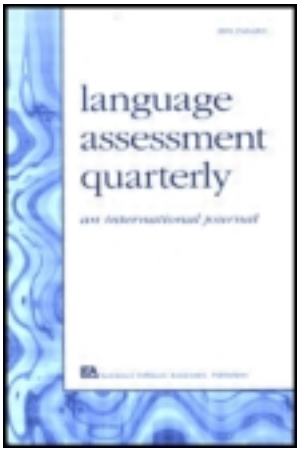

Language Assessment Quarterly
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlaq20

Investigating the Reliability of the Civics Component of the U.S. Naturalization Test

Paula Winke [a]
[a] Michigan State University

Available online: 01 Dec 2011

To cite this article: Paula Winke (2011): Investigating the Reliability of the Civics Component of the U.S. Naturalization Test, Language Assessment Quarterly, 8:4, 317-341

To link to this article: http://dx.doi.org/10.1080/15434303.2011.614031

Language Assessment Quarterly, 8: 317–341, 2011

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.614031

Routledge
Taylor & Francis Group

ARTICLES

Investigating the Reliability of the Civics Component of the U.S. Naturalization Test

Paula Winke

Michigan State University

In this study, I investigated the reliability of the U.S. Naturalization Test's civics component by asking 414 individuals to take a mock U.S. citizenship test comprising civics test questions. Using an incomplete block design of six forms with 16 nonoverlapping items and four anchor items on each form (the anchors connected the six subsets of civics test items), I applied Rasch analysis to the data. The analysis estimated how difficult the items are, whether they are interchangeable, and how reliably they measure civics knowledge. In addition, I estimated how uniformly difficult the items are for noncitizens ($N = 187$) and citizens ($N = 225$) and how accurate the cutoff score is. Results demonstrated the items vary widely in difficulty and do not all reliably measure civics knowledge. Most items do not function differently for citizens and noncitizens. The cutoff is not as accurate as applied in the operational test. The data revealed that test scores contain construct-irrelevant variance that undermines the overall reliability and validity of the instrument. I discuss these results not only to better understand the civics test but also to recommend how United States Citizenship and Immigration Services could conduct a similar study with the goal of raising the reliability and validity of the test.

[1]According to the USICS website (http://www.uscis.gov), the application fee of $675 comprises a general fee of $595 plus a biometrics fee of $80. Applicants who are 75 years or older are not charged the biometric fee. Military applicants filing under Section 328 and 329 of the INA are not charged an application fee.

[2]USCIS is vague in describing how English proficiency is evaluated. There are three English proficiency tests—speaking, reading, and writing. Speaking is assessed by the USCIS officer over the course of the interview. For the reading test, the candidate must read out-loud one of three written English sentences. For the writing test, the candidate must correctly write down one of three dictated English sentences. Key vocabulary used in the reading and writing tests are provided as study materials by USCIS. Information on how the tests are scored is not provided. The official description of the English tests and the official USCIS study materials for the tests can be found on the USCIS website (http://www.uscis.gov) and particularly at the following website: http://www.uscis.gov/newtest.

Correspondence should be sent to Paula Winke, Michigan State University, Department of Linguistics and Languages, Second Language Studies Program, Wells Hall, East Lansing, MI 48824. E-mail: winke@msu.edu

For those not born in the United States, the path to United States citizenship has many requirements (United States Citizenship and Immigration Services [USCIS], 2009b). In most cases, prospective citizens must have been permanent residents for five years, have "good moral character," swear an oath to support the Constitution, and pay a $675 application fee.[1] With some exceptions, they must prove that they can read, write, and speak English.[2] This study examines a final requirement: the civics test.

The civics test is an open-ended, oral test administered during the naturalization interview with an immigration officer, who also evaluates English proficiency over the course of the interview. The immigration officer asks 10 of 100 possible civics questions.[3] A passing score is 6. The questions are designed to test "knowledge and understanding of the fundamentals of the history, and of the principles and form of government, of the United States" (Immigration and Nationality Act of 1952). Applicants may be asked questions such as why the flag has 13 stripes, what Susan B. Anthony did, how long senators serve, and where the Statute of Liberty is located.

The test has been controversial (i.e., Allen, 2006; Preston, 2007; Rothstein, 2006; Tomson, 2007). Some see it as a barrier to citizenship (Han, Starkey, & Green, 2010; Winn, 2005), others as part of a beneficial process that ensures that naturalized citizens can participate fully in U.S. society (see Cameron, 2002, and Blackledge, 2005, for impassioned debates on this issue). Because the questions, answers, and official study materials are publically available (USCIS, 2009a), it can be seen merely as an achievement test (see Brown, 2005) that measures how well applicants have mastered the content of the study guides. For its part, USCIS states the test is "an important instrument to encourage civics learning and patriotism among prospective citizens" and requires only a low level of English proficiency (USCIS, 2008, p. 3). Scholars debate both those assertions (Cameron, 2002; Etzioni, 2007; Kunnan, 2009; Piller, 2001; Shohamy & McNamara, 2009).

Whatever the merits of those arguments, the Naturalization Test is unquestionably—in the government's understated phrase—an "important instrument." Without answering 6 of 10 civics questions correctly, applicants cannot become citizens. And only citizens can vote, get a U.S. passport, apply for federal aid, apply for government jobs, seek official election, and provide family members with a pathway toward permanent U.S. residency or citizenship. Applying for citizenship, then, "is more than an administrative process; it is a life-changing event" (Naturalization Delays, 2008, p. 3). And although data are scarce, about one applicant in three fails the civics test, the English test, or both.[4]

The test is also important because it is expensive and widely administered. In fiscal year 2008, USCIS received $375 million in fees from naturalization applicants (USCIS, 2009c). In the same period, USCIS administered 1,652,468 naturalization interviews, approving 1,050,399 applicants for citizenship, denying 121,283, and giving 480,786 pending status (USCIS, 2010). In 2009, USCIS processed 570,442 applications for naturalization (United States Department of Homeland Security, 2010), but this number does not include the applications held in USCIS

[3]To my knowledge, no publicly available information explains how USCIS officers select a 10-question form or states whether their discretion to select a form is at all limited. No information is available as to form equality or composition.

[4]A 1998 Immigration and Naturalization Service–commissioned study found that out of 7,843 naturalization applications, 34% were denied due to failure on the English, the civics test, or both (del Valle, as cited in Kunnan, 2009, p. 113).

mailrooms and not processed (Naturalization Delays, 2008; USCIS, 2010). Whatever the exact numbers, the 10 civics questions in the Naturalization Test change millions of people's lives.

Any high-stakes test used to make important decisions should be both reliable (meaning that its scores are consistent and as free from measurement error as possible) and valid (meaning that the decisions made on the basis of the scores are, as emphasized by Bachman, 1990, p. 25, *meaningful*, *appropriate*, and *useful*; see also Alderson, Clapham, & Wall, 1995; Chapelle, 1999; Kunnan, 1998; and Messick, 1995, for discussions on reliablity and validity). Reliability and validity are especially crucial when the test is widely used and has significant impacts on the test takers' lives (Kane, 2002; McNamara, 2006; Norris, 2008; Ryan, 2002; Shepard, 1993; Shohamy, 2000, 2001). Yet USCIS has never shown that the civics test is either reliable or valid. Not only has it not released any reliability or validity studies, it has apparently never examined basic questions such as whether the civics items truly measure civics knowledge. Nor has it released, or apparently created, a technical test manual (Kunnan, 2009), a document that provides a scientific basis for score interpretation.

According to the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME; 1999), a technical test manual should describe how items were developed, how the test was piloted, when and how the test was normed, and what population was used to norm the test. If a cutoff score is used, it should document how it was established. The documentation should report the standard error of measurement (SEM), a reliability statistic that describes a band around a test taker's raw score within which that test taker's score would probably fall if he or she took the test many times. The size of that band—the expected variability in a test taker's scores—provides information on how precise a cutoff score should be (see Brown, 2005, p. 188, for more information on SEM). Finally, if a test has more than one form, a test manual should describe how the multiple forms were equated (see pp. 35, 68–70 of AERA et al., 1999). All examinees should be presented with equally difficult test forms; otherwise, test scores are not comparable. In the case of the civics test, different examinees take different standardized forms of the test (USCIS Office of Public Engagement, personal communication, March 17, 2011). Each immigration officer asks a set of 10 items from the pool of 100. If the forms are of unequal difficulty, the test is not a consistent measure of the underlying construct (the test is unreliable; scores depend on the form given). A single cut-score applied regardless of a form's difficulty would be unfair. In such a case, score interpretations (test validity) would be highly questionable.

More detailed definitions of reliability and validity may be helpful at this point. Reliability is an empirical, statistical measure of how consistent test scores are (Bachman, 1990; McNamara & Roever, 2006). According to Bachman (1990), reliability "has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context" (p. 25). Reliability can be determined by testing individual test takers more than once and measuring differences, by calculating item difficulty and discrimination indices,[5] or by investigating item performance in comparison to test taker performance (Bachman, 2004; Brown, 2005). Estimating reliability is an essential part of developing a test and normally is done before using the test (Bachman, 2004; Davidson & Lynch, 2002; Fulcher & Davidson, 2007, 2009).

[5]As explained by Brown (2005), an item's difficulty index is the percentage of test takers who correctly answered the item. An item's discrimination index is a statistic that indicates the degree to which the item separates the test takers who performed well overall on the test from those who performed poorly.

Validity is a less precise term. It can be viewed as a unitary concept (in his seminal 1989 paper on validity, Messick, for example, referred to all validity as *construct validity*) or as a construct-independent framework (AERA et al., 1999) or assessment-use argument (Bachman & Palmer, 2010) that researchers develop through gathering, analyzing, and documenting evidence that supports test scores uses for particular purposes (see Sireci, 2007, for an easy-to-follow commentary on validity theory and its history). But no matter how validity is viewed, everyone agrees that reliability is a minimum requirement: for a test to be valid, both its individual items and the test as a whole must be reliable (Alderson et al., 1995; Bachman, 1990; Chapelle, 1999, p. 258; Lado, 1961). Thus, reliability is at the core of any test's validity argument. Most people would also say that a test can be reliable but still invalid (i.e., Bachman & Palmer, 2010; Brown, 2005; Chapelle, 1999; Fulcher & Davidson, 2007; Kane, 2002; McNamara & Roever, 2006). For example, a grammar test could have high statistical reliability, but scores from it would be inappropriately interpreted and applied if they were used as an indication of speaking proficiency. In other words, validity is commonly used to describe the contextual value of the test beyond its statistical reliability. As explained by Bachman (1990), "the investigation of validity is both a matter of judgment and of empirical research, and involves gathering evidence and appraising the values and social consequences that justify specific interpretations or uses of tests" (p. 26). Validity is sometimes used broadly to describe whether test scores are an appropriate tool for making particular decisions, whether the test beneficially impacts the curriculum leading up to the test (which is also known as test washback—see Cheng, Watanabe, & Curtis, 2004), and whether the test's effects on stakeholders are fair, beneficial, and not detrimental (see McNamara & Roever, 2006; Nichols & Williams, 2009; Norris, 2008; and Winke, 2011, for more on the *broad* concept of validity).

So far, critiques of the U.S. Naturalization Test have been qualitative (more related to the broad validity of the test) and have not focused squarely on the test's reliability. For example, researchers have claimed that the English and civics portions of the citizenship test are subjective (Elliott, Chudowsky, Plake, & McDonnell, 2006), contain inappropriate content (McNamara & Shohamy, 2008), and discriminate against the poor and uneducated (Etzioni, 2007). Specifically, McNamara and Shohamy (2008) reported that the civics test, which they referred to as a "knowledge of society" test (p. 94), makes "substantial and unacknowledged literacy demands on those applying for citizenship" (p. 94). Researchers' opinions did not change after the revised test was released in 2008. Kunnan (2009) described the revised civics test items as unfair and irrelevant to the naturalization process. He wrote that "the Naturalization Test cannot claim that it can assess the English language ability and knowledge of U.S. history and government of applicants for citizenship as the qualities of test construct, content, administration, scoring and reporting are all questionable" (p. 95). These critiques all broadly pertain to the Naturalization Test's validity, that is, whether the test's scores adequately reflect what is needed to confer citizenship upon someone and whether the test's effects on stakeholders are fair, beneficial, and not detrimental. But these critiques do not address the core of the test's validity argument, the test's reliability. This study aims to look at the U.S. Naturalization Test civics component in terms of its reliability.

Studying the reliability of the civics test is difficult because USCIS has not released any data on test performance (Kunnan, 2009). But estimating the civics test's reliability is still possible because the USCIS website (http://www.uscis.gov/citizenship) provides the civics questions and answers and some information about how the test is administered. I could therefore accurately

reconstruct the test and collect data from a new test-taker group and analyze their data by using Item Response Theory (IRT). IRT is often used to investigate the reliability of items that appear on a test. In a nutshell, IRT can be used to assess the soundness of a proposed, ideal model of the test items and test takers, in which the responses of a test taker of a given ability are compared to the difficulty level of each item on the test. (For overviews of IRT, see Bachman & Palmer, 2010; de Ayala, 2009; Hambleton & Jones, 1993.) IRT is unique because its "item parameter estimates are independent of the group of examinees used" (Bachman, 2004, p. 142). Data from the real test-takers are therefore unnecessary as long as the official test questions are given to a similar population in a way that mimics the administration of the real test.[6]

This study has four research questions:

RQ1: Are the 100 civics items on the U.S. Naturalization Test interchangeable?
RQ2: Are all the items reliable? That is, do they reliably measure civics knowledge?
RQ3: Would U.S. citizens who have not prepared for the test mostly pass?
RQ4: Are the individual items easier for citizens than for noncitizens?

The hypothesis for Research Question 1 is that the items are interchangeable. The hypothesis for Research Question 2 is that all of the items are of high quality. In other words, the individual items on the test will be reliable. The hypothesis for Research Question 3 is that U.S. citizens who have not prepared should pass because they know the civics information that the test is meant to measure. The basis for this hypothesis stems from Emilio Gonzalez, the director of USCIS, who stated the civics test "genuinely talks about what makes an American citizen" (Preston, 2007, p. A24). He further stated that the test encourages "citizenship applicants to learn and identify with the basic civic values that unite us as Americans" (USCIS, 2007, p. 2), implying that U.S. citizens possess the knowledge that the citizenship test requires. The hypothesis for Research Question 4 is that the items should be more difficult for noncitizens because citizens presumably know more about civics than do noncitizens who have not prepared for the test.

METHODS

Participants

The study sample comprised 414 adults, of whom 225 were citizens, 187 were noncitizens, and two did not identify their citizenship status. There were 181 male participants, 227 female participants, and six participants who did not identify their sex. Because participants were recruited mainly on the campus of Michigan State University in the United States, most were pursuing a degree in higher education: Six had completed only high school and were not studying for a further degree, 230 were undergraduates, 132 were graduate students, 39 had 4-year college degrees and were professionals, and seven did not identify their educational background. The average age was 32 years, with a standard deviation of 14.65.

[6]An anonymous *LAQ* reviewer duly noted that this study's test takers may differ from the true test takers because "motivation to pass the test may affect engagement with the test materials and hence performance." However, in rebuttal, test developers, out of practical needs, often pilot test items on experimental test takers who are similar to real test takers but whose scores are not used for decision-making purposes. Data from pilot testing are regularly used to develop norming criteria and/or to validate future uses of tests.

Materials

To collect civics test performance data, I created six forms (Forms A–F) of a mock U.S. Citizenship Test, each with 20 items from the pool of 100 USCIS civics test questions (USCIS, 2009a), which are considered public information and can be distributed or copied without alteration unless otherwise specified. First, from the pool of 100 questions, I randomly selected four as anchor items (Questions 3, 12, 66, and 96), which would appear on each of the six forms. I split the remaining 96 civics test items systematically across the six forms (Item 1 from the USCIS pool is Question 1 on Test Form A, Item 2 from the pool is Question 1 on Test Form B, etc.). Thus, each form had 20 items: four anchor items and 16 individual, nonoverlapping items. The answer key was the list of correct responses published by USCIS. A sample form (Form A) appears in Appendix A.

Procedures

The six forms of the test were administered to people who passed by test administrators (17 students in my graduate-level, language testing class) at various places on or near the Michigan State University campus, such as dining halls, classroom building hallways, and sidewalks. Each test administrator held a clipboard and asked passersby if they would like to take a mock U.S. Citizenship Test to test their U.S. citizenship knowledge. The test administrators told volunteers that the test would take about 10 minutes and that they would earn a small piece of candy. The test administrators asked the questions orally to mimic how USCIS officers administer the test. The test takers responded orally (they did not see the written questions), and the test administrators marked the responses as right (1) or wrong (0) on the test form, which also contained the correct responses for comparison. If a response was wrong, the test administrator wrote down the wrong response on the form; I later checked those responses for accuracy. During the test, test takers were not told if they got an individual item right or wrong. After a response, the administrator would say phrases like "Okay, thanks," "Great," or "That's okay" to mimic the responses given in the sample *USCIS Naturalization Test and Interview Video* (available at http://www.uscis.gov/citizenship).[7] At the end of the test, test takers were told how many out of 20 they got correct and whether they "passed" (correctly answered 12 or more items). I asked each test administrator to collect data from 24 individuals as part of a class project. Each test administrator rotated through the six forms four times, asking the first test taker the questions on Test Form A, the second test taker the questions from Test Form B, and so on. The goals were met. However, four of the administrators went beyond the call and gave the test to one or two extra people (six extra test takers total); thus data were collected from 414 individuals.

Analysis

Data from the 414 individuals were submitted to a Rasch IRT model analysis (Bond & Fox, 2007) using Winsteps version 3.70.0.3. Many different Rasch models are used to measure item and test

[7]To access the video, go to http://www.uscis.gov/citizenship and then look under Learners > Study For The Test > Study Materials For The English Test.

performance; I used the simplest one, which is meant for analyzing dichotomous data with the values 0 or 1, where 1 indicates a correct answer, or, in Rasch terms, better ability on the underlying latent trait (Bond & Fox, 2007). The mathematics behind the Rasch dichotomous model has been presented elsewhere (i.e., Beglar, 2009, pp. 104–105; Bond & Fox, 2007, pp. 277–280), but the premise of the mathematical model is that if all items on a test function perfectly, a given test taker will have a higher probability of answering easier items and a lower probability of answering harder items, with what is "easier" or "harder" being relevant to the test taker's ability on the underlying latent trait. Thus, if a group of test takers with varying levels of the underlying latent trait are administered the test items, we can assess how closely the items "fit" along an imaginary linear scale with person ability and test item difficulty matched along the scale.

RESULTS

Before presenting the results, I first provide a brief introduction to the process of interpreting the output from Rasch analysis. Winsteps produces a map of person measures and item calibrations (the Wright map) that shows the abilities of the test takers and the difficulty of the items on the same logit (log odds unit) scale. Bond and Fox (2007) explained,

> The logit scale is an interval scale in which the unit intervals between the locations on the person-item map have a consistent value or meaning. The Rasch model routinely sets at 50% the probability of success for any person on an item located at the same point on the item-person logit scale. (p. 38)

Thus, the map in Figure 1 shows that 12 of the test takers in this study had a 50% chance of passing Items 13, 35, 4, and 81. Their chance of success increased to about 75% for Item 38, which is one logit easier, and decreased to about 25% for Items 14, 37, 65 and 71, which are one logit harder. The map, however, does not tell us the precision of these item and person estimates. For that, we turn to the other data from Winsteps.

RQ1: Are the Items on the Test Interchangeable?

To answer the first research question, I first used Winsteps to create an item separation index, a kind of index that compares the "true" spread of item difficulty to the test's measurement error (Fisher, 1992). Here, the item separation index was 4.52, meaning that the 100 items could be separated into 4.52 statistically distinct levels of citizenship knowledge. The reliability of the item separation index was .95, indicating that the items were very reliably separated in terms of their levels of citizenship knowledge. I therefore split the items into five levels of citizenship knowledge, which I called common knowledge, basic knowledge, general knowledge, expert knowledge, and technical expert knowledge. (See the right-hand column in Figure 1.) The item separation analysis thus indicates that the items have different levels of difficulty and are not interchangeable. Appendix B lists the individual difficulty measures of the items along the logit scale.

I also performed a parallel-forms reliability analysis to see if items are interchangeable (see Bachman, 1990, p. 168; Brown, 2005, p. 176, for more information on this reliability statistic). Each test taker took 20 civics test items. For each test taker, I divided those 20 items into two tests of 10 items by selecting the odd items for one set and the even items for the other. I then calculated

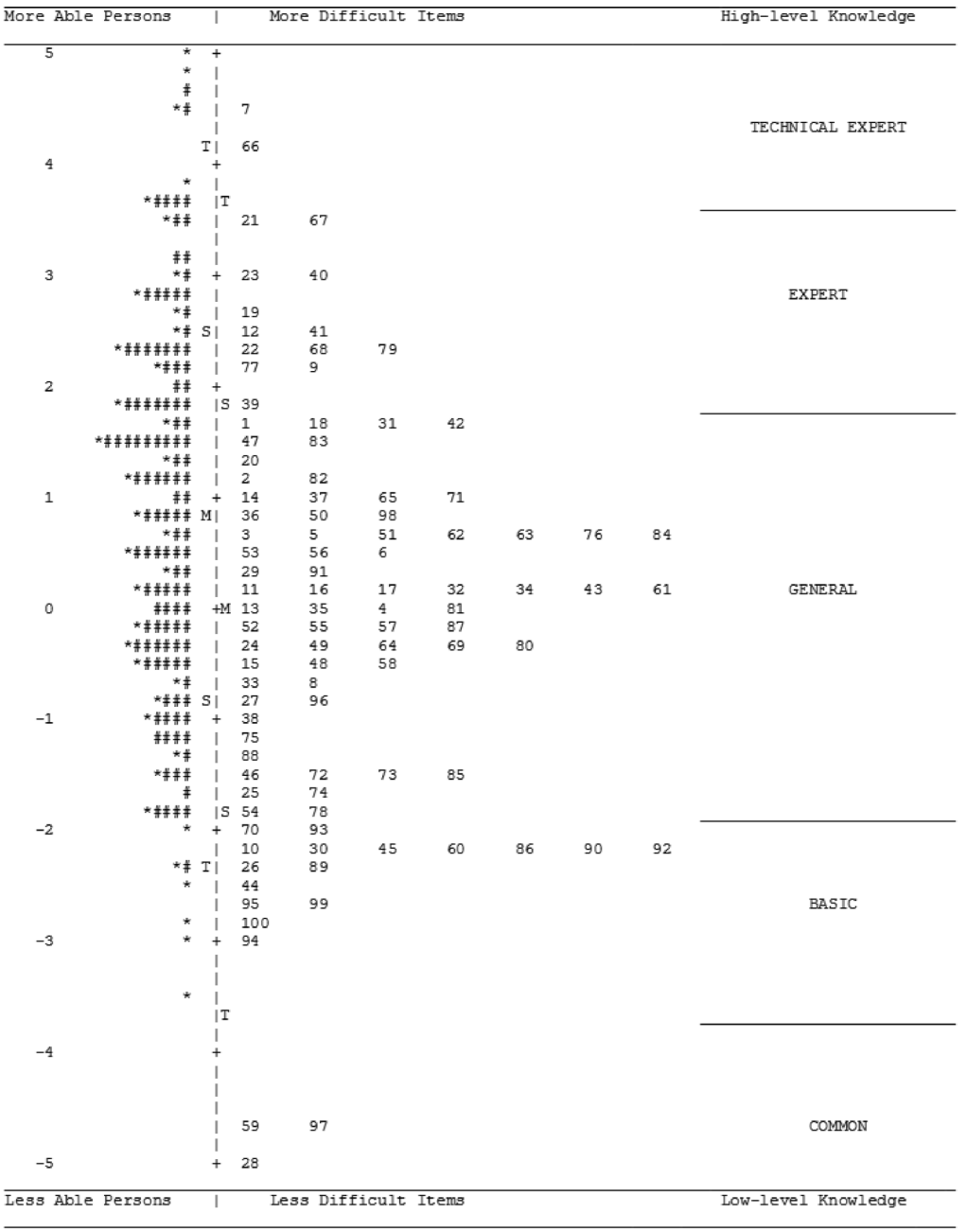

FIGURE 1 Wright map of person measures and item calibrations. *Note.* Each hash symbol is three persons. Each asterisk is one or two persons. M = the mean of the person or item estimates. S = 1 *SD* from the mean. T = 2 *SD* from the mean.

a correlation coefficient (Pearson's r) to determine the degree of relationship between the two sets of scores. The odd and even test forms correlated at $r(412) = .71$, $p = .00$, demonstrating that 29% of the variance across the two forms was unrelated to the underlying test construct. Another way to investigate whether two forms of a test are measuring the same construct is to see if their means are *not* statistically significantly different (Bachman, 1990, p. 168). Therefore, I conducted a t test on the two test's means using SPSS (version 18). There was a significant difference in the scores between the odd ($M = 5.33$, $SD = 2.48$) and even ($M = 6.29$, $SD = 2.16$) tests, $t(413) = –9.04$, $p = .00$, demonstrating the forms are not measuring the same ability. Further demonstrating this are the data on passing. Using a passing score of 6 out of 10, 136 (30%) of the 414 test takers failed both forms (117 were noncitizens, 18 were citizens, one was unidentified concerning citizenship); 181 (44%) passed both forms (19 noncitizens, 162 citizens);[8] and 97 (23%) had conflicting outcomes (51 noncitizens, 45 citizens, and one unidentified), meaning that they passed either the odd or even test form but not both. These findings additionally suggest that the individual items on the test are not interchangeable, as the interchange of items creates varying results (passing or failing) for almost a quarter of the test takers.

RQ2: Are All Items Reliable? That Is, Do They Reliably Measure Civics Knowledge?

To answer the second research question, I evaluated the reliability of the items using Rasch standardized item, weighted mean-square fit statistics. Rasch analysis produces infit and outfit mean-square statistics for each item. By looking at the fit statistics, researchers can monitor how compatible the data are with the model. The expected infit and outfit values are 1, but the fit statistics can range from 0 to infinity. Acceptable mean-square fit values for high-stakes tests are between .80 and 1.20 (Bond & Fox, 2007, p. 243). Infit is high or low when persons have unexpected patterns of scores on items that are targeted roughly at their level. Outfit is high for an item when persons have unexpected scores on items that should be relatively easy or hard for them. Low outfit indicates too little variation in the observed response pattern. Eliminating items that have high misfitting values will normally make most low misfitting items less misfitting. Thus, low infit and outfit values provide less motivation for data editing than do high values. That is, low values (less than .80) are less of a concern than high ones (more than 1.20).

To explain this better and to use very simplistic terms, an item's infit may be large (greater than 1.20) and problematic when, for example, the item is at the intermediate level but it is not being answered correctly by intermediate-level test takers. An item's outfit may be large (greater than 1.20) and problematic when, for example, the item is at the novice level but is being answered incorrectly by advanced-level test takers, or when it is at the advanced level but is being answered correctly by novice-level test takers. Normally, researchers place "more emphasis on infit values than outfit values in identifying misfitting persons or items" (Bond & Fox, 2007, p. 286) because

[8]When you look at the passing rates, the test appears to be working at the extremes ends of ability (the double-passes are mostly citizens, and the double-fails are mostly noncitizens), but almost any test discriminates at a fairly high level at the extremes. What is more difficult to achieve is a test that reliability and accurately measures the ability of test takers who possess average-skill levels. And this test does not do well in reliability measuring those with average civics knowledge. This is additionally problematic because the cutoff is at the average level of ability. The area of skill measurement in which the test may be the least reliable is where the test's responsibility in decision making resides.

outfit is very susceptible to outlying cases (very unexpected patterns of responses). (See the special topic of *Misfit diagnosis* in Linacre, 2011, for clear explanations of infit and outfit.) For example, a strong contribution to outfit may be a novice-level test taker who gets an extremely difficult item correct or a very skilled test taker who gets a very easy item wrong. But infit is less influenced by outliers and reflects more accurately the overall pattern of responses. Thus, I monitored this data set for items with high infit values greater than 1.20 and looked to see if these items also had high outfit. Fourteen of the items had infit mean-square statistics greater than 1.20. Thirteen of these same 14 items had outfit mean-square statistics greater than 1.20. These high fit values suggest underfit to the model and indicate unmodeled noise or other sources of variance in the data; in other words, these items degrade measurement because they are not reliable, they do not produce consistent scores, and they reduce the overall reliability and validity of the instrument. Thus, the answer to Research Question 2 is no, not all items on the civics test are reliable; 14 appear to be unreliable and do not measure civics knowledge. These 14 items and their fit statistics are presented in Table 1.

RQ3: Would U.S. Citizens Who Have Not Prepared for the Test Mostly Pass?

To answer the third research question, I looked at the passing rate of the 225 citizens who took the test, with the cutoff score for passing set at 60% (12 items or more on the test of 20 items). Out of 225 citizens, 187 (83%) passed, indicating that yes, most U.S. citizens who do not prepare for the test would pass. The noncitizens, who had not prepared either, were less likely to pass. Out of 187 noncitizens, 27 (14%) passed. Using Winsteps, I conducted a Welch's t test to ascertain whether there were statistical differences between the mean scores of the two groups. The noncitizen and citizen average test measures were different, $t(405) = 18.64$, $p = .00$, with citizens receiving considerably higher raw scores (average = 14.34) than noncitizens (average = 8.54). The descriptive data pertaining to Research Question 3 are in Table 2.

RQ4: Are the Individual Items Uniformly Difficult for Noncitizens and Citizens Alike?

To investigate the fourth research question, I performed a differential item functioning (DIF) analysis using Winsteps. As explained by Zumbo (1999), DIF indices are interpreted as measuring item *impact* when there are "true differences between the groups in the underlying ability of interest being measured by the item" (p. 224). In such cases DIF as item impact is expected because "impact occurs when construct-relevant dimensions differentially affect the tests scores for different groups of examinees" (Gierl, 2004, p. 3). DIF as item impact is a measure of the test's underlying latent trait; it reflects the groups' true differences on that construct. In other words, "impact refers to a difference in performance between two intact groups" (Dorans & Holland, 1993, p. 36). DIF indices are indications of impact when intact groups *should* differ in terms of the ability being measured.

In this test, the two groups (citizens who are native speakers of English and noncitizen who are nonnative speakers of English and who have not prepared for the U.S. Naturalization Test) should differ in their civics knowledge as indicated by USCIS's director Emilio Gonzalez. The more *impact* (i.e., DIF) the items have, the better. In this case, DIF signals an item mostly likely measures the ability of interest, which theoretically is civics knowledge, although knowledge of

TABLE 1
Unreliable Items on the U.S. Naturalization Mock Civics Test, Ordered by Infit Mean Square

Item		N	*Observed Raw Score*	*Item Difficulty Measure (in Logits)*	*SE*	*Infit Mean-Square Index*	*Outfit Mean-Square Index*
83.	During the Cold War, what was the main concern of the United States?	68	26	1.50	0.30	1.58	1.74
50.	Name one right only for United States citizens.	68	36	0.82	0.30	1.55	1.57
95.	Where is the Statue of Liberty?[a]	68	63	−2.62	0.52	1.54	2.76
53.	What is one promise you make when you become a United States citizen?	69	40	0.46	0.32	1.48	1.87
41.	Under our Constitution, some powers belong to the federal government. What is one power of the federal government?	69	19	2.52	0.33	1.35	1.43
81.	Who did the United States fight in World War II?	68	44	0.06	0.31	1.32	2.88
11.	What is the economic system in the United States?[a]	70	43	0.15	0.30	1.29	1.3
86.	What major event happened on September 11, 2001, in the United States?	70	63	−2.24	0.44	1.25	2.33
33.	Who signs bills to become laws?	69	52	−0.73	0.32	1.25	1.32
55.	What are two ways that Americans can participate in their democracy?	70	45	−0.16	0.29	1.24	1.83
12.	What is the "rule of law"?	414	103	2.51	0.14	1.23	2.27
51.	What are two rights of everyone living in the United States?	69	36	0.73	0.29	1.23	1.32
42.	Under our Constitution, some powers belong to the states. What is one power of the states?	70	25	1.72	0.30	1.22	1.55
54.	How old do citizens have to be to vote for President?[a]	70	61	−1.88	0.40	1.22	0.87

[a]Items that are preferentially asked to individuals older than 65 and who have been a legal, permanent resident of the United States for 20 or more years.

English most probably is intertwined with this construct. In recent years, the more common use of DIF indices has been to investigate item *bias*—such analyses are conducted with the groups of interest being *matched* with respect to the ability that the items are supposed to measure—in such cases, DIF is unexpected and indicates bias relating to the test-taker trait that separates the two groups. (Many studies have separated test takers by their gender, their ethnic origin, or

TABLE 2
Average Scores on U.S. Naturalization Mock Civics Test by Citizenship Status

Citizenship Status	N	*Observed Raw Score*	*Observed Raw Score Average (*SD*)*	*Average Test Measure (in Logits)*	*SE* M
Noncitizens	187	3226	8.34 (3.21)	−0.35	0.09
Citizens	225	1560	14.34 (2.97)	1.91	0.09
All	412	4786	11.62 (4.28)	0.88	0.66 (Model *SE*)

Note. Two test takers out of the 414 participants did not identify their citizenship status. They were excluded from this analysis.

their first language to understand if test items are biased along those characteristics; see Ferne & Rupp, 2007; Flowerdew & Shehadeh, 2008; Park, 2005, 2008, for examples of DIF as a measure of item bias.)

To summarize, in this data set, in which the two groups should have veritable differences in ability concerning their civics knowledge, DIF *should* be present. As stated by McNamara and Roever (2006, p. 84), if the two groups who should differ in their underlying ability on the construct being tested do not score differently on the items, there is "something wrong with the test!"

The Rasch model expects all DIF indicators to be neutral (zero) or uniform, because each item is imagined to impact all ability levels and all groups equally. The output from Winsteps displays the statistical significance (*t* value) associated with each item's difference in difficulty between groups. A general rule of thumb is that if the size of the DIF (the difference between the two groups' individual, absolute difficulty level on the item) is .5 logits or more and if the probability that the item exhibits no DIF beyond statistical randomness is less than .05, then the item does differ in how it functions between groups (Linacre, 2011).

I found that 77 of the items on the civics test were uniformly difficult for noncitizens and citizens alike and 23 were not. Of those 23 items, 10 were significantly easier for the citizens than the noncitizens (as one would expect) and 13 were significantly harder (which is counterintuitive—this is discussed later). These 23 items and their raw-score computations for DIF are displayed in Tables 3 and 4, with the 10 items that were easier for citizens in Table 3, and the 13 items that were easier for noncitizens in Table 4.

DISCUSSION AND CONCLUSIONS

In 2009, *Language Assessment Quarterly* published a special issue (Vol. 6, Issue 1) on language tests for citizenship, immigration, and asylum. The editors of that volume, Shohamy and McNamara, opined that language testers rarely become involved in significant political controversies such as immigration and naturalization policies because testers are mostly involved in developing and researching second language tests (Shohamy & McNamara, 2009). But this is changing, they wrote, because language testers are becoming increasingly involved and concerned with how tests are being used by governments as "instruments of power, as in the case of using tests as tools for setting educational agendas and exerting influence on the political order" (p. 1). Although not a direct test of language, the civics component of the U.S. Naturalization

TABLE 3
Items That Favored Citizens Over Noncitizens, Ordered by DIF Size

Item		*Noncitizens*	*Citizens*	*Noncitizens Minus Citizens*	*Welch*		
		Difficulty Measure (SE)	*Difficulty Measure (SE)*	*DIF Size*	t	df	p
47.	What is the name of the Speaker of the House of Representatives now?	3.37 (1.05)	.86 (.41)	2.52 (1.12)	2.24	50.00	.03
20.	Who is one of your state's U.S. Senators now?[a]	3.11 (.82)	.61 (.43)	2.51 (.93)	2.70	56.00	.01
5.	What do we call the first ten amendments to the Constitution?	2.02 (.63)	−.39 (.51)	2.41 (.82)	2.96	62.00	.00
3.	The idea of self-government is in the first three words of the Constitution. What are these words?	1.89 (.23)	−.25 (.20)	2.13 (.31)	6.90	397.00	.00
76.	What did the Emancipation Proclamation do?	1.83 (.56)	−.52 (.52)	2.35 (.76)	3.07	64.00	.00
43.	Who is the Governor of your state now?	.97 (.44)	−.82 (.52)	1.79 (.68)	2.62	67.00	.01
34.	Who vetoes bills?	.85 (.43)	−.75 (.56)	1.6 (.71)	2.25	64.00	.03
87.	Name one American Indian tribe in the United States.	.65 (.43)	−3.18 (1.42)	3.84 (1.49)	2.58	55.00	.01
52.	What do we show loyalty to when we say the Pledge of Allegiance?	.50 (.41)	−1.58 (.76)	2.08 (.86)	2.42	60.00	.02
57.	When must all men register for the Selective Service?	.46 (.42)	−1.17 (.64)	1.64 (.77)	2.13	65.00	.004

Note. Standard errors are in parentheses. Difficulty measures and differential item functioning (DIF) size are in logits.
[a]Items are preferentially asked to individuals older than 65 and who have been a legal, permanent resident of the United States for 20 or more years.

Test is administered alongside a language component, and it is *the* civics part of a larger measure meant to ensure that prospective citizens are proficient in English and have knowledge of U.S. civics. The civics component is a very strong instrument of power used for discriminating among a significantly large group of, primarily, second-language learners who take the test in their second language. Thus, the reliability of the civics test should be of interest to politically active language testers, especially considering the recent debate in language testing circles over the use of tests in the context of citizenship.[9]

At first, my interests in conducting this study were purely pedagogical. In the spring of 2008, I helped conduct a U.S. citizenship test preparation class at a local public library to nonnative, English-speaking immigrants and refugees who were applying for naturalization. I wondered which of the 100 civics questions we as a class should focus on most. Which were the most

[9]I am thankful to an anonymous *LAQ* reviewer for this argumentation.

TABLE 4
Items That Favored Noncitizens Over Citizens, Ordered by DIF Size

Item		*Noncitizens*	*Citizens*	*Noncitizens Minus Citizens*	*Welch*		
		Difficulty Measure (SE)	*Difficulty Measure (SE)*	*DIF Size*	t	df	p
7.	How many amendments does the Constitution have?	2.57 (.68)	5.51 (.76)	−2.94 (1.02)	−2.87	66.00	.01
66.	When was the Constitution written?	2.82 (.32)	4.56 (.24)	−1.73 (.39)	−4.39	387.00	.00
12.	What is the "rule of law"?	1.32 (.20)	3.09 (.17)	−1.77 (.26)	−6.74	394.00	.00
41.	Under our Constitution, some powers belong to the federal government. What is one power of the federal government?	1.37 (.54)	2.92 (.37)	−1.55 (.65)	−2.37	60.00	.02
83.	During the Cold War, what was the main concern of the United States?	.17 (.40)	2.61 (.42)	−2.44 (.58)	−4.23	64.00	.00
1.	What is the supreme law of the land?	.84 (.45)	2.11 (.37)	−1.26 (.58)	−2.16	63.00	.03
50.	Name one right only for United States citizens.	−.38 (.41)	1.86 (.37)	−2.24 (.55)	−4.05	64.00	.00
53.	What is one promise you make when you become a United States citizen?	−.30 (.42)	1.17 (.39)	−1.47 (.57)	−2.59	65.00	.01
81.	Who did the United States fight in World War II?	−.89 (.42)	1.12 (.40)	−2.01 (.58)	−3.49	65.00	.00
11.	What is the economic system in the United States?[a]	−.53 (.40)	.92 (.38)	−1.44 (.56)	−2.60	64.00	.01
78.	Name one war fought by the United States in the 1900s.[a]	−2.37 (.50)	−.51 (.58)	−1.86 (.76)	−2.43	66.00	.02
86.	What major event happened on September 11, 2001, in the United States?	−3.12 (.64)	−.68 (.56)	−2.44 (.85)	−2.88	63.00	.01
95.	Where is the Statue of Liberty?[a]	−3.58 (.78)	−1.11 (.63)	−2.47 (1.01)	−2.45	64.00	.02

Note. Standard errors are in parentheses. Difficulty measures and differential item functioning (DIF) size are in logits.
[a]Items are preferentially asked to individuals older than 65 and who have been a legal, permanent resident of the United States for 20 or more years.

difficult? Would some questions I thought were easy actually be hard for the nonnative, English-speaking immigrants and refugees in the class? In the class we used the USCIS's educational materials to prepare for the civics test, but time in class was limited. I felt I needed information on the individual items' difficulty levels. It surprised me that such information was not available. As I looked into the matter, I uncovered citizenship testing controversies in the language testing and educational measurement literature (Cameron, 2002; Cooke, 2009; Elliott et al., 2006; Etzioni, 2007; Gysen, Kuijper, & Van Avermaet, 2009; Han et al., 2010; Kunnan, 2009; Laversuch, 2008; Milani, 2008; Mollering, 2009; Piller, 2001; Shohamy & McNamara, 2009; White, 2008) and, more specifically, in the U.S. media (i.e., Allen, 2006; Preston, 2007; Rothstein, 2006; Tomson, 2007), which directly concerned the newly revised civics component of the U.S. Naturalization

Test. These studies and reports underscored how the development, implementation, and motivation for using language tests in awarding citizenship is highly tied to current political agendas. Furthermore, the use and interpretation of such tests are contested and debated worldwide. It struck me that statistical information on the civics items on the U.S. Naturalization Test would be of value not only to teachers of U.S. Naturalization Test preparation courses (for information on the widespread initiatives in teaching civics to adult English-language learners in the United States, see Terrill, 2000) but also to researchers and theorists in the field who are concerned with citizenship testing.

Because the civics test so strongly impacts the lives of so many test takers and their families, it is paramount that it be valid and reliable (Alderson et al., 1995; Chapelle, 1999; Kane, 2002; Kunnan, 1998; McNamara, 2006; Messick, 1995; Norris, 2008; Ryan, 2002; Shepard, 1993; Shohamy, 2000). Recognizing the importance of the test, USCIS revised its 100 civics test questions to make the naturalization process more "standardized, fair, and meaningful" (USCIS, 2009b, p. 1). (The revised questions were released in 2008.) Several researchers have discussed the objectivity, appropriateness, and fairness of the civics test (Elliott et al., 2006; Etzioni, 2007; McNamara & Shohamy, 2008), but this is the first study to measure its reliability. As explained by Alderson et al. (1995), all evaluations add to our overall knowledge of a test and are therefore important. But what matters most is that a test yields scores that accurately reflect the knowledge or ability being tested.

Measurement Error Stemming from Multiple Forms

Perhaps the greatest problem with the civics test is that it is possible that not everyone takes an equally difficult test. For each test administration, a USCIS officer selects a form of 10 items from the pool of 100 (the selection procedure has not been made public). But the items are not interchangeable. The item separation analysis showed that the 100 items on the civics test have 4.5 distinct levels of difficulty. The range of difficulty is tremendous: For example, 68 of 69 test takers in this study knew why the flag has 50 stars, but only five of 69 knew how many amendments the Constitution has. Depending on the procedure for selecting the 10 questions that make up a test form, the wide range of item difficulty may also create a risk of bias. If USCIS officers have unfettered discretion in choosing forms, they could select an unfairly easy or hard test by choosing particularly easy or hard forms, whether consciously or not. USCIS's procedure for creating test forms should be made public. (For example, does each form have the same number of items from each of the different levels of item difficulty?) Or at the very least evidence suggesting the forms' equality in difficulty should be provided (AERA et al., 1999).

The problem is worse if USCIS randomly combines 10 questions to make a test form. The parallel-forms reliability analysis in this study demonstrated that forms with items randomly selected will vary in terms of difficulty. In this case, each participant answered 20 questions, which were later treated as two, separate 10-question forms of the civics test. One way to read the results of the parallel-forms reliability analysis is to consider the reliability coefficient of .71, which shows that 29% of the variance in scores across forms is construct irrelevant. Another way to view the data is to consider that of the 414 test takers in this sample, 97 (51 noncitizens, 45 citizens, and one who did not indicate citizenship status) received a passing score on one test form but not the other. In other words, if these had been real citizenship-test forms, nearly one

applicant in four would have passed or failed the civics component, depending on which test form they happened to draw.

The validity of a test form is especially suspect if it contains one of the 14 unreliable items (those with infit mean-square statistics greater than 1.20—items listed in Table 1), because those items do not tap into any underlying latent trait. Because unreliable items tend to produce inconsistent results, those 14 items should be eliminated or replaced. If for some reason a test candidate were to have all 10 items from the pool of 14 unreliable items (a worst-case scenario in terms of test form reliability), interpreting the meaning of his or her civics test score would be extremely problematic.

The reliability data, however, appear to support the test in two respects. First, citizens answered more questions correctly (14.34 of 20) than did the unprepared noncitizens (8.34 of 20). Second, if the existing cutoff score of 6 out of 10 is meant to differentiate citizens from unprepared noncitizens in terms of their civics knowledge, the cutoff seems appropriate. But these data may be deceiving (see footnote 8). The SEM is far more important in determining the appropriateness of the cutoff score because the SEM can be used to determine the cutoff score's accuracy (Brown, 2005). By having standard deviation and reliability estimates, I was able to calculate the test's SEM, the estimation of how far test takers' scores will vary by chance alone if the test takers were repeatedly administered the test. The standard deviation of 2.31 was calculated by averaging the standard deviations of the odd ($SD = 2.48$) and even ($SD = 2.16$) test forms. Reliability was .71, the correlation between the odd and even forms. Using these numbers, the test's SEM is 1.28.[10] This means that a test taker who scored 5 on the civics test will probably (68% of the time) repeatedly receive a 4, 5, or 6 (within a band of one SEM) if he or she takes the test over and over again and does not learn in the process. An SEM of 1.28 indicates that the 1-point cut between award and denial of citizenship is subject to chance. According to the SEM, when someone receives a 5 on the civics test and is therefore denied citizenship, that decision is arbitrary and not based on the test taker's civics knowledge. If someone receives a 6 and passes, that decision too is arbitrary. The SEM strongly implies that many people's naturalization decisions are made completely at random.

The Items' Inconsistency in Measuring Civics Knowledge

In this study, I used DIF to calculate item-level score differences between the citizens and the noncitizens. Only 10 of the 100 items favored citizens, 77 items equally favored citizens and noncitizens, and 13 of the items favored noncitizens. Those results are surprising. One would expect that citizens know more than unprepared noncitizens about the "fundamentals of the history, and of the principles and form of government, of the United States," the underlying construct that the law stated the test should measure. Because citizens do not find most items easier than unprepared noncitizens, the items may be flawed (McNamara & Roever, 2006), especially the 13 items on which unprepared noncitizens outperformed citizens. The fact that the 100 items mostly do not function differently for citizens and noncitizens undermines the construct validity of the

[10]The formula for calculating the SEM is the following: SEM $= S\sqrt{1 - rxx'}$ where S is the standard deviation of the test and rxx' is the reliability estimate for the test. This formula and an explanation of SEM can be found on page 189 of Brown (2005).

test and questions what, exactly, the underlying construct is. USCIS (2006) stated that through the civics test and study materials it promotes "a common civics-based version of American identity to immigrants and citizens alike" (p. 37). But from these data one could speculate that the civics test questions do not tap into real citizens' civics knowledge. It is therefore not certain that the test preparation materials promote the common civics knowledge that citizens possess.

Implications and Limitations

The results of this study suggest that USCIS needs to change the civics component of the Naturalization Test. The data appear to support Elliott et al.'s (2006) allegation that USCIS did not give sufficient attention to the civics test's technical quality in the process of its redesign—a process that began in 2001, took 6 years, and cost $6.5 million (Preston, 2007). But USCIS should not rely on this relatively small-scale study. It should conduct a larger study along similar lines using data from some of the hundreds of thousands of actual applicants who take the test each year. Once it has analyzed the results, it should omit any unreliable items and ensure that all forms of the test are of equal difficulty. USCIS should also create a test manual if none yet exists. That manual should explain how the test was developed and normed and how the passing score was chosen. It should also explain the procedures for establishing the multiple test forms and justify the equality of the various forms. Other developers of large-scale, influential tests adhere to standards (AERA et al., 1999) concerning test evaluation and documentation. For example, Educational Testing Services (ETS; http://www.ets.org) provides technical test manuals and research results regarding the reliability and validity of its Test of English as a Foreign Language, a high-stakes, college entrance examination that measures the English-language-ability of international applicants (see ETS, 2011). Given the broad use and even higher stakes of the Naturalization Test, USCIS should do no less. At the very least, it should release raw data on public test performance that would enable others to investigate the reliability of the test (Kunnan, 2009). If national security concerns require that test procedures, validity studies, or raw data be kept secret, USCIS should say so and explain why.

This study has pedagogical implications. Classes preparing test takers should focus on the more difficult items (see Appendix B). Likewise, teachers should realize that some items that are easy for citizens may be difficult for noncitizens, such as naming an American Indian tribe or knowing when men must register for the selective service (see Table 3).

This study has limitations. Most important, the test takers were a self-selected group, not a random sample. Nor do the noncitizens in this study adequately represent those who take the U.S. Naturalization Test, for two reasons. First, they had not prepared for the test, which may have skewed the DIF analysis; the performance of unprepared applicants may not predict the performance of applicants who have seriously studied. Second, almost all of the test takers in this study were highly educated, probably more so than the average applicant for naturalization. It is likely that their English language skills were far higher than many of the individuals who take the U.S. Naturalization Test each year. Because the sample was neither random nor representative, the study's results may not be typical of actual naturalization applicants. USCIS could easily overcome this problem by using data from the actual test population. Future academic studies could use several approaches to achieve a more representative sample. A study comparing the performance of citizens and noncitizens could be conducted off campus so that study participants

have an educational level more typical of the general population. A study focusing on noncitizens might be conducted as part of preparation classes for the Naturalization Test. That group would be much more representative of the actual test-taking population. Testing them before the course begins and at the end of the course would also enable a researcher to compare the performance of unprepared and prepared applicants, as well as the effectiveness of the study materials. Finally, this study did not investigate the impact of English language proficiency on test scores. Future studies could do so.

ACKNOWLEDGMENTS

I presented an earlier version of this article in 2010 at the Midwest Association of Language Testers conference in Dayton, Ohio. I thank the students in my graduate-level language testing class for their help with data collection and for their lively in-class discussions on the issues brought up in this article. I thank Spiros Papageorgiou, Ching-Ni Hsieh, and three anonymous *LAQ* reviewers for their comments. Any mistakes, however, are my own.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Allen, R. (2006). *Is the U.S. citizenship test about to get tougher?* NBC News. Available from http://www.msnbc.msn.com/id/15940161/.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice, revised edition* (2nd ed.). Oxford, UK: Oxford University Press.

Beglar, D. (2009). A Rasch-based validiation of the vocabulary size test. *Language Testing*, *27*(1), 101–118.

Blackledge, A. (2005). *Discourse and power in a multilingual world*. Amsterdam, the Netherlands: John Benjamins.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Cameron, D. (2002). It's just not cricket. *Critical Quarterly*, *44*(2), 69–72.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

Cooke, M. (2009). Barrier or entitlement? The language and citizenship agenda in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 71–77.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: Guilford.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Educational Testing Service. (2011). TOEFL technical reports. Available from http://www.ets.org/toefl/research/technical_reports.

Elliott, S., Chudowsky, N., Plake, B. S., & McDonnell, L. (2006). Using the standards to evaluate the redesign of the U.S. naturalization tests: Lessons for the measurement community. *Educational Measurement: Issues and Practice*, *25*(3), 22–26.

Etzioni, A. (2007). Citizenship tests: A comparative, communitarian perspective. *The Political Quarterly*, *78*, 353–363.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148.

Fisher, W. P. J. (1992). Reliablity statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, *6*(3), 238.

Flowerdew, J., & Shehadeh, A. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly*, *42*(1), 115–123.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, *26*(1), 123–144.

Gierl, M. J. (2004, April). *Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA. Available from http://www2.education.ualberta.ca/educ/psych/crame/files/mirt2d_aera2004.pdf.

Gysen, S., Kuijper, H., & Van Avermaet, P. (2009). Language testing in the context of immigration and citizenship: The case of the Netherlands and Flanders (Belgium). *Language Assessment Quarterly*, *6*, 98–105.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Han, C., Starkey, H., & Green, A. (2010). The politics of ESOL (English for speakers of other languages): Implications for citizenship and social justice. *International Journal of Lifelong Education*, *29*(1), 63–76.

Immigration and Nationality Act, 8 U.S.C. § 1423 (1952).

Kane, M. J. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41.

Kunnan, A. J. (2009). Testing for citizenship: The U.S. Naturalization Test. *Language Assessment Quarterly*, *6*(1), 89–97.

Kunnan, A. J. (Ed.). (1998). *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. New York, NY: McGraw-Hill.

Laversuch, I. M. (2008). Putting Germany's language tests to the test: An examination of the development, implementation and efficacy of using language proficiency tests to mediate German citizenship. *Language Planning*, *9*(3), 282–298.

Linacre, J. M. (Ed.). (2011). *A user's guide to Winsteps (program manual 3.72.0)*. Available from http://www.winsteps.com/a/winsteps.pdf.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

McNamara, T., & Shohamy, E. (2008). Viewpoint: Language tests and human rights. *International Journal of Applied Linguistics*, *18*(1), 89–95.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Milani, T., M. (2008). Language testing and citizenship: A language ideological debate in Sweden. *Language in Society*, *37*(1), 27–59.

Mollering, M. (2009). Citizenship testing and linguistic integration in Australia and Germany. *Zeitschrift fur interkulturellen Fremdsprachenunterricht*, *14*(2), 13–27.

Naturalization Delays: Causes, Consequences And Solutions: Hearing Before the H. Comm. on the Judiciary, Subcomm. on Immigration, Citizenship, Refugees, Border Security, And Int'l Law, 110th Cong. (2008) (statement of Emilio T. Gonzalez, Dir., U.S. Citizenship And Immigration Servs.).

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3–9.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.

Park, G.-P. (2005). Differential item functioning: Current issues. *Language Research*, *41*(4), 949–962.

Park, G.-P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly*, *42*(1), 115–123.

Piller, I. (2001). Naturalization language testing and its basis in ideologies of national identity and citizenship. *International Journal of Bilingualism*, *5*, 259–277.

Preston, J. (2007, September 28). Tough questions for a new test: What does "American" mean? *The New York Times,* pp. 1, A24.

Rothstein, E. (2006, January 23). Refining the tests that confer citizenship. *The New York Times*, pp. B1–B2.

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice*, *22*(1), 7–15.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Vol. 9, pp. 15–19). Cambridge, UK: Cambridge University Press.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Essex, England: Pearson Education Limited.

Shohamy, E., & McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly*, *6*(1), 1–5.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477–481.

Terrill, L. (2000). *Civics education for adult English language learners*. Washington, DC: Center for Adult English Language Acquisition, Center for Applied Linguistics. Retrieved from http://www.cal.org/caela/esl_resources/digests/civics.html.

Tomson, E. (2007, May 10). New citizens, tougher tests. *Pioneer Press,* p. A1.

United States Citizenship and Immigration Services. (2006). *Building an Americanization movement for the twenty-first century: A report to the President of the Untied States for the Task Force on New Americans*. Available from http://www.uscis.gov/files/nativedocuments/M-708.pdf.

United States Citizenship and Immigration Services. (2007). *USCIS Montly, October 2007*. Retrieved from http://www.uscis.gov/files/nativedocuments/USCIS_Monthly_Oct07.pdf.

United States Citizenship and Immigration Services. (2008). *USCIS New Naturalization Test Fact Sheet 3*. Retrieved from http://www.uscis.gov/files/nativedocuments/New_Test_Fact_Sheet.pdf.

United States Citizenship and Immigration Services. (2009a). *Learn about the United States: Quick civics lessons for the Naturalization Test*. Retrieved from http://www.uscis.gov/files/nativedocuments/M-638_red.pdf.

United States Citizenship and Immigration Services. (2009b). *Pathway to U.S. citizenship*. Available from http://www.uscis.gov/USCIS/Office%20of%20Citizenship/Citizenship%20Resource%20Center%20Site/Publications/PDFs/M-685.pdf.

United States Citizenship and Immigration Services. (2009c). *USCIS annual report for 2008*. Retrieved from http://www.uscis.gov/USCIS/Resources/Reports/uscis-annual-report-2008.pdf.

United States Citizenship and Immigration Services. (2010). *Applications for immigration benefits and naturalization monthly statistical reports*. Retrieved from http://www.uscis.gov/portal/site/uscis/.

United States Department of Homeland Security. (2010). *Yearbook of immigration statistics: 2009*. Washington, DC: U.S. Department of Homeland Security, Office of Immigration Statistics. Available from http://www.dhs.gov/xlibrary/assets/statistics/yearbook/2009/ois_yb_2009.pdf.

White, P. (2008). Immigrants into citizens. *The Political Quarterly*, *79*(2), 221–231.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, *45*(4).

Winn, M. (2005). Collecting target discourse: The case of the US naturalization interviews. In M. Long (Ed.), *Second language needs analysis* (pp. 265–304). Cambridge, UK: Cambridge University Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available from http://www.educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf.

APPENDIX A

Mock U.S. Citizenship Test (Test Form A)

Age: _____ Gender: ☐ Male ☐ Female Current profession/job: _____	Highest level of education: ☐ GED ☐ High School Diploma ☐ Currently in college ☐ BA/BS ☐ JD ☐ MA/MS ☐ Ph.D. ☐ Other: _____	Are you a U.S. Citizen? ☐ Yes ☐ No If yes, since when? ☐ Birth ☐ Other:_____ If no, explain: _____

DIRECTIONS: This is a mock U.S. citizenship test. These are real questions from the U.S. Naturalization test. We are going to use this data to investigate how difficult the test questions are.

Questions	Correct Answers	0/1
1. What is the supreme law of the land?	• the Constitution	
2. What did the Declaration of Independence do?	• announced our independence (from Great Britain) • declared our independence (from Great Britain) • said that the United States is free (from Great Britain)	
3. Who is in charge of the executive branch?	• the President	
4. The House of Representatives has how many voting members?	• four hundred thirty-five (435)	
5. In what month do we vote for President?	• November	
6. Who signs bills to become laws?	• the President	
7. How many justices are on the Supreme Court?	• nine (9)	
8. What are the two major political parties in the United States?	• Democratic and Republican	
9. What are two rights of everyone living in the United States?	• freedom of expression • freedom of speech • freedom of assembly • freedom to petition the government • freedom of worship • the right to bear arms	
10. When must all men register for the Selective Service?	• at age eighteen (18) • between eighteen (18) and twenty-six (26)	
11. When was the Declaration of Independence adopted?	• July 4, 1776	
12. Who was the first President?	• (George) Washington	
13. What did the Emancipation Proclamation do?	• freed the slaves • freed slaves in the Confederacy • freed slaves in the Confederate states • freed slaves in most Southern states	

(*Continued*)

Questions	Correct Answers	0/1
14. Before he was President, Eisenhower was a general. What war was he in?	• World War II	
15. Name one of the two longest rivers in the United States.	• Missouri (River) • Mississippi (River)	
16. What is the capital of the United States?	• Washington, D.C.	
17. The idea of self-government is in the first three words of the Constitution. What are these words?	• We the People	
18. What is the "rule of law"?	• Everyone must follow the law. • Leaders must obey the law. • Government must obey the law. • No one is above the law.	
19. Why does the flag have 13 stripes?	• because there were 13 original colonies • because the stripes represent the original colonies	
20. When was the Constitution written?	• 1787	

APPENDIX B

The 100 Civics Test Questions by Difficulty (from the Most Difficult to the Easiest)

Item		N	*Observed Raw Score*	*Item Difficulty Measure (in Logits)*	*SE*
7.	How many amendments does the Constitution have?	69	5	4.56	.51
66.	When was the Constitution written?	414	39	4.10	.19
67.	The Federalist Papers supported the passage of the U.S. Constitution. Name one of the writers.	70	9	3.54	.41
21.	The House of Representatives has how many voting members?	69	9	3.51	.40
40.	Who is the Chief Justice of the United States now?	68	12	3.07	.39
23.	Name your U.S. Representative.	69	15	2.97	.34
19.	We elect a U.S. Senator for how many years?	70	15	2.62	.36
41.	Under our Constitution, some powers belong to the federal government. What is one power of the federal government?	69	19	2.52	.33
12.	What is the "rule of law"?	414	103	2.51	.14
79.	Who was President during World War I?	70	18	2.39	.32
68.	What is one thing Benjamin Franklin is famous for?	70	17	2.38	.34
22.	We elect a U.S. Representative for how many years?	68	18	2.30	.33
9.	What are two rights in the Declaration of Independence?	68	20	2.09	.32
77.	What did Susan B. Anthony do?	68	20	2.09	.32
39.	How many justices are on the Supreme Court?	69	24	1.78	.30
42.	Under our Constitution, some powers belong to the states. What is one power of the states?	70	25	1.72	.30
31.	If both the President and the Vice President can no longer serve, who becomes President?	70	24	1.65	.31

(*Continued*)

APPENDIX B
(*Continued*)

Item		N	*Observed Raw Score*	*Item Difficulty Measure (in Logits)*	*SE*
18.	How many U.S. Senators are there?	70	26	1.63	.30
1.	What is the supreme law of the land?	69	26	1.60	.30
83.	During the Cold War, what was the main concern of the United States?	68	26	1.50	.30
47.	What is the name of the Speaker of the House of Representatives now?	69	30	1.43	.31
20.	Who is one of your state's U.S. Senators now?[a]	68	30	1.37	.30
82.	Before he was President, Eisenhower was a general. What war was he in?	69	30	1.24	.29
2.	What does the Constitution do?	68	30	1.14	.30
65.	What happened at the Constitutional Convention?	69	34	1.05	.31
71.	What territory did the United States buy from France in 1803?	68	31	1.05	.30
14.	What stops one branch of government from becoming too powerful?	69	34	1.02	.30
37.	What does the judicial branch do?	70	32	.93	.29
36.	What are two Cabinet-level positions?	70	35	.84	.29
98.	What is the name of the national anthem?	70	35	.84	.29
50.	Name one right only for United States citizens.	68	36	.82	.30
51.	What are two rights of everyone living in the United States?	69	36	.73	.29
62.	Who wrote the Declaration of Independence?	68	37	.73	.30
63.	When was the Declaration of Independence adopted?	69	36	.73	.29
3.	The idea of self-government is in the first three words of the Constitution. What are these words?	414	216	.72	.12
5.	What do we call the first ten amendments to the Constitution?	70	37	.67	.29
84.	What movement tried to end racial discrimination?	69	38	.66	.31
76.	What did the Emancipation Proclamation do?	69	37	.64	.29
53.	What is one promise you make when you become a United States citizen?	69	40	.46	.32
56.	When is the last day you can send in federal income tax forms?[a]	68	40	.45	.31
6.	What is one right or freedom from the First Amendment?[a]	70	38	.43	.29
91	Name one U.S. territory.	70	40	.41	.30
29.	What is the name of the Vice President of the United States now?	69	42	.26	.32
16.	Who makes federal laws?	68	41	.18	.30
34.	Who vetoes bills?	68	41	.18	.30
43.	Who is the Governor of your state now?	70	41	.17	.29
32.	Who is the Commander in Chief of the military?	68	43	.16	.31
11.	What is the economic system in the United States?[a]	70	43	.15	.30
17.	What are the two parts of the U.S. Congress?[a]	69	43	.15	.32
61.	Why did the colonists fight the British?	70	42	.09	.29

(*Continued*)

APPENDIX B
(*Continued*)

Item		N	*Observed Raw Score*	*Item Difficulty Measure (in Logits)*	*SE*
81.	Who did the United States fight in World War II?	68	44	.06	.31
4.	What is an amendment?	69	44	.05	.32
35.	What does the President's Cabinet do?	69	44	.05	.32
13.	Name one branch or part of the government.[a]	70	43	.01	.29
57.	When must all men register for the Selective Service?	69	46	−.15	.30
55.	What are two ways that Americans can participate in their democracy?	70	45	−.16	.29
52.	What do we show loyalty to when we say the Pledge of Allegiance?	68	45	−.18	.30
87.	Name one American Indian tribe in the United States.	68	47	−.24	.32
49.	What is one responsibility that is only for United States citizens?[a]	70	46	−.25	.30
64.	There were 13 original states. Name three.	68	46	−.27	.31
24.	Who does a U.S. Senator represent?	70	48	−.32	.31
80.	Who was President during the Great Depression and World War II?	70	47	−.34	.30
69.	Who is the "Father of Our Country"?	68	48	−.35	.33
48.	There are four amendments to the Constitution about who can vote. Describe one of them.	70	49	−.42	.31
58.	What is one reason colonists came to America?	68	48	−.46	.31
15.	Who is in charge of the executive branch?	69	50	−.53	.31
8.	What did the Declaration of Independence do?	69	51	−.63	.32
33.	Who signs bills to become laws?	69	52	−.73	.32
96.	Why does the flag have 13 stripes?	414	313	−.81	.14
27.	In what month do we vote for President?[a]	69	53	−.84	.33
38.	What is the highest court in the United States?	68	52	−.97	.36
75.	What was one important thing that Abraham Lincoln did?[a]	68	55	−1.17	.36
88.	Name one of the two longest rivers in the United States.	69	57	−1.30	.36
85.	What did Martin Luther King, Jr. do?[a]	70	58	−1.47	.35
46.	What is the political party of the President now?	68	57	−1.50	.38
72.	Name one war fought by the United States in the 1800s.	69	57	−1.50	.38
73.	Name the U.S. war between the North and the South.	70	59	−1.57	.38
25.	Why do some states have more Representatives than other states?	70	59	−1.60	.37
74.	Name one problem that led to the Civil War.	70	60	−1.74	.38
78.	Name one war fought by the United States in the 1900s.[a]	69	59	−1.80	.40
54.	How old do citizens have to be to vote for President?[a]	70	61	−1.88	.40
93.	Name one state that borders Mexico.	68	60	−1.93	.42

(*Continued*)

APPENDIX B
(*Continued*)

Item		N	*Observed Raw Score*	*Item Difficulty Measure (in Logits)*	*SE*
70.	Who was the first President?[a]	69	62	−2.05	.43
10.	What is freedom of religion?	69	61	−2.15	.43
90.	What ocean is on the East Coast of the United States?	69	61	−2.15	.43
92.	Name one state that borders Canada.	70	63	−2.22	.43
30.	If the President can no longer serve, who becomes President?	70	63	−2.24	.44
45.	What are the two major political parties in the United States?[a]	69	63	−2.24	.45
60.	What group of people was taken to America and sold as slaves?	70	63	−2.24	.44
86.	What major event happened on September 11, 2001, in the United States?	70	63	−2.24	.44
26.	We elect a President for how many years?	68	62	−2.33	.47
89.	What ocean is on the West Coast of the United States?	68	62	−2.37	.48
44.	What is the capital of your state?[a]	68	63	−2.57	.51
95.	Where is the Statue of Liberty?[a]	68	63	−2.62	.52
99.	When do we celebrate Independence Day?[a]	70	65	−2.65	.50
100.	Name two national U.S. holidays.	68	64	−2.84	.55
94.	What is the capital of the United States?[a]	69	66	−3.05	.61
59.	Who lived in America before the Europeans arrived?	69	68	−4.67	1.04
97.	Why does the flag have 50 stars?[a]	69	68	−4.67	1.04
28.	What is the name of the President of the United States now?[a]	68	68	NA[b]	NA

[a]Items that are preferentially asked to individuals older than 65 and who have been a legal, permanent resident of the United States for 20 or more years.

[b]Difficulty could not be calculated through Winsteps for Item 28 because everyone who received the item answered it correctly. Its item facility, however, denotes that it was the easiest item on the test.