URPY#2379262, Vol 0, ISS 0

Unhinged: Reading Comprehension Tests as Gatekeepers to Teaching

John Wesley White and Daniel Dinsmore

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left are hyperlinked to the location of the query in your paper.

The title and author names are listed on this sheet as they will be published, both on your paper and on the Table of Contents. Please review and ensure the information is correct and advise us if any changes need to be made. In addition, please review your paper as a whole for typographical and essential corrections.

Your PDF proof has been enabled so that you can comment on the proof directly using Adobe Acrobat. For further information on marking corrections using Acrobat, please visit https://authorservices.taylorandfrancis.com/how-to-correct-proofs-with-adobe/

The CrossRef database (www.crossref.org/) has been used to validate the references.

AUTHOR QUERIES

Q1 Please provide full reference details for (Churchman, 1971) following journal style. If a reference is not supplied, we will delete the unwanted citation.

Q2 There is no mention of Reference (Cronbach 1951, Florida Department of Education 2019) in the text. Please cite the reference in the text. If no citation is supplied, we will delete the uncited reference from the list.

Q3 Please note that the ORCID section has been created from information supplied with your manuscript submission/CATS. Please correct if this is inaccurate.

READING PSYCHOLOGY
https://doi.org/10.1080/02702711.2024.2379262

Routledge
Taylor & Francis Group

Unhinged: Reading Comprehension Tests as Gatekeepers to Teaching

John Wesley White and Daniel Dinsmore Q3

College of Education and Human Services, University of North Florida, 1 UNF Drive, Jacksonville, Florida, USA

ABSTRACT

A teacher's ability to read effectively is critical to that individual's ability to teach reading skills. Correspondingly, most state departments of education require that prospective teachers earn a passing score on a standardized reading comprehension test before they can enter university-based teacher education programs or otherwise get a professional teaching license. Having witnessed quality candidates get pushed away from teaching due to poor performance on the our state's reading comprehension measure and given that previous studies have shed doubt on the construct validity of major standardized assessments (e.g., the SAT and ACT), we examined the validity of our state's standardized reading assessment for teachers. Using data generated by 115 college-aged participants in a prerequisite course for our teacher education programs, we found that our state's assessment did little to measure reading comprehension. Instead, it measured students' test-taking skills. This is exceptionally problematic because tests like this one keep significant numbers of qualified and motivated individuals from entering the teaching profession. Worse, due to the oft-researched relationship between test-taking skills to the socioeconomic background of the test-taker, these impacts may be exponentially worse for individuals from minority and lower socioeconomic backgrounds, thereby further reducing their opportunities to teach.

ARTICLE HISTORY
Received 17 April 2024
Accepted 9 July 2024

KEYWORDS
Teacher education preparation; assessment; equity; reading comprehension; standardized testing

With the importance of reading to greater personal and societal growth, it comes as no surprise that public and private entities have attempted to

CONTACT John Wesley White j.white@unf.edu College of Education and Human Services, University of North Florida, 1 UNF Drive 57/2321, Jacksonville, FL 32233, USA.

devise efficient (aka, inexpensive) means of assessing an individual's ability to read and comprehend a variety of texts. Because reading is foundational for almost all K-12 learning, states and the federal government have mandated that K-12 students be assessed at regular intervals (in the current era, this has come to mean yearly if not more often). Entire industries have arisen around the need for institutions, and particularly educational institutions, to measure their constituents' respective reading abilities. In turn, educational institutions use reading comprehension data for myriad purposes (e.g., to measure and foster student growth, for institutional improvement, to control access to programs and jobs, etc.). In almost all cases, these industries have sought to measure reading comprehension through large-scale, standardized, multiple-choice tests in which the test taker reads a passage and then answers a series of associated questions (Betebenner & Linn, 2010). The goals of standardized reading assessments are arguably noble; to gauge our students' academic progress and to ensure that those entering new arenas (e.g., college, graduate programs, and professions) have the abilities they need to thrive therein.

There is ample evidence that reading is critical to educational and career success (e.g., Snow, 2010). Additionally, the use of assessments is critical to track students' abilities to read (Ortlieb, 2012). Taken together, it is important for assess if students are able to read effectively and, when that is not the case, for those educators to provide them with appropriate means for remediation. However, there are times when assessments of reading may fail to meet these purposes and actually introduce problems into the educational context in which they are used. One example of this is the largely unquestioned use of large-scale, multiple-choice assessments as the primary means of assessing students' reading abilities and the use of such tests as gatekeepers for controlling access to higher education and to careers. Replicating the earlier work of Katz, Lautenschlager, and colleagues (e.g., Katz & Lautenschlager, 1994), we document how one commonly used reading assessment (in this case a high-stakes test required of prospective teachers) fails to adequately measure reading comprehension. Rather, we posit, this high-stakes test measures something entirely different and we forward two possibilities: test taker's prior knowledge and test-wiseness. Because this measure is not unique, we call into question the validity of generalized high-stakes reading comprehension assessments and their extensive use in educational policymaking and as gatekeepers to programs and careers.

Context

There is little debate that the ability to read different kinds of texts effectively and efficiently is requisite for success in many if not most fields of study. Reading is so closely tied to general educational success that reading scores have, at least since the 2001 passage of the *No Child Left Behind Act* (a major revamping of the Elementary and Secondary Education Act), become the dominant measure of individual American public schools' progress or lack thereof (Lee & Reeves, 2012; Reback, 2008). American schools use reading data for multiple purposes: to assess students' reading levels throughout the school year, to tailor reading interventions for struggling readers and to create individualized education plans (IEPs), to make decisions on student advancement to the next grade level, to hire new teachers and specialists, and to gauge their teachers' pedagogical effectiveness. Nations not only test their students' reading proficiency on a yearly basis, they compare the relative strength or weakness of their overall educational systems by using reading comprehension test data (see for example the Programme for International Student Assessment). Reading levels, it would seem, have become the barometer of K-12 educational success.

However, this measuring of a student's ability to read does not stop with her or his completion of primary and secondary school, where reading is actively taught. Rather, reading is so central to success in college and graduate programs that numerous measures have been created to ensure that prospective enrollees in these programs are proficient readers. Passing scores on the Scholastic Aptitude Test (SAT) and the American College Testing (ACT) exam, both with reading comprehension measures, are required for entry into the vast majority of colleges and universities in the United States—so much so that the list of schools that make these tests optional is far shorter at 1070 (Fairtest.org, 2020), and most recently the state university system of California will not even look at optional SAT or ACT scores (Nieto del Rio, 2021). The general portion of the Graduate Record Examination (GRE), also with a reading comprehension measure, is the most widely used assessment required for entry into graduate programs in the United States and in many foreign countries (Educational Testing Service). The Graduate Management Admission Test (GMAT), the Law School Admission Test (LSAT), and the Medical College Admission Test (MCAT)—required for entry into most American business schools, law schools, and medical schools respectively—each include sections that assess test takers' reading comprehension.

The ability to read effectively is so central to education *writ large* that state legislatures and state departments of education have mandated that prospective K-12 school teachers must not only pass the SAT or ACT for acceptance into a baccalaureate program (and pass all of the courses therein), they must also pass additional tests—all of which include a reading comprehension assessment—in order to obtain their professional state teaching license. Forty-five states and the District of Columbia mandate that prospective teachers pass the Praxis Exam while the five other states use a variety of different assessments termed *general knowledge*. All of these measures include a reading comprehension section (sometimes labeled "verbal reasoning") and all serve as gatekeepers to the teaching profession. An inability to pass the reading comprehension portion of these tests—regardless of one's success in college-level courses, one's content and pedagogical knowledge, and one's experience in K-12 classroom-based assignments—means that one is excluded from the teaching profession. Thus, measures of assessing one's reading ability are high stakes for the test-taker. These measures are also, however, high stakes for American students and their schools; they exacerbate the teacher shortages and they discriminate against the kinds of teachers our schools most need: minorities.

The United States is well into a decade in which the demand for highly effective teachers has far exceeded the supply (Sutcher et al., 2015). The demand for new teachers was expected to grow by 1.6 million between 2010 and 2020 alone (US Bureau of Labor Statistics, 2012). Yet despite this robust demand for new teachers in all regions and all types of public schools, too few people are entering the profession (Ingersoll & May, 2011; Sutcher et al., 2015). Colleges of education—long the primary pipeline to teaching—have endured a decade-long decline in enrollment (Westervelt, 2015). Alternative pathways to teaching and the use of so-called emergency teaching certifications (e.g., little to no formal preparation for teaching), have done little to meet the demand for certified teachers (Westervelt, 2015). At the same time that the dominant pipelines to teaching are drying up, increasing numbers of classroom teachers are choosing to leave the profession. Roughly 8% of teachers leave the profession every year and that number grows to 20% or greater in high needs schools (Aragon, 2016; Ingersoll & May, 2011; National Center for Education Statistics, 2016), resulting in a 50% teacher attrition rate within the first five years (Smith & Ingersoll, 2004). The attrition problem is so significant that the National Commission on Teaching and America's Future (NCTAF) notes that "some school districts report a higher dropout rate for teachers than students" (NYU Steinhardt School of Culture et al., 2017). This situation is a crisis for the teaching profession but even more for the nation's K-12 students; the schools most in need of high quality and culturally-competent teachers struggle desperately to find them and, when they do, to keep them. As a result, the nation's most vulnerable students–a group that is growing rather than shrinking–suffer even more.

As of 2012, 49% of the students enrolled in public school were minorities and that number is expected to be at least 54% in the next two years alone (US DOE, 2016). At the same time, however, the teaching force is becoming increasingly homogeneous and thus less representative of—or as understanding of—the students they are charged with teaching (Cushner et al., 2014; National Center for Education Statistics, 2009). According to US Department of Education data (2016), only 18% of current teachers in our public schools are minorities. And while diversity in the teaching force is rising overall—albeit at a glacial pace—the number of Black and Hispanic teachers is decreasing (US DOE, 2016).

This lack of diversity in the teacher workforce is itself lamentable; it is most problematic, however, because it has a significant detrimental impact on high needs students and their educational outcomes. The seminal work of Shirley Brice Heath (1983) and Michelle Foster (1997), as well as an abundance of newer research (see for example Cushner et al., 2014) demonstrate that a lack of cultural-congruence between students and teachers proves especially harmful to minority students, who crave the kinds of stability often lacking in their homes and who desperately need culturally-similar classroom mentors who can readily employ culturally-responsive classroom strategies (Athanases & Martin, 2006; Khalifa et al., 2016). Instead, the nation's high needs students experience a revolving door of teachers and a system in which they are disproportionately taught by a district's least experienced and least culturally aware teachers (Darling-Hammond, 2004). Darling-Hammond's research highlights that the harm to these students is reflected in academic disengagement, high dropout rates, low literacy levels, future low-wage employment, and high rates of incarceration. Further, this damage is cyclical because it proves toxic to the culture of the school well into the future. In summary, standardized reading comprehension measures are not only of questionable validity, they may be serving to keep the very types of teachers our students most need out of our classrooms. Here, we aim to take a relatively unexamined questions—the reliability and validity evidence for a standardized reading examination for entrance into a teacher education program—to determine whether this tests the focal construct of reading comprehension, or whether it measures irrelevant constructs to reading comprehension that might be culturally biased (e.g., background and linguistic knowledge; Chen & Henning, 1985).

In what follows, we examine one of the reading comprehension measures used to assess prospective teachers' reading abilities—a measure that also serves as a gateway to the profession of teaching. Our primary research question is whether or not participants' scores and item characteristics on standardized reading passage items differ depending on whether or not they actually read the text. Given the findings of Katz and colleague's examinations of the reading comprehension portion of multiple high stakes exams (e.g., SAT, GRE, ACT), we predict there to be little to no difference between conditions for at least some of the test items. If our hypothesis is correct, we hope that this work will help to open a new discussion about the validity and uses of these measures.

Methods

Participants

Participants for the study were 115 undergraduate students enrolled in teacher preparation courses in a College of Education in the Southeastern United States. These participants were predominately female (87%), white (78%), with a majority in their junior year (63%), and an average age of 21.75 year (*SD* = 5.69). Additionally, 97% of the sample reported English as their first language with an average GPA of 3.40 (*SD* = .41). With regard to their previous experience with the General Knowledge Test used in this study–which is compulsory for completion of the teaching certificate in the state in which certification is granted– 52% had taken and passed the test previously, 19% had taken and failed the test previously, and 29% had not yet attempted the test. These students completed the research tasks and were provided extra credit in their respective courses for their participation.

Materials and Measures

The materials and measures for this experiment consisted of two text passages with an accompanying set of reading comprehension questions. Since we were interested in using naturalistic passages, we chose to use passages released by the company that creates and assesses the standardized reading comprehension examination used in the state for teacher licensure. The first passage, the Hernando Cortéz (HC) passage, was about the Mexican conquest by Hernando Cortéz. It was 475 words in length with a Flesch Reading Ease score of 50.3 and a Flesch-Kincaid Grade Level of 14.0. The second passage, the Background Music (BM) passage, was about the use of background music for various purposes. It was 503 words in length with a Flesch Reading Ease score of 48.5 and a Flesch-Kincaid Grade Level of 11.9. Both passages in their entirety can be found at (http://www.fl.nesinc.com/studyguide/TIG_GK_Reading/01.asp).

Each passage was accompanied by a set of multiple-choice questions about that passage. For the HC passage there were seven items that consisted of two items that purported to measure *key ideas and details*, three items purported to measure *knowledge of craft and structure*, and two items that purported to measure *integration of information and ideas*. For the BM passage there were three items of each type—*key ideas and details*, *knowledge of craft and structure*, and *integrations of information and ideas*. An example of an item purported to measure *knowledge of craft and structure* follows:

The organizational plan used by the author in paragraphs 2–4 can best be described as

a. Order of importance
b. Spatial order
c. Comparison and contrast
d. Chronological order.

Procedures

For this experiment we used a counterbalanced randomized control trial. After consenting to participate in the study, participants were randomly selected to either answer the questions without having been given the associated reading passage (the experimental group) or to answer the questions after having read the associated passage (the comparison group). Individuals were counterbalanced across passages, meaning that they read the passage and answered the associated questions for one passage while only answering the questions for the other passage.

We used Qualtrics to administer the demographic questions (reported in the *Participants* section), the passage itself, and the questions. Participants were emailed a link to the study consent form, and if they consented were directed to the study materials. Responses to the multiple-choice questions were scored via the scoring guide provided by the state on its website associated with the released passages (the site used past test passage/answer combinations as examples for practice for test-takers). Correct responses were scored a "1" and incorrect responses were scored a "0".

One assumption of this design (and reading comprehension tests more generally) is that when students were presented with the passage that they actually read that passage, or at that very least used the passage in some way to answer these questions. Qualtrics data regarding the length of time spent on the research indicated that the average participant spent 7.62 minutes ($SD = 4.42$) on the research task, indicating that some reading was likely occurring rather than randomly answering questions. We did remove five outliers from this time stamp data as it appeared they may have either not completed the two tasks in one sitting or they left the survey active after completion.

Analysis

To investigate differences in item characteristics across the two groups (i.e., experimental and control), we used both observed and latent approaches. For the observed approaches we analyzed these items using *item difficulty* (i.e., the percentage of participants across the groups that answered the items correctly) and the *index of discrimination*. The index of discrimination is the difference between the item difficulty for the group that *did* read the passage to that of the group that *did not* read the passage. Thus, positive values would indicate that participants who read the passage got that particular item correct at higher rates and negative values would indicate that participants that did not read the passage actually scored better than those who did. In our analyses we relied on Ebel's Ebel (1954) guidelines that items with an index of discrimination greater than .40 were good, those greater than .20 were marginal, and those below .20 were poor. Although we use these guidelines, we do believe in this instance that items should be highly discriminatory (e.g., >.50) due to the extreme condition (i.e., not reading the text) of the experimental group.

With regard to overall scores for the passages (i.e. how many items participants answered out of the seven and three items on the HC and BM passages respectively) we analyzed these in two ways. First, we undertook an independent samples t test to examine if the scores between the two groups (i.e., read the passage and did not read the passage) were different. These were run for each passage as well as summed across both passages. Finally, for the observed analyses we ran an ordinal logistic regression to compare how many individuals got a set number of questions correct, versus how many would have been expected to get that correct by chance. For example, we tested whether the predicted 25 individuals that would be expected to get two items correct by chance was significantly different than what we observed in this sample who did not read the passage.

In addition to the observed analyses, we also relied on latent analyses to dig deeper into the reliability and validity evidence of these items across the two groups. These latent analyses allowed us to parse these data by using these latent approaches to disaggregate error from the item characteristics. In this regard, we relied on both exploratory factor analysis (EFA) and latent reliability indices. For the EFAs we used both the total variance explained by components (i.e., how much variance could be explained by one or more components that could represent the total number of items in the scale) as well as the loadings of each item on those respective components. The higher the loading, the more variance that item contributed to that particular component.

Results

Observed Analyses

Item difficulties and indexes of discrimination for the items from both passages are presented in Table 1. According to Ebel's Ebel (1954) index of discrimination guidelines, only one item (HC1) would be described as *good*. Three items (HC2, HC5, HC6, and BM1) would be described as *marginal* and five items would be described as *poor* (HC3, HC4, HC7, BM2, and BM3). There did not appear to be any pattern of whether the category of these items (*key ideas and details*, *knowledge of craft and structure*, and *integration of information and ideas*) were better or worse. For the three items that were labeled *key ideas and details*, these items were situated across the spectrum with one each being good, marginal, and poor. Similarly, the other two categories—*knowledge of craft and structure* and *integration of information and ideas*—also spanned both the marginal and poor categories in Ebel's scheme.

Next, we tested the scores on the items within the passages (i.e., total score for the HC passage and total score for the BM passage) and across the passage (i.e., total combined score on these passages) across the two groups (i.e., those who read the passage versus those that did not read the passage). To do this, we ran three independent samples T tests. Table 2 presents the mean scores for participants that *did* read the passage, mean scores for participants who *did not* read the passage, standard deviations for those that *did* read the passage, standard deviations for

Table 1. Categories, item difficulties, and indexes of discrimination for the passage items.

Item	Category	Item difficulty for those that *did* read the passage	Item difficulty for those that *did not* read the passage	Index of discrimination
HC1	KID	.77	.29	.48
HC2	KCS	.54	.34	.20
HC3	KID	.43	.16	.16
HC4	KCS	.42	.52	−.10
HC5	KCS	.60	.24	.36
HC6	III	.75	.47	.29
HC7	III	.30	.33	−.03
BM1	KID	.60	.40	.20
BM2	KCS	.57	.39	.18
BM3	III	.69	.72	−.03

Note. HC = Hernando Cortez passage; BM = background music passage; KID = key ideas and details; KCS = knowledge of craft and structure; III = integration of information and ideas.

Table 2. T test results for the Hernando Cortéz, background music, and combined passage scores.

Passage	Means for *did* read	SDs for *did* read	Means for *did not* read	SDs for *did not* read	t	p	Cohen's D
HC	3.70	1.71	2.34	1.61	4.39	<.01	.81
BM	1.86	.93	1.50	.78	2.21	.02	.41
Combined HC & BM	2.77	1.65	1.93	1.32	4.27	<.01	.56

Note. HC = Hernando Cortez passage; BM = background music passage. The combined scores are an amalgam of participants across the conditions, thus one participant's score will show up in the *did* read group and their other score will show up in the *did not* read group.

those that *did not* read the passage, t-values, *p*value, and effect sizes (Cohen's D) for those three tests.

First, we were concerned in these analyses that the mean scores for the passages were lo; however, of the students who took the test previously in our sample, 73% of those students reported passing it. Second, the *p*value for these tests indicate that our sample size was large enough to detect a stable difference in these overall scores. Thus, this provides some evidence that these differences are not due to random fluctuations in our sampling method. Third, and most relevant to our guiding research question for the study was the overall effect of the intervention (i.e., *not* reading the passage) had not only on individual items, but the overall scores themselves. Regarding effect size, there have been many cautions as to how to interpret these effect sizes and that context should play a key role here. So while generic effect size indices (see Fritz et al., 2012) would indicate that these were *moderate* (the BM passage) and *large* (the HC and Combined Scores) effects, in the context of the intervention of *not* reading the passage, these scores did not appear to be significantly different. In other words, we would have expected the differences in these groups to be much larger given the extreme differences in the groups (reading versus not reading a passage before answering questions.

To put this in perspective, we graphed the number of participants along the number of correct response by passage who *did* and *did not* read the passage alongside what would be expected an individual would correctly answer an item by chance (i.e., random guessing). These are included in Figure 1.

As is evident for these charts, the *did not* read group (represented by the light gray bars) outperformed what one would expect them to answer correctly by chance (represented by the striped bars). Additionally, we checked to see if these differences between the number of items correct by chance and for those that did not read the passage were significantly different. Logistic regression—which is appropriate for ordinal level dependent variables—revealed that overall there were significant

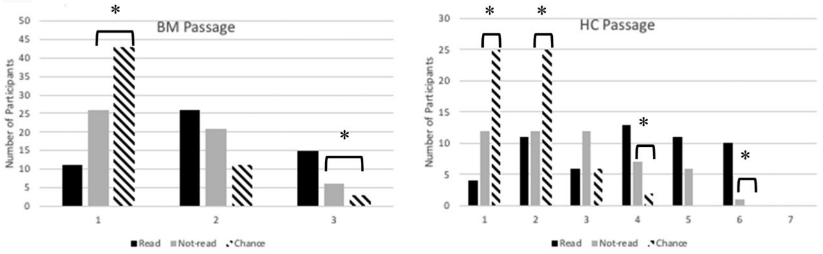

Figure 1. Number of correct responses by passage for the *did* read and *did not* read groups versus chance guessing.

differences in the score outcomes from chance to those that did not read passage.

For the HC passage, there was an overall significant difference ($Wald = 4.60$, $df = 1$, $p = .03$). Further, at each score level, there were significant differences between the groups, except for those that answered three questions correctly ($Wald = 3.64$, $df = 1$, $p = .06$). For the BM passage, the overall analysis was not quite significant ($Wald = 3.50$, $df = 1$, $p = .06$). However, score totals for the BM passage of 1 and 3 were both significant with lower and upper bounds of −4.85, −2.64 and 1.41, 2.90 respectively.

Latent Analyses

For the exploratory factor analyses (EFA), we examined both the variance explained by each component as well as the loadings of the items onto those components. We did both of these analyses by examining the groups separately, since we expect the processes (i.e., a reading comprehension process versus some other process for the group that did not read the passage) to differ. First, the scree plots which show the relative variance explained by each component are presented in Figures 2 and 3 for the HC and BM passages respectively. As is evident from these four plots, the relative variance of the items explained by these components did not differ in any appreciable way. Further, the component loadings were examined to see what differences emerged. For the HC passage, we rotated the first two components for the clearest distinction between those loadings using direct oblimin rotation which allows the two components to be correlated (in the case of these two components $r = .07$). For the BM passage we only retained the first component. Given the lack of clarity of component structure here, there could of course be arguments for differing numbers of factors.

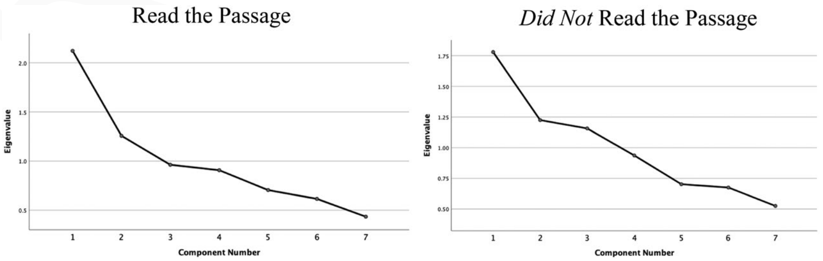

Figure 2. Scree plots (i.e., relative amount of variance in the items explained by each factors) for the HC passage among the group that read the passage and group that did not read the passage.

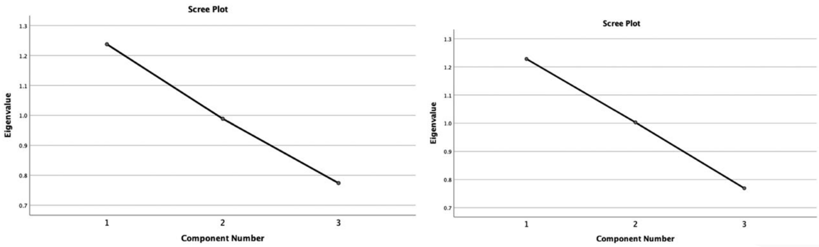

Figure 3. Scree plots (i.e., relative amount of variance in the items explained by each factors) for the BM passage among the group that read the passage and group that did not read the passage.

Table 3. Rotated component loadings for the items on the HC passage across the two groups.

	Read the passage		*Did not* read the passage	
Item	Component 1	Component 2	Component 1	Component 2
HC 1	.73	−.03	.77	−.04
HC 2	.72	.25	.71	−.09
HC 3	.23	.47	.44	.01
HC 4	.61	−.56	.26	.54
HC 5	.26	.70	.42	.15
HC 6	.70	.07	.14	.59
HC 7	−.11	.48	−.30	.83

For both the HC and BM passages, the component loadings (Tables 3 and 4 respectively) were not similar across the two groups. For example, in the BM passage one would expect items to load similarly across components, however, while items one and three loaded strongly on the component for those that read the passage, only item three loaded strongly on that component while the second items loaded strongly in a negative direction. This is further evidence in our view that the underlying processes of responding to these items are quite different. This makes sense as one group read the passage and the other did not. But, more to the point here, suggests that the group that read the passage did not simply use their background knowledge or test-wiseness (or lack thereof) to respond to the items.

Finally, we calculated latent reliabilities for each of the factors described previously. These reliabilities are presented in Table 5, with values of *H* greater than .70 considered to be good (Hancock & Mueller, 2001). In addition to running these reliabilities within each group (i.e., those that read the passage and those that did not), we also ran these reliabilities with these groups combined to see if the reliabilities were affected. With regard to these latent reliabilities, unlike the component structures we *do not* see a discernable pattern of difference among these three groupings of score responses. This is particularly surprising when the assumption in

Table 4. Component loadings for the items on the BM passage across the two groups.

Item	Read the passage	*Did not* read the passage
BM 1	.78	.09
BM 2	.31	−.78
BM 3	.74	.79

Table 5. Latent reliabilities among the passages by group.

	HC Passage		
Group	Comp 1	Comp 3	BM Passage
Did Read	.80	.68	.74
Did Not Read	.76	.76	.76
Combined	.80	.73	.73

the combined group is that some read the passage and some were not, that it was as consistent or better than compared to the reliabilities when the groups were separated. Some of this could be attributed to sample size, but given the size of the samples relative to the number of items, particularly for the BM passage, we would not suspect this to be the case. Given that it is often observed reliabilities that are reported, we also calculated these to see if there were different patterns. The number of items for the BM passage were small, so they yielded very low observed reliabilities—and in one case a negative reliability coefficient, however, for the HC passages the alpha values were again quite consistent across the groups with alpha values of .55, .49, and .57 for the *did* read, *did not* read, and combined groups respectively.

Discussion

The findings for this study concern us for two reasons. First, from this evidence it appears that the degree to which these items measure reading comprehension differs dramatically across items. Each item *should* be quite sensitive to whether or not the individual read the accompanying passage no matter which definition of reading comprehension one chooses (e.g., Cromley & Azevedo, 2007; Kintsch, 1988; Van den Broek et al., 1996). In this regard, our findings here mirror those of Katz, Lautenschlager, and colleague's findings regarding a similar approach to evaluating the validity of the GRE reading comprehension exam.

Second, and the larger of the two concerns, is that these test items do not seem to be measuring reading comprehension, the targeted focal construct of interest. Specifically, while reliability measures across the groups—both latent and observed—appear to be similar, the structures of the constructs as evidenced by the EFAs (i.e., construct validity evidence; Messick, 1980) are not. In other words, there appears to be no problem with the measures when examining reliability evidence only. However, when examining the validity evidence, there appear to be different processes at work here entirely. While it is clear that those participants who did not get the passage did not use the passage itself to answer the question, we wonder if indeed one explanation for these findings is that those that were provided with the passage did not need to use the passage, or perhaps relied upon some of the processes that those who did not have the passage relied on as well. While these are difficult to pin down from these data presented here, we forward to possibilities.

The first of these possibilities is the use of background or prior knowledge rather than the text passage itself. Given the influence of prior knowledge on learning generally (Murphy & Alexander, 2002), and reading comprehension more specifically (McNamara & Kintsch, 1996), it is likely that this construct played a role here as well. Whether or not the passage is presented, if background knowledge—rather than reading comprehension—play a key role in the responding to these test items, this creates an issue for equity and access for minority populations into teacher training programs that has existed for some time (US DOE, 2016).

Second, with regard to testing, there is a real possibility that test-wiseness is playing a role here. This has been a known issue in the reading comprehension testing literature for some time as P. David Pearson (1978) described the "search-and-destroy" testing strategy whereby students match words in the test items to words in the passages themselves. Here, since one group did not have the passages, it would certainly be other testing strategies that they would be relying on. These might include such strategies as eliminating unlikely distractors and using grammatical clues in the items themselves (Dolly & Williams, 1986). Like prior knowledge, there is also evidence that these testing strategies are less available to students from minority groups and these issues certainly do not help minority students (Madaus & Clarke, 2001). If indeed these testing strategies play a role, this creates yet another barrier to the teaching profession.

Thus, the most important takeaway from this area of research is that the ubiquitous use of large-scale reading comprehension assessments are unnecessarily—and we believe unfairly—hindering test-takers from the opportunities for which the tests were developed as a form of gatekeeping. Katz, Lautenschlager, and colleague's work has suggested that the reading comprehension portion of the Graduate Record Examination fails to accurately assess test-takers' ability to read and fully understand a given passage. Because the GRE is required for entry into countless graduate programs, many otherwise qualified students are unable to enroll in these programs. Similarly, in our state, teacher licensure—and acceptance into university-based teacher education programs—require passage of all parts of the General Knowledge Test (GK). In 2017, the passage rate for the English Language Arts portion of the General Knowledge Test (of which reading comprehension is a part) was only 57% (LaGrone, 2018). Thus, many college students with otherwise stellar academic records are denied the chance to teach due to poor performance on a measure with questionable validity evidence. Increasingly, even highly skilled classroom teachers who are working on a temporary license (e.g., have not yet passed the GK) face the loss of their jobs because they cannot pass the GK (LaGrone, 2019). Evidence has suggested that this phenomenon is not unique to our state. Forty-five states and the District of Columbia require prospective teachers to pass the Praxis examination, which also includes a reading comprehension measure (https://www.ets.org/praxis/states). The four remaining states, like ours, use their own tests that include reading comprehension measures.

Central to our study is a concern that the reliance on poorly constructed reading comprehension measures deny otherwise qualified students from entering teaching and are unnecessarily contributing to an increasingly chronic nationwide teacher shortage (Ingersoll & May, 2011; Sutcher et al., 2015). More nefariously, because success on large scale reading comprehension measures may be significantly affected by test takers' prior knowledge and test-wiseness, these measures may be serving as an added barrier to minority students and students from lower socioeconomic backgrounds entering the teaching profession. This, in turn, further contributes to the widening divide between an ever-more-diverse PK-12 student body and an increasingly homogeneous teacher workforce, a phenomenon that has widespread implications for teachers' cultural competence and students' buy-in to schooling (US Department of Education, 2016; Cushner et al., 2014). While it is important to assure that future college students, future graduate students, and future teachers (among others) can read and understand a variety of texts, our data confirms those reported by Katz, Lautenschlager, a colleagues: large scale, multiple-choice reading passages lack the validity required to accurately measure test takers' reading abilities; instead, they measure other things entirely. In short, while large scale reading comprehension measures do serve as gatekeepers to programs and opportunities, they keep people out for the wrong reasons.

Future Directions for Research

While we are comfortable with the conclusions drawn from these data, there are some limitations to this dataset that need to be addressed in future studies. For one, these data were drawn from one university.

Due to the standardized requirements across the state for acceptance into teacher preparations programs, we do not think this is a specific problem per se, we do believe the ability to replicate these findings, especially in other states with different tests is critical.

Second, we drew here on analyses and indices that utilized both observed and latent analysis. While the shift from observed to latent analyses has been upon us for the last few decades, reporting—especially by state agencies—has lagged behind contemporary practice in the research literature. Thus, we think it is incumbent upon researchers to explore new ways to reexamine existing data provided by testing companies. Each of these issues—our smaller sample here and the data testing companies provide—could be solved at least partially through the creation of data sets that companies should provide to the state, which in turn could be available to researchers. This would unleash the vast amount of expertise in our research community to tackle these types of problems. The fact that these testing companies hold state contracts should be good leverage to require that these data be available to state-funded agencies for further analysis. Additionally, this trend would follow the American Educational Research Association's call for transparency in the use of data (AERA, 2016).

Future Directions for Practice

Practically speaking, at the heart of the issue here is whether one could separate the "good comprehenders" from the "poor comprehenders" such that only the "good comprehenders" are admitted into teacher education programs. Due to the low ceiling of these data (i.e., those that read the passage did not score particularly well despite a majority of those that took the test previously having passed it) and the high floor (i.e., those that did not read the passage scored well above chance), there appears to some difficulty in setting cut scores that would adequately separate these "good" and "poor" comprehenders.

Thus, in practice there is a thin line—too thin in our view—between being able to accurately assess reading comprehension and limiting access to the quality teacher training. While we are not in charge of setting policy related to these exams, we would recommend that those that do engage in two types of arguments described by Messick (1980) when considering the ethical imperatives of testing. These two arguments lay bare the potential social consequences of engaging in assessment or engaging in a certain type of assessment. These two types of arguments were described as Kantian inquiry (i.e., comparing a proposed test against an alternative proposal; Churchman, 1971) and Hegelian inquiry (i.e., the social consequences of not testing at all). Our assumption is that the lay person—which may include those making policies such as those examined here—*assume* the benefits of testing for reading comprehension, without considering the potential ramifications of these tests. Thus, it is incumbent on all of us to work together to be sure that the *testing itself* does not do more harm than good or whether other testing procedures may yield better social outcomes.

Concluding Thoughts

For us, the evidence here was both surprising and not. Given previous findings with the GRE, we hypothesized that tests designed to measure something as complex as reading—and to do so across a giant spectrum of test-takers—may continue to provide data of limited validity. We were thus not particularly surprised that many of Hall's findings remain true today. We were surprised, however, in that we continued to hold out hope that standardized assessments of reading comprehension may have improved in the intervening 30 years. We were even more surprised by the degree to which the associated items for each passage functioned so poorly across the two test-taking groups—even those that purported to measure what should be higher-level comprehension processes (e.g., *integration of information and ideas*) that one would think would be difficult to answer correctly without reading the passage.

If we care about equity and access in our teaching workforce, this issue of entrance examinations and the potential that construct-irrelevant items are so prevalent, is more than concerning. Similarly, we are concerned that the massive amounts of hours and monies spent on assessing our students' reading levels—at the K-12 levels, in college, and beyond—are possibly being misspent on tests that are assessing something else entirely.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

John Wesley White http://orcid.org/0000-0002-2588-5787

References

American Educational Research Association (AERA). (2016). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40. doi:10.3102/0013189X035006033

Q1
Q2

Aragon, S. (2016). *Teacher shortages: What we know. A report of the Education Commission of the States*. Retrieved from https://files.eric.ed.gov/fulltext/ED565893.pdf.

Athanases, S. & Martin, K. (2006). Learning to advocate for educational equity in a teacher credential program. *Teaching and Teacher Education*, *22*(6), 627–646. doi:10.1016/j.tate.2006.03.008

Betebenner, D., & Linn, R. (2010). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda. Retrieved from https://www.ets.org/Media/Research/pdf/BetebennerandLinnPresenterSession1.pdf.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155–163. doi:10.1177/026553228500200204

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*, 311–325. doi:10.1037/0022-0663.99.2.311

Cronbach, L. J. (1951) Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Cushner, K., McClelland, A., and Safford, P. (2014). *Human diversity in Education: A Intercultural Approach*. McGraw Hill Education.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, *106*(6), 1047–1085. doi:10.1177/016146810410600602

Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement*, *46*, 619–625. doi:10.1177/0013164486463014

Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, *14*, 352–364. doi:10.1177/001316445401400215

Fairtest.org. (2020). More than 1070 accredited clleges and universities that do not use ACT/SAT scores to admit substantial numbers of students into bachelor-degree programs. Retrieved from https://www.fairtest.org/university/optional.

Florida Department of Education. (2019). *Florida teacher certification examinations (FTCE) and Florida educational leadership examination (FELE) 2018 annual administration and technical report*. Pearson. Retrieved from: http://www.fl.nesinc.com/FL_ScoringReporting.asp.

Foster, M. (1997). *Black teachers on teaching*. The New Press.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology*, *141*, 2–18. doi:10.1037/a0024338

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future—A Festschrift in honor of Karl Jöreskog*. Scientific Software International, Inc.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge University Press.

Ingersoll, R., & May, H. (2011). Recruitment, retention, and the minority teacher shortage. Consortium for Policy Research in Education (Research Report # RR-69). Retrieved from https://www.cpre.org/sites/default/files/researchreport/1221_minorityteachershortagereportrr69septfinal.pdf.

Katz, S., & Lautenschlager, G. J. (1994). Answering reading comprehension items without passages on the SAT–I, the ACT, and the GRE. *Educational Assessment*, *2*(4), 295–308. doi:10.1207/s15326977ea0204_2

Khalifa, M., Gooden, M., & Davis, J. (2016). Culturally responsive school leadership: A synthesis of the literature. *Review of Educational Research*, *86*(4), 1272–1311. doi:10.3102/0034654316630383

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163–182. doi:10.1037/0033-295x.95.2.163

Lee, J. & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, 34(2), 209–231. doi:10.3102/0162373711431604

LaGrone, K. (2018). Failing & frustrated, Florida teachers still flunking state exam at high rates. *ABC Action News* (Tampa, FL). Retrieved from https://www.abcactionnews.com/news/local-news/ i-team-investigates/failing-frustrated-florida-teachers-still-flunking-state-exam-at-high-rates.

LaGrone, K. (2019). Florida school districts respond to bill taking aim at teacher certification (FTCE) controversy. *ABC Action News* (Tampa, FL). Retrieved from https://www.abcactionnews.com/news/local-news/i-team-investigates/florida-school-districts-respond-to-bill-taking-aim-at-teacher-certification-ftce-controversy.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data. ED 450 183. Retrieved from https://files.eric.ed.gov/fulltext/ED450183.pdf.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288. doi:10.1080/01638539609544975

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Murphy, P. K., & Alexander, P. A. (2002). What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *The Journal of Experimental Education*, *70*, 197–214. doi:10.1080/00220970209599506

National Center for Education Statistics. (2009). *Characteristics of public, private, and Bureau of Indian education elementary and secondary school Teachers in the United States: Results from the 2007–08 Schools and Staffing Survey*. Retrieved from https://nces.ed.gov/pubs2009/2009321.pdf.

National Center for Education Statistics. (2016). *Teacher turnover: Stayers, movers, and leavers*. Retrieved from https://nces.ed.gov/programs/coe/pdf/coe_slc.pdf.

Nieto del Rio, G. M. (2021, May 15). Title. *The New York Times*. Retrieved from https://www.nytimes.com/2021/05/15/us/SAT-scores-uc-university-of-california.html.

NYU Steinhardt School of Culture, Education, and Human Development. (2017). *Keeping the teachers: The problem of high turnover in urban schools*. Retrieved from https://teachereducation.steinhardt.nyu.edu/high-teacher-turnover.

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Ortlieb, E. (2012). The past, present, and future of reading diagnosis and remediation. *Journal of Language Teaching and Research*, *3*, 395–400. doi:10.4304/jltr.3.3.395-400

Pearson, P. D. (1978). Some practical applications of a psycholinguistic model of reading. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about reading instruction. IRA*.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5-6): 1394–1415. doi:10.1016/j.jpubeco.2007.05.003

Smith, T. & Ingersoll, R. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, *41*(3), 681–714. doi:10.3102/00028312041003681

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*, 450–452. doi:10.1126/science.1182597

Sutcher, L., Darling-Hammond, L., Carver-Thomas, D. (2015). A coming crisis in teaching? Teacher supply, demand, and shortages in the U.S. *The Learning Policy Institute*. Retrieved from https://edworkingpapers.com/sites/default/files/A_Coming_Crisis_in_Teaching_REPORT.pdf.

US Bureau of Labor Statistics. (2012). *Kindergarten and elementary school teachers*. Retrieved from https://www.bls.gov/ooh/education-training-and-library/kindergarten-and-elementary-school-teachers.htm.

US Department of Education (US DOE). (2016). *The state of racial diversity in the educator workforce*. Retrieved from https://www2.ed.gov/rschstat/eval/highered/racial-diversity/state-racial-diversity-workforce.pdf.

Van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165–187). Psychology Press.

Westervelt, E. (2015). Where have all the teachers gone? *NPR News*. Retrieved from https://www.npr.org/sections/ed/2015/03/03/389282733/where-have-all-the-teachers-gone.