Use of Normative Data and Measures of Performance Validity and Symptom Validity in Assessment of Cognitive Function

David Freedman and Jennifer Manly

Disclaimer: The authors are responsible for the content of this article, which does not necessarily represent the views of the Institute of Medicine.

INTRODUCTION

Formal assessment of cognitive function can provide critical insights into the strengths and weaknesses of a person's memory, perceptual, information, motor, social, language, learning, and executive processes. Cognitive test data provides accurate information for diagnostic and clinical decisions, as well as prediction of real-world functioning. Any assessment in which a person's medical, neurological, or psychiatric diagnosis is under consideration and the relationship of that condition to functional capacity is being determined, cognitive assessment should be a fundamental component of the standard of care. There are many standardized and well-validated cognitive ability measures available for use, but their validity and the interpretation of the results depend, in part, on the question and context of the evaluation. In order to provide meaningful information, an individual's performance is typically compared against other people using normative data. Test standardization helps to establish the parameters of this comparison. Yet, caution is needed to ensure that an individual's performance is compared to the appropriate standard in order to avoid inaccurate characterization of cognitive ability, over- and under-pathologizing, or poor prediction of every-day abilities and capacity.

An adult is eligible for disability benefits under the Social Security Act if he or she has a medically determined physical or mental impairment which severely limits the person's ability to perform "substantial gainful activity," and when that impairment has lasted, or is expected to last, for a continuous period of at least twelve months. A child under age 18 is eligible if she or he has a medically determined physical or mental impairment which causes a marked or severe functional impairment, and which has lasted, or is expected to last, twelve months, or is expected to result in death (SSA DI 00115.015 Definition of Disability). Thus, eligibility for both Disability Insurance and Supplemental Security Income require medical determination of an underlying condition, and also whether and to what extent that condition impairs functioning. Those conditions can include physiological, anatomical, or psychological abnormalities assessed by clinical or laboratory diagnostic techniques.[1] Psychological tests are currently required by SSA for six conditions which are at least partially diagnosed by low intellectual function, and may be used to inform decision making for other conditions.[2]

While the use of cognitive testing for determination of eligibility for both Disability Insurance and Supplemental Security Income is currently limited, cognitive testing could significantly inform the diagnosis and functional capacity associated with other conditions. Since SSA requires a clinical determination which considers the claimant's statements in the context of objective medical assessment, psychological and cognitive testing should be a critical part of that objective evidence. The claimant's symptom presentation, and consideration of collateral information when the claimant's self-report and the medical evidence are not consistent, are also to be considered. All of these areas of behavior and functioning can, and should, be informed by comprehensive cognitive assessment.

This paper focuses on the appropriate use (or non-use) of normative data in cognitive testing and in assessments of effort, feigning and malingering in psychological and cognitive evaluations. It presents a review of the available research on these issues as they relate to the diagnosis of acquired cognitive impairment which affects ability to perform substantial gainful activity, or determination of whether a child has a marked or severe functional impairment.

Uses of cognitive testing

The assessment of psychological and cognitive ability must be based on reliable and valid test instruments that have ecological validity. As explained in a widely used textbook on health measurement, reliability is a property of a particular test instrument when it is used with certain populations under certain conditions: "Reliability is not an immutable, inherent property of a

[1] See, www.ssa.gov/disability/professionals/bluebook/general-info.htm
[2] Intellectual disability (12.05 and 112.05), cerebral palsy (11.07 and 111.07), convulsive epilepsy (111.02), meningomyelocele and related conditions (111.08).

scale; it is an interaction among the instrument, the specific group of people taking the test, and the situation (p.172-3)" [1]. The concept of reliability is a way to understanding the amount of error, both random and systematic, in measuring something (Id. at p.167). Technically, reliability is the ratio of subject variance divided by subject variance plus measurement error, where "the reliability coefficient expresses the proportion of the total variance in the measurements which is due to ‘true’ differences between subjects" (Id. at pp169-70). Reliability is not a characteristic of a test, it is a measure which describes the variance attributable to error and that which is attributable to individual difference. Reliability is the extent to which a test obtains the same result each time it is administered in a specific context to a specific group of people, and the degree to which that result reflects error. Critically, this means that the reliability of a test must be established for the population and in the setting for which it is used.

Validity is the degree to which a test measures what it purports to measure. That is, while different scores may be observed when administering a test, whether that test is measuring the construct it claims to measure is a question of the test’s validity. As with reliability, validity can be ascertained only in a specific context and with specific groups in that context. Thus, test instruments are not valid, but a test may have a level of validity in a specific context when given to a specific group of people (Id. at p.250). To be valid, a test must be reliable (but not vice versa). Reliability can be said to place an upper limit on validity: the higher the reliability of a test, the higher the potential validity. Random error affects both reliability and validity, whereas systematic error affects only validity. A test can be no more valid than it is reliable.

These concepts are critical to understanding cognitive test standardization, norms, demographic corrections, and inferences that can be drawn from cognitive and psychological testing. The importance of context and the comparative components of reliability and validity point to the need to carefully choose which cognitive tests are used, what groups the examinee will be compared to, and how to draw inferences from the results. In turn, those questions must be informed by the purpose of the assessment: is the use of the test instrument to arrive at a diagnosis of acquired impairment, is it to determine overall level of intellectual or cognitive function as compared to age-matched peers, or is it to describe the every-day ability and functional capacity of a person in the real world? Cognitive testing can accomplish each of these goals using the same or similar well validated measures, but the approach to interpretation and norming of raw scores differs.

Test standardization and norms provide a framework within which an individual's scores can be located and against which they can be interpreted [2]. The standardization of a test is an important part of its development, used to align scores (along the often assumed normal distribution of cognitive abilities in the population of interest, the Gaussian curve) and to establish the standard deviation and standard errors of the test. Standardization is a necessary, but not sufficient, part of establishing reliability and validity because it helps to establish the context in which a test is administered and interpreted. The parameters developed through standardization, assuming that they are properly specified and clearly delineated, permit an individual's performance to be compared to a reference group's performance. The composition of the reference group should be determined by the research question which is being addressed [3].

Diagnostic uses of cognitive testing

Cognitive testing can be used for diagnostic purposes, and historically, most neuropsychological tests were developed to provide diagnostic data when few other tools (such as MRI, EEG) were available. For instance, a person presenting at an evaluation with symptoms of aphasia can be tested to determine the level and breadth of impairment, to assess whether the aphasia is receptive or expressive, or whether the observed impairment is an isolated impaired domain or one part of a broader dysfunction. In determining the cause of the aphasia, and concomitantly the onset, course, and treatment response needed, such a diagnostic determination utilizes standardized test parameters and normative "cut-offs" which have been developed with groups of healthy subjects, and validated among both healthy subjects and patients typically known to have lesions localized to specific areas of the brain [4]. In this situation, lesion evidence served as a gold standard from which the cut scores were developed and validated. Determining whether the subject has aphasia or some other condition represents a diagnostic use of cognitive testing. When cognitive tests are used to determine whether there is acquired impairment, an obtained score on a test, no matter how perfect the measurement of the construct of interest, has little meaning without normative data against which to compare that score and the use clinical judgment to draw inferences [2]. For instance, it does little good to have a machine which precisely calculates blood pressure if nothing is known about the range of blood pressures which increase the risk of mortality.

If the goal is to determine if an adult has an acquired impairment, demographic adjustments are appropriate. This is because the comparison being made is between the subject before the acquired illness or injury and the subject after. Because the person cannot be returned to the pre-injury or pre-illness state for assessment, demographic characteristics are an appropriate technique to estimate how that person would have performed, but-for the injury or illness. The more similar the normative sample is to the examinee, the more precise the estimate will be, and the better the accuracy of the measures.

For example, research on mild cognitive impairment in aging has included few non-white subjects [5]. Yet, neurologically healthy African American elderly, on average, tend to perform worse than white elderly on tests of episodic memory [6]. If the question is whether an 85 year-old African American woman with 8 years of education has mild cognitive impairment, the proper comparison group would not be 85 year old white women with 8 years of school, which may lead to incorrect diagnosis of impairment in the African-American woman. While it is not "race" itself that causes the observed effect, it is serving as a proxy for underlying sociocultural variables that do [5, 6], demographic corrections serves as a proxy for experiences and conditions which are not typically measured in the standardization process for tests.

The proper comparison group in this case would also not be the general population of 85 year-olds (which includes people with high school and college degrees, as well as 85 year olds with dementia), or 25 year old African Americans with 8 years of school (age is a major predictor of performance on cognitive tests). The proper normative group would be neurologically and functionally normal 85 year-old African American women from the same geographic region with 8 years of school who do not develop dementia.

Thus, cognitive tests, when used for diagnostic purposes, report on deficits or strengths and use terms such as “moderately impaired” or “within normal limits.” Diagnosis of impairment is based on established cut-offs that maximize sensitivity and specificity of the measure to detect significantly poor functioning in a particular cognitive domain. The normative standard for cognitive tests when used diagnostically is not population based, it is criterion or deficit based and uses the best estimate of the examinee’s premorbid function as the normative standard. Using diagnostic norms, "we are not so much interested in how much of an ability a patient has or where the patient falls within the reference group; rather we seek to know whether a performance is more likely to be characteristic of the reference group or one that is outside of it (p.137)" [7].

Descriptive uses of cognitive testing

Descriptive norms are population-based comparisons, predicated on the assumption that population performance on a test will be normally distributed (Gaussian) and that all scores reside under the curve [7]. The descriptive property of an observed score arises because it is situated within a normally distributed set of scores with known psychometric properties: mean and standard deviation. An observed score on a test, therefore, can be described using relational terms such as average, below average, above average, and also in terms of standard deviation units such as one standard deviation below mean, but does not use terms such as “impairment.” This also means that the examinee must belong to the "population" which defines the normative sample: the examinee must be part of the reference group [7]. For example, IQ scores provide a summary measure of how an individual performs in comparison to a census-matched population of people of the same age. An IQ score approximately two standard deviations below the mean of the normative group positions the person's performance within the lowest three percent of the population overall. While this level of performance is one of the diagnostic criteria for intellectual disability, along with adaptive behavior assessment [8], the normative reference for interpreting IQ tests is population-based.

In the context of a learning disability, for instance, a student may have both strengths and weaknesses in academic functioning. Determining which academic domain (i.e., reading, writing, spelling, math) is below expectations for age and educational level requires normative comparison which sets the student's ability against, not themselves before some acquired injury, but others of her or his age and grade. The “universe” or population of peers is defined depending on the question at hand. For instance, achievement tests could determine how well other students in the same school or school district perform on math, or how the school district's average performance compares to statewide performance. The results are descriptive because they locate the person in relation to the mean and along the standardized distribution of the population.

As with all cognitive testing, the context and purpose of the administration matters. Manly and Echemendia (2007) wrote: "When an individual’s performance is compared to a single reference group – for example, a group of people matched to the US population on age, sex, geography, race, and years of schooling – the normative data are being used descriptively. As long as the methods used to select that population are clearly delineated and properly operationalized, any population can serve as a reference if appropriate to the question at hand (p.320)" [3]. Thus, school performance tests which are used to rank student performance are descriptive because they compare any given student to all other students taking the test in the US. The rank order allows for a descriptive placement of that student in comparison to other students in the US.

The distinction between descriptive and diagnostic settings is crucial for determining how and what norms are appropriate. Age norms tend to make intuitive sense to most people. A ten year old child who obtains an IQ score of 145 would still not be allowed or expected to conduct medical research or pilot a jet. Cognitive functioning follows a trajectory from conception to death and the age comparisons reflect an effort to most accurately describe variation from typical across that developmental and degenerative trajectory [9]. Yet, for neurodevelopmental disorders, cognitive trajectory may differ from typical before the onset and diagnosis of the condition [10]. In the context of developmental disorders which have a substantial effect on school attendance, normative corrections based on years of education will fail to account for the effect that the illness had on school performance [11], thereby inappropriately comparing the individual to a dissimilar group, likely resulting in under-recognition of impairment and functional ability in the diagnostic context.

Estimating every-day functioning

Determination of whether an individual possesses the cognitive ability to function in their surroundings, and to what capacity, is not concerned with premorbid function. The proper comparison group to determine capacity to function in the world is people who are currently performing those tasks without limitations or problems/disability.

Limited research suggests that demographic adjustments reduce the power of cognitive test scores to predict every-day abilities. Silverberg and Millis (2009) tested whether using demographically adjusted scores or absolute scores (i.e., scores which are not adjusted for demographics or premorbid estimated function) better predicted real-world functioning following traumatic brain injury (TBI) [12]. For most functional areas: daily living, community mobility, employment, and global functioning, demographically adjusted scores had lower correlations with capacity. For both employment status and employability, demographically adjusted scores correctly predicted capacity 63.6% and 54.8% and absolute or raw scores correctly predicted 75.8% and 63.5% of the time. The authors concluded that absolute scores had better ecological validity than demographically adjusted scores.

Similarly, a study of non-demented people with Parkinson's disease used a complex measure of cognitive flexibility, sequencing, motor speed, and visual scanning to predict activities of daily living (ADL), and compare raw cognitive test scores with age and education adjusted scores. The results showed that only the raw test scores significantly predicted ADL, but the demographically corrected scores did not [13]. Barrash et al (2010) also found that demographic adjustments decreased the accuracy of neuropsychological testing to predict real-world capacity [14]. Assessing driving ability in people with Alzheimer's and Parkinson's diseases, they observed that raw scores on testing better predicted driving ability compared to demographically adjusted scores.

Summary

Thus, cognitive and psychological assessment can be concerned with diagnosis of impairment, decline, or change in cognitive functioning, determining where a person is functioning along the normal distribution of a cognitive ability within the general population, and/or capacity to perform real-world tasks. Diagnosis of impairment, decline or change is typically measured against pre-injury or pre-decline scores (if one is lucky enough to have them), or our best estimate of what those scores are expected to have been, given the demographic features of the examinee. Descriptive questions about normally distributed abilities such as intellectual functioning or academic achievement use population-based norms that are usually matched to the US Census, and are not concerned with impairment or pre-morbid function. Cognitive measures have been shown to accurately predict real-world functioning or capacity when the normative standard is the general population who are performing the task without limitations or problems/disability, but the normative standard for daily functioning should not include adjustments for age, education, sex, ethnicity, or other demographic variables.

Problems that reduce the effectiveness of all normative corrections

The validity and reliability of cognitive tests must be established through a rigorous development process. This is an assumed predicate to establishing normative adjustments. However just because a test manual or scientific manuscript makes normative adjustments available does not mean that they are appropriate for use for every person, or even that they have been collected and presented in a way that meets the standard of the field. We now raise some issues that are commonly overlooked when practitioners use published norms for cognitive instruments.

Cohort effects and norms

Cognitive test norms become less accurate because at the population level, performance on any particular set of items increases over time [15-17]. Therefore, comparison of observed scores to outdated norms results in an overestimation of true ability and functioning. The Flynn effect, which is the name given to the observed increase in IQ over time at the population level, is consistently estimated at 0.3 points per year [15-17]. The observed rise in scores over time increases the likelihood of under-diagnosing low cognitive functioning [18-20]. Although most commonly studied in relation to IQ, obsolescence in normative data over time may also occur with neuropsychological tests, and the evidence of this continues to accumulate [21-25]. Applying Flynn corrections is a process of choosing the correct normative group against which to compare observed scores [26-28]. In effect, Flynn has identified a cohort effect: as testing occurs further in time from the point at which the test was standardized, the population appears to perform better. Because this cohort effect is systematic, it is not accounted for by the standard error of the test, which only considers the effect of random error at the time of standardization [26].

Although Flynn corrections to IQ scores have an established, robust numeric value, less is known about the precise adjustments which should be made to other types of cognitive testing such as neuropsychological measures. However, the cohort effect appears to operate in the same way: the more time since the standardization of the test, the observed score is more likely to over-estimate true ability [18, 20]. At the present time, additional research is needed on the rise in scores over time for testing beyond IQ, but it is an important consideration which should be addressed and considered even if no specific corrective numeric adjustment can be made.

Practice effects

In a similar vein, people who have been exposed to test material more than once have been found to perform better when re-tested. Practice effects on IQ tests are large and have been reported to last many years [29-33], and are also observed on neuropsychological testing [24, 34-36]. Practice effects have implications for inferences drawn on repeated testing. Many school districts require annual (or tri-annual) re-testing of students who are provided services as children with learning disability or intellectual disability. If the re-testing uses the same test instruments, an artifactual increase in score is likely to be observed. This is especially important when assessing children and adolescents because variability on tests is higher among those age groups [37, 38] and the effect may be larger at the lowest range of scores [39]. Understanding practice effects and considering their impact on observed scores when comparing to norms constructed without accounting for repeated exposure should be routinely discussed as part of interpreting cognitive data.

Inappropriate use of demographic adjustments

Demographically adjusted norms, such as those that correct for age, education, sex, and race/ethnicity, can improve specificity and sensitivity of neuropsychological measures to cognitive impairment. However, it is not uncommon for researchers and practitioners to use demographic corrections for cognitive tests without knowing that they are inappropriate or unsupported by scientific evidence.

A key issue about the use of demographic adjustments is that they are not appropriate when the question is descriptive (e.g., estimation of intellectual function, comparison to grade-matched peers, or determination of every-day functioning). Not surprisingly, misuse can become common when, in order to facilitate the use of IQ measure subtests as neuropsychological (diagnostic) instruments, data from the standardization sample of a census-matched standardization sample is analyzed to create demographically adjusted norms for those subtests. Demographic adjustments are not available for IQ index scores because this would be inappropriate for the estimation of intellectual function.

Use of demographic adjustments also would be inappropriate for people with neurodevelopmental disorders whose prodromal period interferes with educational attainment or the expected relationship of age to cognitive function. In the case of psychotic disorders, because cognitive impairment and developmental trajectory are atypical many years prior to diagnosis, education-adjusted norms would confound the “expected” performance with a variable that represents an outcome of the disorder itself [40-43].

Used correctly, demographic norms are an important tool, when the normative cohort includes a sufficient number of people with the same background as the examinee, because they allow for an estimate of premorbid function, increasing the ability to accurately detect subtle cognitive impairment.

However, it is important to understand the underlying conditions and caveats that are frequently present when normative data is collected. Recruitment of participants for a normative sample, while meeting goals for numbers of people within multiple race/age/sex/education cells, can be difficult and take many years. In order to recruit enough people to stratify by a number of demographic characteristics (i.e., fill cells), researchers sometimes resort to recruitment of people who are unusual for their background, or outliers. Thus the normative cohort is not representative of healthy people within the subgroup. For example, if testing for the establishment of test norms for people age 25 to 60 only occurs on weekdays from 9 am to 5 pm, the people included are less likely to be employed full time than the population of healthy people within this age range. This can also happen when norms are collected in an area where there is not a trusting relationship between the minority community and the research institution. Researchers may reach and recruit only non-representative minorities on the fringes of the community who are willing to provide informed consent. In the same vein, normative data that relies on word-of-mouth to recruit the cohort will not be representative of the community to which the norms would later be applied.

Commonly considered demographic variables that are used in developing normative adjustments are age, race, ethnicity, sex, and education. Many other social, political or cultural differences between groups of people might also be "demographic" conditions for which adjustments could be considered, although generally they are not available for study. For instance, differences based on geographic region, culture, language, economic status, neighborhood conditions, migration or immigration status all have been shown to be significantly related to performance on cognitive tests [44-46].

The fact that traditional racial and ethnic categorization is often too broad (ignoring subgroups), is based on social norms and not stable or characteristics that are correlated with cognitive function, and is fluid over time and place is a problem for cognitive test norms that use race and ethnicity. For example, “Asians” as a group combine many cultures, languages, religions, geography, political and historical divisions, nationalities, and educational experiences, which may be more dissimilar than similar, and yet get grouped into a single “racial” category. Application of norms for Hispanics (either for English- or Spanish-speakers) must always take into account the background of the people included in the normative cohort, including nationality, country of residence/years in the US, language use history, and level of bilingualism.

Generational shifts should be a major consideration, especially among immigrant groups. First generation Latino immigrants have, on average, better general health than many age and economic status matched peers, but their children, the second generation, have worse health outcomes than their age and economic status matched peers. This is often referred to as the Latino paradox because the combined racial discrimination and low economic status of Latinos on average would be expected to increase the risk of mortality and morbidity, but does not necessarily act in the expected way [47]. Assuming that this applies to cognitive function as well, norms may be more effective if they take into account factors such as generational status.

Racial and ethnic classifications are becoming more “blurred” over time as the population of the US becomes more multiracial. Demographic adjustments that include only people of one specific racial or ethnic background may not be appropriate for biracial or multiracial people. The statistical limits to the collection of norms for each racial and ethnic subgroup are readily apparent, with so few people populating each group as to make statistical measurement meaningless. Because racial and ethnic classifications are social classifications, their utility may be extremely limited by time and place.

Often, the effect of "race" on cognitive outcomes is more readily explained by other factors [48-50]. Controlling for single word reading recognition, a proxy for of quality of education, explained an observed difference in test performance which has sometimes been regarded as a difference based on race in studies of moderate to severe acquired head injury. [51]. This research is confirmed in many settings and with several outcomes, including TBI, MCI, hypertension, dementia, Alzheimer's, and among community residents without a specific illness [52-56], and strongly suggests that the variable "race" as a demographic characterization may be more accurately understood as a proxy for other more meaningful social and environmental factors. These factors may not only explain between racial group differences on cognitive tests, but also significantly relate to performance within racial/ethnic groups.

The background of consideration of race and cognitive testing has a long history of discriminatory practices and oppression based on perceived differences, but similar issues might be relevant when considering gender, gender identity, poverty and deprivation, cultural and linguistic differences regardless of race.

A final consideration is that appropriate norms may yet not be available for many groups of people who present for cognitive testing. As a simple example, consider the WAIS-III Spanish language norms. Three sets of norms were developed: Mexican, Puerto Rican, and Spanish. The Mexican and Puerto Rican versions, including the norms, have been critiqued for a variety of problems with the standardization process, drawing into question the validity of those norms for any purpose [57, 58]. Assessing Spanish-speaking or bilingual people who reside in the continental U.S. (and not Mexico or Puerto Rico) also is reported to raise difficult questions about proper interpretation [59, 60]. If a Spanish language instrument is available and Spanish language norms have been collected, they should not automatically be applied to a Spanish-speaker on the basis of their dominant language alone. No scientific basis exists for giving primacy to language differences over the cultural, national, and educational differences.

To summarize, demographic norms are useful for diagnosis of acquired impairment, but not for describing where an individual’s cognitive ability sits within the general population. The use of demographic norms is not appropriate when we wish to predict functional ability in the real-world. In addition, they are not recommended when characterizing an acquired impairment in someone with a neurodevelopmental disorder or whose social/cultural background is different than the normative group [61].

Cultural bias

Many common cognitive tests, for instance IQ tests, may also use normative data, but the assessment question is still relevant for interpretation. While typically described as flawed estimates which do not describe specific abilities and disabilities, IQ tests reflects a general consensus as to how the construct of “intelligence” should be measured, typically based on formulations that derive from the Cattell-Horn-Carroll hypothesis [62]. "Individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought. Although these individual differences can be substantial, they are never entirely consistent: A given person's intellectual performance will vary on different occasions, in different domains, as judged by different criteria. Concepts of 'intelligence' are attempts to clarify and organize this complex set of phenomena. Although considerable clarity has been achieved in some areas, no such conceptualization has yet answered all the important questions and none commands universal assent (at p.77)" [63].

The cultural bias inherent in IQ tests is also a bias in the real-world. Indeed, the normative sample against which individual performance is compared for IQ measures is a cohort that is representative of the US population in general. In other words, if an IQ test is valid, the obtained score (and the associated 95% confidence interval) represent the best estimate of where an individual’s true IQ lies with respect to the general population and the “real-world.” IQ scores are not adjusted based on demographic or background factors such as occupation, sex, or race. Instead, IQ scores represent where the individual stands not with respect to people just like them, but with respect to their age-matched peers across the US, and thus the world in which they are expected to function.

Performance Validity and Symptom Validity Tests

Determining whether optimal effort was put forth and detection of possible feigning of impairment is a critical aspect of clinical cognitive assessment. Proper interpretation of all cognitive measures depends on valid administration of properly normed tests, and also on the examinee being both able and willing to “play the game” of cognitive testing. Most cognitive tests cannot be administered in a standardized manner if the examinee cannot understand the basic instructions for the task. This invalidates the testing. Furthermore, cognitive functioning cannot be determined if there is variable or consistent poor effort on the tasks – if a person does not put forth adequate effort, the scores will not accurately reflect ability. How to assess whether poor effort or feigning has impacted the validity of a cognitive assessment remains debated and unsettled, with limited scientific support for many of the available options. Performance validity tests (PVT) assess the effort that a person gives on a particular test. Symptom validity tests (SVT) assess endorsement of unusual symptoms or the accuracy of self-reporting of symptoms.

Both types can be embedded in other measures or stand alone instruments. These instruments use cut-offs to determine poor effort or feigning. Many instruments are available and the professional neuropsychological organizations recommend the use of PVTs or SVTs, but leave it to the clinician to determine the method and how to interpret the results in the context of the overall assessment, including determining if the cognitive battery is also invalid [64, 65].

A recent review of PVTs reported sensitivity and specificity rates for different test instruments for which data were available [66]. The authors note that they were not reporting positive and negative predictive power because those two measures vary depending on base rates of the condition. This is a critical issue for PVT/SVT's in the SSA Disability context, because currently, no scientifically derived base rate evidence is available. We will discuss base rates first because this underlying problem frames all other issues of interpretation of PVT/SVT performance.

The importance of base rates for determining utility of PVTs/SVTs

For this discussion, positive predictive power (PPP) is the proportion of people who fail a PVT/SVT who are actually feigning impairment or not putting forth adequate effort. Negative predictive power (NPP) is the proportion who pass a PVT/SVT and who are not feigning impairment and are putting forth adequate effort. These two measures represent outcome test’s accuracy and are both ratios; the denominator requires knowledge of the number of people who are truly feigning impairment or are not putting forth adequate effort in the population of interest.

Most research studies on SVTs and PVTs estimate base rates but do not have direct evidence to support these estimates. Research studies include populations suspected of having the condition: civil litigants, people with some financial stake or seeking disability, and people facing criminal penalties, for example. Most PVT/SVT's also estimate base rates and develop cut score thresholds by coaching people to feign impairment or by using the characteristic of interest to group them, such as recent acquired head injury.

Coaching may not replicate behavior in the “real” setting. It cannot be established that asking people to malinger or feign, nor coaching them to do so, generates a pattern of responses that accurately and validly captures the construct of feigning/poor effort. In fact, asking people to feign a psychiatric or medical condition may elicit stereotyped responses that reflect assumptions about the presentation of an illness and have little to do with the manifestation of the illness.

Known groups method may over identify poor effort among populations with cognitive impairment [66]. Many SVT/PVT cut-offs have been established using groups litigating or seeking benefits as "known" to be feigning/poor effort groups. It is at least plausible that some of the people in these groups are suing or seeking support because they were truly harmed or truly need benefits, differentially than those who do not seek benefits or litigate. If, as discussed below, this group's symptoms or other characteristic (demographic, cultural, etc.) make them dissimilar to people who do not seek benefits or are not litigating, the two groups would not be comparable. Using these groups on the *assumption* of feigning/poor effort creates a very high base rate which in turn make the tests appear better at differentiating groups than they are, may negatively affect the reliability of cut-offs and increase the number of false positives.

Positive and negative predictive power, sensitivity and specificity

As a result, studies using test instruments developed without known base rates, may report important properties of the tests (including measures such a PPP, NPP, sensitivity and specificity), but these results may not be appropriate for the settings in which these instruments are most commonly used.

Moreover, base rates of feigning and poor effort may or may not be similar across cultures and demographic differences; however, there is inadequate evidence to determine this at this time. People who sue or seek benefits may not be comparable, in the epidemiologic meaning of concept, to those who do not. This could be a legitimate distinction in the severity of injury, cultural beliefs about responsibility, guidance received from family or clergy or other lay advisors, or any number of other social and environmental determinants that make them different on a basis other than feigning/giving poor effort. Determining the base rate is assumption-based rather than science-based in these circumstances, making assessment of the PPP and NPP, the two measures of most relevance, impossible. Base rates of inadequate effort on cognitive tests in different contexts, and within groups of people that differ socioeconomically and culturally, would have to be taken into account in the development of measures, selection of measures within a battery, and interpretation of performance on those SVT/PVT measures.

As a result of difficulty establishing the true occurrence of feigning, the sensitivity and specificity of PVT/SVT's are more often reported. Sensitivity and specificity are also affected by base rates, but because they are measures of how well a test performs in a specific population, the effect of base rate variation differs for these measures. Sensitivity is the probability that a PVT/SVT correctly classifies a person who is feigning or giving poor effort as such. Specificity is the probability that a PVT/SVT correctly classifies a person who is not feigning or is giving good effort as such. Because these are probability measures, they are somewhat less susceptible to base rate variation, but they also answer a different question than the predictive validity measures. While sensitivity and specificity describe the validity of the test instrument; they do not describe whether any given person taking the test has the condition, but the probability that the test will correctly classify a person.

Much of the research in this area has set a specificity of 90 percent as a reasonable specificity threshold. This would mean that a test has a nine out of ten probability of correctly classifying people without the condition. In the review mentioned above, some of the tests reached this level of specificity but not all [66]. Sensitivity was substantially worse than specificity, varying widely by instrument [66]. The review considered five instruments, the Victoria Symptom Validity Test, the Word Memory Test, the Test of Memory Malingering, the Letter Memory Test, and the Medical Symptom Validity Test. Combined, the forty-eight studies included in the review showed mean sensitivity for these instruments to be 0.69 (95% CI: 0.63, 0.75). Mean specificity was 0.90 (95% CI: 0.86, 0.94), but with wide variation from a high of 95.5 for the Victoria Symptom Validity Test based on two samples with a combined n=48, to a low of 69.4 for the Word Memory Test based on 7 studies with a combined n=204. The wide variation and small sample sizes suggest that these sensitivity and specificity may be unstable estimates even after combining samples from different studies.

Partially in response, many authors recommend using multiple PVT/SVT's [67], but this has raised concerns about false positive rates. In a review and simulation study, administration of more PVTs were associated with higher false positive identification of feigning and poor effort [68], although this study has been criticized as well [67, 69]. Those who advocate controlling false positives by using multiple tests have not taken into account the increased rate of finding a positive (failing) result by chance [70-72].

PVT/SVT Standardization and Normative Samples

A related area of concern for PVT/SVT's is the samples with which they are standardized. Using the TOMM as an example, it is useful to examine a number of areas in which PVT/SVT's have not been fully developed, standardized and normed.

i) Standardization, an example: One of the most commonly used tests, the Test of Memory Malingering (TOMM), is a fifty item recognition test designed to assess malingering compared to memory impairment [73]. It was developed in two phases, the first of which was with a preliminary version tested on non-clinical subjects using a four choice, recognition task. Phase two modified the design and used a two choice recognition task, and this latter version has remained. The initial (phase 1) testing included administering the TOMM as part of a study on aging to 405 people living in the community who were recruited in public places and volunteered to participate. They ranged from 16 to 84 years old, had an average education of 13.1, were 47% male, but no race or ethnicity or other demographic data is reported. Following this initial trial, the TOMM was revised and then normed with a group of 70 healthy volunteers (63% male, no indication of race or ethnicity, 12.7 years of schooling but no indication of quality of education). Use with clinical samples followed. First, a group of 138 neuropsychologically referred in- and out-patients at a Veterans Administration facility and 23 head injured subjects were tested with the TOMM: 13 with no cognitive impairment, 42 with cognitive impairment, 21 with aphasia, 45 with TBI and 40 with dementia. Again, only age and years of schooling are reported. Validation with "at-risk" (17 not "at-risk", 11 TBI "at-risk", 11 cognitively intact, 12 patient controls) and simulators (27 college students) were also used for validation but no demographic data is reported.

ii) Cut scores: Based on these studies, cut score values for determining malingering versus not malingering were developed and are still the most widely used cut scores for this measure. Soon after the release of the TOMM manual, a series of validation studies was published [74]. For each study in the series, only age and education were reported. Similar to the normative studies, this series used students, community volunteers, and simulators. In all, the TOMM performed well in differentiating people instructed to malinger and was more accurate identifying people with TBI who were believed to be "at-risk" for malingering compared to those who were believed not to be.

Subsequently, numerous studies have used the TOMM as a measure of malingering based on these cut scores. The question, however, is whether the available normative data, which included community and college volunteers and small numbers of patients, are adequate to support the broad use of the TOMM across cultures and other demographic groups. In effect, there is no normative data describing how the population as a whole performs on this instrument. More importantly, even for the age groups which are included in the original norming sample, because the subjects are not typical and were not selected by a process that supports the conclusion that they are representative, there is no evidence that the cut scores are generalizable. Selection bias remains infrequently considered and more rarely discussed in standardization studies of PVT/SVTs which rely primarily on healthy volunteers, students, and patients [75]. This raises questions as to whether the cut scores are generalizable to anyone, but especially to populations clearly not included in the development of the measure.

iii) Feigning, poor effort, or impairment: The TOMM is used as an example, based on a review of the literature and test manuals, but this criticism applied to all the PVT/SVT's [76-81]. A systematic search through public databases does not return any studies that suggest any PVT/SVT has been normed on a population-based sample, or a sample unbiased by the way in which they were recruited. Many studies rely on college students, volunteers, clinical practice samples or referrals for evaluation, none of these groups adequately represent the population at large, and the false positive rates in these groups remain under debate [82-84]. This raises concerns about the validity of SVTs and PVTs, and whether they accurately detect poor effort and feigning. As expressed by Bigler (2012), whether failure on an PVT/SVT reflects feigning, impaired effort or impaired ability remains unanswered by the research at this time [82]. A recent study reported that only failure on PVT/SVT's predicted community participation among veterans with TBI, suggesting that these instruments may be assessing true impairment rather than poor effort or feigning, although cognitive impairment, as measured in this study, was not associated with community participation [85].

Another recent study on a sample of people with psychogenic non-epileptic seizures found that failure on a SVT was associated with self-reported abuse history but not with seeking financial compensation [86]. Subjects were referred to an academic epilepsy center and there were no differences between those who failed the SVT and those who passed based on demographic characteristics or psychopathology as measured by the MMPI. Patients were labeled as having a financial incentive if they were receiving or seeking disability or workers compensation. History of abuse was obtained by direct question as to whether the patient had ever experienced emotional, physical or sexual abuse. Reporting of any abuse was associated with a more than doubled the likelihood of failing the SVT. While additional research on this issue is needed, the study points to a fundamental question about what is being measured by PVT/SVTs and whether the correct questions are being asked in the standardization process. The implication of this study is that PVT/SVT's may be misattributing the responses on PVT/SVT's to feigning or poor effort, when in fact they are capturing an aspect of the lived experience (experience of exposure to trauma) of the examinee.

Establishing cut scores

The normative and standardization processes result in cut scores: above or below the cut score, a person is or is not said to be faking or giving good effort. Cut scores are malleable depending on the confidence with which the determination needs to be made, and raising or lowering the cut scores affects sensitivity and specificity of the determination. For instance, if we want to be absolutely sure not to falsely identify someone as malingering, the cut score can be set for specificity equal to one. This can be accomplished by examining the distribution of scores on the test, typically with receiver operating characteristic curves, finding the score at which specificity equals 1, and using that as the cut score. That would mean no one is falsely placed in the feigning group when in fact they are not feigning. The consequence of this will be that more people who are or could be faking will be inaccurately classified as not faking (lower sensitivity). Determining cut scores requires both statistical evaluation (identifying the point at which both sensitivity and specificity are maximized), but also judgment about the consequences of being wrong: is it worse to over-include or under-include? When the test is a screening instrument which leads to further, gold-standard assessment, over-inclusion is typically preferred. However, when the test is the primary means of categorizing a person or labeling them, the risk and harm of over-inclusion should push towards under-inclusion.

PVT/SVTs are currently the primary basis for determining valid performance, because no gold-standard follow-up has been developed. Because inaccurate categorization has potentially life changing consequences, we need to be cautious about inaccurate categorization. Categorization is made based on cut scores developed in the standardization process and then often revised when additional validation efforts show likelihood of more accuracy by changing the cut point [87]. In general, scores in neuropsychological testing use ranges or confidence intervals which account for random test error. However, the approach used by PVT/SVT's has been to categorize without confidence ranges or consideration of error. The underlying assumption is that the cut scores have no error and accurately differentiate two populations: those who we are sure feign and give poor effort, and those who do not.

One common embedded measure of effort on IQ testing uses the digit span subtest to assess effort. Performance on digit span is expected to follow a pattern, since the each set of digits which are repeated back to the examiner get longer as the test continues. The test begins with a set of two numbers, then a set of three, a set of four and so on. The number digits correctly recalled for each set, both forward and backward, are added to calculate Reliable Digit Span (RDS). Thus, if a person can recall both 4 digit questions forward and both 3 digit questions backward, the RDS score is 7. Cut scores have been developed which are interpreted as measures of effort, with the expectation that scoring below the cut-off is rare in people who are putting forth adequate effort and not feigning.

Digit span is a test of verbal memory, attention, and working memory, it is often thought to be less susceptible to cultural biases related to race, ethnicity, education, and poverty; However, these variables have been found to have a significant impact on score. Digit span performance, although often described as "language-free," is associated with both education level and culture. Ostrosky-Solis & Lozano (2006) found this to be the case, reporting education level significantly affected performance on digit span, and that a sample of Mexicans performed worse controlling for age and education, indicating cultural variation in performance [88]. Petersson et al (2001) also reported that a group of people in Portugal who, for socio-cultural reasons, did not have an opportunity to become literate, were impaired on visual naming tests [89]. This supports the idea that education and literacy affect neurodevelopment in both verbal and nonverbal domains.

In clinical studies, digit span has been found to indicate left hemisphere brain damage. Digit span is useful for identifying short and long term symptoms associated with TBI, but this also suggests that its use to assess effort and feigning may risk converting true impairment into poor effort [2]. Poor digit span performance is also associated with many serious illnesses [90-93].

As a result of these influences on digit span performance and the relationship between performance on digit span and the calculation of RDS, the appropriateness of the standard cut scores for RDS are in doubt. Limited research has addressed RDS cut scores in different populations. One study which clearly addressed this issue examined the RDS in Taiwan, and reported that the western normative data is not appropriate for use in Taiwan, and further research is needed to use the RDS as an imbedded measure cross-culturally [94].

A recent meta-analysis of RDS found that people with history of cerebrovascular accident, severe memory disorders, intellectual disability, borderline IQ or lower, English as a second language, and patients of Latino and Native American ancestry scored unexpected low on RDS. African Americans, when a cut score of 7 or lower was used, were also more likely to score in the feigning and poor effort range. Even with a lower cut score value, these groups performed poorly, with lower sensitivity and specificity [95]. This meta-analysis concluded: a "cutoff score of ≤7 achieved a global sensitivity rate of 48% when using weighted averages and 58% when using the Bayesian method; however, this cutoff score also produced inadequate specificity rates (i.e., <90%) for the pooled data (using both statistical methods) and for all clinical subgroups (using weighted averages)" (p.25). Using lower cut-off scores of ≤6 resulted in better specificity, as expected, but worse sensitivity. Even with this highly researched tool, the cut scores and inferences about effort and feigning remain unclear, and its application to groups other than average functioning whites, remains questionable.

Measurement error and confidence intervals

How cut scores are established, standardized and applied, and therefore the inferences which can be drawn regarding a person's performance, are critical determinations and remain unsettled, with insufficient scientific evidence available to support conclusions. In addition, while appropriate to screening instruments, point estimates and cut-off scores do not reflect an adequate approach to diagnostic psychometric testing or a scientifically rigorous approach to drawing inferences from testing data, as they fail to account for the standard measurement error (SEM) of the instrument, and therefore also fail to account for the relationship between an observed score and the error parameters within which a true score can be said to lie with 95% confidence [1].

SEM is a psychometric tool which is estimated from the standard deviation of test and the tests reliability. SEM is specific to the test and group with which the test is used. It is the basis for establishing confidence intervals, and is a critical technique to estimate the range within which true scores reside [1]. An interesting study that set up a group which choose between two faces by random selection, demonstrated that "chance" ranged between 32% and 66% correct identification [96]. Specifically seeking to determine how guessing affects below chance performance on forced choice tests, this study is instructive on the question of the range across which "random" occurs, supporting the need for confidence intervals and error bands on PVT/SVT test instruments. This study examined whether guessing resulted in chance responding on a forced choice instrument, the Warrington Recognition Memory Test for Faces (RMTF).

Results on the RMTF for the chance responders, who were told that they would be shown fifty pairs of pictures of faces, and they should choose the one they thought was the target face, reflected a mean of 51.2 percent correct responding. Significantly, that mean is the average of a range of scores between 32 percent and 66 percent correct, and the range of scores overlaps with both the control group scores and the two groups of instructed to malingering subjects. In fact, the range is evidence that, on average, forced choice tests work as expected: given to enough people who guess, the group mean will be approximately chance performance. At the group level, the RMTF performed in this manner. However, for any given individual performance, no conclusion about less than chance performance is supported. Any given individual who was instructed to guess in this study performed between 32 and 66, not at 50 percent. Thus, 51.2 percent was the mean score for the group which randomly selected answers, with a confidence interval around that group mean which ranged from 32 to 66. An individual who scored anywhere within that range is equally performing at the level of chance. This is the point of establishing ranges within which a true score can be said to reside based on the observed score. Moreover, this study found no separation between the four groups, indicating that while group averages differ, the range of scores do not, and performance on the RMTF is not diagnostically indicative of effort or feigning for any individual subject. This points to another gap in the current literature: the need for confidence intervals and ranges of expected answering, rather than single point cut offs.

This reinforces the need for confidence intervals and error bands to be established around observed scores on PVT/SVT's. The range of chance performance is wide, and overlaps with the ranges of instructed faking and healthy controls. This lack of group separation in scores (the overlap of the ranges), leads to a very different conclusion than that those instructed to feign are identified. In this study at least, the RMTF did not achieve a clear distinction between those instructed to feign and other subjects. As a result, no inference should be drawn about any individual who was tested, as the ranges overlap and the distinctions between groups are not clear.

Below chance performance

One common approach to interpreting forced choice or binary-response (true/false) PVTs is to use chance or worse performance as a benchmark for feigning or poor effort. Often, worse than chance, or substantially worse than chance, performance is viewed as sufficient evidence to sustain the conclusion of malingering, reflecting an intent to deceive or falsely gain. For instance, if a test uses fifty true-false questions, chance performance would be set at 25 correct answers. This is because it is assumed that randomly answering each question achieves a fifty-fifty chance of being correct on each response. Given that, it seems reasonable to expect that a score below 25, that is a score below what someone would get if they guessed at random, reflects a poor effort or feigning.

This approach is predicated on a number of assumptions: 1) that a sufficient number of questions are asked to allow random answering to average out to mean; 2) that each question is independent; 3) that all questions are of equal difficulty and interchangeable; 4) that answering patterns which are at chance or below chance reflect intent; 5) that the performance on forced choice or binary-response instruments have sufficiently low cognitive load and are not associated with cognitive impairment; and 6) that forced choice and binary-response instruments are free from the effects of cultural and socio-demographic characteristics, and thus the same cut point can be used for everyone. A review of the available literature suggests that each of these assumptions is either directed violated or insufficiently studied in the context of PVT/SVT testing. As a result, inferences commonly derived about effort and feigning from chance or worse than chance performance appear to be unsupported by the empirical data.

A below chance score on a forced choice or binary-response test refers to a score below that which is expected based on random guessing; therefore, typically below fifty percent correct. If a person were to repeat the test multiple times, and randomly answer each question on each administration, the average number correct would be fifty percent. Performance below chance is considered to be evidence of an intent to deceive and has been used as an adequate basis for determining that a person has malingered despite only limited research on the sensitivity and specificity of below chance and significantly below chance performance for confirmed poor effort on PVTs [97]. In a large study of 1032 patients referred for neuropsychological evaluation, significantly below chance performance was reported for three instruments: the Portland Digit Recognition Test (PDRT), the Word Memory Test (WMT), and the TOMM. Ninety-five percent of subjects had financial incentives and most were TBI referrals, making this a population suspected to be more likely to feign or provide poor effort. Depending on the cut scores used, when all three instruments were given, 11.8 percent (95% CI: 8.8, 15.8) to 13.4 percent (95% CI: 1.1, 17.6) performed significantly below chance. Patients performed worse on the harder portions of the PDRT and WMT compared to the easy parts [97].

This study relied upon the assumptions listed above and observed performance significantly worse than chance for a significant portion of patients. Yet, because the assumptions underlying such studies are violated, the percentage of patients reported to be feigning or not trying or seeking to deceive may be substantial overestimates and potentially unsupportable. For instance, a study of college psychology students' understanding of every-day physics was designed to assess how these students applied knowledge, experience, intuition and beliefs about the physical world to seventeen questions. The test measured their understanding of phenomena encountered in daily life, such as asking: when an object falls off a moving truck, will it roll in the direction the truck was moving or in the opposite direction? Not surprisingly, students performed poorly on the quiz over all. Remarkably, although not a focus of the research, nearly 40 out of the 150 students performed below chance [98]. In a follow-up with the same subjects, the researchers found that prior academic exposure and better grades measured by self-reported GPA were associated with better performance. This study suggests that worse than chance performance is common among people who rely on beliefs and intuition when answering questions because that beliefs and intuition may be incorrect. Of course, the everyday physics questions are difficult in comparison to PVT/SVT questions which are designed to be quite simple.

Although a different design and question, determining why 27% of students performed worse than chance on the physics test may be instructive to understanding how PVT forced choice test performance below chance could be interpreted. First, the likelihood of performing below chance is affected by the number of items on a test. Using coin tossing as the classic example, the 50/50 split in heads or tails is based on the assumption that the coin is tossed equivalently, without environmental conditions that affect its flight and landing, and that sufficient trials are attempted to obtain a true average. If a coin is tossed only ten times, despite the probability for each toss being 50/50, the results may or may not be 50/50. Toss a coin one thousand times, and the heads to tails ratio will almost certainly approach 50/50, all else held constant. But how many forced choice questions are enough to account for chance? Greve et al's (2009) study suggests that more is not necessarily better with respect to multiple tests, since they found the rates of below chance performance to be higher when three tests were given than on any single test itself [97]. No research has been conducted on the optimum number of questions needed to conclude that below chance performance on PVTs is evidence of feigning, poor effort, or malingering.

Second, are the questions independent? In the coin toss example, each toss is independent of every other toss. Few research studies have determined the degree to which items on PVTs are correlated. If the items are not correlated, then the scale is unlikely to be measuring a single construct. The question of independence of items matters when performance below chance is a cut-off. The scales as a whole should have correlated items, but this may make the inferences about chance performance inaccurate. Again, no studies have been located to address this question. As is seen in the physics knowledge study, answering patterns suggest that some information, even when it is inaccurate, may be systematically applied to a set of questions, and where inaccurate information used to answer one question is applied to subsequent questions, this increases the likelihood of below chance performance. Third, as seen in the large study of multiple SVT's and below chance performance, not all questions are equivalent, with more difficult questions having higher below chance performance than easier questions, at least among people who are suspected to be more likely to feign. No research has addressed this question among other populations. In neuropsychological testing, increasing difficulty of items can be an embedded technique for assessing effort because the expectation is that a person should get more of the easier questions correct and fewer as the questions become more difficult. On PVTs the expectation is that a person who puts forth adequate efforts can answer all or most of the questions correctly, which sets an ceiling which has not been validated for the measures and presumes that every question is the same difficulty. This assumption appears not to be supported since at least some PVTs have different levels of difficulty within the test [97].

Fourth, each of these studies raises doubt as to what conclusions can be drawn from below chance performance. Answering patterns which are at or below chance may, if the test instruments work despite violating these assumptions, indicate invalid performance on the test, and suggest invalid performance on other measures within the cognitive test battery. However, failure on PVTs, even on multiple PVTs, does not speak to or reflect a person's intent. The everyday physics study is perhaps the clearest example of this, where students are believed by the researchers to have put forth their best effort, but twenty-seven percent performed below chance. Perhaps most importantly, additional factors specific to the design of the tests and the individual, other than feigning and poor effort, may also result in below chance performance. In this circumstance, educational quality, reading ability, comprehension, discriminatory experiences, or culturally interpretation of the task or question may all be in play without any current understanding of how or to what degree they affect results.

Fifth, another area of concern, and where only very limited research is available, is whether the cognitive load associated with forced choice or binary-response instruments is sufficiently low, and whether higher cognitive load increases the risk of below chance performance. A study of word and picture memory and attention, cognitive domains often drawn upon by PVT/SVT tasks, used randomly paired pictures and words and following that with a repetition phase, and then a recognition phase [99]. Recognition is a commonly used forced choice method for PVTs, and in this study, 50 previously seen words, but now unpaired with the pictures, and 50 not previously seen words were shown. Half of participants were assigned to attend to the pictures, and half were assigned to attend to the words. For those assigned to attend to the pictures, performance on the word recognition task was below chance. In the recognition section of the experiment when the words were not aligned with the images, performance was substantially below chance, at 37 percent correct. This study suggests that because cognitive load is relevant to performance on simple attentional and memory related tasks, below chance performance on similarly designed PVTs may represent cognitive impairment instead of feigning deficit.

Related to this, while a condition such a TBI is diagnosed on established criteria, occurrence of comorbid psychiatric symptoms is often unmeasured and under-valued when determining how a person responds to forced choice and binary response questions, as well as to their likelihood of endorsing unusual symptoms. Somatic symptoms and recognition memory, again common domains drawn upon by PVTs, were measured in 272 subjects with TBI followed for two years [100]. Four patients who consistently performed below chance (biased responders: BR) and eight patient controls were compared. Recognition memory for the controls remained constant and very good across the two years of assessment. However, for the BR group, recognition memory performance at baseline was equal to the controls but dropped below chance at six month and one year follow-up, recovering at twenty four months but still below controls. Interestingly, scores on the somatic questions of the Hamilton Rating Scale for Depression were inversely related to performance on the recognition memory test for the BR group: as somatic symptoms increased, recognition memory decreased. This may affect the way in which symptom validity and tests relying on expected recognition memory ability should be interpreted.

In summary, the available research studies suggest that below chance performance on forced choice measures is more common among normally functioning people, and is more common among people with true neurological impairment, than previously thought. These findings cast doubt on the positive and negative predictive power, as well as sensitivity and specificity, of below chance performance for poor effort or feigning. However, it is possible that regardless of the cause (e.g., poor effort, psychiatric symptoms, inability to comprehend simple tasks, or significant cognitive impairment), we cannot be confident that below chance performance on PVTs indicates invalid performance, and we cannot assume that below chance performance invalidates the cognitive assessment more broadly.

Culture and symptom endorsement

Somatic symptoms related to psychiatric illness have also been found to be differentially endorsed based on culture and socio-demographic characteristics [101-104]. The concern that PVT/SVT failure, in some cases, may be based on cultural differences and not poor effort or feigning is increased by research which indicates that one year after TBI, a subject's race and ethnicity are associated with neurobehavioral symptom severity and type of symptoms endorsed [105]. Similarly, as described above in relation to the RDS, many types of cognitive testing which were once thought to be culture or language free, are in fact vulnerable to effects of culture, education, and language. No evidence is currently available to determine with certainty, or to determine the level of certainty, how below chance performance on forced choice and other binary response tests, like the TOMM, should be interpreted beyond the conclusion that the test is not valid. Reaching beyond that conclusion to infer intent, lack of effort, feigning, or otherwise rule out the presence of true impairment would appear unsupported by the state of the science.

PVT/SVT's and IQ

As a result of the primacy of IQ testing in disability determinations, special attention must be given to the use of PVT and SVT's in determining intellectual disability (ID). Much of the recent research on this question has arisen in the context of capital cases in which determinations of ID can bar a person's execution under *Atkins v Virginia* (536 U.S. 304, 2002). Although far from settled, a number of reviews suggest that it is not appropriate to use PVT/SVT's with people with ID because the instruments have poor specificity and sensitivity with such groups. For instance, Salekin and Doane (2009) reviewed the state of effort tests and malingering instruments in use with people with ID or suspected of having ID, concluding that the instruments should not be used with this population [106]. They reviewed the research on ID and the TOMM, the Dot Counting Test, the Validity Indicator Profile, and the Word Memory Test, concluding that failure on any of these instruments was not adequately informative to determine if a person was feigning ID. They recommended using clinical judgment to assess strengths and weaknesses, and careful integration of information drawn from many sources.

In a study of neuropsychology outpatients who were not in litigation and not seeking benefits, researchers reported that as IQ performance declined, effort test failure increased [107]. Patients with IQ scores between 70 and 79 failed on average one of the nine efforts tests (range 0 to 4); for those with IQ's between 60 and 69, nearly 3 test failures (range 1 to 6); and those with IQ's 50 to 59 averaged 4 test failed (range 1 to 6). Even those with higher IQ's in this sample of neuropsychological clinic patients failed effort tests however, with even the highest IQ performers (above 120), failing five percent of the time [107]. Hurley and Deal, reported similar findings for people with ID living in residential care facilities and having no criminal justice involvement, raising doubts as to effort and malingering instruments ability to correctly classify people with ID [108]. Others have similarly found poor specificity when using PVTs and SVTs among people with ID [95, 109-111].

i) PVT/SVTs and co-occurring illnesses

The difficulty of identifying feigning or lack of effort in specific psychiatric or physiologic conditions is not unique to ID. People with Post-traumatic stress disorder (PTSD) may also be at heightened risk of failing PVT/SVT's because they experience and report unusual combinations of symptoms compared to healthy controls; PTSD is associated with memory impairments which may result in inconsistent recollection and reporting of symptoms; symptoms may fluctuate longitudinally; and people may lack insight about the nature and scope of their condition [112-114]. This difficulty arising from the unusual presentation of symptoms may also be true for co-morbid conditions in which a person has more than one "cause" for symptoms and behaviors, resulting in complicated presentation and self-report. A study of veterans with PTSD who were seeking compensation found that those with comorbid psychiatric illnesses had higher reporting of symptoms overall and also were more likely to appear to be feigning symptoms on MMPI validity scales [114].

PVT failure may be associated with the condition underlying the referral for evaluation. Research on people with cognitive impairment, dementia, Huntington's, and Alzheimer's diseases suggest that the cognitive processes of the disease increase the risk of false positive scores on PVT/SVT's [115-121]. For instance, Davis and Millis (2014) studied patients referred for neuropsychological evaluation who were administered multiple PVTs [69]. They reported that education and activities of daily living (functional status) were significant predictors of failure on PVTs, suggesting that real-world disability raises the risk of failure. They also found that medico-legal status was associated with failure. Using a cutoff of failure on 2 PVTs, the failure rates among a sample of neurologic-no-incentive patients administered 7 or 8 PVTs was 15%, higher than expected based on reported specificities. Although a small patient sample, this study raises a few interesting issues: first, impaired functional status, meaning a person's capacity for performing daily living tasks, is associated with more failure on PVTs. This suggests that inferences drawn from failures on PVTs must be cautious because these instruments may be measuring true mental or physical impairment, not simply effort or feigning. Second, the association with education indicates that PVTs require more complete normative standardization and study because lived experience and environmental factors may increase the risk for failure. Third, reported specificities of at least some of these test instruments may be lower than the actual use of the instruments demonstrates with patient samples, meaning that the risk for misclassification is higher than stated.

A recent study of predictors of PVT/SVT failure which used regression models so that multiple conditions could be allowed to operate simultaneously in the predictive model, found that education, race, ethnicity, immigration status, and self-reported psychiatric illness, in addition to whether the person was seeking compensation, all differed between those who passed and those who failed [122]. Ethnicity was not predictive of failure in a model adjusted for when foreign-born status was also included because of the strength of the effect of being foreign-born. The psychiatric illness measure was a self-report screening which could reflect symptom over-reporting among those who also fail PVT/SVT's, but the authors suggest that it is unlikely to be a sufficient explanation for their finding given the sample size and the post-hoc review of patient files to confirm whether symptoms had been long-standing or newly endorsed. This study raises important questions about what is being measured by PVT/SVT's when an examinee has a psychiatric illness, whether co-occurring or as a primary concern.

j) PVT/SVTs, bias and discrimination

This overlap between symptoms and expectations may also disproportionately affect cross-cultural assessments, increasing the risk of misclassification [123]. Presentation style, cultural styles of expressing symptoms, test administrator bias, stereotype threat, and a host of other issues may each make correct classification less likely in cross-cultural assessment. In addition, diagnosis of underlying, clearly defined medical and psychiatric conditions are not without race and ethnic biases. For instance, in a large study of race and ethnic bias in the diagnosis of confirmed cases of autism, Mandell et al (2009) reported that African Americans, Hispanics, and other non-White children were less likely to have the diagnosis autism documented [124]. They conclude that this indicates a significant disparity in the identification of children with autism by race/ethnicity.

Other research has demonstrated adversarial allegiance in testing, meaning that examiner perception that one side or the other has retained them, increases the likelihood that they score a subject in accord with the outcome that side seeks. For instance, in the criminal prosecution setting, examiners who believed that they were retained by the prosecution scored subjects higher on tests of actuarial risk and compared to examiners who believed they were defense retained, who found lower risk [125]. When this diagnostic bias is combined with presumptions that people with financial incentives and the lack of scientifically supported techniques for assessing feigning and malingering cross-culturally, the risk of misclassification seems unacceptably high.

In general, although few studies have examined the question of educational level and failure on PVT/SVT's, most that have, report differences by educational level [126-128]. Some of the cognitive domains on which PVT/SVT's draw, are reported to be unbiased by educational but when these domains are more fully assessed with standard neuropsychological test instruments, meaningful differences by education and literacy are observed. In turn, this may appear to be racial or ethnic differences when literacy and quality of education are not measured or considered, even on tasks which are not primarily language based [129-131]. For instance, using a cancellation task in which people are asked to visually search for a shape or letter and ignore distractors, researchers conducting a community-based epidemiological study of aging and dementia which included more than 1,400 subjects who were matched by years of education [55]. As expected, differences in test performance by speed and efficiency were observed by race and ethnic group. However, when matched by WRAT-3 reading recognition scores as a measure of literacy and quality of education, the group differences by race/ethnicity were accounted for. Based on our review, no PVT/SVT standardization or validation study has taken into account reading recognition (as a proxy for school quality), or literacy, in determining group differences, raising another set of as yet unanswered questions about the normative samples, and more importantly, about what factors influence performance on these instruments and how that relates to neuropsychological test batteries.

Research on stereotype threat and perceived race discrimination lends a cautionary note as well. Using common neuropsychological test instruments, researchers found significant performance differences in processing speed, learning, and memory testing, the types of cognitive domains which many PVT/SVTs use to measure effort and faking, based on levels of perceived discrimination and stereotype threat in the testing process [132]. Although this study was not considering PVT/SVT performance, the implications of this for such tests, and the conditions of the administration of such testing, are clear indicia of an important gap in the research on the use and interpretation of PVT/SVT's.

Overall, despite the proliferation of instruments, inadequate demographic norms exist for the interpretation of PVT/SVTs administered to non-White, non-English speaking, immigrant, poor, non-healthy, across age span and from non-dominate cultural groups, or those who have experienced poor quality schooling. In an archival record-based study of patients referred for neuropsychological evaluation at a hospital based clinic, the use of PVT/SVT’s with different ethnic groups in Los Angeles was assessed. The subjects included English-as-a-second-Language speakers as well as people with a variety of medical and psychiatric conditions. The study’s purpose was to obtain within racial group norms and propose cut points for each ethnic grouping based on those norms. Even in this small, non-representative patient study, differences based on the demographic characteristics, as well as educational level and primary language, were observed which led to excess false-positives for these groups [87].

Whether or not demographic norms would assist in better identification of effort and feigning cannot be known at this time, and as reviewed above, the field still needs to establish the scientific rigor of PVT/SVT's in general, along with the more specific application of these instruments across demographically diverse populations. The existing evidence tends to demonstrate that true real-world impairment and a number of demographic characteristics, may all increase the risk of misclassification of poor effort by SVTs/PVTs, but the amount of research on this topic is inadequate to draw conclusions with high confidence. Research studies suggest that PVT/SVTs have inadequate construct validity across demographically diverse people, and proposals to shift cut scores and analyze frequency of responding within diverse subgroups groups does not address this concern.

A widely used test instrument with imbedded SVT measures is the MMPI-2 which consists of 338 true-false questions used to create a variety of scales [133]. The MMPI-2 was normed on 2,600 people from many geographic regions in the United States. The reading level required is said to be grade 4.5 (Flesch-Kincaid). Despite the large, representative normative sample, race and culture concerns with the effort and faking scales remain. A recent review reported that despite few studies and study design issues, the MMPI-2 has been found to both under- and over-pathologize African Americans, Native Americans, and Asian Americans, including elevations on scales which are meant to reflect faking and over-reporting of symptoms [134], although others have not reported similar discrepancies [135, 136]. Another study to assess the validity of the MMPI-2, compared white and African American subjects, drawn from a large tertiary care veterans hospital and a large urban community hospital, on clinical and validity scales by race [137]. They found African American men scored higher than white men on a number of scales: Infrequency (F), 6 Paranoia (Pa), 8 Schizophrenia (Sc), and 9 Hypomania (Ma) and (4) psychopathic deviance and that African American women scored higher than white women on scales Paranoia (Pa) and hypomania (Ma); white men and women scored higher on the K (Correction) scale. Taken together, these findings suggest that African Americans are more likely to be negatively characterized by the MMPI-2, and the African American men are more likely to be viewed as deviant and unusual answering patterns. A more recent study of veterans in substance abuse treatment, also found differential score patterns by race on the MMPI-2. This study excluded invalid responders, those with too many blanks or high VRIN or TRIN scales, yet still found over and under-pathologizing of African Americans [138].

Similarly, false positive rates for the MMPI-2 effort and faking scales has been reported to be elevated for people with some psychiatric and neurological conditions, as well as for those with conditions which lead to high somatic complaints. For instance, earlier research with veterans diagnosed with PTSD found that being African American predicted high F scale which would likely lead to determinations of over-reporting and feigning [114]. F scale elevations have also been reported for adult child sexual abuse victims [139] and numerous studies have critiqued specific scales such as the Faking-bad scale [140, 141].

Future directions for research on PVT/SVTs

The quality of research on the development and standardization of PVT/SVTs must improve. Future efforts would benefit from more rigorous research designs, using large, diverse, cohorts within the population of interest (SSA Disability applicants) who are followed longitudinally. The current state of the field collects data on convenience samples who are not representative, and who do not provide basis on which to determine how a measure will function in the SSA Disability context. Furthermore, many cut-offs are based on "known group" designs that classify individuals based on assumptions which cannot be tested and do not provide sufficient support for drawing conclusions about poor effort, feigning or intent.

Because it is a rarity to find PVT/SVT research outside of the US and outside of certain highly selective litigation settings, the way in which feigning and poor effort present across diverse cultural and educational groups and in different assessment settings is unknown. Strategies to determine the base rates of feigning and poor effort in various populations, and the causes and consequences of poor effort during cognitive testing should be developed. Cultural and social-demographic characteristics may be associated with differing base rates of poor effort. The format of each PVT/SVT (e.g., forced choice or not, embedded or stand-alone) may have an effect on sensitivity and specificity of measures across cultures and settings..

Similarly, PVT/SVT research should focus on developing standard error measurements and 95% confidence intervals around observed scores.

Overall Summary

In summary, formal assessment of cognitive function can provide critical insights into the strengths and weaknesses of a person's memory, perceptual, information, motor, social, language, learning, and executive processes. Cognitive test data provide accurate information for diagnostic and clinical decisions, as well as prediction of real-world functioning. Any assessment in which a person's medical, neurological, or psychiatric diagnosis is under consideration and the relationship of that condition to functional capacity is being determined, cognitive assessment should be a fundamental component of the standard of care.

Many individual cognitive measures and comprehensive cognitive batteries have been properly validated and standardized for assessment of people across the life course. However there are groups of people with diverse cultural, racial, linguistic, or socioeconomic experiences, who need cognitive assessment, but who were not included in the process of developing, standardizing, or norming these measures. Cognitive tests are only valid when administered in the standardized way, and among people who were represented in the validation and normative samples. People who are not putting forth adequate effort on the tests, or people who cannot understand test instructions or engage properly in cognitive testing are not represented in standardization or normative cohorts and therefore, the tests are not valid assessments of cognitive function when administered to them.

Normative standards for cognitive measures differ depending on whether the tests are being used to diagnose acquired impairment, to determine where a person stands with respect to the normal distribution of a cognitive ability within the general population, or to determine capacity to perform real-world tasks. While many cognitive measures have rigorously collected, large and representative standardization or normative cohorts, some widely used measures have surprisingly small and non-representative normative samples. Furthermore, use of tests and their normative data must take into account loss of test sensitivity due to secular or cohort effects as the original normative data becomes older, practice effects, and cultural bias have been shown to alter the accuracy of cognitive test scores. While demographic adjustments are a useful tool for estimating premorbid performance on cognitive measures and diagnosis of acquired impairment, they should not be used to describe where an individual's cognitive ability sits within the general population, or for prediction of functional ability in the real-world. Demographic norms are not recommended when characterizing an acquired impairment in people with neurodevelopmental disorder that may have altered the expected relationship between demographic variables (such as educational attainment) and cognitive function.

Our review of the current state of research on PVT/SVTs suggests that there is inadequate empirical data in general, and their reliability and validity can't be determined in settings where the base-rate of poor effort or feigning is unknown. Very few studies of SVT/PVTs have included confidence intervals, estimates of error, or follow-up of participants in order to provide an appropriate gold standard for the determination of sensitivity and specificity. Moreover, almost no data has been developed to demonstrate reliability and validity for the use of PVT/SVTs among non-White and well-educated cultural and socio-demographic groups.

REFERENCES

1. Streiner, D.L. and G.R. Norman, *Health measurement scales : a practical guide to their development and use*. 4th ed. 2008, Oxford ; New York: Oxford University Press. xvii, 431 p.
2. Lezak, M.D., et al., *Neuropsychological assessment*. 5th ed. ed. 2012, Oxford: Oxford University Press.
3. Manly, J.J. and R.J. Echemendia, *Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology.* Archives of Clinical Neuropsychology, 2007. **22**(3): p. 319-325.
4. Reitan, R.M. and D. Wolfson, *The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation*. 1993, Tucson, AZ: Neuropsychology Press.
5. Manly, J.J., et al., *Implementing diagnostic criteria and estimating frequency of mild cognitive impairment in an urban community.* Archives of Neurology, 2005. **62**(11): p. 1739-1746.
6. Fyffe, D.C., et al., *Explaining Differences in Episodic Memory Performance among Older African Americans and Whites: The Roles of Factors Related to Cognitive Reserve and Test Bias.* Journal of the International Neuropsychological Society, 2011. **17**(4): p. 625-638.
7. Busch, R.M., G.J. Chelune, and Y. Suchy, *Using Norms in Neuropsychological Assessment of the Elderly,* in *Geriatric Neuropsychology: Assessment and Intervention,* D.K. Attix and K.A. Welsh-Bohmer, Editors. 2006, Guilford Press: New York.
8. Schalok, R., et al., *AAIDD's 11th edition of Intellectual Disability: Definition, Classification, and Systems of Support.* 2012.
9. Steinerman, J.R., et al., *Modeling Cognitive Trajectories Within Longitudinal Studies: A Focus on Older Adults.* Journal of the American Geriatrics Society, 2010. **58**: p. S313-S318.
10. Dennis, M., et al., *Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders.* Journal of the International Neuropsychological Society, 2009. **15**(3): p. 331-343.
11. Keefe, R.S.E., *The Longitudinal Course of Cognitive Impairment in Schizophrenia: An Examination of Data From Premorbid Through Posttreatment Phases of Illness.* Journal of Clinical Psychiatry, 2014. **75**: p. 8-13.
12. Silverberg, N.D. and S.R. Millis, *Impairment versus deficiency in neuropsychological assessment: Implications for ecological validity.* Journal of the International Neuropsychological Society, 2009. **15**(1): p. 94-102.
13. Higginson, C.I., et al., *The contribution of trail making to the prediction of performance-based instrumental activities of daily living in Parkinson's disease without dementia.* Journal of Clinical and Experimental Neuropsychology, 2013. **35**(5): p. 530-539.
14. Barrash, J., et al., *Prediction of driving ability with neuropsychological tests: Demographic adjustments diminish accuracy.* Journal of the International Neuropsychological Society, 2010. **16**(4): p. 679-686.
15. Trahan, L.H., et al., *The Flynn Effect: A Meta-Analysis.* Psychological Bulletin, 2014: p. No Pagination Specified.
16. Flynn, J.R., *MASSIVE IQ GAINS IN 14 NATIONS - WHAT IQ TESTS REALLY MEASURE.* Psychological Bulletin, 1987. **101**(2): p. 171-191.

17. Flynn, J.R., *THE MEAN IQ OF AMERICANS - MASSIVE GAINS 1932 TO 1978.* Psychological Bulletin, 1984. **95**(1): p. 29-51.
18. Kanaya, T. and S. Ceci, *The Impact of the Flynn Effect on LD Diagnoses in Special Education.* Journal of Learning Disabilities, 2012. **45**(4): p. 319-326.
19. Sanborn, K.J., et al., *Does the Flynn Effect differ by IQ level in samples of students classified as learning disabled?* Journal of Psychoeducational Assessment, 2003. **21**(2): p. 145-159.
20. Truscott, S.D. and A.J. Frank, *Does the Flynn effect affect IQ scores of students classified as LD?* Journal of School Psychology, 2001. **39**(4): p. 319-334.
21. de Rotrou, J., et al., *Does Cognitive Function Increase over Time in the Healthy Elderly?* Plos One, 2013. **8**(11).
22. Dickinson, M.D. and M. Hiscock, *The Flynn Effect in Neuropsychological Assessment.* Applied Neuropsychology, 2011. **18**(2): p. 136-142.
23. Strauss, E., E.M.S. Sherman, and O. Spreen, *A compendium of neuropsychological tests : administration, norms, and commentary*. 3rd ed. 2006, New York: Oxford University Press.
24. Salthouse, T.A., *Within-Cohort Age-Related Differences in Cognitive Functioning.* Psychological Science, 2013. **24**(2): p. 123-130.
25. Ronnlund, M. and L.G. Nilsson, *Flynn effects on sub-factors of episodic and semantic memory: Parallel gains over time and the same set of determining factors.* Neuropsychologia, 2009. **47**(11): p. 2174-2180.
26. Fletcher, J.M., K.K. Stuebing, and L.C. Hughes, *IQ Scores Should Be Corrected for the Flynn Effect in High-Stakes Decisions.* Journal of Psychoeducational Assessment, 2010. **28**(5): p. 469-473.
27. Gresham, F.M., *Interpretation of Intelligence Test Scores in Atkins Cases: Conceptual and Psychometric Issues.* Applied Neuropsychology, 2009. **16**(2): p. 91-97.
28. Flynn, J.R., *Tethering the elephant: Capital cases, IQ, and the Flynn effect.* Psychology, Public Policy, and Law, 2006. **12**(2): p. 170-189.
29. Sirois, P.A., et al., *Quantifying practice effects in longitudinal research with the WISC-R and WAIS-R: A study of children and adolescents with hemophilia and male siblings without hemophilia.* Journal of Pediatric Psychology, 2002. **27**(2): p. 121-131.
30. Waber, D.P., et al., *Four-Year Longitudinal Performance of a Population-Based Sample of Healthy Children on a Neuropsychological Battery: The NIH MRI Study of Normal Brain Development.* Journal of the International Neuropsychological Society, 2012. **18**(2): p. 179-190.
31. Matarazzo, J.D. and D.O. Herman, *RELATIONSHIP OF EDUCATION AND IQ IN THE WAIS-R STANDARDIZATION SAMPLE.* Journal of Consulting and Clinical Psychology, 1984. **52**(4): p. 631-634.
32. Siders, A., A.S. Kaufman, and C.R. Reynolds, *Do practice effects on Wechsler's performance subtests relate to children's general ability, memory, learning ability, or attention?* Applied Neuropsychology, 2006. **13**(4): p. 242-250.
33. Kaufman, A.S., *Practice Effects*, in *Encyclopedia of intelligence*, R.J. Sternberg, Editor. 1994, MacMillan: New York. p. 828-33.
34. Ronnlund, M., et al., *Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study.* Psychology and Aging, 2005. **20**(1): p. 3-18.

35. Van der Elst, W., et al., *Detecting the significance of changes in performance on the Stroop Color-Word Test, Rey's Verbal Learning Test, and the Letter Digit Substitution Test: The regression-based change approach.* Journal of the International Neuropsychological Society, 2008. **14**(1): p. 71-80.
36. Salthouse, T.A., *Why Are There Different Age Relations in Cross-Sectional and Longitudinal Comparisons of Cognitive Functioning?* Current Directions in Psychological Science, 2014. **23**(4): p. 252-256.
37. Deary, I.J., *The Stability of Intelligence From Childhood to Old Age.* Current Directions in Psychological Science, 2014. **23**(4): p. 239-245.
38. Schneider, W., F. Niklas, and S. Schmiedeler, *Intellectual development from early childhood to early adulthood: The impact of early IQ differences on stability and change over time.* Learning and Individual Differences, 2014. **32**: p. 156-162.
39. Whitaker, S., *The stability of IQ in people with low intellectual ability: An analysis of the literature.* Intellectual and Developmental Disabilities, 2008. **46**(2): p. 120-128.
40. Woodberry, K.A., A.J. Giuliano, and L.J. Seidman, *Premorbid IQ in schizophrenia: A meta-analytic review.* American Journal of Psychiatry, 2008. **165**(5): p. 579-587.
41. Quraishi, S. and S. Frangou, *Neuropsychology of bipolar disorder: a review.* J Affect Disord, 2002. **72**(3): p. 209-26.
42. Mesholam-Gately, R.I., et al., *Neurocognition in first-episode schizophrenia: a meta-analytic review.* Neuropsychology, 2009. **23**(3): p. 315-36.
43. Goodwin, G.M., et al., *Cognitive impairment in bipolar disorder: Neurodevelopment or neurodegeneration? An ECNP expert meeting report.* European Neuropsychopharmacology, 2008. **18**(11): p. 787-793.
44. Uzzell, B.P., M. Ponton, and A. Ardila, *International Handbook of Cross-Cultural Neuropsychology*. 2013: Taylor & Francis.
45. Rhodes, R.L., S.H. Ochoa, and S.O. Ortiz, *Assessing Culturally and Linguistically Diverse Students: A Practical Guide*. 2005: Guilford Press.
46. Fletcher-Janzen, E., T.L. Strickland, and C. Reynolds, *Handbook of Cross-Cultural Neuropsychology*. 2000: Springer US.
47. Ruiz, J.M., P. Steffen, and T.B. Smith, *Hispanic Mortality Paradox: A Systematic Review and Meta-Analysis of the Longitudinal Literature.* American Journal of Public Health, 2013. **103**(3): p. E52-E60.
48. Acevedo-Garcia, D., et al., *Toward a policy-relevant analysis of geographic and racial/ethnic disparities in child health.* Health Affairs, 2008. **27**(2): p. 321-333.
49. Morenoff, J.D., et al., *Understanding social disparities in hypertension prevalence, awareness, treatment, and control: The role of neighborhood context.* Social Science & Medicine, 2007. **65**(9): p. 1853-1866.
50. Williams, D.R., H.W. Neighbors, and J.S. Jackson, *Racial/ethnic discrimination and health: Findings from community studies.* American Journal of Public Health, 2003. **93**(2): p. 200-208.
51. Silverberg, N.D., R.A. Hanks, and S.C. Tompkins, *Education Quality, Reading Recognition, and Racial Differences in the Neuropsychological Outcome from Traumatic Brain Injury.* Archives of Clinical Neuropsychology, 2013. **28**(5): p. 485-491.
52. Manly, J.J., *Advantages and disadvantages of separate norms for African Americans.* Clinical Neuropsychologist, 2005. **19**(2): p. 270-275.

53. Manly, J.J., *Deconstructing race and ethnicity - Implications for measurement of health outcomes.* Medical Care, 2006. **44**(11): p. S10-S16.
54. Manly, J.J., et al., *Acculturation, reading level, and neuropsychological test performance among African American elders.* Applied Neuropsychology, 2004. **11**(1): p. 37-46.
55. Byrd, D.A., et al., *Cancellation test performance, in African American, Hispanic, and White elderly.* Journal of the International Neuropsychological Society, 2004. **10**(3): p. 401-411.
56. Shuttleworth-Edwards, A.B., et al., *Cross-cultural effects on IQ test performance: A review and preliminary normative indications on WAIS-III test performance.* Journal of Clinical and Experimental Neuropsychology, 2004. **26**(7): p. 903-920.
57. Suen, H.K. and S. Greenspan, *Serious problems with the Mexican norms for the WAIS-III when assessing mental retardation in capital cases.* Applied neuropsychology, 2009. **16**(3): p. 214-22.
58. Demsky, Y., et al., *Optimal short forms of the Spanish WAIS (EIWA).* Assessment, 1998. **5**(4): p. 361-364.
59. Gasquoine, P.G., et al., *Language of administration and neuropsychological test performance in neurologically intact Hispanic American bilingual adults.* Archives of Clinical Neuropsychology, 2007. **22**(8): p. 991-1001.
60. Renteria, L., S.T. Li, and N.H. Pliskin, *Reliability and validity of the Spanish language Wechsler adult intelligence scale (3rd edition) in a sample of American, urban, Spanish-speaking Hispanics.* Clinical Neuropsychologist, 2008. **22**(3): p. 455-470.
61. Romero, H.R., et al., *Challenges in the Neuropsychological Assessment of Ethnic Minorities: Summit Proceedings.* Clinical Neuropsychologist, 2009. **23**(5): p. 761-779.
62. Flanagan, D.P. and P.L. Harrison, *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. 2012: Guilford Press.
63. Neisser, U., et al., *Intelligence: Knowns and unknowns.* American Psychologist, 1996. **51**(2): p. 77-101.
64. Bush, S.S., et al., *Symptom validity assessment: Practice issues and medical necessity - NAN policy & planning committee.* Archives of Clinical Neuropsychology, 2005. **20**(4): p. 419-426.
65. Heilbronner, R.L., et al., *American Academy of Clinical Neuropsychology Consensus Conference Statement on the Neuropsychological Assessment of Effort, Response Bias, and Malingering.* Clinical Neuropsychologist, 2009. **23**(7): p. 1093-1129.
66. Sollman, M.J. and D.T.R. Berry, *Detection of Inadequate Effort on Neuropsychological Testing: A Meta-Analytic Update and Extension.* Archives of Clinical Neuropsychology, 2011. **26**(8): p. 774-789.
67. Larrabee, G.J., *False-Positive Rates Associated with the Use of Multiple Performance and Symptom Validity Tests.* Archives of Clinical Neuropsychology, 2014. **29**(4): p. 364-373.
68. Berthelson, L., et al., *False positive diagnosis of malingering due to the use of multiple effort tests.* Brain Injury, 2013. **27**(7-8): p. 909-916.
69. Davis, J.J. and S.R. Millis, *Examination of Performance Validity Test Failure in Relation to Number of Tests Administered.* Clinical Neuropsychologist, 2014. **28**(2): p. 199-214.
70. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern epidemiology*. 3rd ed. 2008, Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins. x, 758 p.

71. Gewandter, J.S., et al., *Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations.* Pain, 2014. **155**(3): p. 461-466.
72. Medici, R., *The Use of Multiple Performance Validity Tests.* Journal of Forensic Psychology Practice, 2013. **13**(1): p. 68-78.
73. Tombaugh, T.N., *Test of Memory Malingering*. 1996, North Tonawanda, NY: Multi-health Systems.
74. Rees, L.M., et al., *Five validation experiments of the Test of Memory Malingering (TOMM).* Psychological Assessment, 1998. **10**(1): p. 10-20.
75. Susser, E.S., et al., *Psychiatric epidemiology : searching for the causes of mental disorders*. 2006, Oxford ; New York: Oxford University Press. xxii, 516 p.
76. Slick, D.J., et al., *Victoria symptom validity test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales.* Journal of Clinical and Experimental Neuropsychology, 1996. **18**(6): p. 911-922.
77. Grote, C.L., et al., *Performance of compensation seeking and non-compensation seeking samples on the Victoria Symptom Validity Test: Cross-validation and extension of a standardization study*. Journal of Clinical and Experimental Neuropsychology, 2000. **22**(6): p. 709-719.
78. Slick, D.J., et al., *Victoria Symptom Validity Test Professional Manual*. 1997, Odessa, Florida: Psychological Assessment Resources Inc.
79. Binder, L.M., *ASSESSMENT OF MALINGERING AFTER MILD HEAD TRAUMA WITH THE PORTLAND DIGIT RECOGNITION TEST.* Journal of Clinical and Experimental Neuropsychology, 1993. **15**(2): p. 170-182.
80. Inman, T.H., et al., *Development and initial validation of a new procedure for evaluating adequacy of effort given during neuropsychological testing: The letter memory test.* Psychological Assessment, 1998. **10**(2): p. 128-139.
81. Hilsabeck, R.C., et al., *The Word Completion Memory Test (WCMT): a new test to detect malingered memory deficits.* Archives of Clinical Neuropsychology, 2001. **16**(7): p. 669-677.
82. Bigler, E.D., *Symptom validity testing, effort, and neuropsychological assessment.* J Int Neuropsychol Soc, 2012. **18**(4): p. 632-40.
83. An, K.Y., K.K. Zakzanis, and S. Joordens, *Conducting Research with Non-clinical Healthy Undergraduates: Does Effort Play a Role in Neuropsychological Test Performance?* Archives of Clinical Neuropsychology, 2012. **27**(8): p. 849-857.
84. Silk-Eglit, G.M., et al., *Base Rate of Performance Invalidity among Non-Clinical Undergraduate Research Participants.* Archives of Clinical Neuropsychology, 2014. **29**(5): p. 415-421.
85. Lippa, S.M., et al., *Ecological Validity of Performance Validity Testing.* Archives of Clinical Neuropsychology, 2014. **29**(3): p. 236-244.
86. Williamson, D.J., et al., *Abuse, Not Financial Incentive, Predicts Non-Credible Cognitive Performance in Patients With Psychogenic Non-Epileptic Seizures.* Clinical Neuropsychologist, 2012. **26**(4): p. 588-598.
87. Salazar, X.F., et al., *The use of effort tests in ethnic minorities and in non-English-speaking and English as a second language populations*, in *Assessment of Feigned Cognitive Impairment: A Neuropsychological Perspective*, K.B. Boone, Editor. 2007, Guilford Press. p. 405-27.

88. Ostrosky-Solis, F. and A. Lozano, *Digit Span: Effect of education and culture.* International Journal of Psychology, 2006. **41**(5): p. 333-341.

89. Petersson, K.M., A. Reis, and M. Ingvar, *Cognitive processing in literate and illiterate subjects: A review of some recent behavioral and functional neuroimaging data.* Scandinavian Journal of Psychology, 2001. **42**(3): p. 251-267.

90. Aleman, A., et al., *Memory impairment in schizophrenia: A meta-analysis.* American Journal of Psychiatry, 1999. **156**(9): p. 1358-1366.

91. Bucker, J., et al., *Cognitive impairment in school-aged children with early trauma.* Compr Psychiatry, 2012. **53**(6): p. 758-64.

92. Cukierman, T., H.C. Gerstein, and J.D. Williamson, *Cognitive decline and dementia in diabetes - systematic overview of prospective observational studies.* Diabetologia, 2005. **48**(12): p. 2460-2469.

93. Nishtala, A., et al., *Midlife Cardiovascular Risk Impacts Executive Function Framingham Offspring Study.* Alzheimer Disease & Associated Disorders, 2014. **28**(1): p. 16-22.

94. Yang, C.C., et al., *Cross-cultural Effect on Suboptimal Effort Detection: An Example of the Digit Span Subtest of the WAIS-III in Taiwan.* Archives of Clinical Neuropsychology, 2012. **27**(8): p. 869-878.

95. Schroeder, R.W., et al., *Reliable Digit Span: A Systematic Review and Cross-Validation Study.* Assessment, 2012. **19**(1): p. 21-30.

96. Flowers, K.A., C. Bolton, and N. Brindle, *Chance guessing in a forced-choice recognition task and the detection of malingering.* Neuropsychology, 2008. **22**(2): p. 273-277.

97. Greve, K.W., L.M. Binder, and K.J. Bianchini, *Rates of Below-Chance Performance in Forced-Choice Symptom Validity Tests.* Clinical Neuropsychologist, 2009. **23**(3): p. 534-544.

98. Norvilitis, J.M., H.M. Reid, and B.M. Norvilitis, *Success in everyday physics: The role of personality and academic variables.* Journal of Research in Science Teaching, 2002. **39**(5): p. 394-409.

99. Dewald, A.D., S. Sinnett, and L.A.A. Doumas, *Conditions of directed attention inhibit recognition performance for explicitly presented target-aligned irrelevant stimuli.* Acta Psychologica, 2011. **138**(1): p. 60-67.

100. Bierley, R.A., et al., *Biased responding: a case series demonstrating a relationship between somatic symptoms and impaired recognition memory performance for traumatic brain injured individuals.* Brain Injury, 2001. **15**(8): p. 697-714.

101. Jang, Y.R., K.H. Kwag, and D.A. Chiriboga, *Not Saying I Am Happy Does Not Mean I Am Not: Cultural Influences on Responses to Positive Affect Items in the CES-D.* Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 2010. **65**(6): p. 684-690.

102. Bagayogo, I.P., A. Interian, and J.I. Escobar, *Transcultural Aspects of Somatic Symptoms in the Context of Depressive Disorders*, in *Cultural Psychiatry*, R.D. Alarcon, Editor. 2013. p. 64-74.

103. Das-Munshi, J., et al., *Cross-cultural factorial validation of the Clinical Interview Schedule - Revised (CIS-R); findings from a nationally representative survey (EMPIRIC).* International Journal of Methods in Psychiatric Research, 2014. **23**(2): p. 229-244.

104. Kirmayer, L.J. and A. Young, *Culture and somatization: Clinical, epidemiological, and ethnographic perspectives.* Psychosomatic Medicine, 1998. **60**(4): p. 420-430.

105. Arango-Lasprilla, J.C., et al., *Neurobehavioural symptoms 1 year after traumatic brain injury: A preliminary study of the relationship between race/ethnicity and symptoms.* Brain Injury, 2012. **26**(6): p. 814-824.

106. Salekin, K.L. and B.M. Doane, *Malingering intellectual disability: the value of available measures and methods.* Appl Neuropsychol, 2009. **16**(2): p. 105-13.

107. Dean, A.C., et al., *The relationship of IQ to effort test performance.* Clin Neuropsychol, 2008. **22**(4): p. 705-22.

108. Hurley, K.E. and W.P. Deal, *Assessment instruments measuring malingering used with individuals who have mental retardation: Potential problems and issues.* Mental Retardation, 2006. **44**(2): p. 112-119.

109. Graue, L.O., et al., *Identification of feigned mental retardation using the new generation of malingering detection instruments: preliminary findings.* Clin Neuropsychol, 2007. **21**(6): p. 929-42.

110. Reznek, L., *The Rey 15-item memory test for malingering: A meta-analysis.* Brain Injury, 2005. **19**(7): p. 539-543.

111. Shandera, A.L., et al., *Detection of malingered mental retardation.* Psychol Assess, 2010. **22**(1): p. 50-6.

112. Guriel, J. and W. Fremouw, *Assessing malingered posttraumatic stress disorder: A critical review.* Clinical Psychology Review, 2003. **23**(7): p. 881-904.

113. Taylor, S., B.C. Frueh, and G.J.G. Asmundson, *Detection and management of malingering in people presenting for treatment of posttraumatic stress disorder: Methods, obstacles, and recommendations.* Journal of Anxiety Disorders, 2007. **21**(1): p. 22-41.

114. Franklin, C.L., et al., *Assessment of response style in combat veterans seeking compensation for posttraumatic stress disorder.* Journal of Traumatic Stress, 2003. **16**(3): p. 251-255.

115. Kiewel, N.A., et al., *A Retrospective Review of Digit Span-Related Effort Indicators in Probable Alzheimer's Disease Patients.* Clinical Neuropsychologist, 2012. **26**(6): p. 965-974.

116. Sieck, B.C., et al., *Symptom Validity Test Performance in the Huntington Disease Clinic.* Archives of Clinical Neuropsychology, 2013. **28**(2): p. 135-143.

117. Bortnik, K.E., M.D. Horner, and D.L. Bachman, *Performance on Standard Indexes of Effort Among Patients with Dementia.* Applied Neuropsychology-Adult, 2013. **20**(4): p. 233-242.

118. Teichner, G. and M.T. Wagner, *The Test of Memory Malingering (TOMM): normative data from cognitively intact, cognitively impaired, and elderly patients with dementia.* Archives of Clinical Neuropsychology, 2004. **19**(3): p. 455-464.

119. Merten, T., L. Bossink, and B. Schmand, *On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients.* Journal of Clinical and Experimental Neuropsychology, 2007. **29**(3): p. 308-318.

120. Hook, J.N., M.J. Marquine, and J.B. Hoelzle, *Repeatable Battery for the Assessment of Neuropsychological Status Effort Index Performance in a Medically Ill Geriatric Sample.* Archives of Clinical Neuropsychology, 2009. **24**(3): p. 231-235.

121. Dean, A.C., et al., *DEMENTIA AND EFFORT TEST PERFORMANCE.* Clinical Neuropsychologist, 2009. **23**(1): p. 133-152.

122. Webb, J.W., et al., *Effort Test Failure: Toward a Predictive Model.* Clinical Neuropsychologist, 2012. **26**(8): p. 1377-1396.

123. Ferraro, F.R., *Minority and Cross-cultural Aspects of Neuropsychological Assessment*. 2002: Taylor & Francis.

124. Mandell, D.S., et al., *Racial/Ethnic Disparities in the Identification of Children With Autism Spectrum Disorders.* American Journal of Public Health, 2009. **99**(3): p. 493-498.

125. Murrie, D.C., et al., *Are Forensic Experts Biased by the Side That Retained Them?* Psychological Science, 2013. **24**(10): p. 1889-1897.

126. Greve, K.W., et al., *Detecting Malingered Pain-Related Disability: Classification Accuracy of the Test of Memory Malingering.* Clinical Neuropsychologist, 2009. **23**(7): p. 1250-1271.

127. Stulemeijer, M., et al., *Cognitive performance after Mild Traumatic Brain Injury: The impact of poor effort on test results and its relation to distress, personality and litigation.* Brain Injury, 2007. **21**(3): p. 309-318.

128. Mahdavi, M.E., N. Mokari, and Z. Amiri, *Educational Level and Pseudohypacusis in Medico-Legal compensation Claims: A Retrospective Study.* Archives of Iranian Medicine, 2011. **14**(1): p. 58-60.

129. Byrd, D.A., D. Sanchez, and J.J. Manly, *Neuropsychological test performance among Caribbean-born and US-born African American elderly: The role of age, education and reading level.* Journal of Clinical and Experimental Neuropsychology, 2005. **27**(8): p. 1056-1069.

130. Boone, K.B., et al., *The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population.* Archives of Clinical Neuropsychology, 2007. **22**(3): p. 355-365.

131. Bramao, I., et al., *The impact of reading and writing skills on a visuo-motor integration task: A comparison between illiterate and literate subjects.* Journal of the International Neuropsychological Society, 2007. **13**(2): p. 359-364.

132. Thames, A.D., et al., *Effects of Stereotype Threat, Perceived Discrimination, and Examiner Race on Neuropsychological Performance: Simple as Black and White?* Journal of the International Neuropsychological Society, 2013. **19**(5): p. 583-593.

133. Butcher, J.N., et al., *Development and use of the MMPI-2 content scales*. 1990, Minneapolis, MN: University of Minnesota Press.

134. Hill, J.S., R.R. Robbins, and T.M. Pace, *Cultural Validity of the Minnesota Multiphasic Personality Inventory-2 Empirical Correlates: Is This the Best We Can Do?* Journal of Multicultural Counseling and Development, 2012. **40**(2): p. 104-116.

135. Tsushima, W.T. and V.G. Tsushima, *Comparison of MMPI-2 Validity Scales Among Compensation-Seeking Caucasian and Asian American Medical Patients.* Assessment, 2009. **16**(2): p. 159-164.

136. Castro, Y., et al., *Examination of racial differences on the MMPI-2 clinical and restructured clinical scales in an outpatient sample.* Assessment, 2008. **15**(3): p. 277-286.

137. Arbisi, P.A., Y.S. Ben-Porath, and J. McNulty, *A comparison of MMPI-2 validity in African American and Caucasian psychiatric inpatients.* Psychological Assessment, 2002. **14**(1): p. 3-15.

138. Monnot, M.J., et al., *Racial Bias in Personality Assessment: Using the MMPI-2 to Predict Psychiatric Diagnoses of African American and Caucasian Chemical Dependency Inpatients.* Psychological Assessment, 2009. **21**(2): p. 137-151.

139. Flitter, J.M.K., J.D. Elhai, and S.N. Gold, *MMPI-2 F scale elevations in adult victims of child sexual abuse.* Journal of Traumatic Stress, 2003. **16**(3): p. 269-274.

140. Butcher, J.N., et al., *The construct validity of the Lees-Haley Fake Bad Scale. Does this scale measure somatic malingering and feigned emotional distress?* Arch Clin Neuropsychol, 2003. **18**(5): p. 473-85.

141. Gass, C.S. and A.P. Odland, *MMPI-2 Symptom Validity (FBS) Scale: Psychometric Characteristics and Limitations in a Veterans Affairs Neuropsychological Setting.* Applied Neuropsychology-Adult, 2014. **21**(1): p. 1-8.