Assessing Written Communication: Is Motivation a Concern?

Joseph A. Rios
Educational Testing Service
Background

• Student learning outcomes caught national attention in 2005
  • “remarkable absence of accountability mechanisms to ensure that colleges succeed in educating students” (U.S. Department of Education, 2006)

• Institutions pressured to demonstrate evidence of student learning
Reasons for Using SLO Assessments

Tools to Assess SLO

A New ETS Initiative

- Research Synthesis
- Qualitative and quantitative market research
- Leverage existing R&D capabilities
- Input from HEIs and organizations
- Critical thinking
- Written communication
- Quantitative literacy
- Digital information literacy
- Civic competency and engagement
- Intercultural competency and diversity

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
A Research-Driven Approach to Assessment Design

Review of influential frameworks
Review of existing assessments
Operational definition
Assessment considerations

Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment
Ou Lydia Liu
Lois Frankel
Katrina Crotts Roohr
April 2014

Assessing Quantitative Literacy in Higher Education: An Overview of Existing Research and Assessments With Recommendations for Next-Generation Assessment
Katrina Crotts Roohr
Edith Aurora Graf
Ou Lydia Liu
September 2014

Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment
Jesse R. Sparks
Yi Song
Wynona Brantley
Ou Lydia Liu
December 2014
Importance of Written Communication

• Academic officers have stated that written communication (WC) is critical to academic success.

• Employers agree with the importance of WC, but express that many college graduates are unprepared for the writing tasks in the workforce.

• Institutions face the challenge of accurately evaluating this skill to support demands of accreditation, accountability, and program or curricular improvement.
Existing WC Assessments

Existing WC assessments of student learning outcomes possess a number of limitations:

1. Selected-response items target relatively low-level skills rather than focusing on passage-level organization and development.

2. The constructed-response section is optional, which may allow institutions to opt out from obtaining a direct writing measure.

3. Require at least 60 minutes to administer.
HEIghten™ WC Assessment

**Selected-response items**
- Social and rhetorical situations
- Domain knowledge and conceptual strategies
- Language use and conventions

**Constructed-response item**
- Knowledge of the writing process

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
From The Framework to Assessment

• Pilot test forms consider

- Construct coverage
- Item types
- Item format
- Context
- Accessibility
Significant Issues

- Faking
- Motivation
- Subscore
- Authenticity vs. Psychometric quality
- Automated scoring vs human scoring
- Accessibility
- Score report
Motivation Concerns
Examinee Motivation as a Threat to Validity

- The willingness to engage as well as invest effort and persistence in working on test items
- If the examinee places little subjective value on success, motivation on the task may be low
- Test performance may not be reflective of the underlying skills or abilities of the examinee
Practical Implications

• On average, motivated examinees score .55 SD higher than their unmotivated counterparts (Wise & DeMars, 2007)

• In terms of aggregated-score inferences, careless responding has been found to bias:
  • Treatment effects (Osborne & Blanchard, 2011)
  • Achievement gains (Wise & DeMars, 2010)
  • Teacher evaluation ratings (Wise et al., 2013)
  • Country-level comparisons (Debeer et al., 2014)
Impact on Aggregated Scores

Proportion of Careless Responses in Total Sample

Standardized Difference ($d$) (Observed - True)

-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1

Hard Unrelated
Hard Related

1% 2.50% 5% 6.25% 12.50% 25%
Impact on Aggregated Scores

Proportion of Careless Responses in Total Sample

Standardized Difference ($d$) (Observed - True)

-1 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0

1% 2.50% 5% 6.25% 12.50% 25%

- Moderate Unrelated
- Moderate Related
- Hard Unrelated
- Hard Related

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING is a trademark of ETS.
Impact on Aggregated Scores

Proportion of Careless Responses in Total Sample

Standardized Difference (d) (Observed - True)

0 1% 2.50% 5% 6.25% 12.50% 25%

Impact on Aggregated Scores

Easy Unrelated
Easy Related
Moderate Unrelated
Moderate Related
Hard Unrelated
Hard Related
Approaches to Dealing with Examinee Motivation

- There are two main approaches to dealing with potential motivation issues in the examinee pool

1) Manipulation of examinee motivation
2) Filtering data from unmotivated examinees
Motivation Interventions
Interventions in Research

• Over 60 papers have implemented motivation interventions

• Type of interventions have been quite diverse

• Inconclusive evidence for which type works best
Identifying Unmotivated Examinees
Identifying Unmotivated Examinees

• Three approaches have been suggested for identifying unmotivated examinees

1. Person-Fit Statistics
2. Self-Report Measures of Motivation
3. Response Time Information
Self-Reported Effort

• Administered to examinees following test administration to assess the degree of motivation that the examinee felt was exerted.

• Two popular measures developed:
  • Wolf and Smith (1995) consists of eight five-point Likert-type items.
  • The Student Opinion Scale (SOS) was developed by Sundre and Moore (2002):
    • Includes two factors of motivation: importance and effort.
    • Five, five-point Likert items per factor.
Response Time Effort

• Based on examining the response times collected when administering a computer-based assessment

• This approach is divided into two tasks:
  • Identify careless responses that were provided so quickly that the examinee would be unable to adequately read the item content
  • Filter careless responses
# Comparison of Self-Report and Response Time Effort

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Reported Effort (SRE)</td>
<td>• Not resource-intensive</td>
<td>• Difficult to evaluate response bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Only assess global motivation for a measure or battery of measures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Can only filter out an unmotivated examinees complete data</td>
</tr>
<tr>
<td>Response Time Effort (RTE)</td>
<td>• Examinee behavior is directly evaluated</td>
<td>• Requires CBT administration</td>
</tr>
<tr>
<td></td>
<td>• Can filter out particular items of rapid-guessing behavior</td>
<td>• Determining thresholds of rapid-guessing behavior can be a challenge</td>
</tr>
</tbody>
</table>
Comparative Analysis

The image shows a scatter plot with two red boxes highlighting the distribution of data points in the plot. The x-axis represents the Self-Reported Effort (SRE) and the y-axis represents the Response Time Effort (RTE). The data points are scattered across the plot, with some points enclosed within the red boxes, indicating a subset of data that may be of particular interest or importance.
Identifying Careless Responses via Response Time
# Four Flagging Methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Procedure for Defining Flagging Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Seconds (3Sec)</td>
<td>&lt;3 second response</td>
</tr>
<tr>
<td>Normative Threshold (NT)</td>
<td>% of mean item response time (examined 15%, 20%, &amp; 25%)</td>
</tr>
<tr>
<td>Visual Inspection (VI)</td>
<td>Intersection of bimodal response time frequency distributions</td>
</tr>
<tr>
<td>Words per Minute (WPM)</td>
<td>Expected words per minute read by university students (examined 200, 300, &amp; 400 WPM)</td>
</tr>
</tbody>
</table>
Four Flagging Methods
Cumulative Proportion Correct Procedure

Item 77 Response Time Frequencies with Proportion Correct Values

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
What to do Next?
Comparison of Filtering Procedures
# Procedures for Filtering Careless Responses

<table>
<thead>
<tr>
<th>Filtering Process</th>
<th>Examinee-Level</th>
<th>Response-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listwise deletes data from an examinee deemed to be “unmotivated”</td>
<td>Treats a careless response as a blank response</td>
<td></td>
</tr>
</tbody>
</table>
## Procedures for Filtering Careless Responses

<table>
<thead>
<tr>
<th>Filtering Process</th>
<th>Examinee-Level</th>
<th>Response-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Listwise deletes data from an examinee deemed to be “unmotivated”</td>
<td>• Treats a careless response as not administered</td>
<td></td>
</tr>
<tr>
<td><strong>Advantage</strong></td>
<td>• Can identify motivation using both response time and self-report</td>
<td>• Does not delete possible valid responses</td>
</tr>
<tr>
<td>Filtering Process</td>
<td>Examinee-Level</td>
<td>Response-Level</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>• Listwise deletes data from an examinee deemed to be “unmotivated”</td>
<td>• Treats a careless response as not administered</td>
</tr>
<tr>
<td>Advantage</td>
<td>• Can identify motivation using both response time and self-report</td>
<td>• Does not delete possible valid responses</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>• Discards up to 25% of the sample data</td>
<td>• Assumes non-flagged responses are valid indicators of ability</td>
</tr>
<tr>
<td></td>
<td>• Assumes that ability is unrelated to motivation</td>
<td></td>
</tr>
</tbody>
</table>
Careless Responding is RELATED to Ability EXAMINEE-LEVEL FILTERING

Standardized Difference ($d$) vs. Proportion of Unmotivated Simulees

- Easy
- Moderate
- Difficult

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Careless Responding is **RELATED** to Ability RESPONSE-LEVEL FILTERING

Proportion of Unmotivated Simulees

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Is Careless Responding Related to Ability?

• Rios et al. (under review) found differences on prior achievement (SAT) between motivated and unmotivated test-takers to be as large as .48 SD

• Allen et al. (2016) found differences as large as .55 SD on the ACT

• This is an issue that has largely gone unnoticed
Conclusions for Motivation on SR Items

• Careless responding may be impactful based on: a) the difficulty of the test, and b) the overall percentage of careless responses in the sample
  • Previous research has focused on individual-level careless responding when evaluating the impact on aggregated scores

• If possible, response time should be used as a proxy of motivation as opposed to self-report measures
Conclusions for Motivation on SR Items

• There are multiple procedures for identifying rapid guessing, but procedures that combine both response time and accuracy may be best.

• In terms of filtering, practitioners should evaluate whether unmotivated examinees tend to be of lower ability.
  • If one is unable to do this, one should use response-level filtering (Rios et al., under review).
    • This requires IRT so it may be unfeasible for some testing programs.
Next Steps

The iterative examinee-level filtering procedure

1. Sort examinees by the percentage of careless responses

<table>
<thead>
<tr>
<th>ID</th>
<th>Careless responses</th>
<th>Cumulative careless responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examinee 1</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Examinee 2</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Examinee 3</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Examinee 96</td>
<td>50%</td>
<td>11%</td>
</tr>
<tr>
<td>Examinee 97</td>
<td>50%</td>
<td>12%</td>
</tr>
<tr>
<td>Examinee 98</td>
<td>77%</td>
<td>13%</td>
</tr>
<tr>
<td>Examinee 99</td>
<td>78%</td>
<td>14%</td>
</tr>
<tr>
<td>Examinee 100</td>
<td>80%</td>
<td>15%</td>
</tr>
</tbody>
</table>
Next Steps

The iterative examinee-level filtering procedure

2. Remove examinees with the most careless responses until the cumulative percentage of careless responses is below the critical value

<table>
<thead>
<tr>
<th>ID</th>
<th>Careless responses</th>
<th>Cumulative careless responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examinee 1</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Examinee 2</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Examinee 3</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Examinee 96</td>
<td>50%</td>
<td>11%</td>
</tr>
<tr>
<td>Examinee 97</td>
<td>50%</td>
<td>12%</td>
</tr>
<tr>
<td>Examinee 98</td>
<td>77%</td>
<td>13%</td>
</tr>
<tr>
<td>Examinee 99</td>
<td>78%</td>
<td>14%</td>
</tr>
<tr>
<td>Examinee 100</td>
<td>80%</td>
<td>15%</td>
</tr>
</tbody>
</table>
Examining the Relationship between Motivation and Test Performance on Selected-Response and Essay Items on the HEIghten Written Communication Assessment
Objectives

- Establish whether rapid guessing on SR can be used as an indicator of motivation on the essay
- Compare essay characteristics between motivated and unmotivated examinees across two classification methods (response time effort and self-report)
- Highlight possible solutions to improving motivation on the WC essays
Low-Stakes Constructed-Response Items

• Getting students to try their best on low-stakes tests is particularly problematic for constructed-response items (Wainer, 2003)

• On constructed-response items, motivated examinees have been found to score as much as 1.59 standard deviations higher than their unmotivated counterparts (Sundre, 1999)
Identifying Low Motivation on Essays

1. Self-reported effort
   • Assumes:
     • Perceived performance does not impact self-reported effort
     • The examinee was motivated to take the effort survey
   
2. Performance relative to expected
   • Assumes:
     • Motivation is unrelated to ability

Rios, Mao, Guo, & Liu (2015) found this assumption to be untenable in practice
1. Can response time effort on SR items serve as an indicator of motivation on the essay?
   • Is essay response time a significant predictor of essay scores?
   • Does essay response time significantly predict SR response time?

2. Are there significant differences in essay performance between motivation groups based on selected-response and self-report motivation classifications?
Written Communication Pilot Assessment

- Analyses were based on one form (N=491)
- Examinees were predominately:
  - Female (66%)
  - White (49%) or Black (28%)
  - Freshmen (38%) or sophomore (31%)
  - From 4-year institutions (90%)
Can Response Time Effort on SR Items Serve as an Indicator of Motivation on the Essay?
Establishing Relationship between Essay Response Time and Score
### Significant Predictors of Essay Scores

<table>
<thead>
<tr>
<th>DV: Essay Score</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>β</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>SAT or ACT</td>
<td>.40*</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>College GPA</td>
<td>.07</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Response Length (ln)</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Response Time (ln)</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.18</td>
<td>----</td>
<td>----</td>
</tr>
</tbody>
</table>
### Significant Predictors of Essay Scores

**DV:** Essay Score  

<table>
<thead>
<tr>
<th>IV</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT or ACT</td>
<td>.40*</td>
<td>.19*</td>
<td>----</td>
</tr>
<tr>
<td>College GPA</td>
<td>.07</td>
<td>.002</td>
<td>----</td>
</tr>
<tr>
<td><strong>Response Length</strong></td>
<td>----</td>
<td>----</td>
<td>.82*</td>
</tr>
<tr>
<td>(ln)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Response Time</strong></td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>(ln)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.18</td>
<td>.79</td>
<td>----</td>
</tr>
</tbody>
</table>

For every 2.57 words, the essay score increases by .009 points.
## Significant Predictors of Essay Scores

<table>
<thead>
<tr>
<th>DV: Essay Score</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>SAT or ACT</td>
<td>.40*</td>
<td>.19*</td>
<td>.19*</td>
</tr>
<tr>
<td>College GPA</td>
<td>.07</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>Response Length ((\ln))</td>
<td>----</td>
<td>.82*</td>
<td>.75*</td>
</tr>
<tr>
<td>Response Time ((\ln))</td>
<td>----</td>
<td>----</td>
<td>.09*</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>.18</td>
<td>.79</td>
<td>.76</td>
</tr>
</tbody>
</table>
Response Time and Length by Score

$r = .70$

- **Response Time (Seconds)**
  - n=19
  - n=27
  - n=53
  - n=113
  - n=129
  - n=7

- **Response Length (Words)**
  - n=137

Essay Score vs. Mean

- 0
- 1
- 2
- 3
- 4
- 5
- 6
Is Essay Response Time a Significant Predictor of Selected-Response Time?
## Does Response Time on Essay Predict Total Time on SR Items?

<table>
<thead>
<tr>
<th>DV: SR Total Response Time</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>SAT or ACT</td>
<td>.05</td>
<td>----</td>
</tr>
<tr>
<td>College GPA</td>
<td>.08</td>
<td>----</td>
</tr>
<tr>
<td>Essay Response Time (ln)</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.01</td>
<td>----</td>
</tr>
</tbody>
</table>
Does Response Time on Essay Predict Total Time on SR Items?

<table>
<thead>
<tr>
<th>DV: SR Total Response Time</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT or ACT</td>
<td>.05</td>
<td>.002</td>
</tr>
<tr>
<td>College GPA</td>
<td>.08</td>
<td>.04</td>
</tr>
</tbody>
</table>

Essay Response Time (ln)   

- Model 1: ----
- Model 2: .35*  

For every 1 second increase in Essay RT, the selected-response RT increases by .15 seconds.
Examinee Motivation Classification Procedures and Validity Evidence
Classification Rules for SR Motivation

Response Time Effort

• Any examinee rapid-guessing on 20% or more of items was classified as unmotivated

• Any response provided in less than 10% of the mean item response time for item \( i \) was flagged as a rapid guess

Self-Reported Effort

• “Did you try your best?”
  • Dichotomous response option: “yes” or “no”
Response Time Effort for SR Items

9% of sample was classified as unmotivated

Mean = 0.94
Std. Dev. = 0.189
N = 491
## Agreement between Self-Report and RTE

<table>
<thead>
<tr>
<th>Self-Report</th>
<th>YES</th>
<th>NO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>394</td>
<td>48</td>
<td>442</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>412</strong></td>
<td><strong>71</strong></td>
<td><strong>483</strong></td>
</tr>
</tbody>
</table>

- 56% classified as unmotivated using RTE reported not trying their best.
- 15% of the sample reported not trying their best.
SR Response Time Differences by Motivation Group

- Motivated
- Unmotivated

Mean SR Response Time

Response Time Classification:
- Motivated: $d = 2.72$
- Unmotivated

Self-Report Classification:
- Motivated: $d = .78$
- Unmotivated
SR Total Score Differences by Motivation Group

\[ d = 1.81 \]

\[ d = 1.10 \]

Classification Procedure

- Response Time Classification
- Self-Report Classification

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Essay Characteristic Differences by Motivation Group and Classification Procedure
Differences in Essay Response Time by Motivation Classification

- Motivated: $d = .99$
- Unmotivated: $d = .69$

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Differences in Essay Response Length by Motivation Classification

- **Response Time Classification**:
  - Motivated: $d = .88$
  - Unmotivated:

- **Self-Report Classification**:
  - Motivated:
  - Unmotivated: $d = .62$
Differences in Essay Score by Motivation Classification

- **Response Time Classification**
  - Motivated: Mean = d = 1.03
  - Unmotivated: Mean = d = 0.75

- **Self-Report Classification**
  - Motivated: Mean = d = 1.03
  - Unmotivated: Mean = d = 0.75

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Possible Solutions and Next Steps
Possible Solutions

• Addressing motivation concerns *a priori* may be a better approach compared to making score adjustments *post hoc*

• Removing unmotivated examinees, can actually lead to more biased scores than not implementing score adjustments (Rios et al., under review)

• Previous research has demonstrated that changing the value associated with the test can improve performance (Penk & Schipolowski, 2015)
Essay Score Differences by Reason for Taking Test

<table>
<thead>
<tr>
<th>Reason for Taking Test</th>
<th>Mean</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Course Requirement (n=96)</td>
<td>2.5</td>
<td>.65</td>
</tr>
<tr>
<td>University Requirement (n=66)</td>
<td>3.0</td>
<td>.76</td>
</tr>
<tr>
<td>Extra Credit (n=126)</td>
<td>3.5</td>
<td>.67</td>
</tr>
<tr>
<td>Financial Incentive (n=58)</td>
<td>4.5</td>
<td>.41</td>
</tr>
<tr>
<td>Volunteer (n=100)</td>
<td>3.0</td>
<td>.67</td>
</tr>
<tr>
<td>Other (n=31)</td>
<td>2.5</td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 2015 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). MEASURING THE POWER OF LEARNING and HEIGHTEN are trademarks of ETS. 33024
Essay Score Differences by Reason for Taking Test

\[
d = 0.24 \quad d = 0.30 \quad d = 0.23 \quad d = -0.41 \quad d = 0.21
\]
Next Steps

• Need to develop technology within delivery system to capture key-stroke information

• Allen et al. (2016) found that median latency between keystrokes was the strongest predictor of test-taking engagement on a written assessment for college students