HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2018 August 08.

Published in final edited form as:

Stud Health Technol Inform. 2017 ; 240: 17–33.

The State of the Science of Health Literacy Measurement

Tam H. Nguyen[a], **Michael K. Paasche-Orlow**[b,1], and **Lauren A. McCormack**[c]

[a]William F. Connell School of Nursing, Boston College, Chestnut Hill, Massachusetts

[b]Boston University School of Medicine, Division of General Internal Medicine, Boston, Massachusetts

[c]RTI International, Division of Public Health Research, Research Triangle Park, North Carolina

Abstract

Advancing health literacy (HL) research requires high-quality HL measures. This chapter provides an overview of the state of the science of HL measurement: where the field started, currently is, and should be going. It is divided into eight key sections looking at (1) the history of HL measurement, (2) the relationship between HL definitions and measurement, (3) the HL conceptual domains most and least frequently measured, (4) the methods used to validate HL measures, (5) the characteristics of the participants in the measurement validation studies, (6) the practical considerations related to administering HL measures, (7) the advantages and disadvantages of using objective versus subjective HL measures, and (8) future directions for HL measurement.

Based on the material presented in this chapter, the following conclusions can be drawn. First, there is an enormous proliferation of HL measures and this growth presents both opportunities and challenges for the field. Second, to move the field forward, there is an urgent need to better align HL measurement with definitions of HL. Third, some HL domains, such as numeracy, are measured more often than others, such as speaking and listening. Consequently, it is important to think about novel mechanisms to measure HL domains that are rarely measured. Fourth, HL measures are most often developed, validated, and refined using classical measurement approaches. However, strong empirical and practical rationales suggest making an assertive shift toward using modern measurement approaches. Fifth, most HL measures are not well validated for use in minority populations; consequently, future validation studies should be mindful of validation samples. Sixth, HL measures can be administered using multiple modes, most frequently via paper-and-pencil surveys. Identifying which mode of administration is most suitable requires reflecting on the underlying measurement purpose and the characteristics of the participants being measured. These considerations should also be made when deciding between a subjective versus objective HL measure.

Cumulatively, this chapter provides tools to help readers select and use the most appropriate measures of HL for their needs. It also provides rationale and strategies for moving the science of HL measurement forward.

[1]Corresponding author: Boston University School of Medicine, 801 Massachusetts Avenue, 2nd Floor. Boston, MA 02118, USA; tam.nguyen@bc.edu.

Keywords

Health literacy; measurement; conceptual domains; validation; psychometrics

> "We owe all the great advances in knowledge to those who endeavor to find out how much there is of anything."
>
> — James Maxwell, Physicist (1831–1879)

1. Overview and History of Health Literacy Measurement

The 2004 seminal Institute of Medicine report *Health Literacy: A Prescription to End Confusion* identified the development of new measures of health literacy (HL) as a key priority for the field [1]. Numerous scientific calls and proposals followed to develop and test HL measures in support of that recommendation [2]. Now, more than a decade later, over 150 HL measures exist, demonstrating a sweeping response to the scientific calls and reflecting tremendous productivity in this area [3–5]. This growth presents both opportunities and challenges for the field [6, 7]. In one respect, each of these measures provides a degree of utility and valuable lessons as the field moves forward. Despite these efforts, however, no "gold standard" measure for HL has emerged, and the variety of measures has made comparing results across studies and populations a serious challenge.

This chapter distills the state of the science of HL measurement by addressing several key areas. In section 2, we examine the relationship, or lack thereof, between definitions and measurement. In section 3, we summarize the conceptual domains most and least frequently measured, using the most current and well-accepted definitions and theoretical frameworks guiding the field. In section 4, we synthesize current methods used to validate HL measures and discuss their relative strengths and limitations. In section 5, we evaluate characteristics of the participants in the validation studies used to establish HL measures. In section 6, we discuss the practical considerations related to the various modes of administering HL measures. In section 7, we identify measures that assess HL subjectively versus objectively and discuss the relative advantages and disadvantages. Finally, in section 8, we provide a road map for the future of HL measurement.

The overarching aims for this chapter are to help readers identify factors that contribute to stronger HL measures, provide tools to help readers select and use appropriate HL measures, and put forth a rational strategy for advancing HL measurement. An inventory of all the existing HL measures available has been omitted intentionally as an easy-to-use, publicly available Health Literacy Toolshed (www.healthliteracybu.edu) serves this purpose.

2. Relationship between Definitions and Measurement

In general, the HL field has promulgated separate discussions regarding the *definition* of HL and the *measurement* of HL [7]. Despite promising recent work developing tools in tandem with definitions, the disconnect between definitions and what the tools measure has been a persistent conceptual stumbling block, which has led to several conundrums that will need to be solved for the field to progress in a coherent manner.

Multiple definitions of HL exist; however, some are vague or inadequately specified to allow measurement. There are even more tools to measure HL, but many are loosely related to a definition [3–5, 8]. This section examines the interplay between HL definitions and measurement from a legacy perspective verses more recent HL tools. It also looks at the paradox and barriers that exist to more fully integrating HL definitions with measurement.

2.1. Legacy Tools

The large majority of empirical HL research has used the Rapid Estimate of Adult Literacy in Medicine (REALM) [9] and the Test of Functional Health Literacy in Adults (TOFHLA) [10], or some variant of these tools. These instruments, however, only align superficially to a definition of HL.

The REALM is a word pronunciation test that uses medical words, an extremely narrow lens through which to view the concept of HL phenomena. In fact, in English, because of the high level of grapheme-phoneme discordance, the REALM provides valuable information; for example, it is quite difficult to pronounce a term like *vitiligo* without any prior familiarity with it. Consequently, as pronunciation in English incorporates a vague notion of understanding, the REALM is slightly more complex than initially apparent. Yet, words can certainly be pronounced without being understood.

Alternatively, the full TOFHLA includes reading, numeracy, and document literacy, and the modified cloze approach to ensure that the TOFHLA tests a person's understanding. TOFLHA takes a broader view of HL, but with distinct limitations. For example, the TOFLHA numeracy testing items require reading skills, making it quite difficult to disentangle numeracy dimensions from reading. Over the span of many projects conducted using the TOFHLA, presentation of differential results for the three subscales is incredibly rare. It is possible that users did not consider examining the separate scales in their projects or that the results were always consistent, making the presentation of separate analyses uninteresting. It is more likely, however, that the test scales are insufficiently distinct at the fundamental level of the cognitive processes involved.

2.2. A Way Forward

A portion of the newer instruments that measure HL have been developed in a manner that was explicitly linked to a specific definition and theory. Investigators for tools such as the Health Literacy Skills Instrument (HLSI) [11] and the Numeracy and Understanding in Medicine Instrument (NUMI) [12] approached their work by using an explicit operational HL definition that would motivate the purpose and scope of the tool. All items were designed to fulfill these specifications. Consequently, for these instruments, it is clear how to map results from test items to HL dimensions.

Interestingly, this has presented a paradox. Although the most commonly used legacy instruments are not based on a particular definition and relate only in general terms to the concept of HL, they nonetheless have demonstrated a high volume of predictive validity. However, newer tools developed in explicit relationship to a specific HL definition have not *yet* demonstrated predictive validity. The lack of extensive evidence exhibiting predictive validity has, in turn, caused some users in the field to be reticent about shifting to the use of

newer HL tools. Until there is a shift to using newer tools, the field will not be able to advance and determine, for example, which HL dimensions are critical and in which contexts.

Overall, the development of tools aligned with specific HL definitions has begun. However, now these tools will need to be used to help refine observational research and guide efforts to design interventions that align more specifically with the challenges faced by people with limited HL.

3. Conceptual Domains Measured and Those Rarely Measured

Conceptual frameworks can help formulate research questions and examine relationships among predisposing variables, mediators, moderators, and other relevant outcomes [13, 14]. Frameworks can also inform an understanding of the domains that comprise a complicated construct. Similar to the fact that the term “health literacy” is widely used but not always well understood or applied consistently—as suggested by the sizeable number of different definitions that exist [8, 15]—multiple HL conceptual frameworks have been put forth [8, 16–18]. Yet, no single framework has gained significant traction or is viewed as the gold standard.

Examining the conceptual domains included in existing HL measures can offer insight into the construct as a whole. The National Assessment of Adult Literacy (NAAL) [19], for example, included three broad domains when examining HL using a more skills-based approach than had been used previously. Specifically, the NAAL used health-related stimulus materials that reflect the type of materials adults encounter in real life to examine three domains: (1) prose literacy, measured as the knowledge or skills needed to search, comprehend, and use information from text organized into sentences or paragraphs; (2) document literacy, defined as the knowledge and skills needed to search, comprehend, and use information from noncontinuous text in various formats (such as job applications, payroll forms, transportation schedules, and maps); and (3) quantitative literacy, measured as the knowledge and skills needed to identify and perform computations using numbers embedded in print materials (such as balancing a checkbook, calculating a tip, or figuring the amount on an order form).

The Health Literacy Toolshed website launched in 2015 and houses over 120 instruments to measure HL, with a goal of increasing the number of instruments available over time. The Toolshed developers organized the initial set of instruments into the following domains: prose (both pronunciation and comprehension); numeracy; communication (both speaking and listening); information seeking in documents, in addition to interactive media navigation; and skills related to the application and function of health information.

Most of the instruments include more than one domain when measuring HL, illustrating the overall complexity of the construct (Table 1). About half of the instruments included the numeracy domain, suggesting a general consensus that it is a core HL component, although instruments that measure only numeracy also exist. A lower number of instruments include a pronunciation domain, reflecting some of the more historical approaches to measuring

literacy and perhaps indicating a movement away from this measurement approach. The communication components of HL—including both listening and speaking—are rarely included.

A small but growing number of instruments (n=19) include an application or functional component. About one quarter (n=26) examine information seeking in documents, and fewer (n=13) examine information seeking via interactive media navigation, including websites. Taken together, the relatively large number of domains used across the array of existing instruments to measure HL confirm that it is a multidimensional construct that must be measured carefully and completely using multiple items and stimuli in which a user can demonstrate their ability to interact effectively.

4. Limitations of Validation Methods

The limited utility and predictive validity of some HL measures may reflect the methods used to develop and validate them. Consequently, exploring how existing measures were validated will provide insight into potential limitations and directions for future work.

At its core, measurement consists of rules for assigning numbers to objects, or concepts, in such a way as to represent quantities of an attribute [14]. The term “rules” indicates that the method of assigning numbers to attributes must be stated explicitly. The construction, scoring, refinement, and validation of latent scales are most commonly guided by psychometric methods associated with Classical Test Theory. However, Modern Measurement Theories offer practical solutions for measurement problems found in health-related research that have been difficult to solve using classical approaches [20–22]. This section first examines the advantages of Modern Measurement Theory, then reviews methodological approaches (i.e., Classical Test Theory vs. Modern Measurement Theory) used to develop and validate current HL measures, and finally suggests approaches for moving HL measurement forward. For brevity, it is assumed that most readers have a basic understanding of Classical Test Theory. A short description of Modern Measurement Theory is provided. Resources are available for those who wish to learn more about both approaches [14, 23].

4.1. Advantages of Modern Measurement Theory

Modern Measurement theories include Item Response Theory (IRT) and Rasch modeling. IRT, which focuses on the item-level rather than the scale-level, is a general statistical theory that uses mathematical models to describe the relationship between an individual’s trait level and how they *respond* to an *item* [23]. This relationship can be described using two main parameter estimates: the discrimination parameter and the location/difficulty parameter. The discrimination parameter, often denoted ***a***, reflects the ability of an item to discriminate between different levels of underlying traits; higher ***a*** values indicate better discrimination. The main difference between IRT and Rasch modeling is that the discrimination parameter across all items is set to the same value when using Rasch models, whereas this parameter is allowed to vary by item when using IRT models. While Rasch modeling provides stronger measurement properties, the fit of real-life data to Rasch models is not often suitable. The location/difficulty parameter, often denoted *b*, indicates the location of the item on the

underlying construct; higher *b* estimates indicate that greater amounts of the underlying trait are needed to answer the question correctly (i.e., harder questions).

The advantages that IRT and Rasch modeling confer over Classical Test Theory are well documented [20, 24, 25]. In particular, there are four key advantages to using these methods to construct and refine HL measures [26]. First, by evaluating the location/difficulty parameter estimates across all items, IRT provides the opportunity to examine the level of HL skills measured (e.g., low, medium, high) and where efforts to develop new items should be focused. Second, the precision, or reliability, of measurement tools can be more accurately modeled using IRT. Specifically, instead of assuming that a tool has equal reliability across the trait continuum (i.e., a Cronbach's alpha=0.98), IRT can be used to identify the variability in measurement precision for individuals of differing trait levels of HL. This can be done by evaluating the test information function curve, which ideally should take on the shape of a horizontal line and be associated with a low standard error value. Third, an underlying assumption of IRT and Rasch models is that the estimated item parameters values (i.e., ***a*** and ***b***) should be consistent for different groups, such as females or males (i.e., population invariance). This is in contrast to Classical Test Theory where scale properties are sample dependent. Although strong evidence supports this property, it does not hold in all cases [20, 24]. When estimated item parameters are different across groups after controlling for ability, an item is considered to have differential item functioning (DIF). IRT-based analyses can help identify items with DIF that may need to be rewritten or excluded. Additionally, even if significant DIF has been identified in certain items, those items can be retained if a model that incorporates the identified DIF is used to mathematically correct for item bias when estimating scores. Fourth, IRT can be used to build and validate item banks, which can subsequently facilitate computer adaptive testing (CAT). IRT does this by calibrating all items within a bank onto the same underlying trait scale. Once items are mapped onto a common scale, it does not matter that different people take different sets of test items.

Given that there are now over 150 HL tools, this can help address the lack of standardization in the measurement of HL and facilitate the comparison of scores and results across studies [5]. Because of these measurement properties, an assertive shift toward these methods would be highly advantageous.

4.2. Methods Used to Develop and Validate Health Literacy Measures

Consistent with the pattern observed in the larger scientific community [22], an analysis of HL measurement validation studies found that the methods used to develop and test HL measures were primarily guided by Classical Test Theory [26]. Specifically, among the 109 measures identified by Nguyen et al. [5], 88% (n=96) used Classical Test Theory and 12% (n=13) used IRT or Rasch modeling [11, 12, 27–37]. For a more up-to-date list of HL measurement tools that used modern methods for validation, see the HL Tool Shed and use the "Modern Approach for Tool Development" option to filter the list accordingly.

Among the measures that used IRT or Rasch modeling, most (n=9) [11, 28–32, 36, 37] used data from estimated parameter values to strategically eliminate items that had low discrimination and items that targeted the same (difficulty) level of the underlying trait.

When reviewing the range in trait levels across the items on measures where the ***b*** parameter estimate was reported, items ranged from ***b***=−6.34 to 2.06 [26]. However, when examining the density of items across this range, most items clustered toward the lower difficulty ranges [26]. This provides strong empirical evidence for developing more items in the higher difficulty ranges.

Seven HL measures reported assessing for DIF across various groups. Three HL item banks were identified in the literature, each with varying levels of complexity and domains of HL measured [12, 27, 37]. Among the three item banks, only one uses CAT; however, it is proprietary [27].

4.3. Strengthening the Validity of Health Literacy Measurement

The vast majority of existing HL measures uses Classical Test Theory to construct and validate scales. While this method has led to useful HL measures, there are limitations to relying heavily on this approach; notably, the lack of item-level data to meaningfully assess the difficulty of items in a scale across the latent trait, the need to revalidate measures when using them in different populations, and the challenge of comparing results across studies that used different measures. When evaluating HL measures that use Classical Test Theory, examining their reliability and validity estimates (i.e., psychometric properties) can be used to judge their strength. Commonly used reliability and validity estimates and their interpretation have been well summarized [14, 38]. Readers can then search the Health Literacy Toolshed to compare and contrast the psychometric properties of existing HL tools. It is worth mentioning that tools are valid to the extent that they are consistent (i.e., reliable) and useful in uncovering relationships (i.e., concurrent and predictive validity). In other words, for an HL tool to be “strong,” it should be both reliable and valid. A systematic literature review by Nguyen et al. [5] found that the evidence supporting the validity of HL tools was weaker than the evidence supporting reliability, which reinforces that caution should be taken when using tools that are developed, refined, and validated using Classical approaches.

Because of these limitations, assertively shifting toward Modern Measurement approaches would be highly advantageous. Early efforts to use this approach have yielded valuable insight. However, to date, the application of Modern Measurement among HL measures has not fully leveraged the advantages of this methodological approach. Building item banks that include an equal density of questions across a wider range of HL trait levels that can be used in CAT applications will strengthen this body of literature. Items included in test banks should ideally demonstrate adequate discrimination parameter estimate values (i.e., $a>1$). Including DIF free items will also improve measurement validity across more populations.

It is important to note that Modern Measurement approaches will not solve all of the issues in HL measurement. Ongoing work is critically needed to refine and align the definition of HL within a conceptual framework and to accurately measure the concept. Furthermore, expanding the focus of HL measurement into the healthcare context (i.e., the communication skills of providers and the complexity of health systems and public health systems) is an important and necessary evolutionary step. While expanding the consideration of HL into these arenas will likely be complex, understanding these elements will not only help move

the HL field forward but also will provide critical insight into how IRT or Rasch models can inform measurement development and refinement. Efforts toward these achieving these goals will necessarily require strategic collaboration. Additionally, designing high-quality mixed-methods research studies that meaningfully integrate qualitative and quantitative findings will be essential.

5. Limitations of Validation Samples

Given that the vast majority of HL measures were validated using Classic Test Theory, examining the samples from which they were validated is necessary to understand how these findings may or may not be reasonably generalized. It is easy to forget about such limitations and then make inaccurate conclusions. For example, it would not make sense to develop and validate a tool exclusively with female participants and then to assume that it will perform the same way with males.

A systematic review of HL measures examined the racial and ethnic composition of participants in validation studies for 109 tools to measure HL [5]. Of the 72 English-language measures examined in this review, 17 did not specify the racial/ethnic characteristic of their sample. Of the remaining 55 measures, 10 (18%) did not include blacks, 30 (55%) did not include Hispanics, and 35 (64%) did not include Asians in their validation sample. When Asian Americans and Hispanic Americans were included, they accounted for small percentages and numbers in the overall sample; interquartile range=10%–34% (n=13–154) and interquartile range=3.5%–16% (n=5–36), respectively.

Consequently, it is likely that inappropriate assumptions have been made when using these tools in other contexts. Additionally, the nature of the bias introduced by such assumptions cannot be estimated. For example, if a tool misclassifies Hispanics because it was developed with a sample that did not have enough Hispanic participants to ensure validity, subsequent analyses for Hispanic participants in studies using this tool could be misinterpreted. Therefore, it is important to interpret much of the HL literature with caution. If a classical approach is used to validate future HL measures, it is imperative that sampling strategies reflect the needs of high-risk groups. Correspondingly, the characteristics of the sample should be described in sufficient detail because this information has implications for the generalizability of a given measure.

Among the 37 non-English-language measures, only two specified the racial/ethnic characteristics of their sample beyond simply describing the general population in which the measure was being validated. For instance, Ko et al. [39] specified that the sample used to validate their “Health Literacy Test for Singapore” was 52% Chinese, 22% Malay, 24% Indian, and 10% Other. In comparison, most other non-English-language measures were similar to the “Hebrew Health Literacy Test,” which reported simply that 119 Israeli participants were sampled to validate their scale; it is unclear which ethnolinguistic groups were represented from this highly multicultural society. The ethnolinguistic and cultural diversity of a specified population will influence the extent to which this may be problematic. For example, this issue is less relevant for Korean-language HL measures because the population is relatively linguistically homogenous, whereas it may be a greater

concern for the use of Hebrew-language or Hindi-language HL measures, as these populations are more linguistically and culturally diverse. Future validation efforts should take into account the cultural diversity of the target population and include such details in reports.

Using Modern Measurement approaches to validate HL measures will reduce some of the challenges associated with the need to revalidate measures when used in different populations. While this property of population invariance found in modern approaches is robust, it does not always hold across all items and populations. Consequently, it will remain important to characterize validation samples to allow for DIF testing. When using Modern Measurement approaches, however, the sampling goal should be to obtain an equal distribution of participants with varying abilities across a latent trait; for example, the sample should include a sufficient number of participants with low, medium, and high levels of HL. This will lead to more stable parameter estimates for each test item. Once the parameters are estimated using an ideal "reference" group, differential item functioning can be tested against any number of different "focal" groups. Additionally, when DIF is identified (e.g., by race), a particular item may be excluded or a correction factor can be introduced. For tools developed with classical methods, if an item does not work the same across groups of participants, investigators do not have recourse. Such results cannot be corrected *post facto*. Consequently, it would be better to extend the validation cohorts for these tools or to abandon these tools for new data collection *ex ante*. Racial and ethnic data for the validation samples for each tool listed in the Health Literacy Tool Shed can be reviewed by choosing the "Read all details" option for any specific tool.

6. Practical Considerations When Using Health Literacy Measures

Researchers and practitioners use HL measures for various reasons, including patient-level assessment, intervention activities, and surveillance. Each situation may invoke the need for a different type of HL tool. For example, clinicians and healthcare professionals may want to assess the HL level of a sample of their patients to understand the general needs of the population they serve, or they may want to assess all new patients and need a tool that is easy to implement in a clinical setting (though clinical screening has not been shown to benefit patients) [40]. Also, researchers may be implementing a public health or community-based intervention and need to measure HL before and after implementation, or use HL as a control variable in an analysis examining a specific health outcome. In both clinical and research settings, HL measurement may be used to trigger specific interventions or to ascertain the possible differential impact of various interventions across the HL strata. In some instances, large-scale periodic HL assessments are conducted at the health system level or even at the national level.

Selecting the right HL tool is critical because different HL measures and different data collection strategies may be needed in a given situation. For example, what is the age and racial/ethnic diversity of the target population, what languages are spoken in the target population, what resources are available, and what are the measurement goals? These factors may influence the data collection method used and should be considered carefully at the outset of any program and in conjunction with the decision about which HL tool is used.

Data collection methods range from mail, telephone, web-based or computer-based, mobile device-based, to in-person, or some combination thereof. In-person data collection could use paper-and-pencil or computer self-report, interviewer-facilitated approaches, or face-to-face verbal communication. Based on the 128 measures currently in the Health Literacy Tool Shed, more measures were developed and validated using paper-and-pencil and in-person strategies. A limited number of measures have been validated for web-based data collection, and very few have been validated for telephone administration (Table 2).

Each mode of administration offers strengths and limitations that are also related to the design of the project or study. For example, with web-based data collection, tools can include visual and/or interactive stimuli as part of the HL assessment process, such as food labels, health insurance forms, or health-related websites. With computer-based data collection and computer-assisted telephone interviewing (CATI), responses are recorded automatically into a dataset for analysis and interpretation, making data entry unnecessary. Development and programming expenses reflect certain fixed costs for computer-based and telephone strategies, and items can be modified relatively easily. Print-based HL tools typically require fewer data-collection fixed costs but involve costs associated with mailing surveys to study participants and follow-up and data entry. Mail-only surveys generally have lower response rates, but they can be improved with telephone or other types of follow-up [41]. Neither mail nor telephone surveys allow for the use of interactive stimuli to assess HL. With mail surveys, participants can be asked to read and interpret text and visuals, which is not the case with telephone administration. Web-based and in-person data-collection modes can use aural approaches. Costs, including staff training, can vary greatly depending on the data-collection mode and instrument used. There can also be an impact on response rates and data quality depending on the data-collection method used [41].

Researchers and practitioners carefully consider the needs of the study population when measuring HL. Reading ability, visual and hearing abilities, computer skills, and access to computers should all be factored into the choice of an HL tool and addressed in data-collection planning. Staff training is required not only to ensure data quality, but also to reduce potential harms. For example, individuals collecting data should be sensitive not to give the impression of testing subjects, as this could promote shame or stigma, especially in lower HL populations [42–45].

In sum, researchers and practitioners may be interested in measuring HL for different reasons. Additionally, researchers need to be mindful of the increasing diversity within populations when measuring HL. Also, having a clear understanding of the measurement goal and the target audience will help identify the best data-collection mode and type of HL tool to use.

7. Subjective Versus Objective Health Literacy Measures

A complex phenomenon that has developed in HL measurement is the elaboration of objective versus subjective measures. In objective measurement, people are challenged by standardized test stimuli to measure an underlying trait; in subjective measurement, people

self-report their responses to questions about their experience, typically on Likert scales. There are distinct benefits and limitations to each of these approaches.

One benefit of subjective measures is the ease of testing because these measures do not require in-person testing and typically involve less cognitive effort than objective measures. This may mean that the risk for stigma is lower for subjective measures than for objective measures. Similarly, most HL measurement tools have been developed for research purposes, but some institutions have implemented subjective HL testing in clinical care. Subjective measures are typically easier to work into the flow of clinical care because they survey peoples' opinions. Also, subjective HL measures have the potential for rapid application. Indeed, some of the most commonly used subjective measures comprise three questions; some use just a single question [46, 47]. At the same time, more elaborate versions of subjective measurement have been developed. For example, the European Health Literacy Questionnaire was developed with 47 subjective items evaluating the three domains of healthcare, disease prevention and health promotion, and a four-component structure reflecting the four dimensions of accessing, understanding, appraising, and applying health information [48, 49]. It is possible that with repeated measurement over time, subjective measurements such as this could provide a different judgment by showing at a broad societal level to what extent the healthcare system is meeting the needs of the population.

The main challenge with subjective measurement is that there is no ground truth; meaning there is no way to know how a person's responses relate to their actual skill level. This is most relevant for certain groups of people who are likely to systematically rate their experiences at a higher level than other people in a manner that does not relate to their actual HL skills. For example, people who have not had much exposure to the health system may not appreciate the high degree of complexity they may encounter, so they may have inflated responses. Alternatively, for example, if male respondents have better scores on a subjective measure in a given project, it would not be clear if this is because men truly have an easier time with the activities being reported or if this difference reflects a subjective phenomenon in the cohort whereby the men in that cultural setting express a higher degree of self-confidence than the women [50]. In some cases, however, subjective measurement may provide the information that is needed. For example, subjective measurement is likely to be more successful in predicting outcomes for populations that have enough experience and enough insight in their HL ability. However, empirical testing is needed to support this perspective, and without additional data it is difficult to interpret results within a given cohort.

The main benefit of objective testing is that it results in a direct measure of the person's skill. There is an inherent value to having empirically grounded data. While this is often useful, there are multiple complexities with this approach. First, objective testing can feel like a test. People know that their skills are being evaluated; this can cause stigma, especially for people who struggle with the test items. Second, these tests typically require in-person testing. Third, the test items may not directly relate to the HL skills needed for a given scenario; a person's test score in one domain or content area may not reflect their skill in another aspect of HL. For example, it would be a mistake to assume that getting a perfect score on the TOFHLA means that a person knows how to use an inhaler. Lastly, given the

limitations in methodological approaches used to develop and validate most HL measures (i.e., Classical Test Theory vs. IRT), there are concerns around meaningful interpretation of scale scores. For example, under Classical Test Theory it is assumed that score intervals across the scale are equal (i.e., on a measure with scores that can range from 0–10, an individual who scores a 5 has half the ability of an individual who scores a 10). However, when applying IRT to test data, what is often revealed is that items cluster around certain areas of the latent trait; most often the middle region (i.e., items with moderate difficulty). Without modeling the density of items across a latent trait, the value of objective measurement is reduced; for example, if an individual's HL score improves from 5 to 8, and the 3- point gain comes from items of the same difficulty level, it would be important to ask how much did that individual's HL ability really improve.

The empirical relationship between objective and subjective HL testing has received limited attention [51]. Kiechle et al. reviewed papers that concurrently used both types of measures and related them to various outcomes. They identified four studies they rated to be fair-quality studies with pertinent data. Among these studies, one reported no difference between objective and subjective HL measures for a rheumatoid arthritis disease severity score; one showed no difference between objective and subjective HL measures for a range of self-reported disease states; one exhibited a difference between objective and subjective HL measures for a patient's ability to interpret their prescription medication name and dose from a medication bottle; and one provided mixed evidence about the consistency between objective and subjective measurements of numeracy for predicting colorectal cancer screening utilization. While insightful, these studies do not provide adequate reassurance to support the assumption that conclusions from objective and subjective measures could be interchanged. At a conceptual level, these tools measure different constructs. Though it would add a layer of complexity, ideally the HL literature that derives from objective measures should be interpreted separately from reports from studies that used subjective measures.

Overall, the choice to use an objective or subjective HL measure depends on the goals and structural parameters of the work. Currently, most phone-based survey research will need to use subjective measurement, as it is difficult to facilitate current objective tests over the phone. When the goals of testing relate to phenomena that are better served with objective testing, this should be done if feasible. Finally, objective tests (e.g., the NAAL) may be better suited for estimating an individual's skills, whereas subjective measures (e.g., the European Health Literacy Questionnaire) may be better suited to assess if the healthcare system is serving the population well.

8. Future Directions

Much has been learned from the progress made in HL tool development since the 2004 seminal Institute of Medicine Report. This chapter has provided a critical review of the state of the science of those measures across several dimensions, including (1) the relationship between HL definition and measurement, (2) the conceptual domains of HL most and least frequently measured, (3) the methodological approaches used to develop and validate HL measures, (4) the characteristics of the participants in the validation studies, (5) the practical

considerations when using HL measures, and (6) the use of subjective versus objective HL measures. Important patterns emerged from this critical review that can be used to help set future directions for HL measurement.

First and foremost, we need to better align HL measurement with definitions of HL. Given the number of different HL definitions that exists, it is imperative that tool developers are clear about what definition they are using and how that definition influenced the operationalization of HL in the tool. This will not only help guide the purpose and scope of the tool, but also help end-users interpret scores. Additionally, as more people use tools guided by a clear HL definition, it will help the field to further identify which definitions and theoretical frameworks are useful for understanding the mechanisms through which HL operates and how they impact outcomes.

An evaluation of existing HL tools demonstrates that some HL domains are measured more often than others. Specifically, prose/pronunciation and numeracy are the most commonly measured domains, whereas listening and speaking are the least measured domains. Consequently, it will be important to think about novel out-of-the-box mechanisms to measure these rarely measured HL domains. Novel approaches could also be used to address common issues of shame and the lack of time that is often associated with HL measurement. An example of an innovative approach is the use of gaze tracking technology while participants read a standard document [52]. It is hypothesized that the gaze patterns of individuals with high HL differ from individuals with moderate and low HL. Ongoing research is testing this hypothesis.

When reviewing the methodological approaches used to develop, validate, and refine HL measures, classical measurement approaches dominate the literature. This is consistent with patterns seen in measurement development and validation for most other patient reported outcomes [22]. However, there are strong empirical and practical rationales for making an assertive shift toward using modern measurement approaches to develop, validate, and refine HL tools. It is critical that the refinement and alignment between the definition of HL and measurement comes first before the full benefits of Modern Measurement approaches can be realized. Once refining and aligning the definition of HL is achieved, there is a valuable opportunity to build a robust HL item bank for CAT applications given the number of HL measures that exists. Creating a robust item bank has tremendous potential for addressing the lack of standardization in HL measurement, reducing participant burden, and addressing the challenge of making comparisons across studies. Achieving this will require skillful coordination, cooperation, and political will among the developers of HL tools and key stakeholders.

As HL tools continue to be developed, validated, and refined, it is critical to be mindful of validation samples. In situations where classical approaches are used, it is imperative that populations at highest risk for low HL are included in the validation sample because this has implications for the generalizability of the tool. A review of most existing HL measures demonstrates that Hispanic Americans and Asian Americans are rarely included in the validation samples of English-language HL measures, despite the fact that these groups have among the highest rates of low HL. In situations where modern approaches are used, the

validation sample should aim to have an equal distribution of participants with varying abilities across a latent trait. This will ensure more stable item parameter estimates. Once a strong focal validation sample is obtained, testing for DIF across a number of different reference groups can be done.

Given the stock of currently available HL tools, a question commonly posed is, what tool should be used? Ultimately, the answer depends on the context of why HL is being measured. This chapter highlighted the differences between subjective and objective measures. Objective measures are generally better for situations when it is important to have a reliable estimate for an individual or set of individuals that are “grounded” in some verifiable way. Subjective measures are more strategically feasible for large-scale measurement of populations or systems over time. Various modes of measurement were also highlighted, including mail, telephone, web-based or computer-based, mobile device-based, in-person, or some combination thereof. Each mode of administration offers strengths and limitations. Identifying which administration mode is most suitable requires reflecting on the needs of the study population, including reading, visual and hearing abilities, and computer skills and access to computers. Considerations should also be made based on cost, desired response rate, and data quality. The Health Literacy Tool Shed is a valuable resource that can be used to filter existing tools based on their mode of measurement and whether they are subjective or objective in nature.

Finally, two parting reflections warrant brief mention. First, it is important to recognize that HL is a dynamic concept, and the rate at which this concept evolves is affected by language, culture, an increasingly global and mobile world, and sweeping health system changes taking place in many countries. Consequently, it will be necessary to have a more informed and sophisticated understanding of ethnolinguistic nuances and changes that occur naturally in most languages; particularly among languages spoken by people who are highly mobile and global. Ignoring these ethnolinguistic changes may decrease the content validity of HL measures over time. Likewise, to expand the focus of HL measurement into the healthcare context, a more informed and sophisticated understanding is needed of how health systems are evolving, as well as the complexities within and across different health systems that may influence HL and patient outcomes.

Second, many of the measurement-related issues identified in this chapter are not unique to HL. It is quite common for conceptual and operational definitions to evolve for emerging concepts of high scientific and social value. The increased recognition and inquiry often leads to a proliferation of new measures. However, even after several decades of study, this has not resulted in a gold standard.

For example, coping as a construct has a long history. Initial studies focused on psychopathology. Later, researchers moved toward positive behavior and the role of emotions. Significant concerns were identified in clarifying the concept and matching it with measurement [53]. After providing an updated review of the swiftly widening literature on stress and coping [54], the authors noted that measurement was still the most controversial issue in the field. One way researchers addressed this problem was to develop coping

measures specialized by situations; for example, coping with sexual trauma. While measures multiplied in number, clarity of the concept of coping in health did not advance.

Lengthy theses can be written to discuss solutions for overcome these challenges. Two broad suggestions are offered. First, it is vital to overcome disunity by moving toward unified, integrative work; both in conceptualizing the definition of HL and operationalizing it into measurement tools. Second, it is imperative that we continue to build from where we are toward higher quality psychometric studies that include both classic and modern measurement approaches. Continuing to engage in the status quo will do more harm than good for the field.

In closing, much progress has been made in HL tool development, which has led to a number of useful HL tools and to measurement challenges for the field. It is important to contextualize the commonality of the measurement problems unearthed in this chapter rather than be discouraged by them. Learning from past lessons provides a hopeful path forward.

Acknowledgments

We would like to thank Rebecca Moultrie for her assistance with Table 1 and Jeffrey Novey for his careful review of this chapter.

References

1. Nielsen-Bohlman LN, Panzer AM, Kindig DA, editorsHealth literacy: a prescription to end confusion. Washington, DC: National Academies Press; 2004.
2. National Institutes of Health, Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention. Understanding and promoting health literacy (R01). Washington, DC: National Institutes of Health, Agency for Healthcare Research and Quality, and Centers for Disease Control and Prevention; http://grants.nih.gov/grants/guide/pa-files/PAR-07-020.html. Retreived January 12, 2012.
3. O'Neill B, Goncalves D, Ricci-Cabello I, Ziebland S, Valderas J. An overview of self-administered health literacy instruments. PLoS One. 2014; 9(12)doi: 10.1371/journal.pone.0109110
4. Haun JN, Valerio MA, McCormack LA, Sorensen K, Paasche-Orlow MK. Health literacy measurement: an inventory and descriptive summary of 51 instruments. J Health Commun. 2014; 19(Suppl 2):302–33. DOI: 10.1080/10810730.2014.936571 [PubMed: 25315600]
5. Nguyen TH, Park H, Han HR, Chan KS, Paasche-Orlow MK, Haun J, et al. State of the science of health literacy measures: validity implications for minority populations. Patient Educ Couns. 2015; doi: 10.1016/j.pec.2015.07.013
6. McCormack L, Haun J, Sorensen K, Valerio M. Recommendations for advancing health literacy measurement. J Health Commun. 2013; 18(Suppl 1):9–14. DOI: 10.1080/10810730.2013.829892 [PubMed: 24093340]
7. Pleasant A. Advancing health literacy measurement: a pathway to better health and health system performance. J Health Commun. 2014; 19(12):1481–96. DOI: 10.1080/10810730.2014.954083 [PubMed: 25491583]
8. Sorensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health. 2012; 12:80.doi: 10.1186/1471-2458-12-80 [PubMed: 22276600]
9. Davis TC, Long SW, Jackson RH, Mayeaux EJ, George RB, Murphy PW, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument. Fam Med. 1993; 25(6):391–5. [PubMed: 8349060]

10. Parker RM, Baker DW, Williams MV, Nurss JR. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. J Gen Intern Med. 1995; 10(10):537–41. [PubMed: 8576769]
11. Bann CM, McCormack LA, Berkman ND, Squiers LB. The Health Literacy Skills Instrument: a 10-item short form. J Health Commun. 2012; 17(Suppl 3):191–202. DOI: 10.1080/10810730.2012.718042 [PubMed: 23030570]
12. Schapira MM, Walker CM, Cappaert KJ, Ganschow PS, Fletcher KE, McGinley EL, et al. The numeracy understanding in medicine instrument: a measure of health numeracy developed using item response theory. Med Decis Making. 2012; 32(6):851–65. DOI: 10.1177/0272989X12447239 [PubMed: 22635285]
13. Glanz K, Rimer BK, Viswanath K. Health behavior and health education: theory, research, and practice. 4. San Francisco: Jossey-Bass; 2008.
14. Nunnelly J, Bernstein IH. Psychometric theory. 3. New York: McGraw-Hill; 1994.
15. Berkman ND, Davis TC, McCormack L. Health literacy: what is it? J Health Commun. [Historical Article]. 2010; 15(Suppl 2):9–19. DOI: 10.1080/10810730.2010.499985
16. Manganello JA. Health literacy and adolescents: a framework and agenda for future research. Health Educ Res. 2008; 23(5):840–7. DOI: 10.1093/her/cym069 [PubMed: 18024979]
17. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav. 2007; 31(Suppl 1):S19–26. DOI: 10.5555/ajhb.2007.31.supp.S19 [PubMed: 17931132]
18. Squiers L, Peinado S, Berkman N, Boudewyns V, McCormack L. The health literacy skills framework. J Health Commun. 2012; 17(Suppl 3):30–54. DOI: 10.1080/10810730.2012.713442
19. Kutner M, Greenberg E, Jin Y, Paulsen C. The health literacy of America's adults: results from the 2003 National Assessment of Adult Literacy (NCES 2006-483), U.S. Department of Education. Washington, DC: National Center for Education Statistics; 2006.
20. Embretson SE. The new rules of measurement. Psychol Assess. 1996; 8(4):341–9.
21. Gulliksen H. Theory of mental tests. New York: Wiley; 1950.
22. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care. 2000; 38(Suppl 9):II60–5. [PubMed: 10982090]
23. Hambleton RK, Swaminathan H, Rogers WH. Fundamentals of item response theory. Newbury Park: Sage Publications; 1991.
24. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Instructional Topics in Educational Measurement. 1993:38–47.
25. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient. 2014; 7(1):23–35. DOI: 10.1007/s40271-013-0041-0 [PubMed: 24403095]
26. Nguyen TH, Paasche-Orlow MK, Kim MT, Han HR, Chan KS. Modern measurement approaches to health literacy scale development and refinement: overview, current uses, and next steps. J Health Commun. 2015; 20(Suppl 2):112–5. DOI: 10.1080/10810730.2015.1073408
27. ETS. [Retrieved September 23, 2016] Health Activities Literacy Scale. 2012. http://www.ets.org/literacy/about/content/health_activities_content
28. Guttersrud O, Dalane JO, Pettersen S. Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. Public Health Nutr. 2014; 17(4):877–83. DOI: 10.1017/S1368980013000530 [PubMed: 23472785]
29. Lee SY, Bender DE, Ruiz RE, Cho YI. Development of an easy-to-use Spanish health literacy test. Health Serv Res. 2006; 41(4 Pt 1):1392–412. DOI: 10.1111/j.1475-6773.2006.00532.x [PubMed: 16899014]
30. Lee SY, Stucky BD, Lee JY, Rozier RG, Bender DE. Short assessment of health literacy-Spanish and English: a comparable test of health literacy for Spanish and English speakers. Health Serv Res. 2010; 45(4):1105–20. DOI: 10.1111/j.1475-6773.2010.01119.x [PubMed: 20500222]
31. Leung AY, Cheung MK, Lou VW, Chan FH, Ho CK, Do TL, et al. Development and validation of the Chinese Health Literacy scale for chronic cre. J Health Commun. 2013; 18(Suppl 1):205–22. DOI: 10.1080/10810730.2013.829138 [PubMed: 24093357]

32. Nakagami K, Yamauchi T, Noguchi H, Maeda T, Nakagami T. Development and validation of a new instrument for testing functional health literacy in Japanese adults. Nurs Health Sci. 2014; 16(2):201–8. DOI: 10.1111/nhs.12087 [PubMed: 23991825]
33. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health. 2013; 13:658.doi: 10.1186/1471-2458-13-658 [PubMed: 23855504]
34. Sauceda JA, Loya AM, Sias JJ, Taylor T, Wiebe JS, Rivera JO. Medication literacy in Spanish and English: psychometric evaluation of a new assessment tool. J Am Pharm Assoc (2003). 2012; 52(6):e231–40. DOI: 10.1331/JAPhA.2012.11264 [PubMed: 23229985]
35. Steckelberg A, Hulfenhaus C, Kasper J, Rost J, Muhlhauser I. How to measure critical health competences: development and validation of the Critical Health Competence Test (CHC Test). Adv Health Sci Educ. 2009; 14(1):11–22. DOI: 10.1007/s10459-007-9083-1
36. Stucky BD, Lee JY, Lee SY, Rozier RG. Development of the two-stage rapid estimate of adult literacy in dentistry. Community Dent Oral Epidemiol. 2011; 39(5):474–80. DOI: 10.1111/j.1600-0528.2011.00619.x [PubMed: 21592170]
37. Yost KJ, Webster K, Baker DW, Choi SW, Bode RK, Hahn EA. Bilingual health literacy assessment using the Talking Touchscreen/la Pantalla Parlanchina: development and pilot testing. Patient Educ Couns. 2009; 75(3):295–301. DOI: 10.1016/j.pec.2009.02.020 [PubMed: 19386462]
38. Netemeyer RG, Bearden WO, Sharma S. Scaling procedures. Thousand Oaks, CA: Sage; 2003.
39. Ko NY, Darnell JS, Calhoun E, Freund KM, Wells KJ, Shapiro CL, et al. Can patient navigation improve receipt of recommended breast cancer care? Evidence from the National Patient Navigation Research Program. J Clin Oncol. 2014; 32(25):2758–64. DOI: 10.1200/jco.2013.53.6037 [PubMed: 25071111]
40. Paasche-Orlow MK, Wolf MS. Evidence does not support clinical screening of literacy. J Gen Intern Med. 2008; 23(1):100–2. DOI: 10.1007/s11606-007-0447-2 [PubMed: 17992564]
41. Dillman DA, Smyth JD, Leah MC. Internet, phone, mail, and mixed-Mode surveys: the tailored design method. 4. New York: Wiley; 2014.
42. Easton P, Entwistle VA, Williams B. How the stigma of low literacy can impair patient-professional spoken interactions and affect health: insights from a qualitative investigation. BMC Health Serv Res. 2013; 13:319.doi: 10.1186/1472-6963-13-319 [PubMed: 23958036]
43. Farrell TW, Chandran R, Gramling R. Understanding the role of shame in the clinical assessment of health literacy. Fam Med. 2008; 40(4):235–6. [PubMed: 18382832]
44. VanGeest JB, Welch VL, Weiner SJ. Patients' perceptions of screening for health literacy: reactions to the newest vital sign. J Health Commun. 2010; 15(4):402–12. DOI: 10.1080/10810731003753117 [PubMed: 20574878]
45. Wolf MS, Williams MV, Parker RM, Parikh NS, Nowlan AW, Baker DW. Patients' shame and attitudes toward discussing the results of literacy screening. J Health Commun. 2007; 12(8):721–32. DOI: 10.1080/10810730701672173 [PubMed: 18030638]
46. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Fam Med. 2004; 36(8):588–94. [PubMed: 15343421]
47. Morris NS, MacLean CD, Littenberg B. Literacy and health outcomes: a cross-sectional study in 1002 adults with diabetes. BMC Fam Pract. 2006; 7:49.doi: 10.1186/1471-2296-7-49 [PubMed: 16907968]
48. Sorensen K, Pelikan JM, Rothlin F, Ganahl K, Slonska Z, Doyle G, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health. 2015; 25(6):1053–8. DOI: 10.1093/eurpub/ckv043 [PubMed: 25843827]
49. Sorensen K, Van den Broucke S, Pelikan JM, Fullam J, Doyle G, Slonska Z, et al. Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). BMC Public Health. 2013; 13:948.doi: 10.1186/1471-2458-13-948 [PubMed: 24112855]
50. Lee SY, Tsai TI, Tsai YW. Accuracy in self-reported health literacy screening: a difference between men and women in Taiwan. BMJ Open. 2013; 3(11):e002928.doi: 10.1136/bmjopen-2013-002928

51. Kiechle ES, Bailey SC, Hedlund LA, Viera AJ, Sheridan SL. Different measures, different outcomes? A systematic review of performance-based versus self-reported measures of health literacy and numeracy. J Gen Intern Med. 2015; 30(10):1538–46. DOI: 10.1007/s11606-015-3288-4 [PubMed: 25917656]

52. Mele ML, Federici S. Gaze and eye-tracking solutions for psychological research. Cogn Process. 2012; 13(Suppl 1):S261–5. DOI: 10.1007/s10339-012-0499-z [PubMed: 22810423]

53. Schwarzer R, Schwarzer C. A critical survey of coping instruments. In: Zeidner M, Endler NS, editorsHandbook of Coping: Theory, Research, Applications. New York: John Wiley & Sons, Inc.; 1996. 107–32.

54. Aldwin CM. Stress, coping, and development: an integrative perspective. New York: The Guilford Press; 2007.

Table 1

Domains of health literacy assessed in the 128 instruments in the Health Literacy Tool Shed

Health Literacy Domain	HL Tool Shed measures assessing this domain
Prose: pronunciation	20
Communication: listening	6
Communication: speaking	3
Numeracy	63
Application/function	23
Information Seeking: document	31
Information Seeking: interactive media navigation	14

Source: Analysis of the 2015 Health Literacy Toolshed data http://healthliteracy.bu.edu/

Table 2

Mode of administration used by 128 measures in the Health Literacy Tool Shed

Mode of Administration	HL Tool Shed measures using this mode
Computer-based	22
Face-to-face	82
Mail survey	5
Paper-and-pencil	60
Phone-based	3

Source: Analysis of the 2015 Health Literacy Toolshed data http://healthliteracy.bu.edu/