FLIGHT/VIDAS

User Manual

Fort Lauderdale, Florida

Enalan Communications, Inc.

2015

FLIGHT/VIDAS User Manual

Enalan Communications, Inc., Fort Lauderdale Florida.

For the most up-to-date information on the FLIGHT/VIDAS project and to learn how to access assessment materials, go to:

www.flightvidas.org

Acknowledgments

The development of a new measure of health literacy required a coordinated team effort without which the project would not have been successful. The contributions of the FLIGHT/VIDAS team members are gratefully acknowledged:

Amarilis Acevedo, PhD

Drenna Waldrop-Valverde, PhD

Sara Czaja, PhD

David Loewenstein, PhD

Ana-Maria Homs, PsyD

Rosemary Davenport, RN, MSN, ARNP

Robin Jacobs, PhD

Joshua Caballero, PharmD

Lillian Valiente

Jamie Mazzurco, DO, MPH

The FLIGHT/VIDAS project was supported by a grant from the National Heart Lung and Blood Institute (R01HL096578) to Dr. Ownby

Contents

1 Introduction

Ongoing attention to health literacy as a factor in health has highlighted a number of issues about the concept. While a number of influential conceptual definitions of health literacy have been advanced such as Healthy People 2010 and its follow-up Health People 2020, the original IOM report definition, and the Calgary Charter, how make these definitions operational in creating assessment tools and interventions to improve health literacy has not been clear. Statements defining health literacy as "the ability to obtain and use health information" are a critical first step in defining health literacy, but continuing efforts to measure how effectively patients can do the things specified in a definition are important as well.

Coincident with ongoing efforts to understand the impact of health literacy on health, a cadre of dedicated educators and clinicians has worked to fill the void in knowledge by creating patient education materials that are designed to provide critically important information in ways that patients can understand and use.

All these efforts have depended on the ability to define and measure health literacy, for if a person's level of health literacy cannot be reliably determined, research on the effects of health literacy on their health or the impact of an intervention on health literacy cannot be evaluated.

The limitations of existing measures of health literacy have been extensively discussed elsewhere and will only be summarized here. The most influential measure of health literacy has probably been the Test of Functional Health Literacy in Adults (TOFHLA) and its shorter version,

the S-TOFHLA [refs]. While research with the TOFHLA and S-TOFHLA has been critically important, both have been criticized for tapping a limited range of content. Their response format may confound age-related in cognitive abilities with functional health literacy. Further, while published cutoff scores for the TOFHLA defining "adequate," "marginal," and "inadequate" health literacy have often been used in research and to validate other measures of health literacy (the S-TOFHLA and the NVS as well as others, they are not anchored to external criteria to establish their validity (personal communication, JoAnn Nurss, date). A final issue with the use of the TOFHLA and many other measures is the fact many people achieve very high scores on it but still have difficulty in understanding health information. This issue, called a "ceiling effect" by those who study test development, mean that the time patients spend answering questions that are too easy is wasted, while the actual information provided by the last few most difficult items of the test is minimal.

The Rapid Estimate of Adult Literacy in Medicine (REALM) has also been extensively used. Its score categories are referenced to a well-known reading test, but it relies only on a person's ability to pronounce health-related words to assess health literacy. This means that a person's ability to understand or apply information isn't evaluated. The REALM also suffers from ceiling effects, meaning that while it can detect individuals with very low health literacy, it provides little information about many persons' health literacy.

A discussion of health literacy assessment would not be complete without mentioning the 2003? National Assessment of Adult Literacy, or NAAL (White & Dillow, 2005). It provided valuable information on how people in the general population perform on health literacy tasks [insert ref from J geront ed], but because of concerns about item security it is not available for widespread use. One additional contribution of the NAAL was to provide operational criteria for levels of health literacy based on the earlier strategy of defining levels of literacy used in the National Adult Literacy Survey [ref].

[Insert description of levels of literacy from NALS and NAAL and application to F/V].

It was in this context that in 2010 we set out to develop a new measure of health literacy. We chose to address perceived weaknesses of existing measures, to make a new measure functional in diverse groups (Spanish and English, older and younger, lower and higher levels of education), and one that could be computer administered and scored. The remainder of this manual describes the process by which we developed the new measure of health literacy and how we have worked to make it valid and useful to others.

Overview

The intent of this manual is give those interested in the assessment of health literacy the information needed to make informed decisions about how to use the scales developed in the FLIGHT/VIDAS project. Although we've published information about its development, reliability, and validity elsewhere (Ownby et al., 2013; Ownby, Acevedo, Waldrop-Valverde, Jacobs, & Caballero, 2014; Ownby, Acevedo, Jacobs, Caballero, & Waldrop-Valverde, 2014b; Ownby, Acevedo, Jacobs, Caballero, & Waldrop-Valverde, 2014a; Ownby, Acevedo, Goodman, Caballero, & Waldrop-Valverde, 2015), we hope that this manual will bring all this information together in one place. In addition, since some of the publications referenced were based on interim data analyses, those presented here should be viewed as authoritative as they are based on our final dataset created at the end of the project.

In addition, this manual presents two previously-unreported scales that have been created primarily in response to feedback and questions from potential users. While our primary focus in creating FLIGHT/VIDAS (F/V) was to develop a computer-administered and –scored measure of health literacy that could be used to tailor computer-delivered patient education interventions, it has become clear that while computer administration may be helpful in many contexts, many potential F/V users still work in

situations where computers and Internet access are cumbersome or simply not available. As explained in Chapter 3, Development, in initial item development we created more than 225 items, and in the final development phase we tested 98 of them. As a number of these items could be transferred to paper and pencil format, we were able to develop a 20-item scale that can be used to assess persons' health literacy. Further, in response to many questions about the use of F/V to quickly identify individuals who are likely to have limited health literacy skills, we have created a 10-item screening scale that will require only a few minutes to administer and score. We present information on the reliability and validity of these scales in Chapter 4.

The overall plan of this manual is first to explain in this introduction why we created F/V, then to go on and detail the conceptual model that underlies F/V in chapter 2, The ASK Model. Chapter 3 reviews in greater detail the steps taken to develop F/V items include preliminary psychometric assessments, while Chapter 4 provides more detail on the measure include final estimated of reliability and validity. The Appendix presents normative data across language, age, and education groups.

Why another health literacy measure?

Health literacy, defined as an individual's ability to obtain health-related information and use it to make decisions, (Nielsen-Bohlman, Panzer, & Kindig, 2004) is increasingly recognized as an important factor in patient health. Several reviews show that individuals' health literacy is related to their health status, function, and use of services (Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011; Dewalt, Berkman, Sheridan, Lohr, & Pignone, 2004) and it has even been related to increased risk of mortality (Bostock & Steptoe, 2012; Sudore et al., 2006). The existence of effective interventions to improve health literacy (Ownby, Waldrop-Valverde, Jacobs, & Caballero, 2012; Sheridan et al., 2011) highlights the possibility that improving it may be a strategy for improving health outcomes and addressing race- and ethnicity-related health disparities (Osborn et al.,

2011; Waldrop-Valverde et al., 2010).

Commonly used measures of health literacy include the Test of Functional Health Literacy in Adults, or TOFHLA (Parker, Baker, Williams, & Nurss, 1995), the Rapid Estimate of Adult Literacy in Medicine, or REALM (Murphy, Davis, Long, Jackson, & Decker, 1993), and the Newest Vital Sign (Weiss et al., 2005). Each measure has strengths and weaknesses. The TOFHLA, for example, assesses a patients' ability to understand what they read as well as their numeracy skills. A limitation of the TOFHLA, however, is the requirement that the clinician administering it be trained and the time required for the clinician to individually administer and score it, typically at least 30 minutes. The time required for administration thus limits its use in clinical and research settings. A shorter version, the S-TOFHLA, is available (Baker, Williams, Parker, Gazmararian, & Nurss, 1999) but suffers from ceiling effects (many people achieve high scores) that limit its use in research since a limited range of scores affects the ability to detect its relations to other variables.

The REALM also must be administered, scored and interpreted by a trained clinician. This measure only assesses health literacy as patients' ability to correctly pronounce a series of health related words (e.g., anatomical terms and the names of diseases and condition) and thus does not directly assess their ability to understand what they read. The REALM does not assess numeracy skills, consistently shown to be an important aspect of health literacy. The Newest Vital Sign only assesses patients' comprehension of a single food label, and thus only taps a very narrow range of skills. It may have limited use except perhaps for the purpose of detecting whether a patient has poor reading comprehension skills.

A number of years ago, David Baker noted the limitations of existing measures of health literacy (2006). Problems encountered in assessing health literacy are summarized more recently by Pleasant and McKinney (Pleasant & McKinney, 2011) and in another review by (Jordan, Osborne, & Buchbinder, 2011). Previously available measures have been criticized for assessing a limited set of skills and for their development using patients

drawn from single racial, ethnic, age, or socioeconomic groups. Other criticisms have noted the limited content and face validity of the measures and limited demonstrations of the measures' construct validity (Jordan et al., 2011). Further, although both Spanish and English versions of several measures are available, they were not developed using psychometric procedures that establish their equivalence across languages making comparisons difficult.

An issue limiting the usefulness of the TOFHLA is the response format it uses in evaluating reading comprehension. The TOFHLA uses the cloze procedure (Ackerman, Beier, & Bowen, 2000) to assess reading comprehension. In this approach, comprehension is tested by asking the person assessed to supply a word missing in a sentence (e.g., "The sky is _____"). This approach may create items that are differentially more difficult for older persons. Cloze procedure performance has been related to information processing speed and verbal fluency, reduced in older persons (Ackerman & Cianciolo, 2000), and our own data indicate that age-related differential item functioning exists on a significant number of items from the reading comprehension subtest of the TOFHLA (Ownby & Waldrop-Valverde, 2013). Item DIF occurs when individuals from different groups such as men or women or racial groups, who have the same level of ability, have different probabilities of answering an item correctly. The empirical finding of this kind of difference is usually interpreted as evidence that some factor besides the person's actual ability affects their performance, perhaps cultural, linguistic, or some other bias (Embretson & Reise, 2000). The finding of age-related DIF on the TOFHLA reading comprehension subtest suggests that other item formats (e.g., multiple choice questions) may be more appropriate when assessing health literacy in older persons.

Almost all existing paper-and-pencil measures require hand scoring, making them time- and effort-intensive. Clearly, a computer-administered and scored measure of health literacy would make is assessment more accessible in both clinical and research settings by reducing demands on clinician or researcher time while better standardizing the measure's administration. Integration of such a measure into an electronic health record

might allow for inclusion of health literacy scores into patients' health records. This would transmit information about patients' level of health literacy directly to treating clinicians, allowing them to better understand patients' information needs. The automated assessment of health literacy might also allow for the automated tailoring of disease-related information, a strategy previously shown to be effective in influencing patient behavior (Noar, Benac, & Harris, 2007; Ownby, Hertzog, & Czaja, 2012b), and which may be effective in reducing health disparities (Jerant, Sohler, Fiscella, Franks, & Franks, 2011).

Pleasant et al.(2011) argue that new measures of health literacy should be multidimensional and assess health literacy as a latent construct. A multidimensional approach would recognize that functional health literacy comprises a number of distinct skills or abilities, such as reading, listening, and performing quantitative operations (Nielsen-Bohlman et al., 2004). Evaluation of latent constructs is frequently used in psychological assessment to study an ability or trait that cannot be directly measured. Multiple test items believed to be related are administered and then what they have in common is statistically extracted, usually with factor analysis. Many item response theory (IRT) models approach the measurement of abilities as latent constructs and have been used to develop assessments of health literacy (Hahn, Choi, Griffith, Yost, & Baker, 2011; Lee, Stucky, Lee, Rozier, & Bender, 2010; Yost et al., 2009). Pleasant et al.(2011) also suggest that assessments should recognize that measures are most likely to be accurate when they are similar to the context in which the actual behavior occurs. Assessment using a video simulation of a clinical encounter, for example, may be more accurate than asking for responses to written questions.

Jordan et al. (2011) reviewed existing health literacy measures and find many of them lacking in important measurement characteristics. These authors note the great variability in the content assessed by measures and the lack of a coherent conceptual model underlying them. Although measures such as the TOFHLA and the REALM provide descriptive score categories such as "adequate" or "inadequate" to assist in interpretation,

the rationale for them is not clear. Jordan et al. also report limitations in the construct validity of measures, noting that the correlations of the measures with other measures of health literacy and reading are quite variable. These findings imply that different measures may actually evaluate different abilities and skills, calling into question their typical interpretation of measures of the same construct, an issue also raised by Haun et al.(Haun, Luther, Dodd, & Donaldson, 2012) Finally, Jordan et al. assess the feasibility of actually using the measures they review, noting that the need for time, individual administration, and scoring is a substantial limitation that may limit their use.

Several researchers have addressed these limitations in developing new assessments of health literacy. Hahn et al.(Hahn et al., 2011; Yost et al., 2009) created a health literacy assessment using a touch screen computer format they call the "Health Literacy Assessment Using Talking Touchscreen Technology," or Health LiTT. Their measure, developed in Spanish and English (Yost et al., 2009) allows for automated administration and scoring and was developed using IRT methods. Data on this measure's development in Spanish is limited, however, and the sample of Spanish-speaking adults used in development efforts was not clearly characterized as to bilingualism or linguistic preference. How the Spanish-speaking participants were chosen to be tested in Spanish or English is not clearly described, nor is their level of acculturation. The measure is not equivalent in Spanish and English as test stimuli differ in the two versions of the measure, limiting its usefulness in research, and both Spanish and English-speaking groups were patients in primary care with low levels of educational attainment. Although clearly a relevant population, the development of the measure with persons likely to have a limited range of ability may indicate that the measure will not function well in assessment of persons with higher levels of ability. The importance of understanding patients' English competence when information is delivered to non-native speakers has been shown in studies by several investigators.(Aguirre, Ebrahim, & Shea, 2005; Zun, Sadoun, & Downey, 2006) The measure may thus replicate the commonly-observed ceiling effect from the TOFHLA. The Health LiTT measure is not based on a coherent theory or conceptual

model of health literacy, although the authors link its development to an existing descriptive definition of health literacy (Yost et al., 2009). Further, the reading comprehension section of this measure continues to rely on the cloze procedure which, as noted above, may result in items that are differentially more difficult for older participants. Finally, in a 2011 publication the authors report that the Health LiTT will be available through The Assessment Center (http://www.assessmentcenter.net), a free online resource that allows investigators to access a standardized set of measures for use in research. At the time of this writing, however, it is not available on this site. The actual availability of this measure for use is not clear.

Lee et al.(2010) developed an instrument based on the REALM, selecting items based on analyses of DIF between Spanish- and English-speaking patients seen in a primary care clinic. As with other measures of health literacy, this measure continues to tap a narrow range of content and has limited demonstrations of its relation to other measures that might help establish its validity. It does not assess numeracy at all, and only provides a limited assessment of comprehension. It thus suffers from the other of the limitations others have criticized, including sampling a limited range of content, uncertain relation to actual health behaviors, and development on a small population of clinic patients.

A group at the Research Triangle Institute led by Lauren McCormack has also developed a new measure of health literacy, the Health Literacy Skills Instrument (McCormack et al., 2010). This measure was developed using a rigorous psychometric approach and can be computer administered and scored. The development population was broader than that used to create most other measures (research volunteers vs. clinic patients in many other studies), and the development population was large (several thousand individuals). Analyses of its validity have been presented (McCormack et al., 2010; Bann, McCormack, Berkman, & Squiers, 2012). The manual for this measure, however, does not provide directions for administering the measure either in person or by computer, raising questions about whether the measure would be reliable without a standard approach to administration. The authors suggest that the measure can be administered as a

paper and pencil test, but it uses an audio recording to assess listening comprehension and requires access to several web pages to answer two questions, thus raising questions about how this might be possible. Such an administration would again not be standardized, raising questions about the validity of this form. Finally, the HLSI is not available in Spanish so that it may have limited usefulness with the most rapidly growing minority population in the United States.

This measure thus addresses many of the issues previously raised in critical evaluations of measures of health literacy. Its psychometric characteristics have been established, and it taps a wider range of content than other measures. It includes two items tapping listening comprehension, although it appears likely that these items cannot be used as a separate scale. The measure can be computer administered and scored, but no standard format for this administration is provided raising questions about the reliability of test scores that might result from diverse approaches to administration. It can be administered by computer, but as described it appears likely that this administration format requires the use of a computer mouse to answer questions and to navigate hyperlinks. Given the difficulty many elders and others with little computer experience have in using a computer mouse (and especially the fine psychomotor skills required to click the small dots used by this instrument to a preferred answer), it is likely that the format of administration as developed may place the very groups in which health literacy is most important at a disadvantage when responding to its questions. Although it includes items that mimic health-related calculators that might be found on the Internet, they do not actually assess Internet information search. As the web pages are on an external server, the ability to administer them requires and Internet connection and limits the ability of users to administer the measure by paper and pencil. Finally, the measure is brief, does not have separate subscales for skills such as reading, numeracy, and listening, and is not available in Spanish.

As summarized by Pleasant and McKinney (2011) and suggested by Jordan et al. (2011) workers in the field have argued that new measures of

health literacy should be developed that broaden the range of content assessed, are based on diverse groups, and have better-demonstrated psychometric characteristics. We have developed a new measure of health literacy that addresses these criticisms. It samples a wide range of content chosen from the domains listed in the 2004 IOM report (Nielsen-Bohlman et al., 2004) on the competencies needed for adequate health literacy. It is based on a coherent conceptual model of health literacy (further discussed in Chapter 2) that has been developed based on our own and others' research. It includes items that assess prose, document, and quantitative literacies in each of the domains. It has been developed through a rigorous two-stage process in which items have been pilot tested, assessed for equivalence in both Spanish and English as well as in younger and older individuals, and has been subjected to assessments of construct and concurrent validity. It assesses not only reading and quantitative skills but also uses video simulations of healthcare related encounters to assess listening comprehension and to provide test stimuli that bear a close relation to the actual situations in which health literacy skills might be applied. By asking questions that assess expressive writing skills (by asking, for example, where certain kinds of information would be placed in a form) the measure indirectly assesses expressive written language skills. The purpose of this manual is to describe the initial development and testing of this new measure and provide preliminary data on its validity and reliability. The new measure utilizes a broad range of item formats and contents, includes listening comprehension, and has been developed in Spanish and English. The English project has been named *Fostering Literacy for Good Health Today* (FLIGHT) and the related Spanish project has been named *Vive Desarollando Amplia Salud* (VIDAS).

The measure that eventually became FLIGHT/VIDAS was first conceptualized in 2005 as a result of our experiences in researching medication adherence and especially in creating an intervention to improve adherence to medication in elderly patients diagnosed with memory problems. One of the reviewers of the grant application for this study fortuitously recommended that we address our participants' health literacy, a suggestion that we readily adopted. In pilot testing the intervention and the

assessment battery that included the Test of Functional Health Literacy in Adults, or TOFHLA (Parker et al., 1995), it became clear that our participants varied widely in how well they understood how to take their medications and in level of health literacy. Two things were clear from the pilot testing: (1) health literacy was an important factor that appeared related to medication adherence, and (2) our potential participants had difficulty understanding the cloze format of the TOFHLA. We addressed the first problem by creating a semi-automated tailored information intervention that provided information to participants at two levels of difficulty (Ownby, 2005). We addressed the second problem by creating materials that slowly introduced our elderly participants to the cloze format. Results of this study, although based on a very small group, suggested that the patients who received the tailored information intervention had better adherence as measured by electronic pill bottle caps (Ownby, Hertzog, & Czaja, 2012a).

A third problem emerged in evaluating the participants in this study. Even when our participants were able to understand the respond meaningfully to the items of the TOFHLA, the process of administering it was time consuming. This is a problem noted by a number of other authors, and has generated responses based in two general strategies. Some have worked to develop shorter measures of health literacy, such as the Newest Vital Sign (Weiss et al., 2005). Briefer measures can be administered and scored more quickly, but at the cost of a narrower range of health literacy skills assessed. This may limit the reliability and validity of the measure. An alternative strategy has been to develop measures that can be computer administered and scored. This is the approach we have taken in creating FLIGHT/VIDAS. As described more extensively in Chapter 3, in developing F/V we explicitly worked with a broad range of health-related content and item formats while creating a measure that could be administered and scored with minimal intervention from a clinician. A key focus of the project has always been to create a measure that could feed forward information about a person's level of health literacy to clinicians and to downstream interventions to improve their understanding of health conditions.

Summary

In this introduction, the issues that led us to undertake the development of a new measure of health literacy have been reviewed. Chapter 2 explains the conceptual model on which the new measure is based, and Chapter 3 discusses the development process. Chapter 4 presents information on the reliability and validity of the F/V. Finally, in Chapter 5 we discuss ways to use the F/v scales in practice and research.

2 The ASK Model

Overview

Health literacy, defined as an individuals' ability to obtain and use health information to make choices about health care, is related in many way to their health (Berkman et al., 2011; Dewalt, Berkman, Sheridan, Lohr, & Pignone, 2004), health status, service utilization, self-care behaviors, and even risk for death (Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011; Dewalt et al., 2004). It has also been tied to race- and ethnicity-related disparities (Osborn, Paasche-Orlow, Davis, & Wolf, 2007; Osborn et al., 2011; Paasche-Orlow & Wolf, 2010). Even with ongoing research, though, important questions about health literacy remain. One key question is how health literacy can be operationally defined so it can be measured and inter-ventions to improve can be developed.

This problem arises from the diverse ways health literacy has been defined and the varying content and format of widely-used health literacy measures. For example, in most studies, health literacy has implicitly defined as a person's performance on a test of health literacy. Since each of the most commonly-used health literacy tests measures it in a different way, it is difficult to know exactly what these studies mean. In some cases, health literacy has been assessed as a person's ability to demonstrate that he or she understands health-related in-formation. This strategy underlies the Reading subtest of Test of Functional Health Literacy in Adults, or TOFHLA (Parker, Baker, Williams, & Nurss, 1995a). In other studies, health liter-acy is defined as a person's ability to correctly pronounce health-related words, as the Rapid Estimate of Adult Literacy in Medicine,

or REALM (Murphy, Davis, Long, Jackson, & Decker, 1993) or to identify synonyms of health-related words as in the Short Assessment of Health Literacy for Spanish Adults, or SAHLSA (Lee, Bender, Ruiz, & Cho, 2006)). Each of these strategies assesses something related to health literacy, but their diversity leaves open the question of what each has in common with the "social construct of health literacy." (Pleasant, McKinney, & Rikard, 2011)

Each measure samples different content, uses different response formats, and has been developed on different populations (Pleasant & McKinney, 2011). The TOFHLA, for example, evaluates reading comprehension by asking a person to supply words eliminated from the text (the cloze procedure), while its numeracy scale asks that he or she explain how to take medications. The REALM requires the person to correctly read aloud words related to healthcare. The need for a similar measure for Spanish speakers led to the development of the SAHLSA, but the low frequency of orthographically-irregular words in Spanish meant that it was necessary to develop a different response format. The SAHLSA asks the person tested to view a stimulus word on a card and choose which of two other words is most similar in meaning.

Performance on these measures requires basic reading skills and conceptual health knowledge, and several, especially the TOFHLA subtests, also require reasoning, problem solving and numeracy skills (Ownby & Waldrop-Valverde, 2009). The variety of contents and formats, however, suggests that the abilities required for successful performance on each are different (Haun, Luther, Dodd, & Donaldson, 2012; Jordan, Osborne, & Buchbinder, 2011). Griffin et al., for example, showed substantial differences in which patients were identified as having limited health literacy by different measures (2010), and similar findings have been reported in other studies (Haun et al., 2012; Osborn et al., 2007).

Because of these issues, in developing F/V we hypothesized that more clearly establishing the relations of health literacy measures to other factors might provide a better understand-ing of what each measures. We

also hypothesized that a better understanding would also provide a clearer picture of health literacy by defining its component skills. In the study we review in this chapter, based on reviews of empirical research and conceptual models of health literacy, we hypothesized that after taking into account individuals' social and cultural contexts, the variables most relevant to health literacy would be their general cognitive abilities, academic skills, and health-related knowledge.

Social Context of Health Literacy

Studies have shown that health literacy is related to age, race, ethnicity, and socioeco-nomic status. For example, persons older than 65 years of age performed at lower levels on the Health Literacy scale of the National Assessment of Adult Literacy (Kutner, Greenberg, Jin, & Paulsen, 2006). Blacks and Hispanics have also been shown to perform at lower levels on measures of health literacy. Closely intertwined with other demographic characteristics is socioeconomic status, itself related to health literacy (Paasche-Orlow & Wolf, 2007). The finding that English-speaking Hispanics may be at a disadvantage to non-Hispanics when their health literacy is assessed in English (Aguirre, Ebrahim, & Shea, 2005) suggests that preferred language may also be a key characteristic. Gender may also be related to perfor-mance on tests of health literacy (Aguirre et al., 2005). For purposes of measuring the broader context in which individuals exist, we evaluated factors such as race, ethnicity, gender, accul-turation, education, income, and occupational status.

Cognitive Abilities

Understanding the relation of tests of health literacy to basic cognitive abilities (or general intellectual abilities, often assessed by IQ tests) may be especially important since research has shown that both general intellectual abilities and health literacy are related to health (Mottus et al., 2013; Chin et al., 2011; Wolf et al., 2012; Baker et al., 2002). General intellectual ability can be defined as reflecting a person's acquired knowledge

and communication ability (crystallized ability) and capacity to reason and solve novel problems, often referred to as fluid ability (Cattell, 1963; Horn & Cattell, 1966; Carroll, 1993). Baker et al. (Baker et al., 2002) showed that overall performance on the Mini-mental State Exam (MMSE; (Folstein, Folstein, & McHugh, 1975)) was related to S-TOFHLA scores. Levinthal et al. (2008) evaluated the rela-tion of demographic and cognitive variables to performance on the S-TOFHLA and found that both were related. Chin et al. (Chin et al., 2011) found that age, education, basic cognitive abilities, and disease-related knowledge were related to performance on the S-TOHLA and REALM. Others have shown that performance on tests of health literacy is related to various abilities including memory, verbal fluency, reasoning, and general intellectual functioning (Federman, Sano, Wolf, Siu, & Halm, 2009; Wolf et al., 2012; Yost, DeWalt, Lindquist, & Hahn, 2013).

Academic Skills

By its very nature, health literacy is related to academic skills such as reading and mathe-matics (Zarcadoolas, Pleasant, & Greer, 2005; Nutbeam, 2008; Baker, 2006). Academic skills can be distinguished from basic cognitive abilities by their acquisition via formal instruction during schooling. While basic cognitive abilities are thought to be stable over time (Deary, Pattie, & Starr, 2013), academic skills such as reading, writing, and arithmetic are amenable to change through formal interventions well into adult life (Kruidenier, MacArthur, & Wrigley, 2010). While it is important to distinguish between general reading skills and health literacy (Sørensen et al., 2012; Nielsen-Bohlman, Panzer, & Kindig, 2004), the correlation between pa-tients' performance on measures of academic skills and health literacy has also been used to validate measures (Parker, Baker, Williams, & Nurss, 1995b; Bass, III, Wilson, & Griffith, 2003).

Conceptual Knowledge

In addition to social context, cognitive abilities and academic skills,

health knowledge is related both to performance on tests of health literacy and health. Disease-specific knowledge, for example, has been linked to health literacy in diabetes (Rothman et al., 2005), hypertension (Chin et al., 2011; Gazmararian, Williams, Peel, & Baker, 2003), HIV infection (Hicks, Barragan, Franco-Paredes, Williams, & Del, 2006), asthma, and congestive heart failure (Gazmararian et al., 2003). This makes sense, since even persons with excellent reading skills may have difficulty understanding material based on unfamiliar concepts. Again, conceptual health knowledge can be differentiated from basic cognitive abilities because it clearly can be taught. It is also different from academic skills as it only requires a person to demonstrate understanding of specific facts.

Health Literacy as Expertise

Several authors have argued that health literacy is not a distinct ability due to the large contributions of general intellectual ability (especially verbal ability) and reading skills to performance on health literacy measure [refs]. While it is true that general ability and reading skills are important in understanding health literacy, our strategy has been to model health literacy as a form of expertise. Expert performance has been studied in a number of domains, ranging from playing chess to sight reading music. We think this approach is apt because, like other forms of expertise, understanding and applying health information to make healthcare choices requires a combination of general intellectual abilities, specific skills, and task-related knowledge. For example, general intellectual ability is important in becoming proficient at chess, and the specific ability of visual working memory may be particularly important (Reingold et al., 2001). Still, a high IQ and good visual working memory do not make a chess master—skilled performance in playing chess requires knowledge of rules and the capacity to recognize sequences of moves that constitute recognizable patterns (Roring & Charness, 2007). In a similar way, obtaining, understanding, and using health information requires basic cognitive abilities, academic skills in reading and numeracy, and conceptual knowledge related to specific issues, whether disease pathophysiology and treatment or health promotion through diet and exercise.

The ASK Model

The studies reviewed above illustrate the core abilities and skills related to health literacy. Conceptual knowledge as well is a key factor. Basic cognitive abilities, academic skills, and health knowledge are each related to performance on measures of health literacy. Few stud-ies, however, have included measures assessing all of these domains. Further, studies that have used variables from multiple domains have shown that variables from one domain may affect the importance of others, as when inclusion of a measure of general intellectual ability reduces the importance of other factors in health. In creating F/V, we hypothesized that understanding health literacy as the property of an individual required taking into account a person's social context, as emphasized by many authors. This aspect of individuals' health literacy was assessed as their age, gender, race, and socioeconomic status (a composite of education, income, and occupational status).

Given the body of research showing the importance of general intellectual abilities in health literacy and health outcomes, participants in F/V completed a brief battery of cogni-tive measures that allowed us to take crystallized and fluid intellectual ability into account in understanding health literacy. We also recognized, as do others, that academic skills are im-portant in understanding health literacy. In order to take this into account, participants in F/V completed a standardized measure of academic skills that was developed in both English and Spanish, either the Woodcock-Johnson or Woodcock-Muñoz reading comprehension and applied mathematics problems subtests.

Finally, based on our analysis of health literacy as a form of expertise, we believed that conceptual health knowledge (knowledge of specific facts about disease and health) would be important aspects of health literacy. While there are a number of disease-specific knowledge measures, no general health knowledge measure that was both brief and validat-ed was readily available. As part of creating F/V we therefore included a number of conceptual health knowledge questions. We hypothesized that, as

Table 2-1. Final regression analysis for ASK model

	B	SE	Beta	*t*	*p*
Intercept	7.51	2.35	n/a	3.20	0.001
Language	-1.58	0.40	-0.14	-3.96	< 0.001
Gender	0.06	0.33	0.01	0.18	0.86
Race	-0.95	0.43	-0.08	-2.22	0.03
Age	-0.14	0.01	-0.45	-15.86	< 0.001
SES[a]	0.67	0.18	0.12	3.74	< 0.001
Social Status[a]	-0.07	0.08	-0.02	-0.88	0.38
Crystallized[a]	0.10	0.02	0.21	5.12	< 0.001
Fluid[a]	0.00	0.00	0.01	0.37	0.71
Reading[a]	0.09	0.02	0.18	4.44	< 0.001
Knowledge[a]	0.68	0.07	0.34	9.80	< 0.001

[a]SES = socioeconomic status index derived from principal components analysis of education, income, and occupational status; Social Status= Participant self rating on McArthur Foundation subjective social status ladder; Crystallized = crystallized intellectual ability; Fluid = fluid intellectual ability; Reading = Woodcock-Johnson or Woodcock-Muñoz Passage Comprhension subtest; Knowledge = F/V conceptual health knowledge scale.

with general information as assessed in the Wechsler intelligences scales (Wechsler, 1944; Wechsler, 1958), it might be possible to identify a group of health-related facts that, in a scale of 10 to 15 items, be a useful measure of general health knowledge. As is further discussed in Chapters 3 and 4, we were successful in creating such a scale.

Table 2-2. Variability in health literacy related to social context, cognitive ability, reading, and knowledge

Model	R	R^2	Adj R^2	SE	Δ[a]	*p*
1[b]	0.66	0.43	0.43	4.28	0.43	<.001
2[b]	0.76	0.58	0.57	3.70	0.15	<.001
3[b]	0.78	0.60	0.60	3.58	0.03	<.001
4[b]	0.82	0.67	0.67	3.26	0.07	<.001

[a]Δ = change in R^2 value related to each group of variables in Models 1 to 4.

[b]Model 1 = social context variables only; Model 2 adds crystallized and fluid general ability; Model 3 adds reading comprehension; Model 4 adds conceptual health knowledge.

The final conceptual model thus specified that after taking social context into account, the key factors that define an individual's health literacy are basic cognitive abilities, core aca-demic skills, and health-related knowledge (ASK). In a published study, we examined how well this model predicted our participants' performance on the measures of health literacy in-cluded in our assessment battery. All participants completed either the English or Spanish version of the TOFHLA, with English-speaking participants completing the REALM and Spanish-speaking participants completing the SAHLSA. Participants also completed the final 98-item set of F/V questions.

By examining the relation of these groups of variables to our participants' scores on each measure, we sought to evaluate the contribution of each group to health literacy. Results of the final statistical model for F/V are presented in Table 2-1. The contribution of each group to each measure as R-squared values is presented in Table 2-2.

3 Development

Overview

The measure that has become FLIGHT/VIDAS went through an extensive development process that began in 2008 with pilot testing of health literacy items on one of the first commercially-available touch screen computers in a clinic at the University of Miami. Subsequently, with support from the National Institute of Mental Health, the National Heart Lung and Blood Institute, and several other institutes of the National Institutes of Health, we have been able to develop measures of health literacy specifically useful with patients with HIV infection as well as more broadly with those with diverse educational and health status backgrounds.

The process of developing items for FLIGHT/VIDAS began in earnest in 2008 when our group of collaborators first met (either in person or by telephone) and began to develop a group of items that would have a broad focus on the categories of health literacy advanced in the 2004 IOM report while tapping the dimensions of literacy developed in previous literacy assessments by the Educational Test Service (ETS) and the National Assessment of Literacy Survey (NALS). Our group of researchers developed items designed to assess participants' knowledge in the categories proposed in the IOM report (left column in Table 3-1) in the prose, document, and quantitative formats (columns headings). Examples of item targets are presented in the table. Once written, items were translated into the software program chosen to present items to study participants and record their answers.

In addition to creating items in English, many items were developed in Spanish and then translated into English.. Items were then translated into the other language with the study team evaluating each others' translation and most often adaptation of the item content and format for its appropriateness in each language/cultural group. The initial group of 225 items was first administered to a group of English-speaking participants that allowed us to evaluate the items' quality and difficulty. Participants in this phase of development were systematically interviewed at the end of their participation to elicit their reactions to questions and asked for their feedback about confusing language and the appropriateness of existing response options. Results from this initial pilot were then used to reviese items for use in phase I.

Method

This section provides an overview of study procedures (see Figure 3-1). Test items were developed to sample a broad range of health-related content in Spanish and English. The sample on which the measure was validated was purposefully drawn from a range of abilities and backgrounds as evidenced by participants' occupations and educations. In order to accurately characterize Spanish-speaking participants, we developed a procedure to assess language dominance in Spanish-English bilinguals. Items in both languages were created to minimize the impact of regional usage and data analyses employ a combination of classical test theory (DeVellis, 2006; Gulliksen, 1950; Lord & Novick, 1950) and item response theory (Embretson & Reise, 2000) techniques.

In Phase 1, a pool of candidate items was administered to Spanish and English speakers, with approximately one-half of each group aged 50 years or older. Items were screened for difficulty and discrimination (correlation with total score) and for age- and language-associated DIF. The original pool of items was reduced and some new items were written to enhance the total scale's range of content and difficulty. In Phase 2 , items developed in Phase 1 were administered to an age-stratified sample of community-dwelling Spanish and English speakers along with measures

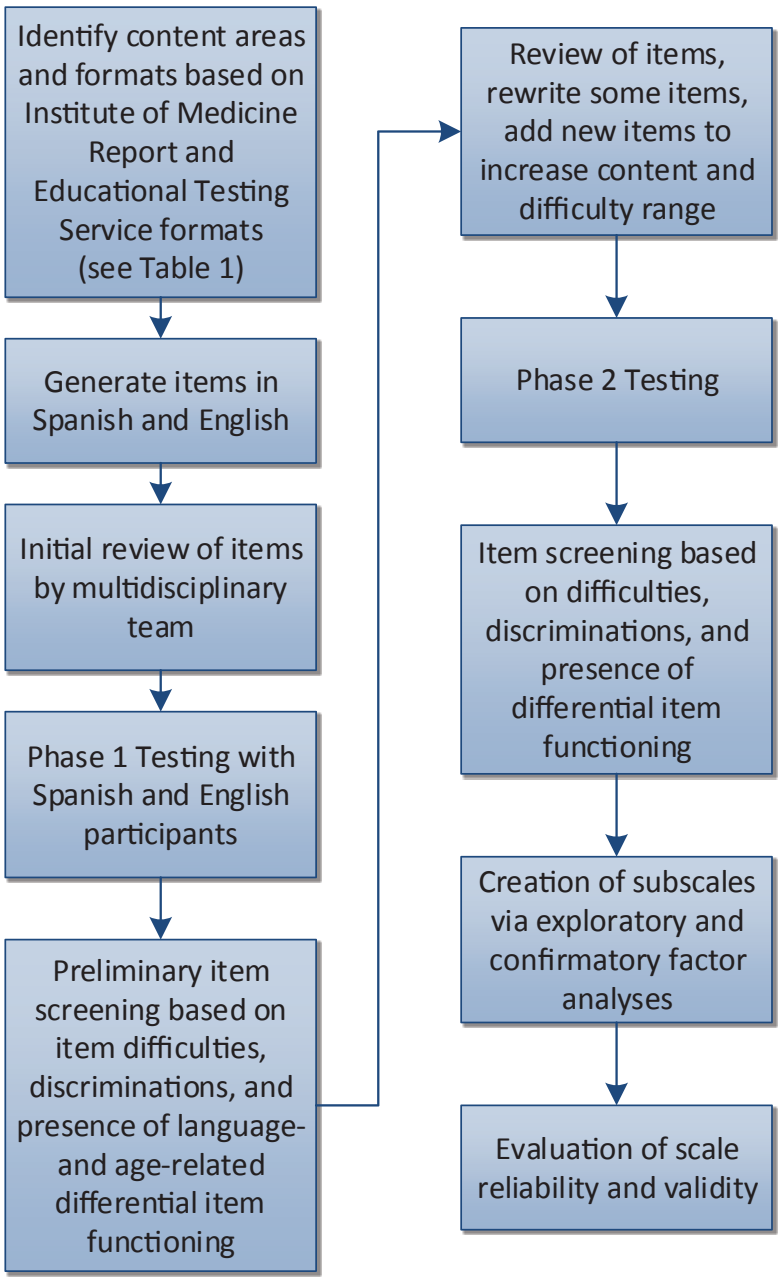

Figure 3-1. Item development process. Reprinted with permission of the publisher from Ownby et al. (2013).

Table 3-1. Item content examples by IOM report domains and ETS formats (from Ownby et al, 2013; reprinted with permission)

Goals of Health Literacy	Prose
Health promotion	Read a passage on exercise and identify desirable duration of exercise
Understand health information	Read a passage on risk factors for diabetes and identify relevant behaviors that would reduce someone's risk
Apply health information	After being provided with information on physical activity guidelines, identify appropriate exercise duration and frequencies
Navigate the health care system	After reading an informational brochure, be able to describe how specific health care services are covered by an insurance program
Participate in encounters with health care professionals	After viewing a video of a person's encounter with a physician providing a new medicine, identify information provided by the physician about dosage and schedule
Give informed consent	After reading information about a colonoscopy, describe the risks and benefits of the procedure

Document	Quantitative
Make menu choices based on fat and sodium guidelines	Calculate the number of grams of fat in a package of a product given a per serving value
Given a checklist of risk factors for diabetes, be able to complete a checklist of risk factors for the disease	Given information on normal and abnormal blood glucose levels, identify normal and abnormal levels
Given narrative information on exercise frequency and intensity, complete an exercise log.	Calculate the number of calories used during exercise give a table of exercises, times, and values;
Review information from a table on dates and times for applying for specific health care benefits	Calculate relative costs of two insurance plans
After viewing a video describing how to apply for long term care insurance, fill out an application	After viewing a video that presents information on desirable weights, calculate one's own body mass index
After viewing a video that presents information on informed conwent for a clinical study, describe its risks and benefits.w	Given a graphical representation of the probability of a medication side effect, correctly identify how likely its occurrence will be.

chosen to establish the new scale's validity, including other measures of health literacy, health-related quality of life, health status, and health service utilization.

Computer Delivered Format

In order to ensure that the resulting measure will be inexpensive and easily deployed, it has been developed using off-the-shelf touchscreen computers that are readily available and reasonably priced (HP TouchSmart®; Hewlett-Packard Corporation, Palo Alto, CA). These computers had touch screens running the Windows® operating system (20 inch diagonal measurement) and include self-contained speakers that allow participants to hear all items as they are presented. In another project, we had done extensive user testing with the general format of touch screen interface, iteratively testing the interface, modifying it in response to user comments, and then retesting it. In Phase 1, we used the previously developed touchscreen format. Interviews after each participant completed the phase 1 focused on possible usability and navigation issues as well as on the format and content of the questions.

Item Development

A framework for item development was created based on the domains of literacy skills needed for health as outlined in the 2004 Institute of Medicine (IOM) report.1 For each of the seven health-related goals listed in the Report (left column in Table 1) items were created in one of the three formats commonly used in assessing literacy--prose, document, and quantitative.40 A 7 X 3 item content matrix was created and used as a guide in item development (Table 3-1). Candidate items were developed by individual team members and reviewed by the entire team. Members represent eda range of health professions including medicine, nursing, social work, pharmacy, and psychology. Each team member had extensive experience in clinical work and thus was familiar with the types of clinical problems encountered by patients in obtaining health care. The lead investigator (RO) had extensive experience with psychometric scale development .

The team included a psychologist with a strong background in multicultural and multilingual assessment (AA) who also has published multiple articles on psychometric assessment. Other members of the team had extensive experience in patient education and assessment.

Some items were first created in Spanish and then translated into English, while others were created in English and translated. A guiding principle in item development was to create items that would be culturally and linguistically equivalent rather than creating word-for-word translations. From the project's inception, word and item selection were focused on use of high-frequency words and terms to ensure that participants would understand all questions. Care was taken to use words in both languages that would be understood by persons of varying socioeconomic and educational levels and that were not region- or nation-specific.

Items developed within the 7 x 3 content matrix (Table 3-1) targeted the component skills of literacy (conceptual knowledge, listening and speaking, writing, reading, and numeracy) as outlined in the IOM report. For example, to assess conceptual knowledge, items that tapped basic health facts were created (e.g., "Hemoglobin A1C measures which of the following?"). Listening comprehension was assessed using 60-90 second videos of simulated interactions with health care providers or presentations of health information. For example, one video showed an encounter in which a patient was given a new medication and directions for its use, and another simulated a TV news presentation on finding health information on the Internet. After viewing, participants responded to multiple choice questions. It was not possible to directly assess participants' oral expression, but questions were created that presented problems that could only be solved by communicating with providers (e.g., "Arthur doesn't understand what the doctor says. What can he do?").

Written expression was assessed as document literacy through questions evaluating participants' ability to complete materials such as insurance forms. Navigating the health care system included interpreting hospital maps; some documents and maps included items that asked

participants to respond by tapping on the appropriate area of the screen (e.g., "Tap on the area where you would find information on how to use toothpaste with a 4 year old"). Reading comprehension was assessed through questions about passages of varying difficulty levels, and numeracy was assessed through items demanding reading, arithmetic computation and decision making based on probabilities. Approximately ten items were created for each element in the item content matrix resulting in 208 candidate items.

Phase 1: Initial Item Testing

This base group of items was administered to 69 Spanish- and 73 English-speaking participants. Language dominance of Spanish-speaking participants who indicated that they also spoke English was assessed by comparing their performance on the relative proficiency indices (RPI) of the reading and listening comprehension subtests of the Woodcock-Johnson (English) and the Woodcock-Muñoz (Spanish) Psycho-Educational Batteries (Rolling Meadows, IL: Riverside Publishing). Level of acculturation was assessed using the Marin Acculturation Scale. Most participants showed clear superiority in one language or the other (i.e., more than one standard deviation difference in RPI scores), including those who indicated they had proficiency in both languages. Only those participants who showed clear evidence of greater proficiency in Spanish completed the Spanish assessment. The importance of actually assessing Hispanic participants' language skills is underscored by a study that showed that Hispanics who state they are fluent in English may function at lower levels compared to native English speakers.

Almost half of each language group was 50 years of age or older (30 of 69 Spanish and 29 of 73 English speakers) allowing for the assessment of language- and age-related DIF. As discussed above, we believed that this issue was important in light of our finding that almost one-half of the items on the reading comprehension scale of the TOFHLA showed evidence of age-related DIF. After responding to all items, participants completed

interviews during which items were reviewed with them for problems in clarity and to assess whether the items actually measured what was intended. Initial analyses were completed using jMetrik (www.itemanalysis.com), assessing item difficulties, discriminations (item-total correlations), and the presence of DIF. The Mantel-Haenszel chi-square statistic as well as the Educational Testing Service grading system were used to evaluate DIF, supplemented by review of nonparametric item response curves.

In creating the final item pool for further study, some items were eliminated due language-related DIF, while other items were rewritten after team consultation. No items showed substantial age-related DIF, supporting our decision to avoid using the cloze response procedure used in other measures, as it may bias items against older individuals (Ownby, Acevedo, & Waldrop-Valverde, 2014). A number of items were either low or mid-range in difficulty; many of these were eliminated when their content or format duplicated other items. Data from interviews were used to rewrite items when participants indicated that an item was confusing or when the interviews showed that the item did not actually assess its target skill. Several new items were created in this phase in order to broaden the range of content covered and provide items with greater difficulties. Although these items were not subjected to the same developmental testing as other from Phase 1, they were be assessed for psychometric characteristics prior to inclusion in the group of items tested in Phase 2.

Phase 2: Further Development

The purpose of phase 2 was to evaluate item characteristics and validate the measure by assessing its relations to other measures of health literacy and participants' health. A purposive sample with a range of abilities (based on education and occupation) was recruited over specific age ranges by decades from the 20 to 70s. We judged that this strategy was most likely to be the most efficient approach to obtain optimal item statistics with a relatively small samples (Orlando, 2004; Wingersky & Lord, 1984). Interested participants were first screened for cognitive status us-

ing the Short Portable Mental Status Questionnaire and paragraphs from the Wechsler Memory Scale using cut off scores previously developed in a study of computer use in elderly participants. Participants were screened for intact vision and auditory abilities using a visual screener and auditory comprehension of material presented over headphones calibrated using a handheld decibel meter (Extech Model 407730, Extech Instruments, Waltham MA). Spanish-speaking participants were recruited from several different national backgrounds including the countries of Central and South America as well as the US and Mexico. The language of assessment was determined as described below, using the procedure developed in Phase 1. Participants first completed the language preference subscale of the Marin Acculturation Scale. When language preference was not clear, the determination was supplemented with additional testing when participants indicated significant use of both languages.

In addition to the new health literacy items, participants completed a battery of established health literacy measures (TOFHLA in both Spanish and English; REALM or SAHLSA, the self-report questions developed by Chew et al. (2008), and literacy- and numeracy-related academic skills and basic cognitive abilities). They also provided information on their health status, health-related quality of life, and health service utilization. Participants completed assessments in two sessions (individually-administered cognitive and health literacy measures in one, questionnaires and the health literacy measures administered by touch screen computer in the other) with order of administration of each session randomly counterbalanced to account for order effects. Because of the length of assessment sessions, participants completed both sessions either in a single day (during which they take at least a one-hour break for lunch) or on two days. Measures were selected to allow the evaluation of the relation of the new measure to existing assessments of health literacy, basic cognitive skills, relevant academic skills such as reading and math skills, and health status variables.

Who Were Our Participants?

Since we wanted to develop a measure of health literacy that would be useful in both Spanish and English speakers, we wanted to better understand how well our bilingual participants spoke English and Spanish. We focused on ensuring that persons assessed in Spanish were actually competent in Spanish through a screening and testing procedure that , when their language backgrounds were not clear, evaluated with objective tests the language proficiency of individuals who indicated that they were proficient in both English and Spanish.

The screening procedure for all participants included a question about whether they spoke any other languages in addition to English. Persons who indicated that they spoke more than one language were asked about their language use at home and with friends using questions from the Marin et al (1987) cultural assessment tool. While the majority of potential participants indicated that they primarily used English or Spanish in their daily lives, some reported a balance. Those whose responses suggested significant competence in both language were tested using a procedure developed during Phase 1 of the project. Individuals whose linguistic status was unclear were therefore asked to complete both timed listening and reading assessments in each language [WJ/WM scores]. The majority of participants, even those who indicated that they felt they could complete F/V in English, clearly felll in a specific language group based on their responses. It is important to note that many individuals who initially indicated that they could complete the measure in English clearly had better Spanish language skills.

Construct Validity

Several strategies were used to assess the validity of the new measure. Construct validity was assessed through a series of exploratory and confirmatory factor analyses that assessed the dimensions of health literacy assessed by the items. The factor structure of the measure was assessed

Table 3-2. Participant Gender, Language and Age

		Age Group						
		18-29	30-39	40-49	50-59	60-69	70+	Total
English	Men	15	13	14	24	14	14	94
	Women	18	22	20	18	26	45	149
	Total	33	35	34	42	40	59	243
Spanish	Men	14	11	21	13	11	24	94
	Women	12	19	34	36	23	32	156
	Total	26	30	55	49	34	56	250
Total	Men	29	24	35	37	w25	38	188
	Women	30	41	54	54	49	77	305
	Total	59	65	89	91	74	115	493

footnote

Table 3-3. Participant Race, Language and Age

		Age in Decades						
		18-29	30-39	40-49	50-59	60-69	70+	Total
English	White	15	9	10	22	22	44	122
	Black	18	26	24	20	18	15	121
	Total	33	35	34	42	40	59	243
Spanish	White	26	30	55	49	34	56	250
	Total	26	30	55	49	34	56	250
Total	White	41	39	65	71	56	100	372
	Black	18	26	24	20	18	15	121
	Total	59	65	89	91	74	115	493

footnote

Table 3-4. Means and Standard Deviations for Age, Education, and Health Literacy Measures by Language and Age

			Age	Education	TOFHLA Num	TOFHLA Read	REALM	SAHLSA
English	18-29 N=25	Mean	23.33	12.82	48.00	47.21	63.18	
		SD	3.92	2.13	2.29	2.04	3.72	
	30-39 N=35	Mean	33.71	13.47	47.74	46.37	61.26	
		SD	2.78	2.20	3.54	3.71	9.24	
	40-49 N=35	Mean	45.00	12.22	46.12	45.00	59.41	
		SD	2.76	2.25	4.01	5.06	11.90	
	50-59 N=42	Mean	54.02	12.90	47.14	44.93	61.24	
		SD	2.97	2.32	3.52	5.93	8.09	
	60-69 N=40	Mean	63.93	13.34	47.32	44.49	61.68	
		SD	2.89	2.82	4.65	8.18	9.26	
	70+ N=59	Mean	77.15	13.79	47.19	45.42	63.12	
		SD	4.79	2.47	4.42	6.16	8.78	
	Total N=243	Mean	52.91	13.16	47.24	45.51	61.78	
		SD	18.91	2.43	3.90	5.70	8.84	
Spanish	18-29 N=26	Mean	24.08	11.92	44.42	46.65		44.38
		SD	3.61	3.17	7.27	2.77		4.63
	30-39 N=30	Mean	34.97	12.07	43.70	45.03		45.17
		SD	2.57	3.48	6.20	7.29		4.85
	40-49 N=55	Mean	44.73	12.65	43.40	44.94		45.91
		SD	2.98	2.38	6.51	5.20		4.82
	50-59 N=49	Mean	54.02	12.02	42.73	42.65		44.96
		SD	2.68	3.29	6.56	9.04		5.50
	60-69 N=34	Mean	64.29	11.18	41.32	39.21		45.74
		SD	2.73	2.97	7.23	10.10		3.45
	70+ N=56	Mean	78.38	9.68	39.70	31.88		43.22
		SD	6.06	3.93	7.15	12.47		5.89
	Total N=250	Mean	53.43	11.52	42.30	40.96		44.86
		SD	17.84	3.39	6.94	10.25		5.09

separately for English and Spanish speakers, and the equivalence of the measure's factor structure in both language groups was evaluated.

Classical test theory item analyses were used to assess item difficulties and their relation to the measure's underlying core dimension. From the initial item pool evaluated in phase 1, a subset of items were chosen for further evaluation in Phase 2 based on a range of item content and difficulty as well as the absence of differential item functioning. Final item analyses after data collection was complete allowed the verification of scales provisionally developed based on phase 1 data and interim analyses of phase 2 data.

Differential item functioning between English and Spanish speakers was assessed at the initial stage of item development, after pilot testing the items, and at the end of data collection. As the first two stages of measure development, items were found that functioned differentially in English and Spanish speakers. At the final analysis, only one item remained with significant DIF; it was eliminated. Scale equivalence in English and Spanish was evaluated using item response theory (IRT) linking and equating procedures available in jMetrik (Meyer, 2014).

Table 3-5. Tests of improvement of model fit for exploratory factor models with 1 to 6 factors

Comparison			χ^2	df	*p*
1-factor	against	2-factor	521.224	97	0.0000
2-factor	against	3-factor	330.075	96	0.0000
3-factor	against	4-factor	227.275	95	0.0000
4-factor	against	5-factor	144.246	94	0.0007
5-factor	against	6-factor	135.577	93	0.0026

Table 3-6. Fit statistics for exploratory factor models with 1 to 6 factors

	χ^2	df	*p*	RMSEA	CFI	TLI
1	5627.32	4655	0.000	0.021	0.936	0.935
2	5077.05	4558	0.000	0.015	0.966	0.964
3	4713.66	4462	0.004	0.011	0.983w	0.982
4	4502.78	4367	0.074	0.008	0.991	0.990
5	4372.94	4273	0.140	0.007	0.993	0.993
6	4247.94	4180	0.228	0.006	0.996	0.995

Exploratory Factor Analyses

Factor analysis is a way to identify dimensions that underlie someone's responses to a group of test questions. For example, we could ask people a group of questions that tap their skills with words and at the same time ask them to answer questions that tap their ability to visualize geometric forms in their head. If we were to factor analyze their responses, we would probably find that the word-related items were closely related to each other, and the items about geometric forms would be closely related to each other, too. The two sets of items would be related to each other as well, but not as closely. If we were to use the statistical technique of factor analysis to look at the date, we could find out how closely each of the questions we asked is related to word-related skills or geometric form-related skills. The results of this kind of factor analysis give us each task's loadings on a dimension of word-related and one of visually-related abilities.

The 98 items administered to participants in Phase 2 were subjected to exploratory factor analyses using routines available for factor analysis of categorical variables available in MPlus (Muthén & Muthén, 2012). The

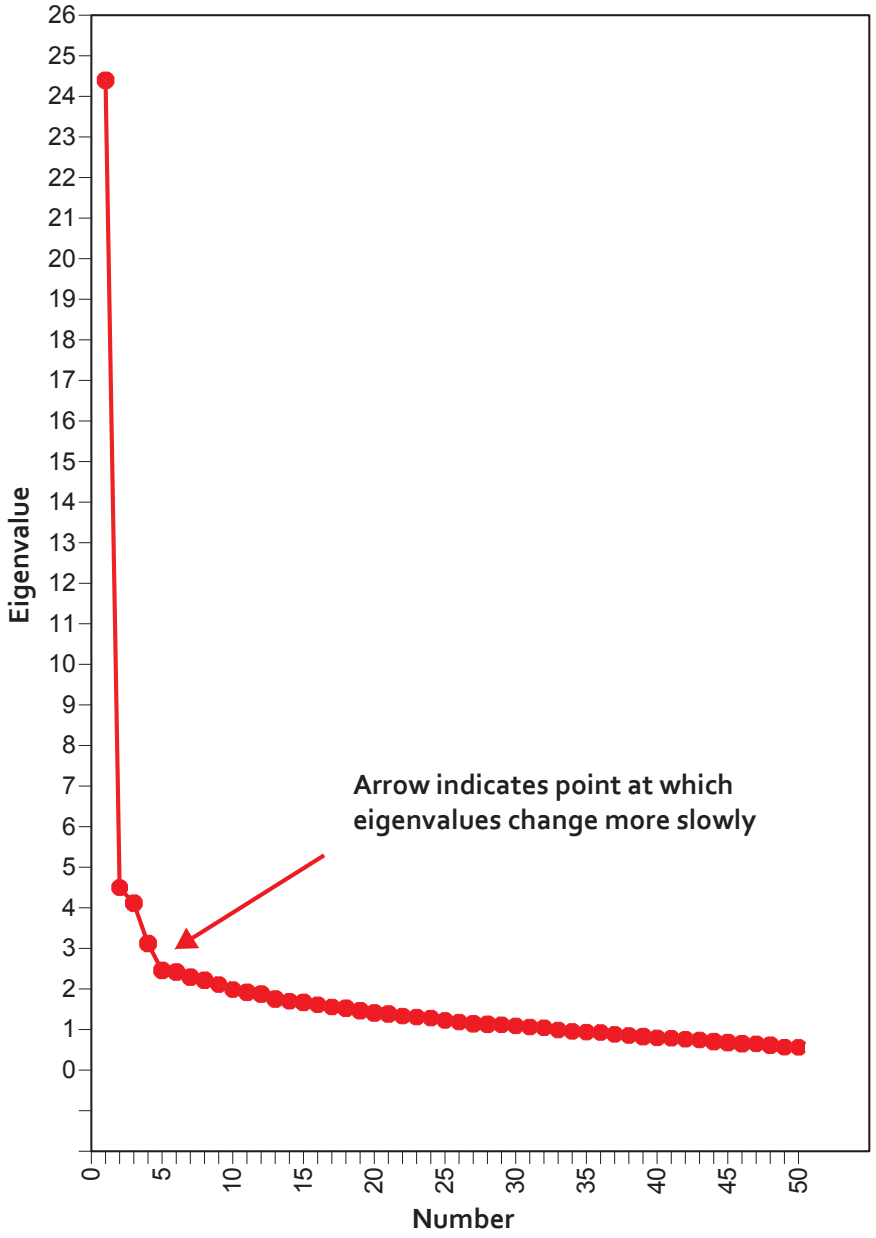

Figure 3-2. Scree plot of eigenvalues for the combined sample of English and Spanish speakers

number of factors in the items was determined by inspection of the scree plot of eigenvalues (Figure 3-2), statistical tests of the improvement of factor model fit with larger number of factors, and assessment of how well factors models with different numbers of dimensions actually fit our data. The scree plot (Figure 3-2) shows that the size of eigenvalues extracted became smaller and tended to change in very small amounts after four factors were extracted (arrow in the Figure).

Statistical tests of the improvement in model fit with increasing numbers of factors are presented in Table 3-5. Model fit improved with each additional factor up to six. Fit statistics for the exploratory factor analytic models are presented in Table 3-6. A four-factor model resulted in a non-significant χ^2 value, indicating that the model with this number of factors fit the data well. Other values listed in the table are the RMSEA (root mean square error of approximation), CFI (comparative fit index), and TLI (Tucker-Lewis index). Values of the RMSEA less than 0.05 are considered to indicated good model fit, while values of the CFI and TLI greater than 0.95 are considered desirable. It can be seen that the four-factor model met these criteria. Finally, even though five- and six-factor models showed improved fit (Table 3-5), evaluation of the content of these two models showed that only a small number of items were included in these factors and that they did not represent important dimensions of FLIGHT/VIDAS.

Exploratory analyses were also completed in the English and Spanish samples separately. Results of exploratory factor models were similar in both language groups on examination of fit and factor loadings. This interpretation was subsequently tested in confirmatory analyses.

Confirmatory Factor Analyses

These exploratory analyses allowed to us to develop a model of the factors underlying the FV variables for further testing in confirmatory factor analysis (CFA) models. EFA results suggested the appropriateness of a four-factor model based on inspection of the plot of the eigenvalues and change chi-square values for models with progressively larger number of

Table 3-7. Fit statistics for confirmatory models by language group and equivalence models

Model	χ^2	df	*p*	RMSEA	CFI	TLI
English	2892.03	2685	0.0028	0.018	0.964	0.963
Spanish	2855.76	2685	0.0110	0.016	0.972	0.971
Configural	5749.42	5370	0.0002	0.017	0.968	0.968
Scalar	5897.31	5442	0.0000	0.018	0.962	0.961

factors. A four-factor model appeared to fit our data well.

These four factors represented **(1) general health literacy** (reading passages and understanding documents), **(2) numeracy** (arithmetic calculations and understanding probability and risk, **(3) conceptual health knowledge,** and **(4) listening**. A four factor model based on these dimensions was therefore created and tested. Tests of this model showed that it fit the data of the combined sample of English and Spanish speakers well, although substantial model misfit was present as indicated by a significant χ^2 value (3159.04, df = 2685, $p < 0.001$). Indicators that suggested the four-factor model fit the data well included the RMSEA at 0.019 (less than 0.05 indicating adequate fit) and the CFI of 0.97 and TLI of 0.97 (greater than 0.95 indicating adequate fit).

English and Spanish Scales

A key goal of the F/V project was to create a measure that would be useful and equivalent in both English and Spanish. Throughout the project this goal was kept in mind, from the initial conception of the project through item development and finally in data analyses. The equivalence of the F/V scales were assessed in three ways: (1) assessment of the equivalence of the meausre's factor struture in English and Spanish; (2) evaluation of all items for indications of differential item functioning; and (3) testing the relation of the HL scale in both languages through item response theory

scale linking. These complementary approaches provided an overall assessment of the extent to which the scale can be used with persons who speak either language and the perhaps most importantly compare scores between groups.

Factor Structure Equivalence

An important question is the extent to which the dimensions of health literacy represented in FLIGHT/VIDAS are the same in English and Spanish speakers. This questions was assessed in confirmatory analyses that evaluated, first, the extent to which the factor structure was the same (even if the language groups were different on level of performance or in the ways that each item measures the dimensions it's related to) and then a stricter comparison that these same issues were the same in the two groups. Fit statistics for the four-factor models for the English- and Spanish-speaking and for the configural and scalar models are presented in Table 3-4. It can be seen that all models were associated with significant χ^2 values, suggesting that they did not fit the data exactly. All models, however, were associated with RMSEA values less than 0.05 (suggesting adequate fit) and with CFI and TLI values greater than 0.95 (again suggesting adequate model fit). The fit of the scalar invariance model was significantly poorer than that of the configural invariance model (χ^2 [df = 72] = 154.80, $p < 0.001$), although it can be noted that the fit of both invariance models was adequate as evaluated by the three fit indexes.

Differential Item Functioning

Differential item functioning (DIF) of items exists when a test item is more or less difficult for members of one group compared to another even though they have the same overall ability. Testing for DIF is important, since when it exists one group's scores on a measure may be different from another's because of the measure rather than because of real differences between the groups. Because of our concern for developing a measure that would be widely useful, we tested for DIF related to language, and to age (because we had found evidence of age-related DIF in the TOFH-

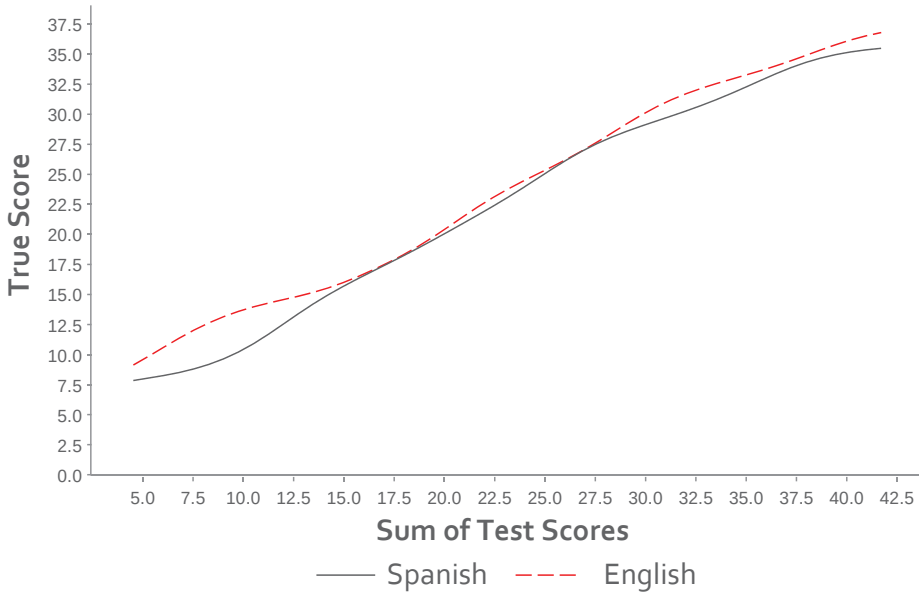

Figure 3-3. Test characteristic curves for language groups

LA; Ownby, Acevedo, and Waldrop-Valverde, 2013; 2014). DIF related to language group was assessed at the initial stage of item development, after pilot testing the items, and at the end of data collection. DIF analyses are usually based in item response theory (IRT) analyses and require large sample sizes. For this development project we used a nonparametric approach to DIF analysis available in the jMetrk software (Meyer, 2014). This strategy allows an assessment of between-group DIF by plotting item characteristics for each group. When DIF exists in an item, there is a discrepancy between the difficulty of an item in each group for individuals of the same overall ability. In F/V development, we assessed language, age, and education related DIF in phase 1 and in selecting items for phase 2. The final scale Eliminated all items with substnatial DIF. The test characteristic curve for the final 40-item scale is presented in Figure 3-3. It can be seen that the overall curve for the two groups is essentially equivalent, although the scale may be slightly more difficult for lower ability Spanish speakers.

Scale Equivalence and Linking

Scale equivalence means that scales measure the same thing in two different groups; in the case of the FLIGHT/VIDAS project, we were primarily interested in evaluating how performance on the scales of FLIGHT/VIDAS was the same among English and Spanish speakers. This assessment was important because it would allow users to assess differences in health literacy between groups of English and Spanish speakers, and to date, no measure of health literacy had been shown equivalent in the two language groups. (Even though the TOFHLA has both English and Spanish versions, as of this date no readily-identifiable study has demonstrated its equivalence in the two language groups. In fact, the reading items of the Spanish version are not the same as in the English version, making the two measure's equivalence unlikely).

Scale equivalence in English and Spanish speakers was evaluated using IRT linking and equating procedures available in jMetrik (Meyer, 2014). In this approach, IRT analyses assessed item difficulties in the two groups and then compared them so that scores on the scales in one language group could predict scores in the other. Various models compare group performances, predicting the mean or combining the mean with the scale variance. All models suggested that the HL scales in English and Spanish were equivalent. Equations to predict scores on one measure based on another each had intercepts of 1.00, consistent with minimal overall difference in level of performance between groups. Depending on the linking method, the coefficient in the regression equation relating one score to another was either zero (mean predicting mean) or very small (-0.02 for the Haebara method and -0.04 for the Stocking-Lord method). While this explanation is technical, the conclusion to be drawn from these analyses is straightforward. Scores on the English HL scale are very similar to those on the Spanish version of the scale.

Table 3-8. Computer Administered FLIGHT/VIDAS scales

Scale	Example
General Health Literacy (HL): The ability to read and complete mental operations on health care information, including identify relevant information in prose, documents, and figures (39 items).	***Prose:*** After reading instructions for laboratory test preparation, correctly identify appointment time. ***Document:*** Correctly identify fields in an insurance form; Use an electronic device on a Web page to calculate body mass index.
Numeracy (NUM): The application of quantitative skills including arithmetic operations and appraisal of relations among numeric concepts such as ratios and percentages (24 items).	Correctly identify meaning of terms related to probability; Correctly identify number of grams of fat consumed in a meal based on values in a table.
Conceptual Knowledge (Experimental Scale; FACT): Demonstrate understanding of specific concepts related to health care (14 items).	Correctly identify the organ treated by a medical specialist such as a cardiologist
Listening Comprehension (Experimental Scale; LIS): The ability to acquire and remember information presented orally (13 items).	After viewing a video of a clinician giving information about participation in a clinical research study, correctly identify treatment alternatives.

This table originally appeared in Ownby et al. (2013). Reprinted with permission of the publisher.

Final Scales

The factor analytic work combined with classical test theory analyses of item characteristics provided the information needed to choose items that would allow us to reduce the 98 items tested to a smaller number that could be assembled into scales measuring specific dimensions of health literacy. The dimensions of health literacy were described above, and the final scales are described in Table 3-8. We also received requests from a number of potential users to create a measure that could be hand administered and scored, and our discussions with a number of persons at various conferences where we presented data on F/V development led us to believe that a short measure that could be used for screening would also be useful to a number of clinicians and researchers. We therefore created two additional subscales of F/V using items that did not require computer administration.

Computer Administered Scales (Table 3-8)

As exploratory factor analyses suggested that the FLIGHT/VIDAS items included a dimension of general health literacy reflecting an individual's ability to extract meaning from written texts and documents. This scale was named General Health Literacy and referred to as HL. Another scale reflected the ability to use quantitative concepts in health care decision making or tracking of health status and is interpreted as reflecting numeracy (NUM). Consistent with our plan to develop a number of items that assessed conceptual knowledge of health, a subgroup of items was judged to reflect conceptual knowledge as contrasted with the ability to use more complex mental operations; this scale is referred to as FACT. Finally, a number of items assessed the ability to acquire information presented in video simulations of health care encounters, health-related TV new story, information related to providing informed consent, and instructions on healthy diet.

Pencil and Paper Scales

As previously noted, although our primary intention was to develop a scale that was computer administered and scored, ongoing feedback from potential users has indicated a strong interest in being able to administered F/V individually without a computer or Internet access. Given the number of questions and the diverse formats of items tested in Phase 2, it was possible create a paper and pencil version of the measure that could be hand scored. In creating this version of the scale, we created test materials that were replicas of the computer screens used in the computer-adminstered version of the scale.

We note that Chesser et al. (2014) compared computer-administered and individually-administered versions of the S-TOFHLA. Results of this study showed that the computer version was essentially identical to that individually-administered version. Results of this study were consistent with results of multiple other studies of computer-administered versions of paper and pencil scales. These have shown that computer-administered versions of measures are, in general, functionally equivalent as long as test materials and order of administration are the same (Gwaltney et al., 2008; Millsap, 2000).

The 20 items of the HL paper scale were selected to reflect a moderate range of difficulties, a range of content, and to consistently have high item discrimations. This strategy would make the scale maximally useful across settings (e.g., healthcare, general health promotion) and include not only reading but also document and numeracy tasks. Information on these scales' reliability and validity in presented in Chapter 4.

Distribution of Scores

An important problem with the use of the TOFHLA, REALM, and SAHLSA in research on health literacy with normal groups is the finding that a very large number of persons will attain perfect or near perfect scores on these measures. For example, in the F/V final sample, the skewness statis-

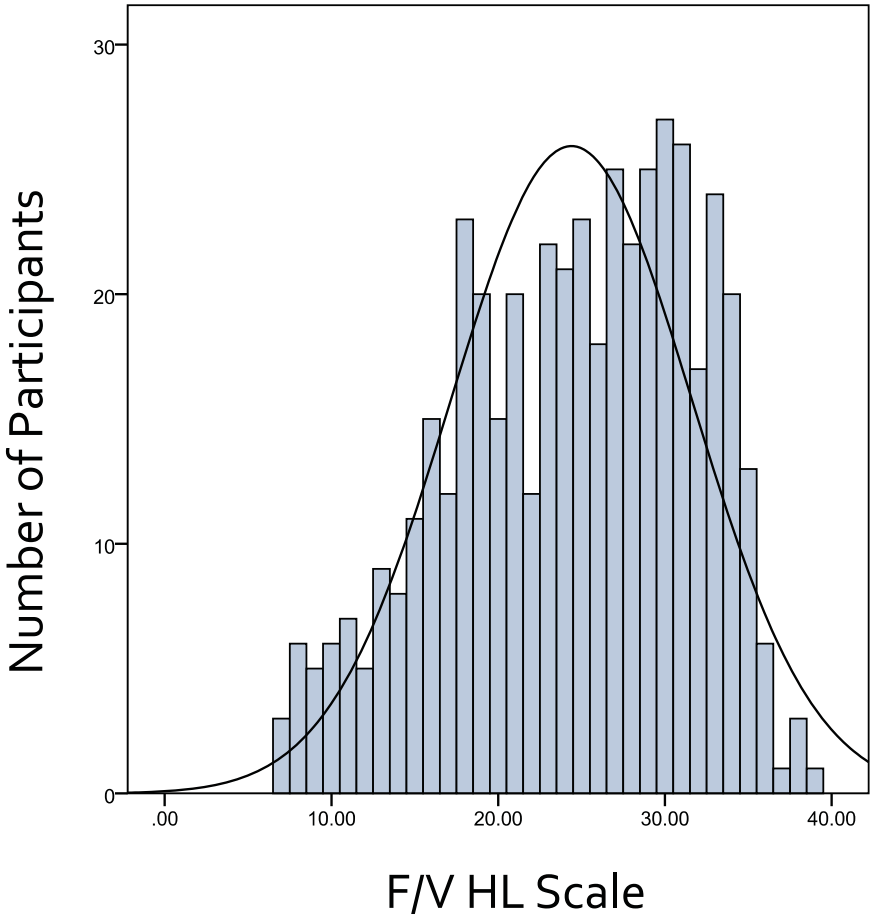

Figure 3-4. Distribution of scores on the 40-item F/V General Health Literacy scale

tic for the TOFHLA total score was -1.88, indicating that the distribution of scores was highly skewed. In fact, nearly 60% of our participants had scores of 90 or greater.

By contrast, the distribution of scores for the F/V scale was approximately symmetric, with a skewness statistic of -0.38. The distribution of F/V General Health Literacy scale scores is presented in Figure 3-4.

4 Reliability and Validity

Overview

Multiple strategies are available to evaluated the reliability and validity of a measure like F/V. In evaluating the F/V scales, we assessed internal reliability. We evaluated validity in several ways, as the scales' correlations with other measures and its ability to discriminate among groups of people who reported difficulties in understanding written health information.

Reliability

A standard strategy to evaluate the reliability of a measure is to calculate an index of the scale's internal consistency. Probably the most common measure of this sort is Cronbach's alpha. Acceptable values for this measure vay depending on the purpose for which a scale is used, but most authors suggest that a value of 0.80 and greater can be considered acceptable. Table 4-1 present the alpha values for each of the FLIGHT/VIDAS scales for each language group and for the combined sample. The standard scales for the measure, HL, NUM, and Paper each have acceptable levels of internal reliability as assess by Cronbach's alpha. The FACT and LIS scales have borderline internal reliabilities and should be considered useful in research and for further development.

Table 4-1. Cronbach's alpha for subscales

	HL40[a]	Paper	NUM	FACT[b]	LIS[b]
English	0.87	0.85	0.88	0.73	0.64
Spanish	0.85	0.84	0.86	0.61	0.56
Combined	0.87	0.86	0.88	0.69	0.58

[a]HL40 = 40-item general health literacy scale; Paper = 20-item paper and pencil version; NUM = health numeracy scale; FACT = conceptual knowledge scale; LIS = listening scale.
[b]Because of low reliability values, the FACT and LIS scales should be considered experimental.

Validity

Face validity

As more extensively discussed in Chapter Three, the items used in creating FLIGHT/VIDAS were developed from a matrix of conents and formats developed by a panel of experts who created a defintion of the goals of health literacy. The interested reader can consult Table 3-1 to further assess the exent so which he or she believes data provided here support its association with a measure of quality of life.

Convergent validity

In this form of validity, a scale is assessed by its relation to other measures of the same construct. In the FLIGHT/VIDAS project, the full version of the Test of Functional Health Literacy in Adults (TOFHLA) was administered in either English or Spanish to all participants. English-speaking participants also completed the Rapid Estimate of Adult Literacy in Med-

Table 4-2. Subscale correlations with TOFHLA

	Paper	NUM	FACT	LIS	TOF R	TOF N
HL40	0.97	0.79	0.76	0.68	0.65	0.49
Paper		0.82	0.70	0.68	0.65	0.51
NUM			0.53	0.57	0.58	0.49
FACT				0.43	0.44	0.36
LIS					0.55	0.37
TOF R						0.53

TOF R = TOFHLA Reading Scale; TOF N = TOFHLA Numeracy Scale;HL40 = 40-item general health literacy scale; HL20 = 20-item general health literacy scale; Paper = 25-item paper and pencil version; NUM = health numeracy scale; FACT = conceptual knowledge scale; LIS = listening scale.

icine (REALM) while Spanish speakers completed the Short Assessment of Health LIteracy in Spanish-speaking Adults (SAHLSA). Correlations of the scales of FLIGHT/VIDAS with these other measures are presented in Tables 4-2 and 4-3

Known groups validity

Known groups validity refers to the ability of a measure to discriminate among groups defined by some external criterion. In the case of F/V, we chose to evaluate groups of participants whose levels of health literacy were defined by their performance on the TOFHLA, since it is one of the most widely used measures of health literacy and a number of other measures have been linked to those categories. In addition, we defined groups based on their own report of difficulties in understanding written health information using one of the questions developed by Chew et al. [ref]. Re-

Table 4-3. Subscale correlations with REALM, SAHLSA, and academic measures

	REALM	SAHLSA	Reading	Arithmetic
HL40	0.42	0.46	0.56	0.52
Paper	0.46	0.43	0.56	0.53
NUM	0.40	0.25	0.56	0.63
FACT	0.36	0.46	0.52	0.41
LIS	0.29	0.36	0.42	0.33

Note: Only English speakers completed the REALM, while only Spanish speakers completed the SAHLSA. Reading = Woodcock-Johnson or Woodcock-Muñoz Reading Comprehension subtests; Arithmetic = Woodcock-Johnson or Woodcock-Muñoz Applied Problems subtests.

sults of these analyses showed statistically significant differences in F/V HL scores across groups defined by TOFHLA categories of "Inadequate," "Marginal," and "Adequate" *(F* [2, 483] = 103.51, $p < 0.001$), and across the three self-report questions asking about difficulty in understanding written information *(F* [4, 468] = 13.66, $p < 0.001$), confidence in filling out medical forms *(F* [3, 472] = 27.24, $p < 0.001$), and reading hospital forms *(F* [4, 471] = 16.02, $p < 0.001$). Mean scores for the F/V HL scale for each level of confidence in filling out medical forms are presented in Figure 4-1.

Scales

The larger number of items evaluated and their diverse content and format allowed us to create a number of subscales from the total. These include the computer-administered scales derived from factor analyses: General Health Literacy (HL), Health Numeracy (NUM), health-related

Table 4-4. Subscale correlations with other health-related measures

	SF Gen Health	SF Well Being	Sx	Condn	EQ5D	CESD
HL40	0.10	0.21	-0.18	-0.15	0.25	-0.26
HL20	0.10	0.16	-0.18	-0.19	0.24	-0.23
Paper	0.11	0.19	-0.17	-0.13	0.24	-0.26
NUM	0.09[a]	0.14	-0.15	-0.14	0.22	-0.22
FACT	0.08[a]	0.16	-0.11	-0.02[a]	0.11	-0.19
LIS	0.11[a]	0.13	-0.09[a]	-0.12	0.08	-0.15

[a] All correlations are significant, $p < 0.05$ except those marked with superscript. SF Gen Health = SF36 General Health; SF Well Being = SF36 Emotional Well Being ; Sx = Self-report of number and frequency of physical symptoms; Condn = number of health diagnoses; EQ5D = health quality of life index predicted from SF36; CESD = Center for Epidemiological Studies Depression Scale

conceptual knowledge (FACT), and understanding auditory health information (LIS). Because of interest in a paper and pencil version of FLIGHT/VIDAS, a scale that included only items that could be presented in paper and pencil format (Paper) was also created. Fiwwnally, the need for a brief measure that could be used for quick screening gave rise to a ten-item screening scale, also available in paper and pencil format.

General Health Literacy (HL)

The general health literacy scale (HL) includes the best 40 items from phase 2 of the FLIGHT/VIDAS project. "Best" in this context means that

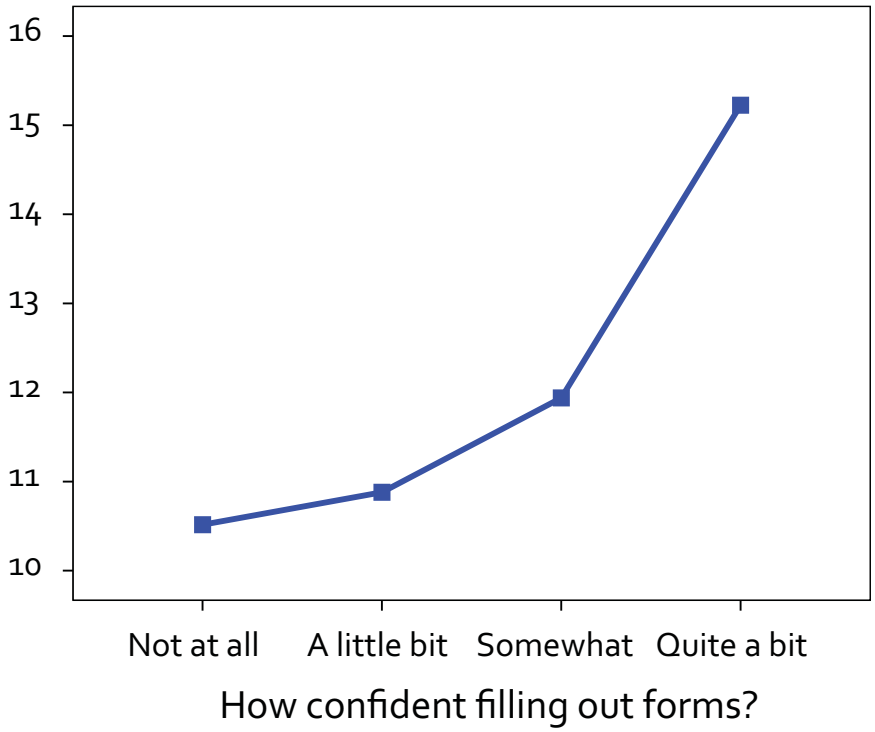

Figure 4-1. Group means for HL paper form scored across responses to question on confidence in filling out forms.

they represented a broad range of content and difficulties while having a strong overall relation to the rest of the items (point biserial correlation with scale total). It includes items that evaluate reading comprehension, numeracy, conceptual knowledge, and listening.

Paper

This scale is composed of items drawn from the HL scale that could be presented in paper and pencil format. A test booklet with these items and an answer sheet are available so that this scale can be administerd without a computer. In lieu of computer-delivered narration of questions, the test booklet provides directions for the individual administration of each item as the examiner reads relevant portions of each question. This pro-

cedure may be especially important for persons with low general literacy levels who may find the audio narration important in understanding questions. Because this format was not specifically validated in the original development project, norms for this scale should be interpreted cautiously, but as noted in Chapter 3, studies suggest that computer and paper and pencil measures are largely equivalent when test materials are as similar as possible in each form.

Numeracy (NUM)

This scale includes items that require mental arithmetic or understanding of probability. Many of the items ask that the person assessed read tables or figures to identify information needed to answer the question.

Conceptual Health Knowledge (FACT)

This subscale includes 14 items that only require knowledge of a specific fact about health or healthcare. None require reading prose or arithmetic operations. This scale has a low level of internal reliability, although it has been clearly related to important health status variables in analyses of the ASK model (see Ownby et al., 20XX and chapter 2 of this manual). Because of its psychometric characteristics, its use should be confined to research use. It may be useful in combination with HL and NUM in understanding the extent to which a patient's functional health literacy is related to their basic academic skills in reading or mathematics or to their general knowledge about health and healthcare.

Listening Comprehension (LIS)

Items on this scale are based on questions related to the three video simulations or the audio visual slide presentation included in the original administration. This scale has a low internal reliability and is undergoing further development. In one published study (Ownby et al, 2015), participants' responses to the items related to a simulated informed consent interview were related to their overall health literacy.

Screening

Scales to measure health literacy can be used for a number of purposes. One person may be interested in understanding health literacy's relation to quality of life or risk of dying or health disparities, while another wants a quick and easy way to identify people who may need extra help in understanding a standard set of medication instructions included in a clinic's electronic health record. The first person may need a measure that assesses a broad range of content and may be able to use a scale that takes more than a few minutes to administer and score, while the second person is likely to want something shorter.

The person who wants to identify people who may have a problem understanding medication directions would be described as wanting a validated **screening** measure. How well screening measures identify people can be described with several indexes, the most important of which are the following. Since no test is perfect, these indexes give you an idea of how confident you can be about making a decision based on a test score.

Sensitivity. This is the number of persons who have a positive test divided by those who actually have a condition (for example, low health literacy). These are considered "true positives." Since almost no test can detect everyone with a problem or condition, this number is usually less than one. Many medical and psychological tests have sensitivity values in the range of 0.70 to 0.80.

Specificity. This is the proportion of people who don't have a condition who have a negative test. These are "true negatives." As with positive test values, few tests are 100% accurate. Many tests have specificity values in the range of 0.70 to 0.80.

Positive predictive value (PPV). This is how likely it is that someone with a positive test actually has the problem or condition, like health literacy. A PPV value of 0.80, for example, would mean that you could be 80% sure that a person with a positive score actually has low health literacy.

Table 4-4. Characteristics of scale cutoff scores predicting low health literacy as reading level less than 8th grade

	Sensitivity	Specificity	Accuracy	PPV[a]	NPV[a]
HL40	0.79	0.66	0.72	0.64	0.81
Paper	0.79	0.64	0.71	0.62	0.81
Screener	0.70	0.70	0.71	0.76	0.63
TOFHLA	0.78	0.65	0.71	0.63	0.80

[a]PPV = positive predictive value; NPV = negative predictive value. Values are based on Youden index for all measures except screener which is based on a minimum PPV of 0.75.

Negative predictive value (NPV). This is how likely it is that someone with a negative test actually doesn't have health literacy—that is, probably has adquate health literacy.

Accuracy. Accuracy is the total number of correct identified individuals (whether they are correctly considered to have or not have the condition) divided by the total number of persons assessed.

These indexes depend on the cutoff score used, so it's possible to have different cutoff scores depending on whether you want a highly sensitive or highly specific test. PPV and NPV also depend on how common the condition (like low health literacy) is in the group of persons tested. In the tables for FLIGHT/VIDAS scales we provide here, we have assumed a prevalence of low health literacy of 50% based the overall estimate of the prevalence of health literacy reported in a review study (Paasche-Orlow et al., 2005).

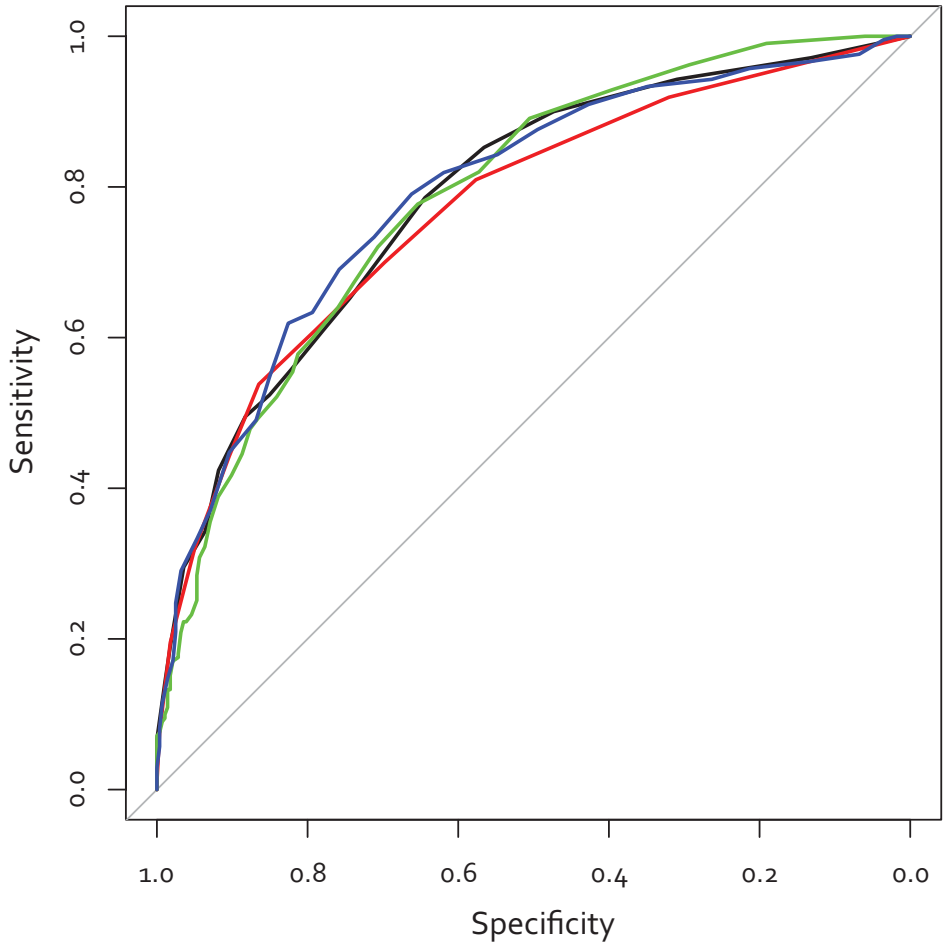

Figure 4-2. ROC Curves for TOFHLA Total Score (green), F/V 40-item general health literacy scale, F/V 20-item paper scale (black), and 10-item screening tool (red) detecting persons with an 8th-grade or lower reading level

Deciding on cutoff scores

If you want to be sure of identifying everyone with low health literacy, even if it means that you think some people need help even when they don't, you would choose a cutoff score with a high sensitivity. You

will be able to identify almost everyone who needs help, but may spend time with some people who may not have needed the help. If you want to besure that someone doesn't have low health literacy, then you would use a test with a high specificity.

One useful strategy for deciding on curoff scores is to emply a technique called receiver operating curve (ROC) analyis. Calculation of sensitivities and specificities for a range of curve proides data to plot various combinations of the two in a curve (Figure 4-2). Figure 4-2 shows ROC curves for the TOFHLA total score, the F/V 40-item health literacy scale, the F/V 20 item scale, and the F/V 10-item screening tool. The criterion used in these analyses was whether someone had a Woodcock-Johnson or Woodcock-Muñoz Passage Comprehension at or below the 8th grade level (suggesting less than adequate level of health literacy).

The performance of a measure in ROC analysis can be characterized in part based on the area under the curve (AUC) associated with its combined sensitivity and specificity. The areas under the curve were: TOFHLA total score = 0.79; F/V HL40 = 0.79; F/V 20 item paper and pencil scale = 0.79; F/V 10 item screening tool = 0.77. None of these values was significantly different from any of the others (Delong's test of correlated ROC curves; all *p*s > 0.05) although the test of the difference between the F/V 20-item paper and pencil scale and 10-item screening tool approached significance ($z = 1.78$, $p = 0.08$).

People are sometimes surprised to find out that a test result doesn't mean something with certainty. This is true not only in the realm of health literacy tests, but many other tests as well. For example, reported sensitivities and specificities for the TOFHLA and the Newest Vital Sign are in the range of 0.70 to 0.75 (Weiss, 2005; Osborn et al., 2008). Combined with internal reliabilities in the same range (Cronbach's alpha), any person's score on these measures should be interpreted cautiously, probably in the context of the person's education and the complexity of the health literacy challenge they are presented in real life.

Types of health literacy

If we want a measure of health literacy to identify those who have a problem understanding health information, we have to decide on some external criterion for what "low health literacy" means. For a number of health literacy measures such as the S-TOFHLA and the NVS, low health literacy is defined based on cutoff scores from the TOFHLA development sample. In that study, the TOFHLA was administered to a large number of people; those who scored below the average were considered to have "Inadequate" health literacy, while those with other scores were considered as having either "Marginal" or "Adequate" health literacy. A key concern in this study is that the way low health literacy was defined was not related to an external criterion. The REALM was developed in comparison to a standard reading measure.

In defining low health literacy, we looked at several different ways of deciding who has adequate or low health literacy. Since all participants completed teh TOFHLA, it was possible to decide who had low health literacy based on the traditional cutoff scores used for that measure. FLIGHT/VIDAS participants also completed an individually-administered and well standardized and normed measure of reading comprehension. While general reading comprehension isn't the same thing health literacy, it seems likely that persons with low levels of academic literacy (below the 8th grade level, for example) might have difficulties in the skills component of the ASK model (see Chapter Two). We thus also explored the extent to which FLIGHT/VIDAS scales can identify individuals who scored below the 8th grade level on this general reading comprehension measure. Finally, we asked participants questions about their level of difficulty in understanding written health information. Those who reported having difficulty in understanding written health information were considered

The cutoff scores used in the National Assessment of Adult Literacy were developed based on analysis of the tasks required in answering the items on that measure. We followed this procedure in examining and categorizing the demands of items on FLIGHT/VIDAS, categorizing items

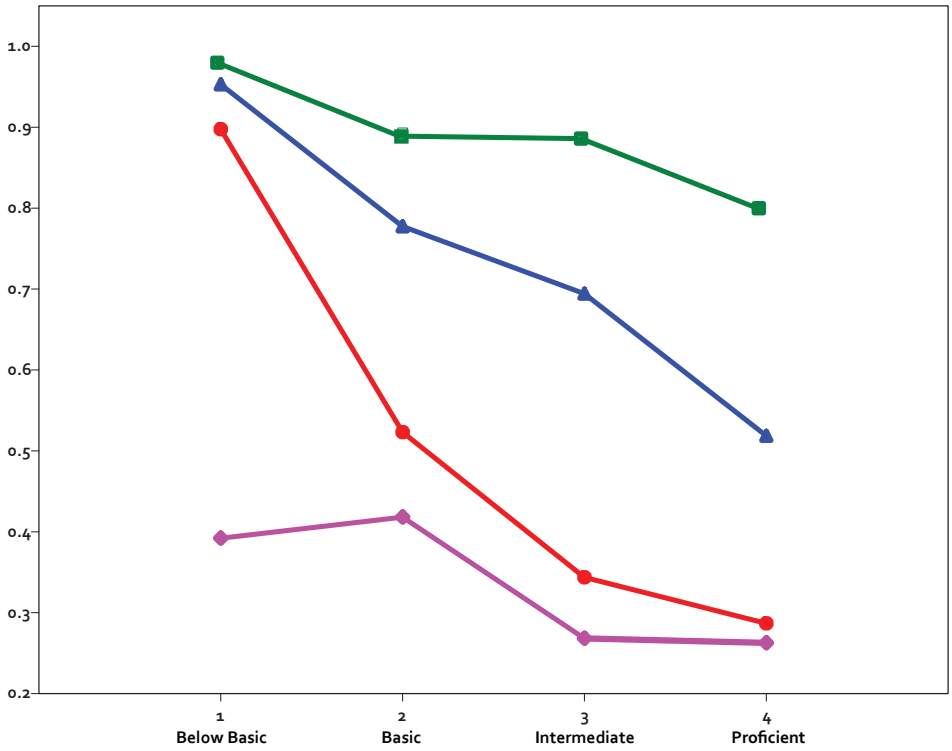

Figure 4-3. Median proportion of items from each category answered correctly by individuals in each group. Group: Purple = Low Health Literacy; Red = Basic; Blue = Averagee; Green = Proficient

as requiring Below Basic, Basic, Intermediate, or Proficient skills. Items in each of the categories were scored, and participants' scores on each group of items were used in a latent class analysis to identify subgroups of participants based on their performance on each group of items (Ownby, Acevedo, Waldrop-Valverde, manuscript in preparation).

We found that some persons had difficulty with nearly all items on all scales, but were more likely to answer the questions in the Below Basic and Basic groups. This group, which had difficulty with even basic and straightforward health literacy tasks, is referred to as having Low Health Literacy. A second group performed fairly well on the Below Basic group

of questions and moderately well on the Basic items, but less well on Intermediate and Proficient questions. These persons can be characterized as having Basic health literacy. A third group did well on items in the Below Basic, Basic, and Intermediate groups of items but not on the Proficient items. These individuals can be characterized as having Average health literacy. Finally, a fourth group performed well on all groups of items and are clearly Proficient.

In examining each group's performances, it became evident that the health literacy of members of the groups could be described in terms of the kind of health literacy tasks the group's members could and could not perform. Of individuals in the Low Health Literacy group , 86% correctly answered a question about the number of persons missing in a picture and 32% correctly answered a question that required a straightforward application of information from a table of body mass index values. Of per-

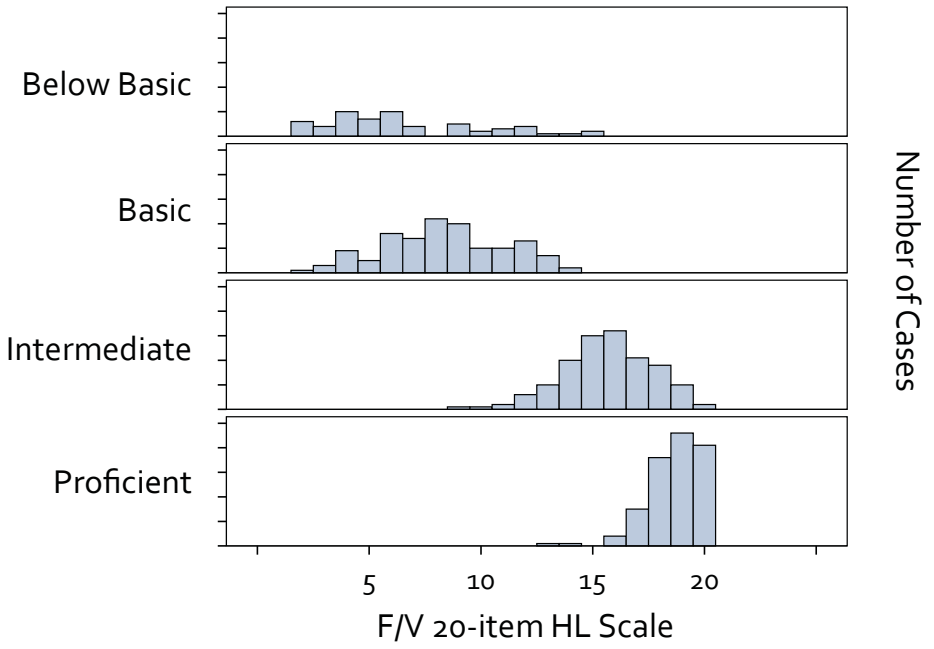

Figure 4-4. Distribution to F/V scores in health literacy proficiency groups.

sons in this group, however, only 45% could correctly answer a question involving mental calculation related to grams of carbohydrate when given information in a simple table. These individuals' performance on an academic reading measure (Woodock-Johnson or Woodcock-Muñoz Passage Comprehension test) was at the late primary grade level (4th-6th grades).

Individuals in the Basic group were likely to answer the questions ranked as below basic (see Figures 4-3 and 4-4) but were less likely to answer more complex questions even when they required extraction of single pieces of data from a table or simple but several stage arithmetic calculations (such as calculating the grams of fat even from several food sources). These individuals' performance on an academic reading measure was at the late primary grade level (about 6th grade).

Individuals in the Average group could perform tasks of moderate complexity, such as extracting relevant information from a prose paragraph or completing an arithmetic calculation that involved several steps. Their performance on the academic reading measure was at about the 8th grade level.

Finally, individuals in the Proficient group were consistently able to extract and use information from complex prose or document sources, and could assess probability even in a complex scenario involving competing risks of treatment outcomes and side effects. This group's performance on the academic reading measure was at the 12th grade level.

Summary

Overall, results of assessments of F/V reliability and validity show that it has adequate reliability and validity for use as an assessment of health literacy. It is particularly noteworthy that the F/V 20 item scale is shorter than the TOFHLA but performs as well with respect to identifying persons with low levels of literacy. Critically important is the wide range of difficulty in F/V scores, eliminating the problem of ceiling effects that plagues most other measures of health literacy.

5 How to Use FLIGHT/VIDAS

Using FLIGHT/VIDAS

As explained in previous chapters, the extensive item development process completed in the F/V project allowed us to identify a number of scales that can be used for different purposes (e.g., diagnostic assessment or screening) and to assess different aspects of health literacy (prose and document, numeracy, listening, conceptual health knowledge). If you're interested in using F/V, this chapter should help you decide which scale is best for your purpose.

The F/V scales can be administered in one of two ways (see Table 5-1, next page). Computer administration requires a tablet or desktop computer and an Internet connection. Paper and pencil administration requires the testing materials (see the F/V website at www.flightvidas.org for information about how to obtain the test). You should read the questions (but not much of the test of the question) to the person being assessed since that it how the computer did it in the process of creating F/V.

In our experience, most people will want to use a measure of health literacy for one of the following reasons:

Screen Someone for Low Health Literacy

If this is what you want to do, use the 10-item screener. This will take 10-15 minutes, and you can score it as you go along. The answer sheet has

Table 5-1. Available F/V scales

Scale	Access	Items	Time	Purpose
Paper	Paper	20	15-20	Assess general health literacy
Screen	Paper	10	10-12	Screen for low health literacy
HL40	Computer	40	20-30	Diagnostic assessment of prose and document health literacy
NUM	Computer	29	20-30	Diagnostic assessment of numeracy skills
LIS	Computer	13	15	Assess listening skills in simulated clinical encounter, patient education, and research settings
FACT	Computer	14	15	Assessment of conceptual health knowlege

Information on how to access computer scales and to obtain test materials for paper and pencil scales is available at www.flightvidas.org.

a table to let you know how the person performed on the measure, and whether they are likely to have low or basic health literacy skills.

The 10-item screener is also available for computer administration. You can find out how to access the computer version on our website at www.fligthvidas.org. The computer version will generate a score and interpretation, but requires a computer or tablet and an Internet connection.

Normative data for the screenint tool is availabe in the form of T scores (see Table 5-4 and 5-5). The scores are based on the average performance of our participants in each age group so that you can understand their performance in relation to their age. It's important to note that these scores

Table 5-2. T Score Equivalents of Raw Scores on the English 20-item HL Scale

Raw Score	18-29	30-39	40-49	50-59	60-69	70 and older
1	2	14	26	18	22	22
2	5	17	28	20	25	24
3	8	19	30	23	27	26
4	11	21	32	25	29	28
5	14	24	34	27	31	30
6	17	26	36	30	33	32
7	21	28	37	32	35	34
8	24	31	39	34	37	36
9	27	33	41	37	39	39
10	30	35	43	39	41	41
11	33	38	45	42	44	43
12	36	40	47	44	46	45
13	39	42	49	46	48	47
14	42	45	51	49	50	49
15	45	47	53	51	52	51
16	48	49	55	53	54	53
17	51	52	57	56	56	55
18	54	54	58	58	58	57
19	57	56	60	60	60	59
20	60	59	62	63	63	61

are not adjusted for participants' level of education. This means that interpreting the score should also take into account someone's level of education. These scores can help you understand how a person's level of health literacy compares to other people in the same age group. If you are only interested in knowing whether the person has low, basic, or higher health literacy, you should use the cutoff scores on the answer sheet.

Assess Someone's Health Literacy

You may be interested in doing a more in-depth assessment of a person's health literacy in a way that takes no more than 20-30 minutes. For this purpose, use the 20-item health literacy scale. It will give you information about the person's overall level of health literacy, a more precise estimate of the person's needs with respect to written health communications, and examination of specific items will give you insight into the person's strengths and weaknesses in understanding health information. This scale includes a number of items that require numeracy skills.

Normative data is available for this scale, too (see Tables 5-2 and 5-3). These data provide T scores that will help you understand how well a person scores on the test compared to other people in their age group.

This scale is available in a paper and pencil version. As with the screener, the paper and pencil version should be administered according to the instructions in the test booklet, with portions of the questions read aloud to the person assessed. This scale is also available for computer administration. As with the screener, you can find out how to access this scale on our website at www.fligthvidas.org. The computer version will generate a score and interpretation, but requires a computer or tablet and an Internet connection.

Do Research on Health Literacy

The F/V project developed data on a longer 40-item general health literacy scale as well as the numeracy (NUM), Listening (LIS), and conceptual health knowledge (FACT) scales described in Chapter 3. Each of these

Table 5-3. T Score Equivalents of Raw Scores on the Spanish 20-item HL Scale

Raw Score	18-29	30-39	40-49	50-59	60-69	70 and older
1	12	13	21	27	33	35
2	15	16	23	29	35	37
3	17	18	25	31	37	40
4	20	21	28	33	39	42
5	23	23	30	35	41	44
6	25	26	32	37	43	47
7	28	28	34	39	45	49
8	31	31	36	40	47	52
9	33	33	39	42	49	54
10	36	36	41	44	51	56
11	39	38	43	46	53	59
12	41	41	45	48	55	61
13	44	43	47	50	57	64
14	46	46	50	52	59	66
15	49	48	52	54	61	68
16	52	50	54	56	63	71
17	54	53	56	58	65	73
18	57	55	58	60	67	76
19	60	58	61	62	69	78
20	62	60	63	64	71	80

scales is currently only available for computer administration, although a paper and pencil version of the FACT scale is currently being developed. For information on how to access and use these scales, please contact us via the F/V website at www.flightvidas.org.

Use F/V Items to Assess Another Aspect of HL

Because of the diverse content and formats of the F/V items, several interested researchers have used them to develop scales to evaluate specific aspects of health literacy for specific purposes. One of the F/V team, Josh Caballero, PharmD, used F/V items to create a scale relevant to medication management[ref]. It can be used in medication management counseling. Another person is working with us in creating a scale that can assess aspects of health literacy relevant to pediatrics.

Summary

This chapter has briefly described how to use F/V for the most common purposes we've encountered. Please contact us for additional information about using F/V or for any questions. The most up to date information on F/V is available at www.flightvidas.org.

Table 5-4. T Score Equivalents of Raw Scores on the English 10-item Screening Tool

Raw Score	18-29	30-39	40-49	50-59	60-69	70 and older
1	10	16	30	25	26	26
2	15	21	33	29	30	29
3	21	26	37	33	34	33
4	26	30	40	37	38	37
5	32	35	44	41	42	41
6	37	39	47	45	45	45
7	43	44	51	49	49	49
8	48	49	55	52	53	52
9	53	53	58	56	57	56
10	59	58	62	60	61	60

Table 5-5. T Score Equivalents of Raw Scores on the Spanish 10-item Screening Tool

Raw Score	18-29	30-39	40-49	50-59	60-69	70 and older
1	11	18	20	30	34	39
2	16	23	24	33	38	43
3	22	27	29	37	42	47
4	27	32	33	40	46	52
5	33	36	38	44	50	56
6	38	40	42	47	54	60
7	44	45	47	51	58	64
8	49	49	52	54	62	68
9	55	54	56	58	66	73
10	61	58	61	61	70	77