitive? Construct underrepresentation? Who defined the construct? Context matters. Even if measurement is actually value-free, and unbiased and, we keep seeing gaps in achievement, why do we continue to use these tests to make high stakes decisions? Why not put the high stakes test on hold while we figure out the
rest of the system, so that we make sure it’s not the test? If anything, COVID-19 has taught us that this is possible.

Wrong Answer is the pilot concept for a large-scale social action campaign. Prof. Al-Bilali claimed Wrong Answer to be “a vision forward.” The project goals are to:

- develop stakeholder driven action steps for “better” testing practices in America
- provide an empowering space related to educational testing for marginalized groups who are characterized by race/ethnicity, economic status, and access
- ignite a grassroots movement that shifts current testing practices and policies to a more equity-centered design in testing

The goal of my dissertation work is to assess the evaluability of the Wrong Answer Project for meeting these project goals and providing validity evidence for the social consequences of testing. My approach will be informed by Strosberg and Wholey’s eight guiding questions for assessing the basic elements of evaluability (1983) and Kane’s argument-based approach to validity (1992).
The “Intervention”
An excerpt, from Ntozake Shange’s, Spell #7:

“my father is a retired magician
which accounts for my irregular behavior
everything comes out magic hats
or bottles wit no bottoms & parakeets
    are as easy as a couple a rabbits
or 3 fifty-cent pieces/1958
my daddy retired from magic & took
up another trade cuz this friend of mine
from the 3rd grade/ asked to be made white
on the spot

what cd any self-respectin colored American magician
do wit such an outlandish request/ cept
put all them razzamatazz hocus pocus zippity-doo-dah
thingamajigs away       cuz
colored chirren believing in magic
waz becoming politically dangerous for the race
& waznt nobody gonna be made white
on the spot just
from a clap of my daddy’s hands
& the reason im so peculiar’s

cuz I been studyin up on my daddy’s technique
& everything I do is magic these days
& its very colored/ very now you see it/ now you
dont mess wit me”
In explaining Aristotle’s principal that “art imitates nature”, Augusto Boal wrote that, “art re-creates the creative principal of created things” (Boal 1985, pg. 1). Thus, imitation in this sense is the action of taking matter (the components of a given substance) and manipulating it into a form with a purpose only as perfect as the matter manipulated and method of manipulation. For example, a dramatized stage play is only as true to nature as the words chosen to form the story and the way in which the words are formed and performed. Art as a vehicle for social transformation invokes this principal of imitation to allow for art experiencers to re-live forms of reality that Boal mentions can intimidate and purify (1985). The art of theater can serve as a mechanism for examining a substance derived of matter based in current social realities. In re-living this substance through theater an audience can be implicated to realize the challenges facing society and desire better societal outcomes.

The NCME Mission Fund Project titled “Wrong Answer: Examining Stakeholder Experiences in High Stakes Testing” is a theater experience that embodies Boal’s teachings dressed in approaches innovated from the project advisor, Theater for Social Change Prof. Judyie Al-Bilali. The community intervention, as it was often referenced, was a devised theater experience that incorporated elements of the U.S. patented applied theater approach of Brown Paper Studio to provide an introspective experience to explore how marginalized groups have been socialized to educational testing and its effects in the home, school, community, and work environment. The goal of the project was to implicate the theater audience with an experience that offers social transformation and
healing in respect to testing and the negative connotation most examinees have with testing.

The specific objectives of this project were:

1) To develop an adaptable theater script that can be manipulated to represent any marginalized groups’ perspectives and experiences in high stakes testing.

2) To facilitate an interactive experience focused on the discourse of high stakes testing perceptions amongst marginalized groups.

3) To develop replication materials that allows for public use of the project as a method for contributing more marginalized perspectives to this discourse.

The project objectives were pursued across three different phases.

Phase 1: Development and Planning

*Script Development*
*Assemble Production Team*
*Casting*
*IRB Application*

Phase 2: Implementation

*Rehearsals*
*The Studio*
*Videography*
*Video Permission*
*Dialogue Facilitator Training*
*The "Intervention"*

Phase 3: Dissemination

*Video Production*
*Public Reporting*
*Website*

The script (Appendix A) was developed using a democratic process that ensured multiple voices contributed to the final product. My first draft of the theatrical experience was developed from rich conversations with a convenient sample of relatives, friends, students, and colleagues. From this point the goal for the pilot of the project was to
represent the perspectives of both Black and Latinx students from low socio-economic communities within the script. My project advisor suggested that I represent the pilot from the perspective of my truest reality. *The Wrong Answer Project* is available to the public for future iterations to represent other marginalized voices, so I used the pilot to represent the low socio-economic African-American experience.

After my second draft, the script was presented to an Honors Brown Paper Studio class of 8 students of various identities in the Spring of 2019. The class performed a cold read of the script then dialogued and offered suggestions for changes to the script. This same process was employed with 10 Upward Bound high school students and staff of various races at the University of Massachusetts -Amherst Summer 2019. After casting for the premier, the cast also dialogued and shared suggested edits for the script before it was finalized mid November 2019. Some manipulation to the script persisted even to the day of dress rehearsal which truly made this script development a devised experience.

*Production Team Assembly*

Majority of the budget was allocated to the team of designers and personnel which made the pilot a success. The team included: a production manager, theatrical performance director, stage manager, project advisor, costume designer, lighting designer, scenic designer, sound designer, videographer, video editor, creative director, and dialogue facilitators. Each of these team members had contracts drawn for the commission of their service. My multifaceted role as the producer, playwright, music director and choreographer was pro bono. The remainder of the grant funds were allocated to other production needs.
Casting

The production team hosted three official auditions between September 30th and October 15th, 2019. Some auditions were submitted virtually. The cast included four principal roles and eight ensemble members. Rehearsals were two times a week starting October 30th, 2019. The production dates were December 6th and 7th, 2019.

IRB Application

The project was submitted to IRB and determined not necessary for review due to the minimal risk and anonymous method of data collection.

Phase Two: Implementation

Rehearsals

Rehearsals continued for twice a week from October 30th until December 1st. The week leading up to production is called Tech Week. It occurred from December 1st to December 5th. Rehearsal was every evening for at least four hours during Tech Week. Activities included an intensive que to que mapping of all production elements (lights, sound, costume, props, scenic, acting, video). Tech week will often include full run throughs of the entire experience from open to close in preparation for the premier date which occurred on December 6th.

The Studio

The New Africa House Studio at The University of Massachusetts-Amherst was used to host rehearsals and premier the production. The studio space includes light installment options, a full sound system, and projection capacities. There is a moveable stage with curtains. The studio seats 150 patrons. There was an array of installments
within the studio. Each installation contributed to the overall intervention experience. These installations included scenic design elements, costumes, lighting, sound, and video. Throughout the rehearsal process commissioned designers visited the space to consult with the theatrical director and producer on the technical elements. The design team met at least monthly as a group and independently with the producer as necessary to make the studio space come alive.

*Videography*

There was photo and video documentation of some rehearsals and meetings, and the intervention experience. A photographer and videographer were commissioned for the work. The film and photo captured contributed to the end product video documentary.

*Video Acquisition*

The permission to use an excerpt from the short film by Firelight Media entitled “Children Left Behind” was acquired for use during part one the theatrical experience. Permission was granted by the advising producer Dr. Amy Stuart Wells of Teacher College in New York. The film is publicly available online through Vimeo.

*Dialogue Facilitator Acquisition*

There were three experienced Inner Group Dialogue facilitators acquired for the intervention. The facilitators were recruited from current graduate students and alumni of the University of Massachusetts – Amherst, College of Education and Amherst College. The dialogue facilitators were trained in the goals of the experience. During the training, the facilitators collectively developed a protocol focused on engaging the audience in a fruitful conversation on the hopes for a newly imagined high stakes testing system.
The Intervention

The intervention consisted of a multi-modal theatrical experience. Tickets were free and reserved online prior to the performance dates. The audience checked-in and gathered in the lobby of the New Africa House (situated on the first floor) awaiting the official descent into the studio space (situated in the basement). Audience that wanted to join the experience but did not reserve tickets were placed on a waiting list. There were 60 seats, plus standing room.

At the appointed start time, from the lobby in a single file line audience descent took them down the stairwell passing actors as they sang the Negro Spiritual “This May Be the Last Time”. (An elevator was available for those that needed it.) Once the audience members reached the downstairs lobby, the single line was instructed to enter the studio space through a metal detector and guards (who served as front of house staff and ushers) directing them to find their seats. The scenic design of the space included elements suggesting an inner-city impoverished school setting and a sterilized testing environment.

Ms. Engel (the teacher) was center stage encouraging everyone to “Find your seat, your exam will begin shortly.” On every seat was a clipboard with a survey on test perceptions (Appendix B) attached and a sharpened number two pencil. Once all the audience members entered from the lobby, Ms. Engel gave the following instructions to the audience (reading from an instruction manual): “At this time please turn off or silence all cell phones. The purpose of the exam in front of you is to capture your perceptions on testing. You will have 5 minutes to read the instructions and complete the exam. Please read the instructions at the top of the page and begin.” Ms. Engel then left the stage and
walked around the room militantly keeping an eye out for cheating and close watch on the clock. There was a clock ticking sound cue at the time of “testing.” At the 5-minute mark a classroom bell sounded, and Ms. Engel then said, “Time is up! Please place your exams under your seat as we prepare to watch a short video on the history of testing.” Two students wheeled out an old school TV cart that displayed a 7-minute excerpt of the short film by Firelight Media entitled “Children Left Behind.”

The end of the video showing led to a school bell with a transition into the 45-minute theatrical performance. After experiencing the theatrical performance, two of the principal actors asked all audience members to take the “test” that they filled out at the beginning of the experience and rip it to pieces. The audience was then thanked, and the cast invited the audience to participate in one of the key practices of Brown Paper Studio. Everyone was invited to reflect on the experience thus far by writing on the brown paper hanging from the walls with colored markers (sort of graffiti style). Some paper was used for free writing for reflection. Other paper had questions stems which included: What is assessment? What is high stakes testing? What is equity? What does your racial/ethnic identity have to do with how you perform on high stakes tests? How could we consider equity in the design of high stakes tests?

The free writing reflections lasted for roughly 15 minutes. Attendees were free to leave if they pleased. The remaining audience members were instructed to join 1 of 3 circles lead by a trained dialogue facilitator to have a conversation about the social issues that result from current testing practices and suggestions on how we can come to a democratic decision about best practices in testing that includes the voices of
marginalized groups. (See Appendix C for the Dialogue Protocol). The dialogue lasted 30 minutes. Refreshments and fellowship commenced after the dialogue.

Phase 3: Dissemination
The final phase of the Wrong Answer Project was dissemination. This phase focused on making the project available for public use and disseminating the results of the pilot using a method accessible to a significant amount of society. A website has been developed as a platform for accessing all the materials for the Wrong Answer Project including the script, this formal report, and two video products amongst other resources. The goal is to ultimately replicate and expand this work including the voices of other marginalized groups and geographic locations. The website serves as a resource for spreading the word about the project to new potential implementation sites.

Additionally, there was a conference presentation on the findings at the 2020 NCME annual conference. After piloting this work, I realized to effectively utilize this product as evidence for the social consequences of testing there should be an evaluability assessment. This is where this proposed dissertation project comes into play.

Analysis
Data was captured from the graffiti writings on the brown construction paper posted on the walls and notes were taken during the semi-structured dialogue circle discussions. The graffiti style writings and circle dialogue are methods used within the Brown Paper Studio framework. The brown paper reflections were used to assess whether the audience was appropriately implicated. Themes from the audience responses on the brown paper were aggregated and mapped within the domains that relate to the root issues of this work. These domains included:
1. Defining Education
2. Defining Assessment
3. Defining High Stakes Testing
4. Defining Equity
5. Intersection of Race/Ethnicity and Test Performance
6. Societal Views of Equity in Testing

There was a note taker designated to each of the three dialogue circles for each night of the experience. Themes from the notes taken during the circles were used to explore two main domains related to the advancement of this work:

1. Current Societal Views of High Stakes Testing
2. Society’s Re-imagination of High Stakes Testing

**Resulting Data**

The premier of the Wrong Answer Project occurred on the 6th and 7th of December 2019. There were 83 of 120 tickets reserved that were claimed and 42 walk-up attendees. This made for a total of 125 participants in the experience. No demographic information was captured from the audience members. From the eye of the observer there was an array of ages, race and ethnicities, genders and educational levels present for the experience.

*Brown Paper Graffiti*

There were 178 graffiti responses that were useable based on legibility. The responses ranged from a single word to full paragraphs. Some of the brown paper had pre-written questions for reflection which included: What is assessment? What is high
stakes testing? What is equity? What does your racial/ethnic identity have to do with how you perform on high stakes tests? How could we consider equity in the design of high stakes tests? The remaining brown paper was left blank for free reflection. The themes drawn from the graffiti are presented in the table below by domains used to confirm that the audience arrived at an appropriate mindset to enter meaningful dialogue focused on re-imagining high stakes testing for society.

Table 1. Graffiti Writing Themes

<table>
<thead>
<tr>
<th>Domain</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining Education</td>
<td></td>
</tr>
<tr>
<td>What does education mean to you?</td>
<td>Risking; Fighting; Growth; Learning; Discovery</td>
</tr>
<tr>
<td>Who is our education system supporting?</td>
<td>White; Upper class; Privileged; People with Access</td>
</tr>
<tr>
<td>Defining Assessment</td>
<td></td>
</tr>
<tr>
<td>What is assessment?</td>
<td>Measurement; Evaluating Ability; Judgment; Comparing to Standard; Unfair; Biased; Subjective; White man's standard</td>
</tr>
<tr>
<td>Defining High Stakes Testing</td>
<td></td>
</tr>
<tr>
<td>What is high stakes testing?</td>
<td>Determination; Progress towards Higher Education; Consequences; Pressure; Unfair; Oppressive</td>
</tr>
<tr>
<td>Defining Equity</td>
<td></td>
</tr>
<tr>
<td>What is equity?</td>
<td>Acknowledging oppression; Systemic; Transforming</td>
</tr>
<tr>
<td>Intersection of Race/Ethnicity and Test Performance</td>
<td></td>
</tr>
<tr>
<td>What does your race/ethnicity have to do with how you perform on high stakes tests?</td>
<td>Access to resources; High Pressure; Lack of Support; Undereducated parents</td>
</tr>
<tr>
<td>Societal Views of Equity in Testing</td>
<td></td>
</tr>
<tr>
<td>What does equity in testing look like to you?</td>
<td>non-Eurocentric; Different subjects; Queer</td>
</tr>
<tr>
<td>How could we consider equity in the design of high stakes tests?</td>
<td>Include Student Voice; Acknowledge Inequalities; Start Over; Include Diverse Voices</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
</tr>
<tr>
<td>Free Reflection</td>
<td>Test disapproval; Defined by more than tests and tests scores; Reform the test; Remove the test</td>
</tr>
</tbody>
</table>
**Dialogue Groups Circles**

There were three dialogue circles split into three different rooms based on a number allocated to the playbill audience members received at the beginning of the experience. One group stayed within the studio space to dialogue while the remaining two groups moved to classrooms located on the same floor of the studio. Almost 75% of audience members (n=93), stayed for the dialogue groups. The dialogue circles each began with gauging the audience’s general reflection of the theatrical experience and brown paper graffiti writing. Themes drawn from the dialogue group circles are presented in the table below by domains used to explore societal views on how we can advance this work.

**Table 2. Group Dialogue Themes**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Societal Views of High Stakes Testing</td>
<td>Anxiety and Stress Inducing; Competitive; Forced Assimilation; Differences in Access to Resources; Assumed to Lead to Equal Opportunities</td>
</tr>
<tr>
<td>Semi-structured Open-Ended Reflection</td>
<td></td>
</tr>
<tr>
<td>Society’s Re-imagination of High Stakes Testing</td>
<td>Collaborative vs. Competitive; Integrate into curriculum; Non-privatized; Centered on Different Student Learning Styles; Embedded with Critical Thinking vs. Knowledge Regurgitation; Embedded with Career and College Standards; Project-based; Essay-based; Pre-Test/Screener of Appropriate Testing Method for Examinee; Testing optional Boycotting; Test Prep for All; Capture Voices from Society; Run for Political Office; Push a change campaign;</td>
</tr>
<tr>
<td>What would re-imaging assessments look like?</td>
<td></td>
</tr>
<tr>
<td>What are next steps for change?</td>
<td></td>
</tr>
</tbody>
</table>
A Vision Forward

Brown Paper Studio is an American patented theater practice for social transformation founded by my project advisor Prof. Judyie Al-Bilali. As a teaching assistant and student of Prof. Al-Bilali’s, I have seen firsthand the benefits of employing these methods for invoking social transformation. It is my hope that the results from the Wrong Answer Project can also be utilized as a part of an empirical argument to validate the Brown Paper Studio practice as a tool for invoking social transformation.

The audience participation for this experience exceeded my expectations. Our ticket reservations sold out both nights and we had a large waiting list even amongst challenges with the weather. Given the size of the studio (and some no shows for ticket reservations) we were able to accommodate our attendees. This is one indicator of the general interest of the community to see real change happen in the realm of high stakes educational testing.
CHAPTER 2

METHODS

Research Questions

My dissertation study is an evaluation research project. I used the Evaluability Assessment framework to explore the following research question:

Does the Wrong Answer Project serve the purpose it is purported to serve?

Who is it serving? Is it serving them in a way that serves them best?

Evaluability Assessment

Evaluability assessment (EA) is a method of exploratory evaluation. The purpose of EA is to generate evaluation findings that aid in further developing a program and advance evaluation readiness. Other forms of exploratory evaluation include rapid feedback evaluation, evaluation synthesis, and small-sample studies (Wholey, 2004 & 2010). Joseph Wholey introduced EA to the field in his 1979 seminal text. It has evolved over the past four decades. Lam and Skinner (2021) in their recent synthesis of evaluability studies found that there is an array of different evaluability frameworks in use within the evaluation published literature since Wholey’s foundation. Majority of the frameworks in use operationalize Wholey’s original 8-step model (1979; 1983) or the more concise 6-step model (1994; 2004; 2010; 2015). Other evaluability models in use included Leviton et al. (2010) systematic screening and assessment method, the 6-step model of Thurston and Potvin (2003), and the 10-step model of Smith (1989). Regardless of the method, the common themes that exist throughout evaluability assessment include engaging program personnel and stakeholders, reviewing program
definition and structure, gathering data, analyzing and using the data to inform program management.

Wholey (2015) posited that, “evaluability assessment answers the question of whether a program is ready for useful evaluation, not whether the program can be evaluated (any program can be evaluated)” (pg. 90). The six key steps involved in EA are to:

1) Involve intended users and other key stakeholders.
2) Clarify the program design.
3) Explore the program reality.
4) Assess the plausibility of the program.
5) Reach agreement on any needed changes in program design or implementation.
6) Reach agreement on the focus and intended use of any further evaluation.

The first step to involve intended users and other key stakeholders means to include in the evaluability assessment process decision makers, influencers and those of whom the program will have the greatest impact. For The Wrong Answer Project (WAP) this would include the Founder (myself), Associate Producer (Prof. Judyie Al-Bilali), Theatrical Director (Jen Onopa), the NCME Mission Fund Program Officer (Michelle Boyer), the past NCME President who supported the funding of the project (Steve Sireci), an Intergroup Dialogue Designer (Sabine Jacques), a Knowledgeable Observer (Jennifer Randall) and an array of student artists who contributed to planning and implementing the pilot experience. These individuals will be kept aware and will help reach agreements during the evaluability assessment process. They will be referenced as my “Insights Team”.

35
The next step of **clarifying the program design** focuses on the logic model or theory of action. It involves interviewing key members of management to clearly define the program implementation model and expected outcomes. We have not formulated a logic model or theory of action for *The Wrong Answer Project*. Our NCME Mission Fund 2018 Application and Implementation Plan are sources for model definition and project intentions. Findings from the evaluability assessment was used to inform the development of a logic model and theory of action. Both are additional products of my dissertation study. I answered the following questions (Wholey, 1994, 2010, 2015) to clarify my project design:

1) From your perspective, what is *WAP* trying to accomplish, and what resources does it have?
2) What results have been produced to date?
3) What results are likely in the next year or two?
4) Why would *WAP* produce those results?
5) What are *WAP*’s main problems? How long will it take to solve those problems?
6) What kind of information do you get on *WAP*’s performance and results?
7) How do you (how would you) use this information?
8) What kind of program performance information are requested from key stakeholders?

The Insights Team and I came together to **explore the program reality**. This step of the EA process involves “comparing the [intended] program design with actual program inputs, activities, outputs, and outcomes; identify feasible performance measures; and identify problems inhibiting effective program performance” (Wholey, 2015 pg. 93). We
examined program artifacts (i.e., the application, Implementation Plan, script, dialogue facilitation protocol, final report, program reviews). Our goal during this step was to answer the following questions (Wholey, 1994, 2010, 2015):

1) What are our goals for the WAP?
2) What are the major project activities?
3) Why will those activities achieve those goals?
4) What resources are available to the WAP? Number of staff? Total budget? Sources of funds?
5) What outputs are being delivered by the project? To whom?
6) What evidence is necessary to determine whether goals are met?
7) What happens if goals are met? What happens if they are not met?
8) How is the project related to local priorities?
9) What data or records are maintained? Costs? Services delivered? Service quality? Outcomes?
10) How often are these data collected?
11) How is this information used? Does anything change based on these data or records?
12) What major problems did we experience? How long will it take to solve those problems?
13) What results have been produced to date?
14) What results are likely in the next two to three years?

Next, we assessed the plausibility of the program using data from steps one through three. Only the Theatrical Director, Associate Producer, and I discussed WAP’s outputs
and the progress towards outcomes. We set the intention of honestly exploring the
capacity of WAP to meet its mission and goals. Continuing with the Wholey way, we
assessed project plausibility guided by the following questions adapted from Strosberg &
Wholey’s (1987) 8-step model:

1) Are program activities and resources likely to meet planned objectives?

2) Are there any major impediments to the achievement of program objectives?

3) Is there any evidence that program outputs will meet expected outcomes?

4) Should we spend money on the WAP even though important pieces have not been
implemented appropriately?

5) Does WAP have well-defined uses for information on progress toward its
measurable outcomes?

6) Do project staff use process data?

7) Are there any clear indicators of model fidelity? And how to enforce fidelity?

8) What portion of the WAP is ready for evaluation of progress toward agreed upon
objectives?

9) What evaluation and management options should the founding team consider?

10) What are recommendations for areas not ready for evaluation?

The cumulative data gathered and discussed during the evaluability assessment
helped us as the key stakeholders and users of evaluation reach an agreement on any
needed changes in the project design or implementation. Up to this point we reviewed
the project definition, design, activities, outputs and outcomes. Throughout this review
the goals of WAP were measured and gaps, shortcomings, and improvements were
identified. This step of the evaluability assessment charged the key stakeholders to
identify and agree upon the changes needed to improve project outcomes. The Insights Team and I explored the following questions:

1) What components of the program design, activities, outputs, and outcomes have gaps/shortcomings, seem unrealistic, or are not working?
2) Which of these components should be removed? Which should be changed?
3) What changes should be made to improve these components?
4) How will these changes shift evaluation and management options?

The final step within the evaluability assessment was to **reach agreement on evaluation focus and intended use**. The results of the evaluation research performed on the WAP informed the development of a logic model, theory of action, and a menu of evaluation options. I consulted with my academic advisors on my dissertation committee on a theory of action and the appropriateness of different evaluation methods for WAP. Their feedback was shared with the theatrical director and associate producer to finalize a logic model, theory of action and an evaluation approach as additional deliverables for the evaluability assessment.

**Validity Evidence**

As the Insights Team and I assess the evaluability of the Wrong Answer Project, I hope that we will keep the value of our project at the forefront of our reflections and responses. Kane’s argument-based validity approach is helpful as a mechanism of understanding how and why this project might meet its expected objectives. According to a recent systematic review of validation methodology (Lavery et. al, 2020), Kane’s argument-based validity is one of the most respected and utilized methods to date. Coupled with the flexibility to validate virtually any measure for a particular purpose, this
method is well fit as a validity approach for my project. It is important to establish the validity argument that WAP is a measurement tool that meets the purpose for which it was constructed.

The focus of Kane’s framework is on the interpretation (meaning making or explanation) of test scores. Argument-based validity theory posits that score value can be measured and judged by providing evidence for the respective “interpretive argument” of the test scores. The interpretive argument is defined by stated conclusions or decisions made from test scores based on a set of assumptions. For the purposes of our project, “test scores” are the artifacts of WAP (data collected from Brown Paper Graffiti writings and dialogue circle notes, reports produced and dissemination, and project reviews).

Within argument-based validity the interpretive argument is adopted as “the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its more questionable assumptions” (Kane, 1992, pg.527). The interpretive argument follows closely to Cluster 1: Establishing Intended Uses and Interpretations of the Standards (2014) for validity. Particularly, Standards 1.1-1.6 which state:

- **Standard 1.1** – *The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The populations(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.*

- **Standard 1.2** – *A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.*
• **Standard 1.3** – If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations.

• **Standard 1.4** – If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence if necessary.

• **Standard 1.5** – When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented together with relevant evidence.

• **Standard 1.6** – When a test use is recommended on the grounds that testing or the testing program itself will result in some indirect benefit, in addition to the test scores themselves, the recommender should make explicit the rationale for anticipating the indirect benefit.

The interpretive argument gets to the root of how and why we can state conclusions and make decisions drawn from test scores by providing meaningful evidence that supports assumptions and inferences. There are different kinds of evidence gathered to support an inference or assumption. Kane (1992) describes using evidence from observations, generalization, extrapolation, theory-based inferences, and decisions.

**Observational** evidence focuses on procedural artifacts. We must ensure that the test score results from a measurement procedure that is followed with fidelity. The measurement process for WAP is unique and draws from arts-based and qualitative approaches. Evidence of the measurement observed gives me the who, what, where, and
When setting the foundation for the how and why within the interpretive argument. A measurement process implemented with fidelity will give me strong support for the interpretive argument. A lack of fidelity could lead to weak evidence thus nullifying the interpretive argument.

**Generalization** evidence supports the idea that a measurement instance is representative of the greater population parameter. Thus, providing proof that a sample could be considered a true estimate of the actual effect within the entire population. Generalization evidence ensures that the measurement process if repeated will continually be an accurate representation of the entire population from which the sample was drawn. For WAP we can consider generalizability internally and externally. Internal evidence will be drawn from identifying reliable processes that have contributed to valid measurement. Future iterations of WAP in different samples will contribute to external generalizability. External evidence will also be about identifying similar arts-based work that has occurred in similar populations and come to similar conclusions to WAP.

**Extrapolation** goes hand-in-hand with external generalizability efforts as it supports the exploration of the results of our measurement procedure to non-measurement focused processes. This piece of validity evidence may take further study beyond the scopes of this evaluability assessment as it requires a more in-depth research design. An example of extrapolation work in the future could be an examination of the association of high stakes testing perceptions of citizens that attended WAP and their perceptions of and social action towards the testing practices of their local community.

**Theory-based inferences** will serve as supporting evidence for our interpretive argument and the development of our theory of action and logic model for WAP. We will
investigate different arts-based heuristics and hermeneutical theories that have been utilized to inform social action. There are other arts-based initiatives like WAP that uses theater as a mechanism for examining the social consequences of high stakes testing or similar social issues. These initiatives are siloed and exist throughout time and across space, but they are all rooted in similar theory. We will investigate the theory within these projects and other evidence-based arts-based practices.

**Decisions** made from the results of the WAP is the final piece of evidence we will use to support the interpretive argument. These decisions are either intended or unintended outcomes of our project. Evidence of decisions made from the test score could be considered an extension of observational evidence that focuses on the resultant of our interpretive argument. What will happen as a result of the test score? Within WAP specifically, did we inform social action? If so, how? The “how” could be a part of the planned purpose of the WAP or a complete surprise. Both pieces of evidence add value to the interpretive argument and expected impact/outcomes for our logic model.

I would like to use Kane’s argument-based lens and Standards 1.1-1.6 of the Standards to formulate a valid, more robust, defendable structure for the Wrong Answer Project. The WAP pilot implemented in 2019 was developed with looser systemization within the project. There was no formal logic model or theory of action created to inform the implementation plan. I used my prior experiences as a theater producer for social change to inform the piloted project approach. My project shows great promise. A deeper formulation of the validity argument will draw trust and buy-in for WAP as a measurement tool and further justify the need for the evaluability assessment. I want to
show that the *Wrong Answer Project* works, here is why and here is how we could manage it.

The general outline of the argument-based framework is to:

1. Decide on the statements and decisions to be based on the test scores
2. Specify the inferences and assumptions leading from the tests scores to these statements and decisions
3. Identify potential competing interpretations
4. Seek evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments

An external product to my dissertation study is manuscript to which I have invited the Insights Team to develop using an argument-based validity framework and the *Standards* to refine value by informing the interpretive argument of the *WAP*. The Insights Team and I will answer the following questions adapted from Kane (1992):

1) What conclusions will be drawn from the results of the *Wrong Answer Project*?
   a. What assumptions or inferences are made about the results to draw these conclusions? What evidence supports these assumptions and inferences? What *Standards* of Validity does this support?

2) What decisions will be made from the results of the *Wrong Answer Project*?
   a. What assumptions or inferences are made about the results to make these decisions? What evidence supports these assumptions and inferences? What *Standards* of Validity does this support?
3) Are the decisions made and conclusions drawn from the results informed by the stated purpose of the WAP? What Standards of Validity does this support?

4) Are there any competing interpretations for the results of the WAP? What evidence can be used to refute the potential counterargument? What Standards of Validity does this support?

5) Are the details of the argument explicitly stated?

The argument-based validity framework can be utilized to detail the value of the Wrong Answer Project. The validity argument lays a clear foundation for developing WAP’s theory of action and logic model.
CHAPTER 3
RESULTS

The Insight Team had five meetings to discuss the evaluability assessment of the Wrong Answer Project. The results from these meetings are presented in thick description with some fragments and some summarized conversation points. All contributes in a unique way to the assessment. There are six bolded subheadings that display the six steps to evaluability assessment referenced within the methods section. There was a total of 10 participants within the Insight Team. Each served as a unique role on the Insight Team either as a key player or expert within the field of measurement.

Insight Team: Key Player & Expert Roles

Afrikah ................................................................. Production Manager.

Angela ................................................................. Performer/Principal Role.

Cassie ................................................................. Performer/Supporting Role.

Darius ................................................................. Founder/Producer.

Iya ................................................................. Advisor/Associate Producer.

Jen O. ................................................................. Theatrical Director.

Jen R. ............................................................... Knowledgeable Observer.

Michelle ............................................................ Grant Program Officer.

Sabine ............................................................... Intergroup Dialogue Designer.

Steve ............................................................... Knowledgeable Reviewer.
Figure 6. W.A.P. Insights Team - Key Players

WRONG ANSWER!

Insight Team - Key Players & Experts: Evaluability Assessment Working Group

Iya

Afrikah

Sabine

Michelle

Cassie

Angela

Steve

Jen R.

WE DA PEOPLE!!!
Involve intended users and other key stakeholders.

If *Wrong Answer!* was living and breathing, she would be a sarcastic person of color who’s tired of being tired. And tired of hearing the same story over and over. She wants dialogue, understanding, educated solutions. And likes claps synced with her words. Think of a mix of Samantha White from “Dear White People”, Lena Waithe, and Amanda Seales. It’s giving Black. Radical. Feminism. Thus, I desired my Insights Team to be mostly people of color and feminine, unapologetically.

I was a participant-observer in the pilot of the *Wrong Answer Project*. I have been intimately connected with this work by conversing, writing, creating, gathering, facilitating, designing, overseeing, co-directing, choreographing, performing, and producing. That’s right, I performed in my own production! One of our performers rolled their ankle while learning choreography and I had to step in. Grateful I was able to fit his clothing! Additionally, I plan to replicate and expand W.A.P. to focus on a different source population, audiences, and geographic areas. **Professor Iya Judyie Al-bilali** advised my project from its infancy through implementation. Prof. Al-bilali’s social capital and joint faculty appointment yielded supports and interest from both the Department of Theater and Department of African American Studies within the College of Humanities and Fine Arts at the University of Massachusetts-Amherst. As project advisor she offered in-kind resources such as access to the New Africa House Theater (including sound, lighting, and project boards) and student recruitment for auditions. Iya’s theater for social transformation method *Brown Paper Studios* is a staple of the W.A.P. experience. This is the first study that has validated Iya’s theater method for research purposes. We plan to use the contribution of my dissertation work in the
developing validation argument for *Brown Paper Studios* as a mechanism for social transformation. **Jennifer Onopa** was another key player who supported the watering of *W.A.P.* into the beautiful flower it has blossomed into today. Jen O.’s background in directing, playwriting, devised theater, community theater, teaching and youth mentorship has played a vital role in the design, development, implementation, and evaluation processes. Her voice is important to the Insight Team as one of the leaders of this work who is all too familiar with the impact high stakes tests have on students.

It was **Dr. Jennifer Randall** who validated me as a Research, Educational Measurement, and Psychometrics student by saying, “Now that’s cool” after sharing with her that I wanted to write a theater show for my dissertation work. Jen R. was one of my special professors during my pre-candidacy phase who encouraged me just by existin-n-bossin. A Black, Female, prior Professor in Research, Educational Measurement, and Psychometrics and Associate Dean of Academic Affairs at University of Massachusetts-Amherst College of Education and current Endowed Professor of Psychometrics and Test Development at University of Michigan Marsal Family School of Education. Can we say: #BlackGirlMagic #BlackGirlsRock I consider her a secondary advisor, mentor, and trusted rectifier. Dr. Randall was able to attend the *W.A.P.* pilot and released a publicly available testimony about the experience. Now, **Afrikah Smith**, was my right hand during the production process. They were initially given the title “Production Assistant” ***annnnnntttt WRONG ANSWER***. Afrikah was an integral part of the administration and management of the production process. I could not have done this without them! They hold a key stake in the success of *W.A.P.* with hopes of seeing the good fruit of their hard labor. **Dr. Michelle Boyer** was in support of my project on the
committee for the National Council on Measurement in Education’s Mission Fund Grant Award. She was our grant officer to which I submitted the contract, implementation plan, budget, and final report. Michelle is an alumnus of the REMP program who offers great insights “from the field”. We crossed paths in Steve’s validity class.

Speaking of Dr. Steve Sireci’s Test Validity course offered me theory and language for what I have been feeling my whole life and seeing happen to my people: the social consequences of testing. Steve has been a wonderful mentor during my doctoral tenure. In 2019, when we were awarded, he was the incumbent president for the National Council on Measurement in Education. “Hardly home but always reppin.” Steve stay pubbin my work in psychometrics spaces and continues to push me to replicate and expand the work of W.A.P. Dr. Sireci is a distinguished professor of Research, Educational Measurement and Psychometrics at the University of Massachusetts-Amherst College of Education. Sabine Jacques is my sister from another mister who I trust to keep it one hunit and center love and community. We have jointly co-facilitated Brown Paper Studios in various student and community groups both in the states and internationally. Sabine’s feedback was useful during script reviews and data collection design. She was a third of the talented trio of intergroup dialogue designer-facilitators who birthed my vision of the community solutions circles. Angela Kwebiiha and Cassie McGrath were student performers in the theatrical and movement components of the Wrong Answer Project. Both contributed to the devising process which included script adjustments, movement discovery, and recurring community feedback. Angela and Cassie complete our Insights Team by including the students’ perspective of the work.
Both were upperclassmen at the time of implementation at the University of Massachusetts-Amherst.

There were various ways that the Insights Team could contribute their voice in responses to the evaluability assessment questions. The preferred space for me were the 5 Zoom meetings. These were scheduled according to the majority availability of the group. Key staff and experts were also allowed to talk to me via phone, send voice notes, text messages, or type email responses to the questions. The versatility in data collection increased feasibility and engagement options for the insight team.

**Clarify the program design.**
(Respondents: Founder/Producer)

Using the NCME Mission Fund 2018 Application, Implementation Plan, and W.A.P. website ([www.wronganswerproject.org](http://www.wronganswerproject.org)) as resources for model definition and project intentions, I responded to the questions that follow to clarify the program design. There were eight evaluability assessment questions that I answered to clearly define the program implementation model and expected outcomes. The following results are organized by question. Six of my responses are first person monologue with formatted emphasis on key words. Two of my responses are a paragraphed list of paraphrases.

1.) *From your perspective, what is the Wrong Answer Project trying to accomplish, and what resources does it have?*

The *Wrong Answer Project* is trying to accomplish democracy within high stakes educational testing. True democracy – for the people by the people – by creating an embodied space, process, and experience: **focused** on minoritized and historically erased Black and Brown people of varying identities; that **invokes** justified **indignation** and **civic**
responsibility; and fuels a grassroots movement towards the change of high stakes testing policies and practices. It is trying to accomplish a new way of gathering authentic data on educational testing consequences. W.A.P. is also trying to raise up the pluralistic voices of the unheard, marginalized, and erased peoples to the heard, hegemonic, unfairly magnified people. Not for the raising up alone, but with hopes that voices and suggestions would be heard and considered for the reimagination and reconstruction of high stakes testing. From the pilot, we have established an array of resources: a script, implementation model, final report, website, intergroup dialogue facilitation design, props, publicity materials, people, and the data we collected. We have my passion, vision, and tenacity. Additionally, connectedness - we have the interest of the pilot team and public audience to see our mission fulfilled.

2.) What results have been produced to date?

Devised scripts. Intergroup dialogue designs. Props. Full house attendance during both piloted production dates. We synthesized and analyzed the graffiti writings and intergroup dialogue notes and developed a final report. The report also included a thorough synthesis of the implementation process. The pilot findings were referenced in detail during the introduction of this manuscript.

3.) What results are likely in the next year or two?

My hope is to replicate and expand the work within a different geographic location of Black students at a high school or college OR an Indigenous community at a high school or Tribal College. I assume the results will be adjacent to those of the pilot but geographically or indigenously specific. This will be a wonderful opportunity to
triangulate results and contribute to the validity argument for W.A.P. as evidence for the social consequences of testing. Additionally, hoping to continue to interrogate the high stakes educational testing community to reimagine equitable decision making and testing approaches by marketing W.A.P. and it results better. The plan is to share with more change makers in testing.

4.) Why would WAP produce those results?

We have a model and artifacts from the pilot of the Wrong Answer Project which includes: an implementation plan embedded with Brown Paper Studios theater for social transformation methods (a community responsive practice), a script that can be adapted and devised for the Black and Indigenous population, and a grant writer (Me) to pitch the work to a few foundations. My plan is to be more intentional about sharing the work of W.A.P. This would happen by applying some of the feedback from the Insights Team to update the website, marketing, and disseminated products.

5.) What are WAP’s main problems? How long will it take to solve those problems?

We must sustain resources and funding. The Wrong Answer Project will always need a space and place to implement. It is imperative that we partner with education institutions that have access to theater space or equipment so that we have in-kind resources. We need to sustain funding to continue to produce. Also, assessing community impact is a bit tough. There needs to be a way to hook audience members so that we can follow-up on how their perspective may have changed. Spaces and audiences that receive the results of the experiences (the grassroots solutions) should receive
follow-up to assess **impact** and lasting change. This would require the collection of more contact information to solve the problem of establishing **measurable** societal impact.

6.) *What kind of information do you get on WAP’s performance and results?*

We received **audible** reactions and social responses (whispers and conversations) that **affirmed** what the audience saw in the show as a part of their reality. Enthusiastic shouts of agreement, excitement, and liberation when it came to ripping up the tests. The **intention** and **focus** during the graffiti writings, the data from our graffiti writings, the **rich** intergroup dialogue discussion, and intergroup dialogue notes all were **information** received from the performance and synthesized within the results. Testimonials from informed observers, the student performers and the production team also provided information on W.A.P.’s performance.

7.) *How do you (how would you) use this information?*

All the information can be synthesized into a report and presentations to **disseminate** to lay audiences (students and families), test constructors and the psychometric community, educational practitioners, policy makers, and administrators to **continue** the social, operational, and political movement. Also, to **inform** changes to the implementation and production of future iterations of W.A.P.

8.) *What kind of program performance information are requested from key stakeholders?*

The National Council on Measurement in Education required the **submission** of a final report given the receipt of the Mission Fund Grant Award. It was a report that I had the **freedom** to develop the format and indicators for program review. It is my desire to
continue reporting on each replication and iteration of the W.A.P. For the usefulness of the information as mentioned above in question seven.

**Explore the program reality.**
(Respondents: Founder/Producer, Associate Producer, Theatrical Director, Production Manager, Grant Officer, Intergroup Dialogue Designer, Student Performers, Knowledgeable Observer, and Knowledgeable Reviewer.)

This step of the evaluability assessment (EA) process involved “comparing the [intended] program design with actual program inputs, activities, outputs, and outcomes; identify feasible performance measures; and identify problems inhibiting effective program performance” (Whooley, 2015 pg. 93). We examined program artifacts (i.e., the application, Implementation Plan, script, dialogue facilitation protocol, final report, program reviews, and website). Our goal during this step was to hold an open discussion and dialogue addressing meaningful exploration questions (Whooley, 1994, 2010, 2015). There are fourteen evaluability assessment questions that were discussed to explore the program reality of the Wrong Answer Project.

Results are organized by EA question. The results from six EA questions are presented as responses from each member from the Insights Team, briefly summarized in a playwright dialogue separated by hard returns. **Key words** and **concepts** have been **formatted** for emphasis for those questions. This aesthetic formatting is used throughout **these** seven sections adding artistic and systematic influence to the results narrative. Red and Blue are mostly a system of highlighting process **insights** and summative **outcomes**. There are eight EA question results for this section that are presented in a continuous list paragraph. Responses are phrases separated by a period punctuation. This writing style follows from the conversation’s popcorn-snowball style of responses offered by the

1.) What are our goals for the Wrong Answer Project?

Iya

To bring voices of the community including students, parents, and educators into the process of evaluating learning. It’s an opportunity for the people to take it into their own hands. The people at the top have no intention of facilitating that. This is the end of their job.

Jen O.

To bring voices into the evaluation process for the purpose of dismantling the one that exists. Creating long lasting change in the voice that has been problematic.

Darius

To move and shift the current high stake testing policy, practices and procedures to more equity centered, anti-racist designs and to encourage and empower our communities that have felt the brunt of high stakes testing. To let them know that testing is an inevitable part of life and that our system is broken, but we can offer a solution and if we are really in a democracy the people at the top should listen to that solution.

Cassie
To give **voice** to people not typically considered. People often know how they learn, and it’s not often tied to testing. That evaluation is not an assessment of their knowledge. It’s a mass acknowledgement that the system is **broken** and in this space, there is a **freedom** and **knowing** to come to solutions.

**Afrikah**

To be a good space to hold **conversations** for diverse learners (like me who had an IEP), a space where people can **talk** about their needs (particularly those with an Individual Education Plan). Maybe a **curriculum** or professional **development** can be created from the results. I cannot help but think about how current testing practices **don’t allow for developments needed** to be had.

**Steve**

*W.A.P.* is **teaching white people** about **issues** that weren’t in our consciousness.

It’s a *positive unintended consequence*.

**Sabine**

To lean into **liberatory consciousness** and elements like that of Barbara Love, awareness, and the vision you are working towards. *W.A.P.* is a visionary project, that presents these are the **issues** how can we come to solutions together. It is **deep reflective listening** and **self-awareness** work. It’s an opportunity for a spectrum of folks to engage in **open dialogue** on how has testing been for you. It’s partnership - bringing students to the center. It’s a very big, intentional street data project because it is **going to the core** and **root experiences** of people.
Michelle

To be a **mechanism** not only for white people to learn but **for the industry itself to understand**, how we have been defining constructs. **Bringing people to the table** to inform **the gap** in those who have **not** had access to the table. To learn how to **make tests better** in a sense that **what** we are measuring and **how** we measure it truly effects the experience of diverse peoples.

2.) **What are the major project activities?**

Planning phase included informal conversations (with family, friends, and community) about high stakes testing experiences, grant writing, script writing, production team development, devising/script development, production designing, fundraising, and implementation. Publicity for an inclusive audience. Graffiti writing on the walls. Intergroup dialogue circles and gathering of community data. Recruitment of cast and designers, auditions, and rehearsals. Social gatherings as a cast. Postmortem discussion – what worked, didn’t work and where we see the project going. The ripping up of tests right before going into self-reflection. The video clip from “Children Left Behind” by Firelight Media (great focal point - explanation helped set the tone very well).

3.) **Why will those activities achieve those goals?**

Jen O.

It’s **building community coalition** through this **embodied** experience of watching and being a part of the performance. We tried out the script with students to help and **affirm the script**, this process will mean that the script makes sense for each

58
community. We capture community data. I am moved by how people could be assessed differently. Students and community were coming up with valuable solutions, really good ideas were developed.

Iya

It's in person engagement at a time when we have been profoundly separated and forced into digital connection…and the idea that we are assessing the whole person we are not just interested in your score. And there is inclusion every step of the way from ripping the test up to devising to all elements of interactive. The publicity brought a multigenerational, multi-everything audience. There are brown paper components, the child emerges when the whole body can say this is me. And we partied our way out of there with fellowship and food at the end.

Cassie

In terms of creating a movement, this gives space to people for convening to ignite the movement. As a performer in the “intervention” it is so clear to me in my mind. The way the show was set up from the music to the video, oh, the opening to the show in the stairway when we were singing folks entered, it really set the stage. It set a strong scene and foundation in the body for the audience, and they felt like they were a part of it and made them open up. All of it, the movement, then the opportunity to talk about it at the end. Sometimes this happens individually in small groups with who you went with, but building that into the show, gives you space to use that part of your brain in a good way. Being an actor in it I certainly learned a lot and educational testing has become a passion of mine. It made me think of what I should be thinking

59
about as a writer. If I could be a qualitative number, I would be. Also, the social gatherings as a cast were great and helped in creating the bond and helps the work live on.

Sabine

Interdependency of the activities – putting people first, reflecting on their own lived experiences, and giving them the agency to express that. An opportunity to be critical of their own experience and the school system. To see themselves from their identity and experience. A mirror and a window. It is the spirit of encouragement. We are going into their lived experiences and see that as valuable.

Steve

Forcing a conversation that would not have happened, and you are doing it in a way that is empowering. (That thing that you accepted that you do not have to bear, let’s talk about it) more like facilitating a conversation that would not happen naturally. And its empowering.

Michelle

Pen to paper can be a barrier because so many things are bubbling up. I haven’t seen a model like this other than a community forum. The reason why all those activities worked because it was multi-modal and accessible.

Angela

The video replay was a great focal point. The explanation helped set the tone very well.
Jen R.

I think this project opens the door for conversations around better testing practices in America. I cannot see how it lays out specific action steps. The free-write graffiti boards may have revealed some action steps in the responses of the participants. (You would know better than I.)

4.) What resources are available to the WAP? Number of staff? Total budget?

Sources of funds?

There were 29 staff: 15 on the production team and 14 performers. National Council on Measurement in Education Fund Grant Award of $10,000. In-kind resources of UMass: New Africa House Studio space, lighting, sound, clean and breathable air. Personnel: designers that helped bring to fruition vision, advising team, Intergroup dialogue team. Scripts hours and hours of volunteers (about 30 student reviews – Upward bound students and undergraduates). An employed Producer that could fill in gaps for financial needs. Scope and deliverables – good timeline and space to adapt. Friends and family – informal conversations and love support (it takes a village). Circles of community and influence. The resources page of the website includes several outputs; and these materials are available to the public. Website tracking, zoom, and internet.

5.) What outputs are being delivered by the project? To whom?

The intervention, final report, website, public facing documents, two NCME presentations for researchers and practitioners, interview preview video. Strengthening of the relationships in the collaborators. Community empowerment. Conversations with people were happening and they brought them into their world and context. Shaken the
field of assessment in a way. The community that was built and friendships. The bonds on the projects that we collaborated on following, those have importance as well. Lots of data over time.

6.) What evidence is necessary to determine whether goals are met?

Iya

We have a blueprint. Action steps of being able to tell people how to do this. Empowering space- people spoke freely in moderated conversations and wrote freely about it on the wall. Word of mouth, social impact.

Jen O.

Generated substantial data from the process and performance. The script is really important. The steps to better education, got feedback with tangible ideas of how testing can work better.

Cassie

Different mediums of sharing the gospel: raising voices on social media, blog posts, memos to policy makers..lot of different mediums but the word necessary is hard to grapple with.

Afrikah

Changes in teacher curricular approaches and the community impact data.

Steve
Positioning the wrong answer project to **partner**. Leverage the **website**. Evidence of webpage hits and contacts.

**Jen R.**

You will need **qualitative** data (in terms of individuals telling you explicitly what their action steps were) or some other **written evidence**. Hindsight is, of course, 20/20, but ideally you would have asked people at the end of the performance or **followed up** with them via survey, text message, phone call, etc. and asked them what Wrong Answer has inspired them to do next. **Testimonials from participants** that the space created was **empowering** (you have some of these on the website). For the third goal, you will need evidence that **someone has taken up** the shell/frame from **W.A.P.** and actually **developed/modified it** to for some other sociocultural group.

7.) **What happens if goals are met?**

**Iya**

It **works** and people are **happy**. Social and system **change**. It’s one of the many steps that are needed to **shift the education** system that has collapsed. We have something. An alternative that we **know** works. People who do love children and education are looking for some **right** answers **right** now.

**Jen O.**

**School practices changing** if they are not bound by high stakes testing. Could **change** the **kinds of thinking** and **projects** in schools. Could **change** a lot of
relationships (trust). Has the potential to restructure education. We are going towards discovering the gifts of collective.

Steve

Professors and teachers adjusted their praxis after the experience.

Cassie

Equity: potential to really become an independent thinker.

Michelle

Tests are no longer tools of oppression. The role of tests in education changed – support student learning and pursuit of goals. People of color are at the table from construction to actualization. A multigenerational reframing of how we understand education.

Jen R.

If the goals are met, you will have contributed an excellent framework for public and widespread use, which would mean widespread impact.

8.) What happens if they are not met?

Jen O.

We have built into this process a constant self-evaluation. It is such an interactive relationship process there is always opportunity for course correction. At all points there is always dialogue in the process. Even if there are upsets, we can always talk through them.
Iya

If the goals are not met, we will keep moving toward death and destruction: Artificial Intelligence, deadening our thinking…. like chat GPT. You won’t be able to have your own thoughts. AI is not creative. Not dialogic and not embodied. Its pushing teachers to go to those creative practices. Tapping into multiple modes/methods of instruction.

Cassie

Hard for the goals to not to be met because the community that was built and the space that was created by others.

Michelle

If we are given a standardized test, we can just check boxes, when you reimagine testing that seemingly more complicated. Makes intelligence more black and white than it actually is. Not standardizing will make it more complicated but more valuable.

Steve

If we are in the schools and implementing W.A.P. and at the end of the year, they are still taking the test.

Jen R.

If the goals are not met, you will simply have provided a smaller community with an impactful experience.

9.) How is the project related to local priorities?
It will strengthen relationships between administrators and educators in every situation. It’s concerned with education and so every locality is concerned with this. It has the capacity to illuminate what is happening locally. There are certain local issues that are not the same everywhere. Education system is hyper local. One thing we learned is if you do better on high stakes test, we will get more money for the community. There is a strong and growing commitment to assessments that value and affirm all students (not just the white ones) if the ultimate goal of this project is to disrupt this harm, that is how it is related to local priorities.

10.) What data or records are maintained? Costs? Services delivered? Service quality? Outcomes?

Brown paper graffiti, notes from dialogue circles, script, implementation plan, pilot report, website. Testimonial – really green actors, keeping the space open for all levels of performers and creative experience, the spin. Experiencing creative churns something beautiful in you.

11.) How often are these data collected?

During implementation, and during design when the project is initiated.

12.) How is this information used? Does anything change based on these data or records?

Activate new implementation initiatives, used as artifacts of the success of the project. Script – fluidity…each time you do a script it’s going to effect the next one. Team will change, other parts of the process. Culture focus will change given the group,
the budget will change…even if the money is not all there… could have stage reading…. or full-on scale up…it will adapt to the resources of the community involved.

13.) What major problems did we experience? How long will it take to solve those problems?

Iya

Snow – and still got good crowds and people mostly stayed…people wanted to talk…majority of the audience stayed to engage in graffiti writing and solution circles like 70-90%. We had great publicity. Word of mouth was also great.

Jen O.

Time (rehearsal time, more and more in public life people don’t have any time). But we can share the project to the time we have available. What can we do with the time we have, its valuable regardless of what it ends up being. With less experienced performers there were accountability issues. Had plans to replicate during the summer but COVID happened. Completely ephemeral, but took a little bit of theater magic: problem solving, having each other’s back, responsibility, accountability, collective hope and faith for the work. Space – we rehearsed in different spaces at different times, and coordinating everyone’s schedule.

Darius

Choreography had to shift a couple times for the learning level of the performers and the devising process. Performer dropped to due to injury. So, the founder/producer had to jump in at the last minute.
Cassie

Sound system went out and Darius had to give the curtain speech. Performers felt uncomfortable with singing but felt safe singing as a collective.

Michelle

The way that information is kept: how are you dealing with the volume of data? Have you thought about ways of distilling.

14.) What results are likely in the next two to three years?

Feasible to have the project get a broader reach on social media or be repeated based on the blueprint that exists. More conversations more awareness. One of the goals to be empowering people and marginalized groups effected. So much potential in the coming years as it reaches more people and will inspire people to investigate the issues years later.

Assess the plausibility of the program.
(Respondents: Founder/Producer, Associate Producer, Theatrical Director).

Only the Theatrical Director (Jen O.), Associate Producer (Iya), and I discussed W.A.P.’s outputs and the progress towards outcomes. We set the intention of honestly exploring the capacity of W.A.P. to meet its mission and goals. Continuing with the Wholey way, we assessed project plausibility guided by the questions adapted from Strosberg & Wholey’s (1987) 8-step model. There are ten evaluability assessment questions that focus on assessing the plausibility. All results are organized by question and presented in dialogue with emphasized formatting.

1.) Are program activities and resources likely to meet planned objectives?
Jen O.

Yes, they already have. We proposed action steps to a national review board, we provided an empowering space where people from several marginalized groups could convene and discuss their experiences with testing and their suggestions for changes and improvements. We stimulated conversations which have still continued around shifting current testing practices.

Iya

I believe almost any future iteration of this project (from staged reading to focused discussion group to full production) would lead to similar outcomes. People, time and money resources – we certainly have these and did have them when we started the project in the New Africa House.

Darius

Yes, we showed what partnership and limited resources could really pull together. A continuation, replication, expansion would mean at minimum partnering with an institution of some sort with in-kind resources and applying to more grant award opportunities. We have a solid model that could work with any of the marginalized groups intended.

2.) Are there any major impediments to the achievement of program objectives?

Jen O.

I think that the major objectives were met. I wonder about how the voices and data gathered from future projects will be captured and shared with key stakeholders
and test designers in the way that this pilot succeeded in doing. It may be that future productions need a connection with a stakeholder organization such as a university or a policy group for it to achieve its full potential.

**Darius**

The major impediment was the snow and time to rehearse. There was a snowstorm in December during the pilot, but we were still able to get a full house both nights. Our actors/performers were fairly busy and some new to theater, so getting folks to show up and be intentional about the work was a bit of a challenge but it all worked out beautifully in the end. I am really busy. There is additional work that needs to be done to really raise voices. There has been dissemination efforts within the research and practice spaces and within the media. There could be more efforts to raise up the voices to educational measurement leaders, administrators and developers.

**Iya**

The most important piece is to allow time for script development and rehearsal. We want the artists involved to feel grounded and confident in the work. It is important not to rush on time. Give it its due gestational period.

3.) Is there any evidence that program outputs will meet expected outcomes?

**Jen O.**

For me, the conversations of audience members and their sharing the impact of standardized tests on their lives as well as their generating alternate forms of assessment for greater equity --these met and exceeded the expected outcomes. The next
step for me is how to convey this information to policy makers, test designers, and high-level educational supervisors.

Iya

The evidence that we did it and it worked beautifully. The data that was collected and the multi-cultural and multi-generational composition of the audience was a huge success. We had high school students, elders in the community, university folk, just a lot of different people. So, this will draw a variety of voices.

Darius

The W.A.P. has caught social traction in positive and negative ways. The expected outcome is to interrogate and encourage reimagining and redesign in respect to standardized high stakes tests. This is currently taking place. It’s an exciting time as we learned from the pandemic that doing away with high stakes tests for a season does not employ much damage to our educational systems. This is a great season to continue to implement this work.

4.) Should we spend money on the WAP even though important pieces have not been implemented appropriately?

Jen O.

I am not sure how to respond...I think we did implement the project pretty closely to the goals that were established. Should it be continued? I think yes. I think there is a huge need for new kinds of thinking and solutions to massive issues
happening in education right now, and a creative, collectively-designed, community-centered project is an ideal way to generate solutions.

Iya

Yeah we should spend money on it. It’s a creative process that requires willingness and expectations of discovery. If it was predictable, it would have been done already. Its innovative. It’s bringing together students, community, and educators in a process.

Darius

Yes, no piece was implemented inappropriately. There does need to be continued pressure on the educational measurement community by disseminating findings and continuing the social action campaign.

5.) Does WAP have well-defined uses for information on progress toward its measurable outcomes?

Darius

We have been mentioning the internal iterative community processing that takes place in the development space for the theatrical work which really contributes to a holistic and pluralistic product. A use of information on progress towards more long-term measurable outcomes has to do with how we position the story that is told based on the current events, practices, and policies. As society changes the narrative told will be updated, transformed, or no longer needed (hypothetically – I mean the system so
broken not sure if “no longer needed” will ever happen in my lifetime, but one can hope!)

Iya

There is a constant conversation about script development, student participation… flexibility and adaptation are required parts of the theater making process. Measurable outcomes are the positive engagement with an alternative evaluation method that centers the students voice, needs, skill level and aspirations as an individual rather than measuring them according to a standard. Everyday young people encounter the mainstream standardized testing system those measurements are becoming more and more obsolete.

6.) Do project staff use process data?

Jen O.

I'm not sure what this means; will our team use process data? I'm not sure. Maybe for feasibility and marketing purposes?

Iya

Yes. The project staff are artists and educators, and we are constantly revising the data to strengthen the process. We expect things to change and are built for adaptation because we are used to collaborating on a project. We like to engage in conversation. We thrive on connectivity.

7.) Are there any clear indicators of model fidelity? And how to enforce fidelity?

Darius
The clearest indicator of model fidelity is to always be community-driven and pluralistic in value. This is enforced by establishing a consciousness to that being the center of all the work. Its an inclusive fidelity in a sense. “Ya’ll think this good? Yeah! Bet, let’s move forward.” OR “Nah. Well, let’s sit with it and discuss.”

Jen O.

I’m not sure if the project will roll out to other communities with one of the original team members acting as a consultant. If so, that could help with some fidelity to the mission. When we did the original project, we were creating our own model of fidelity. Regular production team meetings to connect to mission and goals were crucial for this.

Iya

Ensured by the fact that brown paper studio seven basics are a fundamental structure. We constantly go back to those basics in the work in the studio and our audience. They may grow but they cannot grow and change without consensus and conversation.

8.) What portion of the WAP is ready for evaluation of progress toward agreed upon objectives?

Jen O.

The script, video of production, testimonials, brown paper feedback, discussion feedback. Longer term impacts are perhaps not currently measurable.

Iya
All activities to date are ready for evaluation. We have not tested it in a school. So that is the next step and what we are preparing for.

9.) What evaluation and management options should the founding team consider?

Jen O.

I think if we could develop a project assessment for what happens with all the community data that is gathered that would be a useful tool for continued project sustainability. For example, what action steps is the community taking after this project? And tracking those. Does the community present a list of recommendations to policy makers? Does the community work with local and state education authorities to push for different kinds of assessments (as has happened in NYC with the Consortium for Performance Based Assessment and regular lobbying with the NY State Board of Regents) Can we develop a digital portal to gather testimonials and data to regularly present to policy makers on our side of things?

Iya

Management options: a fluidity of roles in the management/artistic team. Important members of the team should write their evaluations at the same time that we reflect with the community. Journaling and listening circles are good evaluation tools. The results of those have to analyzed and used.

Darius

W.A.P. is a unique creation intersecting storytelling, learning, and empowerment. The founding team should consider an empowerment, educative, and indigenous
evaluation method. It can be managed against the wise of leaning into storytelling, its capacity to gather learnings, and link success to a model that takes into account the relationships within nature.

10. What are recommendations for areas not ready for evaluation?

Jen O.

I believe I shared some recommendations above.

Iya

The results of a school - start in schools that have populations that are recognized in this study. Talking about black and brown youth unfairly stigmatized in testing.

Darius

There needs to be a formalized social action plan that can measure impact to the community and the educational measurement field.

Reach agreement on any needed changes in program design or implementation. (Respondents: Founder/Producer, Associate Producer, Theatrical Director, Production Manager, Grant Officer, Intergroup Dialogue Designer, Student Performers, Knowledgeable Observer, and Knowledgeable Reviewer.)

The cumulative data gathered and discussed during the evaluability assessment helped us as the key stakeholders and users of evaluation reach an agreement on any needed changes in the project design or implementation. Up to this point we reviewed the project definition, design, activities, outputs and outcomes. Throughout this review the goals of W.A.P. were measured and gaps, shortcomings, and improvements were identified. This step of the evaluability assessment charged the key stakeholders to identify and agree upon the changes needed to improve project outcomes. The Insights
Team and I explored four questions. The third and fourth questions were combined as they both related to implementing changes. Thus, there are three sections of results each are presented with emphasized dialogue.

1.) What components of the program design, activities, outputs, and outcomes have gaps/shortcomings, seem unrealistic, or are not working?

**Jen O.**

I think that the capacity of various groups to do a full-fledged production may be unrealistic, especially if the project is happening at a public school or a community center. Potential accommodations or modifications for groups with less staff and budget for a production could be: a list of plot points to consider for the devised script (i.e. establish the main character and their personal and community context, establish the educational environment, establish the conflict (the struggle with the exam, the pressure on teachers, the pressure on students), establish the impacts of the exam on the protagonist and community members, etc.) to help guide those who may be less experienced with playwriting or devising; a list of possible ways to perform or present the material (a staged reading followed by a facilitated discussion); a list of potential action steps.

**Iya**

A requirement of the success of the project is trained facilitators in brown paper studio. The biggest question is how to gather a team and apply brown paper studio methods to the next Wrong Answer Project implementation.

**Darius**
The implementation model works brilliantly. The dissemination phase and public reporting needs some work. It is important to keep the campaign going. To stay unlimited in the movement there needs to be a plan for continued social action efforts that utilizes the website, developed products and social capital so that interrogation and reimagination does not fall mute. This could mean advocacy within the political and operational educational testing spaces. Also, the use of social media for contact points.

Jen R.

The question I have is whether or not this project is intended to focus primarily on the use of theater/dance to reach your goals or is the intent to inspire any creative/artistic approach? And if it is the latter, what would some of those outlets be? And for those who do not engage with the creative arts (because they/we have no talent, for example) – is the shell/template created by the W.A.P. just not for them? Does this kind of project have a particular audience in mind?

Steve

The goals of the project are awesome, and they can’t be accomplished by one project. W.A.P. is a huge part but it’s not going to solve the complete issue. Your staff for going after grants is you. From a formative comment that is a place you can build up. Build your grant team is an important next step.

Michelle

It would be helpful to have some process and focus around designing the specific research questions that are connected to the broader goals of the project. Interrogate the construct and criterion you use? How do you get from A to B? You collect all this
information how do you get to your goals? How do you start to make track with very specific research questions. I would like that more clearly defined in your process.

Cassie

Proved it is realistic. Agree not much shortcoming because it was such a success.

Steve

Curriculum, what can you do in a 40-minute class? How do you break it up? How do you do it for little kids? How do you do it for the adults?

Michelle

A value add could be that connection with industry. How to make the connection tighter? As someone in industry, how do I as a psychometrician consume what you are putting out. Test developers tend to focus on industry. Reality is groups that develop tests are asking for something specific. Be careful about defining who is the audience for this more than developing questions and scales.

Sabine

During the facilitated dialogues some of the sessions didn’t get to all the semi-structured questions. Those dialogue questions were important for collecting more data. Also, more conversations with the facilitators for the overall vision of the dialogue, what are some of the key points.

Steve

When I hear about running out of time in a dialogue, the picture that goes in my head is a really good dialogue. Makes sense that good conversations take a path that is
hard to structure while respecting what people are saying. Maybe following up with a survey for unanswered questions?

Jen R.

There needs to be some component that articulates or allows the participants to articulate action steps. What those steps are does not matter, but the point is to empower. It is certainly possible that someone could leave the experience feeling dejected (e.g., tests are bad and they hurt our kids) but they need to think about what they can do about it. Including the script is, obviously, wonderful and critical – but I also suggest annotating the script for folks, so that they know your thinking during every scene. In this way, they can adapt it “properly,” so to speak, for the groups they are trying to target. Just telling people what the purpose of a particular scene (or a few lines) was; or what was the underlying issue you were addressing with a particular scene/line. Almost like ELA teachers with Shakespeare – make sure people understand the logics behind every decision you made, so that they don’t end up modifying it into a hot mess.

2.) Which of these components should be removed? Which should be changed?

Jen O.

At this stage I don't think any components need to be removed or changed but may need adaptation for subsequent groups.

Darius

The dissemination efforts should be changed.
Cassie

Creative decisions that we made during rehearsals were not set but it also helped for free flowing. More foundation needed to be set. Videography and “the intervention” went very well. Are we setting intentions at the beginning that this is in three parts. Setting the intention of what to expect from the audience. Marketing struggle.

Outreach, if you didn’t know someone in the performance how to get people to come up? Reach to wider audience?

3.) What changes should be made to improve these components? How will these changes shift evaluation and management options?

Jen O.

If a group is adapting the production or presentation format that may impact evaluation. It seems like this project would need to be replicated at least one or two more times to foresee shifts in evaluation or management. The Wrong Answer team may need to consider fundraising for operating support to be able to do the consulting work needed for managing the project. The Wrong Answer team may also wish to be part of collecting documentation and evidence about the projects.

Darius

Letters and other products can be designed and developed for political officials and leaders in the educational measurement spaces, as well as middle management and students to keep the coals hot with change.

Steve
Partners and staff. Need to implement better public reporting to get the word out. We need somebody funding the organization. What can we do to pub it more? Can we get NCME to pub it more? It’s a Mission Fund Project. They may have some ownership in it. Get the word out with the marketing plan, more of word of mouth? Connect to Amy Stuart Wells work? Can she help promote it? Ezekiel Dixon Roman Director of Gordan Ramsey to partner with? Ed Gordan every Friday. Look back at what you considered who your audience is? Pub it on social media. Local reporters could come cover as well to spread the word. People are excited about interactive, who will feel the empowerment and growth? A lot of digital product already. Might need to build a staff.

Michelle

It’s open source with an intention to be a movement that grows. How does that fit in the practical world of licensing? There are a lot of ways to think about licensing. You don’t have to be responsible for every production. But you do have some ideas about good use and bad use. You have some best practices and a vision forward and ways to grow this. You should have some control over inappropriate use.

Steve

Big omission. Contact us! How do people get in contact with you. “Are you interested in doing something like this too? We want to spread this project. Join the movement.” On the website, social media, publicize it man. Talk to Jamila about how she marketing herself.

Jen O.
Thinking about the script being on the website, how do we monitor its usage. Set up a sample but if you want to use the whole script. Let’s talk about it.

Cassie

Little cards to give with a QR code? Add all the information. How to keep in touch, website, forms to submit more of their experience. Something like that.

Sabine

Further the conversation with the young people. Establish more relationship.

Cassie

The graffiti writing time gave folks a good amount of time to get ready to converse. After if you want to hang around and write anything else that you didn’t get to mention in group. Or leave a question or on FB tonight we will post. (Give folks a hook to stay connected).

Angela

Spreading out the methods to others, budgetary constraints? Video permission? Does everyone have rights to the video? Financial barriers?

Reach agreement on the focus and intended use of any further evaluation.
(Respondents: Insights Team, Advisor/Chair & Dissertation Committee.)

The final step within the evaluability assessment was to reach agreement on evaluation focus and intended use. The results of the evaluation research performed on the W.A.P. informed the development of a working logic model, theory of action, and a menu of evaluation options. I consulted with my insight team and my dissertation
committee on my evaluation framework and the appropriateness of different evaluation methods for W.A.P.

Here in a list paragraph of paraphrases and key words is the feedback on the presented framework working title *Evaluation as the Water Cycle*: really intriguing, really important transformative questions. How do you think about engaging in it? Process is cyclical and that’s really beautiful, appreciate it and never thought about it in this way feels disruptive to white dominant norms and space for more time. Appreciate the elegance of the model, what it accomplishes that traditional models cannot accomplish. It encompasses the magnitude of the problem that other models cannot. It gets our arms around the magnitude. Evaporation connected to transformation – feels like it takes away but it goes up which is positive. Ties in that the work takes a lot of time, I still have a strong sense of what the evaluation is and the goal you are trying to receive, its relatable to the lay person. It’s an accessible model. Water connects the continents and allows us to live. Water cleanses and is a rebirth. What you described is really dynamic, constant learning process. You will want to make documented meaning at a certain point. What does this look like specifically?

The evaluation framework is transformative, educative, and imitates nature.

**Transformative Evaluation**

According to Donna Mertens, “Transformative evaluators ask provocative questions.” (Alkin, 2004, pg.229). As we build an evaluation plan for the *Wrong Answer Project*, Mertens encourages us to consider transformative evaluation questions that focus on four philosophical assumptions: Axiological assumptions (the nature of ethics), Ontological
assumptions (the nature of reality), Epistemological assumptions (the nature of knowledge and the relationship between the researcher and the community) and Methodological assumptions (the nature of systematic inquiry). Some pre-formulated questions from Mertens writings that could contribute to the evaluation of W.A.P. are:

- How can this evaluation contribute to social justice and human rights?
- To what extent is the evaluator able to identify cultural norms within communities that are supportive of or deleterious to the pursuit of social justice and human rights?
- How do evaluators demonstrate that they have taken action to support those norms that support human rights and social justice and challenge those that sustain an oppressive system?
- To what extent did the evaluator reveal different versions of reality?
- How did the evaluator determine those versions of reality that have the potential to either support or impede progress toward social justice and human rights?
- How did this evaluation contribute to the change in the understanding of what is real?
- What evidence is there that the evaluator explicitly addressed issues of power differentials and that the voices of the least powerful are accurately expressed and acted on?
- How did the evaluator establish a trusting relationship with the stakeholders?
• How was a cyclical design used to make use of interim findings throughout the study?

• To what extent did evaluators engage with the full range of stakeholders to gather qualitative data that enhance their understandings of the community?

• How were data collection methods used to be responsive to the needs of diverse stakeholders?

Some of these questions will be invoked into the drafted evaluation plan below.

Educative Evaluation

Jennifer Greene writes about evaluation being educative (Akin, 2004, pg.97). Through a summary of the published works on measurement and evaluation by Lee Cronbach, Jennifer Greene gives a perspective of evaluation activities and results being useful if it results in “actionable substantive insights”. For me that means informing policy and practice. Can the data we collect from the Wrong Answer Project inform social justice? Well, there is potential to learn from these evaluation data to inform change strategy. Our voices will be lifted to policy makers, practitioners, interrogators, and agitators to give water to the rain from the reservoir (grassroots).
Indigenous Evaluation

According to Aristotle, art is an imitation of nature - the whole of created things (Boal 1). An indigenous (and biblical) approach to evaluation looks to nature as an evaluation model and has a focus on community and culture (LaFrance & Nichols, 2009). Why did I say biblical? Well, often times the bible will reference nature to support teaching and learning. For example, in the book of Matthew 13:31, Jesus said “The kingdom of heaven is like a mustard seed, which a man took and planted in his field. Though it is the smallest of all seeds, yet when it grows, it is the largest of garden plants and becomes a tree, so that the birds come and perch in its branches (King James Bible, 1769/2008).”

The cycle of water will be the structure of what I would like to consider for the evaluation framework and models. I propose framing all W.A.P. evaluation models in respect to a three-round framework of the water cycle: collection, evaporation (referenced here forward as upliftment), condensation, precipitation.

Recall that the Wrong Answer Project is the pilot concept for a large-scale social action campaign. The overarching project goals are to:
• develop stakeholder driven action steps for “better” high stakes testing practices in America

• provide an empowering space related to educational testing for marginalized groups who are characterized by race/ethnicity, economic status, and access

• ignite a grassroots movement that shifts current high stakes testing practices and policies to a more equity-centered design in testing

The specific objectives of the December 2019 pilot project were:

• To develop an adaptable theater script that can be manipulated to represent any marginalized groups’ perspectives and experiences in high stakes testing

• To facilitate an interactive experience focused on the discourse of high stakes testing perceptions amongst marginalized groups

• To develop replication materials that allows for public use of the project as a method for contributing more marginalized perspectives to this discourse

The following sections describe each of the water cycle stages within the proposed three-round framework. The first two phases have been expanded to highlight how transformative and educative questioning could be invoked within a given evaluation model. For the development of an evaluation plan, it will be important to establish the scope of the evaluation in respect to the three-round framework. For example, the implementation and evaluation scope for the December 2019 pilot would fall between stages 1.1-2.4.
1.1 Collection: During the first collection experience we should be positioning W.A.P. within the culture and testing experience of the source population for the work. It sets the stage for the work. For the pilot this was the high stakes testing experience of low-socioeconomic status Black high school students.

- **Objective**: Set the stage of the Source Population

- **Activities**: Define source population, location, budget, and team

- **Process (inputs)**:
  
  - Who is our source population? (stakeholders)
  
  - Where is source from? Where will we host the experience?
  
  - What is our budget? In-kind resources?
  
  - Who is our production/implementation team?
  
  - How do we engage (collect data and insights) with source population to understand their culture and high stakes testing experience with precision?
  
  - What data are collected during engagement with source population?
  
  - How are data collection methods be used to be responsive to the needs of diverse members of the source population?

- **Impact (outcomes)**:
1.2 Upliftment: This water upliftment stage has to do with preparing and planning the WAP for implementation using the culture and testing experiences in collection. Lifting the voices and stories shared during collection, upliftment invokes these voices into the development of the implementation plan and script devising process.
- **Objective**: Develop culturally responsive implementation plan and script

- **Activities**: Devise (draft, review, edit, adapt cycle) implementation plan and script

- **Process (inputs)**:
  - Who is on the devising team?
  - How was the devising team selected?
  - How do we engage (collect data and insights) with source population to devise products embedded with their culture and high stakes testing experience with precision?
  - What data are collected from source population during devising process?

- **Impact (outcomes)**:
  - To what extent does the devising process represent different versions of reality?
  - To what extent does the products devised address the cultural norms within the source population that are supportive of or deleterious to the pursuit of social justice and human rights within high stakes testing?
  - How do evaluators demonstrate that they have taken action during development of the implementation plan and script to support
those norms that support human rights and social justice and challenge those that sustain an oppressive system?

- How did the evaluator determine which versions of reality to uplift within the devised products that have the potential to either support or impede progress toward social justice and human rights of high stakes testing?

- How does these products contribute to a change in understanding the source population’s reality of high stakes testing?

- What evidence is there that the devisors explicitly addressed issues of power differentials and that the voices of the least powerful are accurately expressed and acted on?

- How did the producers/devisors establish a trusting relationship with the source population during the devising process?

- How was a cyclical design used to make use of interim findings during the devising of the implementation plan and script?

- To what extent did producers/devisors engage with the full range of stakeholders to gather qualitative data that enhance their understandings of the community?

- How were data collection methods used to be responsive to the needs of diverse devising team and source populations?
1.3 **Condensation:** The condensation phase is the rehearsal and recruitment. The amalgamation, coming together to live in the source populations’ culture high stakes testing experience and rehearse reliving a cultural experience that produces truthful responses from the audience.

1.4 **Precipitation:** The precipitation is the implementation of the experience to an audience of students, parents, teachers, professors, administrators, test constructors, policy makers, practitioners, interrogators and agitators.

2.1 **Collection:** We get back to collection as the audience members sit with the culture and testing experience. Reflecting and embodying the “intervention” experience.

2.2 **Upliftment:** This upliftment is the graffiti and intergroup dialogue data collected. Audience members then contribute their voices, stories, and experiences. Subverted by the theatrical experience.

2.3 **Condensation:** Condensation is the synthesis and analysis of the data from the audience and program evaluation data. Answering specific research questions.

2.4 **Precipitation:** Precipitation is the social action campaign and advocacy to educational measurement influencers (students, parents, teachers, administrators, test constructors, policy makers). Multi-modal dissemination and constant contact with source population communities.

3.1 **Collection:** Collection is sitting with the advocacy efforts received. This phase of collection takes some time. Social and political impact over time.
3.2 Upliftment: The final upliftment is the adoption of the advocacy suggestions. What’s this process look like?

3.3 Condensation: Condensation is the synthesis and planning of change efforts in testing.

3.4 Precipitation: The third precipitation is the reimagined high stakes testing models implemented.
CHAPTER 4
DISCUSSION & NEXT STEPS

Discussion

My dissertation study is an evaluation research project. I used the Evaluability Assessment framework to explore the following research question:

**Does the Wrong Answer Project serve the purpose it is purported to serve?**

*Who* is it serving? *Is it serving them* in a way that serves them best?

Using Joseph Wholey’s 6-step method I selected nine valuable and invested key players and experts to join the Insights Team (evaluability assessment working group). Then I reviewed the *Wrong Answer Project* (*W.A.P.*) artifacts to clarified the program design. As a collective the insights team explored the program reality, assessed the plausibility of the program, talked through needed changes, shortcomings and improvements. I also developed an evaluation framework, *Evaluation as a Water Cycle*, to discuss as an evaluation and management option with the insights team. For the discussion, I will focus on answering the research question directly and referencing the results captured during each of the steps of Wholey’s evaluability assessment design in part and aggregate.

Recall that the *Wrong Answer Project* is the pilot concept for a large-scale social action campaign. The overarching project goals are to:

i. develop stakeholder driven action steps for “better” *high stakes* testing practices in America
ii. provide an empowering space related to educational testing for marginalized groups who are characterized by race/ethnicity, economic status, and access

iii. ignite a grassroots movement that shifts current *high stakes* testing practices and policies to a more equity-centered design in testing

The specific objectives of the December 2019 pilot project were:

i. To develop an adaptable theater script that can be manipulated to represent any marginalized groups’ perspectives and experiences in *high stakes* testing

ii. To facilitate an interactive experience focused on the discourse of *high stakes* testing perceptions amongst marginalized groups

iii. To develop replication materials that allows for public use of the project as a method for contributing more marginalized perspectives to this discourse

**Findings Summary**

As the producer and founder of the pilot of this work with the diverse voices, honest feedback, and thorough review of the key players and experts’ insights, I deduce that the *Wrong Answer Project* as a social action campaign is moderately serving the purpose it is purported to serve. The project even in its infancy has begun meeting the goals of (i.) *developing action steps for “better” high stakes testing practices in America* and (ii.) *providing an empowering space related to educational testing for marginalized groups who are characterized by race/ethnicity, economic status, and access*. However, the goal of (iii.) *igniting a grassroots movement that shifts current high stakes testing practices and policies to a more equity-centered design in testing* needs improvements to truly serve the purpose *W.A.P.* is purporting to serve. The December 2019 pilot met all three of
its objectives and was a successful exhibit of the Black low socio-economic status student’s experiences.

**Exploring the Program Reality**

The Wrong Answer Project is trying to accomplish democracy within high stakes educational testing. True democracy – for the people by the people – by creating an embodied space, process, and experience focused on intersecting minoritized Black and Brown identities that invokes justified indignation and civic responsibility and fuels a grassroots movement towards the change of high stakes testing practices. It is trying to accomplish a new way of gathering authentic data on educational testing consequences. *W.A.P.* is also trying to raise up the pluralistic voices of the unheard, marginalized, and erased peoples to the heard, hegemonic, unfairly magnified people. Not for the raising up alone, but with hopes that
voices and suggestions would be heard and considered for the reimagination and reconstruction of high stakes testing.

During the Evaluability Assessment, the Insights Team was asked to examine program artifacts (i.e., the application, Implementation Plan, script, dialogue facilitation protocol, final report, program testimonials, and website) then discuss the goals for \textit{W.A.P} in their own words. An aggregation of their responses using key words and concepts can affirm my definition above. According to the collective Insights Team, the goals are to gather voices, hold conversations, talk, engage in deep reflective listening, have open dialogue, bring people typically not considered and diverse learners to the table concerning problematic, broken, issues and gaps within high stakes testing to dismantle the voice of the majority, create long lasting change, move and shift current policies, practices and procedures, teach white people, encourage and empower our communities that have felt the brunt, and to be a mechanism for industry to understand going to the core and root experiences of people.

\textbf{Assessing the Plausibility}

When discussing plausibility, the insight team focused on how did \textit{W.A.P}. work toward achieving its goals and objectives and why did those activities that led to the goals and objectives succeed in meeting them. We have a model and artifacts from the pilot of the \textit{Wrong Answer Project} which includes: an implementation plan embedded with
Brown Paper Studios theater for social transformation methods (a community responsive practice), a script that can be adapted and devised for the Black and Indigenous population, and a grant writer (Me) to pitch the work to a few foundations.

The insight team offered powerful language to support reviewing the plausibility of the program. W.A.P. is building community coalition, opening the door, a mirror and a window, creating bonds and relationships, by affirming the script, assessing the whole person, free-writing and facilitating a conversation using a multi-modal, in-person, interdependency of activities in the spirit of encouragement and empowerment.

Figure 11. Audience Graffiti Response to Prompts

Figure 12. Audience Facilitated Dialogue
**Improvement Mechanisms**

My plan is to be more intentional about sharing the work of W.A.P. This would happen by applying some of the feedback from the Insights Team to update the website, marketing, and disseminated products. There is a need to establish a connection to the testing industry and co-conspirators in the cause by installing a formal marketing and publicity plan. My next steps are to come together with a marketing team including my Creative Director of the pilot for this work to develop a strategy for the raising up of results like those included below. I imagine taking advantage of the virtual age of social media to develop marketing on Facebook, Twitter, Instagram and TikTok. There are current innovative and trendy ways that we can reach more people and hook more interest.

**Figure 13. Test Time Spoken Word Highlight**

**Figure 14. Audience Solutions from W.A.P. 2019 Pilot**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Society’s Re-imagination of High Stakes Testing</td>
<td>Collaborative vs. Competitive; Integrate into curriculum; Non-privatized; Centered on Different Student Learning Styles; Embedded with Critical Thinking vs. Knowledge Regurgitation; Embedded with Career and College Standards; Project-based; Essay-based; Pre-Test/Screener of Appropriate Testing Method for Examinee; Testing optional Boycotting; Test Prep for All; Capture Voices from Society; Run for Political Office; Push a change campaign;</td>
</tr>
</tbody>
</table>

What would re-imaging assessments look like?

What are next steps for change?
W.A.P. has a blueprint for raising collective voices allowing for the embedding of a multi-generational framing by holding a space for constant self and collective evaluation, an empowering space, pluralistic in value where tangible ideas are developed highlighting community need and impact. The reality is that we must sustain resources and funding. The Wrong Answer Project will always need a space and place to implement. It is imperative that we partner with education institutions that have access to theater space or equipment so that we have in-kind resources. We need to sustain funding to continue to produce. Also, assessing community impact is a bit tough. There needs to be a way to hook audience members so that we can follow-up on how their perspective may have changed. Spaces and audiences that receive the results of the experiences (the grassroots solutions) should receive follow-up to assess impact and lasting change. This would require the collection of more contact information to solve the problem of establishing measurable societal impact.
Validity of Evidence for Social Consequences

Shephard (1997) stated, “most measurement specialists acknowledge that issues of social justice and testing effects are useful ideas, but some dispute whether such issues should be addressed as part of test validity. They worry that addressing consequences will overburden the concept of validity or overwork test makers”. (p. 5) Is this “worry” or “overburdening” a viable excuse for the measurement community to use to defend the appalling lack of work in this area? It is concerning that over the past two decades, evaluating the social consequences of test use has been largely ignored by measurement specialists.

Messick (1989) stated “the terminal value of the test in relation to the social ends to be served goes beyond the test maker to include the decision maker, the policy maker, and the test user” (p. 89). Thus, evaluating the consequences of testing is a community responsibility, but such sharing of responsibility does not mean test developers should have the freedom to wash their hands and be done once a test is assumed to be used for a purported purpose in society. In reference to Koretz, Barron, Mitchell, and Stecher (1996); Pomplun (1997); and Chudwosky and Behuniak, (1997); Lane (1998) mentioned that “researchers are now beginning to examine the consequences of assessment programs by using various procedures such as surveys of principals and teachers and focus groups” (p. 25). She continued, “there is a need for more direct evidence, such as

Figure 17. Movement Expression On Test Day
evidence obtained through the analyses of classroom instructional and assessment activities, in addition to information obtained through surveys and interviews” (p. 25).

My study highlights a project that has the potential to transform how we use validity evidence for the social consequences of testing. From the deep review of the Wrong Answer Project, there was much support for the experience empowering, encouraging, and offering hope for educational change and equity. The U.S. Department of Education recognizes fairness as an important component of technical quality. They recognize that assessments should be “accessible to all students and fair across student groups in their design, development and analysis” (USDOE, 2018, p.18). As we turn to the Standards (AERA et al. 2014) fairness is presented as an evaluation of: the test administration process, measurement bias, access, and individual test score interpretation. From the perspective of the Standards, a test is not unfair simply because subgroups have different testing outcomes. Fairness recognizes these differences may exist and should be contextualized. The Wrong Answer Project shines a light on the path towards justice, not just technical “fairness”.

![Figure 18. Decent into Enlightenment](image)

**Conclusion**

In the assessment community, when we hear the words: value, fairness, equity, marginalization, achievement gap, and social justice, we should think of validity evidence based on the social consequences of testing. We should also think about who will do the
work to ensure these concepts are at the forefront of our thought as we approach test construction, score interpretation and use.

I believe it is time for the assessment community to formally acknowledge the importance of studying the social consequences of testing in developing and evaluating validity arguments for testing programs. I also believe social justice approaches to assessment should be prioritized in modern testing standards. The Wrong Answer Project though not without fault has the great potential to contribute solutions to challenges we see with the consequences of high stakes testing decisions. It’s the first of its kind that moves the measurement community toward a future of research and practice that points to advocacy, with the goal to change value within community and in the political space.

Psychometricians can and should become advocates of social justice, because such advocacy is consistent with valid use of educational assessment data. By focusing on gathering and analyzing validity evidence based on testing consequences, the psychometric community will need to work side-by-side with educators and policy makers, which will have enormous potential to provide better educational experiences for students. The future of psychometricians as advocates of social justice in education has the potential to open the federal government’s pockets and mind to support investigations of the positive and negative consequences associated with high stakes testing programs. By identifying these consequences, we can then work to increase the positive effects and decrease the negative effects, so that educational tests do far more good, than harm.
A

WRONG ANSWER INTERVENTION PROTOCOL

Wrong Answer

FINAL DRAFT

Darius D. Taylor
PART ONE

Audience members will gather in the lobby awaiting the official descent into the space. From the lobby audience descension will take them down the stairwell of the actors singing the hymn *This May Be The Last Time*. Once the audience members reach the downstairs lobby, the line enters the space through a metal detector and guards/front of house/ushers directing them to find their seats, the exam will begin soon. There will be scenic design elements suggesting an inner city impoverished school setting and as the audience enters they will hear the sound of a ticking clock.

Ms. Engel will be stage center encouraging everyone to “Find your seat, your exam will begin shortly.” On every seat will be a clipboard with a survey on test perceptions attached and a sharpened number two pencil. Once the house is closed, Ms. Engel will give the following instructions to the audience (reading from an instruction manual): “At this time please turn off or silence all cell phones. The purpose of the exam in front of you is to capture your perceptions on testing. You will have 5 minutes to read the instructions and complete the exam. Please read the instructions on the first page, sign your name and begin.” Ms. Engel will then leave the stage and walk around the room militantly keeping an eye out for cheating and close watch on the clock. At the 5-minute mark a bell will sound, and Ms. Engel will then say, “Time is up! Please pass your exams to the front with the clipboard and number two pencil as it was received.” She will then collect the exams (maybe with help from a stage hand), then exit. The period bell will ring and all students from the cast will filter in and sit close to the tv screens for the filming.

A short film by Firelight Media entitled “Children Left Behind” will be shown. It highlights issues in assessment. The video will happen before the theatrical performance (PART TWO).

PART TWO

Characters

Danaijah (Nay Nay): Junior in high school. Very smart, type A, gets “the system” only enough to excel, but wants to burn it all down. Black female.

Micah (Mikey): Junior in high school. Also, very smart, doesn’t apply himself and knows that the system wasn’t built for him to succeed so there is no good in even trying. Black male.

Mothers voice: ominous mother voice frequently introduced into the space.

Ms. Engel: 11th grade English teacher and test proctor who has competing values about testing. White woman in her late 30s/early 40s.

Dr. Gray: School Psychologist, Black or brown person in late 30s/early 40s, acts as a voice of reason for testing.

Johnny & Jimmy: Danaijah’s Brothers: Fraternal twins who are 2 years younger than Danaijah. 9th graders.

Chorus: Ensemble of dancers for classroom scenes.
Scene 1

Lights come up on stage left to a single bed, night stand and alarm clock. Danaijah is sleeping, and her clock reads 6:29. It turns to 6:30 and the alarm sounds. She wakes up immediately, turns off her alarm. Gets out of bed and stretches.

Danaijah

Ahhh, this is it, this is the day that I prove my worth to this stupid system.

(As she leaves her room to head to the bathroom to brush her teeth)

Mother (voice over)

NAY-NAYYYYY!

Danaijah

HAHN?!

Mother’s (voice over)

I KNOW THAS YOU! WAKE YO BROTHAS UP!

Danaijah

Okay, mama. (under her voice: ..dang..)

Stops at brothers’ room on the way to the bathroom.

Wake ya’ll ho asses up, its test day bitches!

The brothers moan roasting her under their breath (something like “if you don’t get yo im ready for a test lookin ass”) unexcited and get back under the covers when she leaves.

This is the day, this the day I’m gonna stick it to ‘em…yeah…the man.. The system. Ge ready fa these hands!

Mother’s voice

Naija! I’m off to work. Don’t forget to fix you and yo brothers breakfast before ya’ll set off to school.

Danaijah
Okay mommy - love you! Have a great day at work! (Goes to kitchen to start wippin’ up some bacon, eggs and grits with chocolate milk for breakfast for her and her brothers.)

Mother’s voice

Love you too baby! I’m sure you gonna ace that test.

Danaijah

(in a low voice) You can bet yo ass I will…Jimmy John come get yo breakfast made by Chef-Girl - D

Johnny

Aw man what you made for us today?

Jimmy

Momma said to make us breakfast, bro I’m ‘bout make a fast break on this bowl of/

Danaijah

/ay ay ay that’s enough. We got all the food groups here for a balanced breakfast to ace the test. Oranges gives you energy, grits stick to ya bones for longevity, eggs and bacon give you the protein to power the muscle (pocks Jimmy in the head).

Johnny

ooooo balanced breakfast

Jimmy

maaaann, you too geeked fa this test. We jus goin today for practice anyway. ..bro who you takin to the dance tho?

Johnny

Oh I was thinkin bout takin Samantha...what about you?
What you think about Isabelle?

*improv dialogue about the dance and the girls.*

Danaijah

Who’s Isabelle? Ya’ll gotta get yo mind right and focus on this exam.

Johnny

….But I do hope we do good tho bro. You know we gotta secure that bag to play college ball.

Danaijah

Ya’ll betta eat ya food.

*brothers improv pocking fun again*

Danaijah

Don’t disrespect me.

*Lights down on kitchen as Nay is giving Jimmy a look of disapproval. Lights up on bedroom stage right. Mikey is sleeping, and his clock read 6:37. He has overslept. And is still sleeping. Snoring even.*

Mother’s voice

MIKEY!

*He is still sleeping and snoring.*

Mother’s voice

MIKEY.! *Still sleeping.*

Mother’s voice

MICAH DERRICK JAMES ROBINSON, YOU BETTA GET YO ASS UP OUT THAT BED THIS INSTANT. DON’T YOU GOT THAT TEST TODAY! YOU GONE BE LATE. GET UP.

Mikey ruffles in bed.

Mikey
5 more minutes ma!

Mother’s voice

BOI, DON’T MAKE ME COME UP THERE! AINT THIS TEST SUPPOSED TO GET YOU INTO COLLEGE. YOU WANNA END UP LOCKED UP LIKE YO BROTHA OR NO GOOD LIKE YA DADDY?! GET THAT TAIL UP AND GET YO MIND RIGHT AND READY!

Mikey

huhhhhh…

Mother’s voice:

WHAT WAS THAT?

Mikey

I said, I’m uuuuuup.

Mikey ruffles to get himself together mumbling under his voice...

damn, I hate this, I hate school, and HATE tests…testing just not the mood like this stupid test wasn’t made for me no way. I’ve taken test after, test after, test and realize theyyy (rubs white side of his hand extended in the ai) just don’t want a nigga to succeed. They don’t wanna see me out here shinin’. So let me just get ready to take this “L”.

Lights down.

End scene.

The bell rings and there is an improved hall scene. The bell rings again, the cast sets up chairs in 5 by 2 rows for the classroom then lineup for march in, dance, then chat/scurry before class (Dr. Gray, Danaijah and Micah exit during this).

Sound: Play “All for us” (fade out start ~0:58) when cast is in lineup
Scene 2

Enter Ms. Engel.

Ms. Engel
Okay, okay. Find your seats and clear your desk. Remember only a number 2 pencil and a calculator can be used on this exam.

Danaijah enters looking pressed and prepped for the day.

Danaijah
Morning Ms. Engel.

Ms. Engel
Good morning, Danaijah. You look like a warrior ready for battle.

Danaijah
You got that right Ms. Engel. I’ve had a balanced breakfast. Not too heavy that I get the ‘itis’ but not too light that I can’t think straight. I’m ready to grab this test by the balls and show it who’s boss!

Ms. Engel
Well, I’m sure you will.

The last of the students filter in and Ms. Engel realizes Mikey is missing.

Ms. Engel (to the class)
has anyone seen…

Mikey storms in with a poptart in his mouth.

Mikey
I’m here Ms. E, don’t worry about me!

Mikey turns and notices everyone in rows, rather than the circle seating that he is used to.

Ms. Engel
Find your seat Micah.

Mikey walks to his seat which is upstage stage right. There is a sort of ominous walk that he has to his seat. As he looks around anxiety hits, you can see on his face that he is overwhelmed. When he has made it to his seat an ominous movement starts with the students.

**Sound: Play “Anthem” - fade out start ~1:00**

At the end of the movement Ms. Engel passes out sealed test booklets as she begins reading instructions for the test to the entire class.

As she finishes reading the instructions **Sound: Play “White Noise/clock ticking”**

Mikey

Tick tock
Sharpened # 2 pencil in hand
Startin’ to feel sweat from ma glands
“Is anybody else hot?!”

Tick tock
That look like my 11th grade Teacher Ms. E
But somethin don’t seem to line up for me
somethin is different
Maybe it could be that she hella scripted
“What’s that you readin’?”
“What’s this, whatchu sayin’?”
Ms. E is this another game we playin?

Tick Tock
Desks in parallel rows, facing the front.
“What happened to my spot in the group circle next to Johnny and that funk?”

Tick Tock
The energy in here is varied, this air is stale
My heart won’t slow down
I know I’m black, but I’m turnin pale!

Tick Tock
Block to block, bubble to bubble
Skip, erase, pick “C”, make a tree?
This test gonna be the death of me
Lights come back up and Mikey takes his seat.

Ms. Engel

Alright, now that you are finished filling in your personal information, if there are no more questions. Remember you will have 40 minutes to complete this section. Good luck! You may begin.

Choreographed movement of test performance begins lights fade to a spot light on Ms. Engel.

And it is in this moment that I am forced to push my chicks out of the nest and hope that their wings are strong enough to fly.

But what if the air is not right…they all have different wings.

And faced with foreign terrain …they will all make different decisions.

How do I know they are ready…what does ready even mean?

Readiness is relative.

But that’s not what the numbers will say.

The numbers will say X amount of students passed.

The numbers will say X amount of students are not performing to “standard”.

The numbers will say I’m not an effective teacher.

The numbers will say that I failed and subsequently my school is underperforming.

…I want to do right by my kids…let them achieve in their own lanes.

…I’m pressured by my superiors and my superiors’ superiors to test well.

…because testing well leads to funding decisions and a better reputation.

Here on the ground the work is so.. heavy…

but up at the top they make it seem so …light.

There are so many components that contribute to their weight, that I wrestle with.

…but I’m forced to put a smile on, forget everything that makes my kids unique and

Standardize? This can’t be the right answer.

(continues proctor movements)

Sound: Play “White Noise/clock ticking”

Danaijah is working on her exam in sequence with the movement of the classroom. She begins to veer off into her own distinct movement emotting various feelings come up but specifically dignified anger because maybe she is not as prepared as she thought she was. She stands up and walks down stage center.

Danaijah

Trying to put all of us into a test booklet

Is the wrong answer
It’s the wrong task for assessing what we’ve mastered
If you took the time to look inside, you will find
That our mind processes information differently
That there are all these intricacies
That contribute to our views because of our history
Our ways of knowing and interpretations of reality differ
And it makes us bitter that you don’t consider our values.
In a nation that claims
“Liberty for all”
“For the people by the people”
“Equality” and “Equity”
In an education system
Where multiple intelligences exist
Diverse learners in the classroom exist
And trends over the years prove
achievement gaps are real
We must believe that one test is best for all?
WRONG ANSWER
But I’m gonna act like its right
To win this fight
And hope I don’t lose myself along the way.

Danaijah takes her seat. Spotlight shifts to lights up on the classroom.
Mikey yells, slams his test booklet down on his desk and runs out of class.

music transition with “role modelz” fade out start ~ (0:50)

Scene 3
Stage shifts to (down stage left or center stage) the school psychologist’s office. Mikey runs in, entering from stage left. Dr. Gray is typing on their computer.

Mikey
Man fuck this shit Dr. G! Fuck that test, fuck this school, fuck everything!
Dr. Gray

(sarcastically in response) Hello Micah.

Mikey

I told you to call me Mikey. Don’t nobody call me Micah, but my momma when I’m in trouble.

Dr. Gray

Micah is such a great name but hello Mikey. What’s the problem? Shouldn’t you be sitting for your exam right now?

Mikey

This aint cool, man. How the man gone try to judge me from a test booklet. How they gonna tell me to take 40 minutes to answer these math questions that I haven’t even talked about in class? These tests ain’t right. Somethins gotta give.

Dr. Gray

Micah... Mikey...in every aspect of your life you will be tested. In school, at home, on the streets. Will you always plan to run when it gets hard or when you don’t want to do it?

Mikey

I didn’t say it was hard Dr. G, it just doesn’t make any sense to me why they think that some of these irrelevant ass math questions determine whether or not I got the juice. I know and you know that I got the juice. I don’t need these whypipo’s test to tell me that.

Dr. Gray

You’re right, but you seem real shaken up by a test that you don’t care about.

Mikey

…..mannn Dr. G …my momma countin’ on me to do good on this exam. She think its my ticket to college.

Dr. Gray
Your mom is a wise woman. I understand your frustration, but the test determines your aptitude and potential performance in college. Don’t you want to go to college?

Mikey

I mean I think so Dr. G. ..can I tell you somethin?

Dr. Gray

yeah, anything you know that.

Mikey

…I think my brain is good Dr. G. Like really good, ..I get how these systems work and I’m smart, I consider myself socially and emotionally intelligent. I... I can talk to people.

Dr. Gray

mhnmm, yeah

Mikey

…but I just don’t get how this test measures those aspects within me. I feel like they are trying to measure me with the wrong tools. Like they are mismeasuring me…is that a thing Dr. G?

Dr. Gray

I mean there is certainly implications of using tests for a particular purpose. Testing companies they try to ask you questions on the exam that relate to what you will be tested on in college. But what I am hearing from you is that testing practices here (and maybe globally) don’t recognize everyone’s voices right now.

Mikey

hmmm, you think my voice matters? We watched this video in class, the history of assessment to get us kind of prepared for battle. Or what I thought. I got to that test and froze. I don’t think I can make it in college. Nah, I wanna do something else that don’t require those tests. Uhn unn…

Dr. Gray

Mikey, I think you should take a step back and think about the bigger picture here. In every aspect of your life you are going to have to show worth and merit. Show that you have the capacity to achieve and succeed. You will be judged and evaluated against some metric of predetermined existential worth, so you have to get used to it and show up when life calls for it.
Mikey

I feel trapped, like I’m in a jail cell when I am taking that test. I feel like a sacrifice being made to the gods of capitalism in society… *(Mikey stands and walks down stage center)*

My blood the sweat and tears of thousands
Maybe millions of people who look like me
My people.
Who perpetually endure the toxic education system
rooted in multiple oppressive values
related to Race, Ethnicity, Gender, and Class.
These outcomes, success metrics that supposedly show your value, tell the story;
Developed by power launderers, who position themselves
Purposefully to permit power seekers potential power
Deeming them successful (or unsuccessful) in the launderer’s field
Everybody has to prove themselves...”what’s your worth?”
I know it sure as hell ain’t related to no “systems of equations” or “geometric functions”.
This test is telling me to perform my intelligence in a particular way
A way that makes me remove true aspects of myself and “play the game” to be successful
Will my sacrificial lamb be worthy enough Dr. G?

Dr. Gray

That’s a choice you have to make for yourself Mikey. Trust and believe that I have had many frustrations and upsets due to feeling unworthy and misunderstood.. A word from the wise, it doesn’t get any easier. Yes, we have to play a game that we had no say in the development of but push through. It will be worth it in the end.

Micah

I know my worth Dr. G. I know what I am capable of. These test people and the government think they got it all figured out. They can tell me who I am by putting me in a box based on how I score on some exam that doesn’t measure my true worth. I’m sure based on what they are asking me in there on that test that I will be labelled in a score group that “needs improvement”, not even make it to college and continue to struggle like all Black people do in America. *(He gets up to leave.)*

Dr. Gray

If you give up and believe that, you are letting them win you know. You can’t let them win. Letting them win keeps you enslaved. So what your exam doesn’t work out, there are other
pathways to success. Choose the path of resilience, the path that Black people, our people, have walked on since we were dragged here against our will.

Micah

Aight, I hear you Dr. G. Imma head out. (He exits.)

insert music transition

Scene 4

By the time Micah makes it back to class time is almost up for the exam, some students are finished and ready to leave. Danaijah is still there working, there is 7 minutes left on the exam.

play “air freshener” stop by 0:30

Ms. Engel

Okay class, pencils down in 7 minutes. Try to start thinking about a powerful finish!

There are about two students still testing plus Danaijah. One of the students looks up at Micah as he walks back in. He can see the discontent on his/her face.

Ms. Engel

Micah where have you been you’ve missed your entire exam?

Micah

In Dr. Gray’s office Ms. E. You know this test wasn’t made for me.

Ms. Engel

Micah this test is important for your future. You can’t just leave when you feel like it.

Danaijah

DONE! (She walks her test booklet to the front of class.) With one minute to spare!

Ms. Engel

(Accepting Danaijah’s test booklet.) Thank you Danaijah. I’m sure you killed it.
Danaijah

I know I did Ms. E (flips hair or pantomimes flipping hair. then she walks out.)

Ms. Engel

(Turning back to Micah) Now, you, mister, will need to go straight to Dean Crogg’s office. We need to figure out what to do with you.

Micah hangs out irritated then when bell rings he is fumbling through his back pack. Danaijah returns, she has forgotten her backpack

Micah

huhhhhhh (He walks out.)

Micah is met by Danaijah in the hallway.

Danaijah

You know the test wasn’t that bad Mikey. If you just applied yourself.

Micah

Easy for you to say Nay, this shit is a cake walk in the park for you.

Danaijah

That metaphor doesn’t make sense, but I see what you tried to do there.

Micah

Everybody can’t perform for the whites like you can Nay. It’s a real gift.

Danaijah

Boop, boop! Haters gone hate, have fun in the Dean’s office.

They separate she heads to her locker and he heads to the Dean’s office. She runs into the twins.

What up lil uglies (She puts her arms on their shoulders), how’d ya’ll do on the exam today?

Jimmy

That test was trash.
Johnny

Yeah, why you so happy Nay.

Danaijah

Cuz I know I KILLED IT baby bro. Like I can feel it in my bones. With these test scores, I can go to Yale, No Stanford, No HARVARD.

Jimmy/Johnny:

Yay, for you. (sarcastically)

Johnny

I mean you right it wasn’t that bad.

Jimmy

but the whole experience was trash.

Johnny

yeah throw the whole test away.

Jimmy

like the anxiety in that room gave me anxiety

Johnny

and we don’t get anxiety,

Jimmy

ever.

Danaijah

welp good thing you have two more years to get those emotions in check. Don’t worry, big sis got you. It’s handled. (flips hair or pantomimes flipping hair.) insert music transition “Diva”

Jimmy/Johnny

Greeeat. (sarcastically)
improv about the girls they mentioned at breakfast this morning as they exit

Student (PJ)

Hey Danaijah you know we getting our test results today.

Danaijah

What?! Let me get home!

reprise classroom dance of students getting test results

Scene 5

In Danaijah’s room

Mother’s voice

NAIJAH!

Danaijah

Yeah, mah!

Mother’s voice

You got some mail, I think its yo test results, want me to bring it/

Danaijah

/comingggg…!!

Running out of her room the scene changes and she runs straight into Dr. Gray’s office with test results in hand hysterical.

Danaijah

No, no, no, no, no, no, no, no, nooooooo…. 

Dr. Gray

Good morning Danaijah.
Danaijah

Dr. Gray something aint right! Something is not right. I just…look at this!

*She passes Dr. Gray her score report, Dr. Gray understands her disappointment.*

Dr. Gray

Ahh, I see. This must be hard to/

Danaijah

/Dr. Gray what did I do wrong? I studied. I pay attention in class, I ask questions. Where did I miss the mark? With this score, I will never get into Harvard.

Dr. Gray

Danaijah, please know that this test score does not define you and this is not the end for you.

Danaijah

How do you know that? All these systems do is fail people like me and this is evidence of that. This test was supposed to be my ticket to college. To a better life for me, and my family. This seems like the end. I should just get my drive thru voice ready.

Dr. Gray

Come on, there’s nothing wrong with blue collar work, but do you really want to work in a drive thru for the rest of your life? Do you believe that’s part of your destiny?

Danaijah

My destiny? I don’t know where to start with a response for that, but I know where I’ve come from and I know that statistics show I should be pregnant now (or in the next few years), working at a drive through getting public aid…..until this morning I refused to be that statistic. I ain’t nothing special with a score like that.

Dr. Gray

Throughout life it is important for us to have moments of clarity, milestones, they keep things in perspective for us. Think of this test and this score as a milestone, regardless of what it looks like you cannot give up on the race.
So, what I’m hearing is that this is not completely my fault? So maybe the test is crap, yeah that’s what it is.

Dr. Gray

Well, testing companies do a lot of work to make sure tests are fair for everyone. So when you say the test is crap, I’m not completely sure that is a valid statement. The purpose of the test is to measure your readiness for college.

Danaijah

Dr. Gray, I am ready for college. This test got me fucked up. You know I think they got the purpose all wrong…. (she steps down stage)

Depression is what I feel after receiving my score
Oppression comes to mind as I think of the implications
Seems like dog whistle politics these tests
“To determine which student’s the best”
It aint me! rarely is it me or even people who look like me
Let’s put the whistle down
And call a spade, a spade
The way that these tests are made
Got my people feelin’ played
We been fighting since our ancestors stepped foot on this soil
Not just physically but against principalities
Engrained in systems, organizations and institutions
We fought for our freedom
We fought for our rights
We fought to integrate into THEIR systems
    Because when we tried to thrive on our own they shut it down!
But they still want to keep us in chains (pause.)
The purpose of this test is subjective to the examinee
    For the elite maybe it does measure your college readiness
But for black people it’s a shackle
Few can actually set themselves free and be liberated
But must take caution because with liberation comes separation from those who remain shackled
So we must continue to fight…for justice.
Dr. Gray
Well put Danaijah, but where do you see this going? Where do you see this fight taking place?

Danaijah
We need a new test Dr. Gray that measures ME. All of me and my melonated complicated self. There’s just too many factors that contribute to who I am and how I make decisions in life that maybe this test is missing.

Dr. Gray
ha, you sound like Micah.

Danaijah
wow, Mikey, actually formed an intelligent argument?
(Micah enters Dr. Gray’s office.)

Micah
I heard somebody say my name, ya’ll session over Dr. G?

Dr. Gray
Mikey we were just talking about your love for the exam.

Micah
Mannn, I hate that test. It’s stupid and does a poor job measuring me and all of my intelligences.

Danaijah
huh, I’m impressed, maybe you are right Dr. G.

Dr. Gray
It may be good for you all to share your experiences with each other. You never know what will come out of building community.
*The look at each other, shake hands.*
Danaijah

You right Dr. Gray cuz it takes a community! We need to change this test right Mikey? It needs to represent all of our voices. The voices of you, you, you and you! Its time for us to come together as democracy suggests and make a better test. We breakin down walls and breakin all the rules. In a minute we all are going to literally write our suggests on the walls and there’s questions to ask for you too. But first…we rippin this test to shreds. Come on rip it with me!

Micah

Literally pick up that exam and rip it to pieces everyone!

_play “Before I Let Go”, lights down_

_End scene, End Part 2._
PART THREE

The theatrical experience will turn interactive as the cast then invites the audience to participate in one of the key practices of Brown Paper Studio. The audience will be encouraged to reflect on the experience thus far by writing on brown paper hanging from the walls with colored markers (sort of graffiti style). Some paper will be free writing for reflection. Other paper will have prompted questions for writing which will include:

- What is assessment?
- What is high stakes testing?
- What is equity?
- What does your racial/ethnic identity have to do with how you perform on high stakes tests?
- What does your economic status have to do with how you perform on high stakes tests?
- How could we consider equity in the design of high stakes tests?

After the free writing session, audience members will be instructed to do a quick gallery walk to explore what everyone wrote. Then remaining audience members will be instructed to join 1 of 3 circles that will dialogue on the social issues that result from current testing practices and suggestions on how we can come to a democratic decision about best practices in testing that includes the voices of marginalized groups. Leading questions for the dialogue circles will consist of the following:

- What resonated with you during this this experience?
- What does equity centered design look like in practice?
Perceptions Survey

The following survey questions ask your perceptions on specific aspects of current high stakes testing practices in the United States of America. The National Council on Measurement in Education defines a high stakes test as a test used to provide results that have important, direct consequences (or implications) for individuals, programs, or institutions involved in the testing.

Select one response to how much you agree with the following:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. High stakes tests measure all students equally.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>2. Students’ race influences their performance on high stakes tests.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>3. High achieving students may perform poorly on high stakes tests.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>4. Economic status influences student performance on high stakes tests.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>5. Students with any learning style can excel on high stakes tests.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>6. Students’ ethnicity influences their performance on high stakes tests.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>7. High stakes tests measure a teacher’s ability to teach.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>8. Current high stakes testing policies are fair.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
<tr>
<td>9. High stakes tests take a certain level of preparedness.</td>
<td>(A) Strongly Disagree</td>
<td>(B) Disagree</td>
<td>(C) Agree</td>
</tr>
</tbody>
</table>
FACILITAION PROTOCOL

- Introduce yourself: Name, pronouns (and why we use pronouns)
- Have audience members introduce themselves; name, pronouns

INTRO: Why are we having this conversation?

This performance is a theatrical intervention on high stakes testing. As you may have seen in the show, there is a history and presentation of who benefits and who is disadvantaged from high stakes testing, and how the outcome may affect students from different socioeconomic backgrounds. But how do we change that? We hope that in this conversation that not only do we delve into learning more, but to also create action steps on how current high stakes testing can be reimagined and incite equitable change for all.

Community Guidelines - Share with audience members:

1. Use “I” Statements as opposed to making generalized statements about experiences/perspectives. Example: “In my experience…”
2. Take Space/Make Space. Be aware of how much you are or aren’t taking compared to others in the conversation.
3. Assume Best Intentions. Don’t assume malice, speak with respect.
4. Ouch/Oops = Intent vs Impact.
5. Call In instead of Calling Out. We all have different backgrounds talking through these topics, so we should be patient and learn from one another. Try “how did you come to understand this?”

Facilitation Questions to Consider:

- What were your immediate reactions to the play and/or responses on the Brown Paper?
  - What did you notice or see?
  - What patterns or themes did you observe?
- What is the point of assessment? What is being assessed and for what?
  - Facilitators: Deficit lens - addressing what students lack and the skills/tools that they don’t have vs Asset lens - addressing the tools and skills that students do have/ bring to the classroom
- How have you thought about or interacted with assessment before today?
- How do you see your own assessment experiences reflected in the issues brought up in the play?
- “This test was supposed to be my ticket to college. To a better life for me, and my family. This seems like the end.” - Danaiah
  - What are some of the impacts of assessment of students? Schools? School districts?
  - How does assessment contribute to education inequality?
    - Education inequity - is the unequal distribution of academic resources, including but not limited to; school funding, qualified and experienced teachers, books, 2w2nd technologies to socially excluded communities.
    - What are the larger social impacts of education inequity?
What would a reimagining of assessment look like?
  o What would it feel like for students?
  o What are some ways to make assessment more equitable?
  o What are other more inclusive ways to assess students with varying abilities and identities?

Why are race and other identities emphasized/relevant to this discussion?
  o Facilitators: Think about the eugenics of IQ and who's good at testing, factors that make you good at testing, abilities, etc.

How do we make sure to center factors, like race, when creating assessment tools?

Next steps?
  o What will it take for you to implement some of these ideas/conversations in your spheres of influence?
  o What will it take for equitable institutional changes to take place in standardized testing?
  o What are you willing to let go of to make this happen?
BIBLIOGRAPHY


Enrollment Projection Model, 1972 through 2029. (This table was prepared August 2020.)


