Information Utility: Quantifying the Total Psychometric Information Provided by a Measure

Kristian E. Markon
University of Iowa

Although advances have improved our ability to describe the measurement precision of a test, it often remains challenging to summarize how well a test is performing overall. Reliability, for example, provides an overall summary of measurement precision, but it is sample-specific and might not reflect the potential usefulness of a test if the sample is poorly suited for the test’s purposes. The test information function, conversely, provides detailed sample-independent information about measurement precision, but it does not provide an overall summary of test performance. Here, the concept of information utility is introduced. Information utility provides an index of how much psychometric information a measure (e.g., item, test) provides about a trait overall. Information utility has a number of important applied implications, including test selection, trait estimation, computerized adaptive testing, and hypothesis testing. Information utility may have particular utility in situations where the accuracy of prior information about trait level is vague or unclear.

Keywords: information, reference prior, reliability, computerized adaptive testing

One major achievement of modern measurement theory is greater precision in describing how much information a measure (e.g., a test or item) provides about an underlying latent trait. The test or item information function (IF) specifies how much psychometric information a measure provides about a specific trait level, allowing psychologists to determine how well the measure performs in assessing someone at that level. The IF supplements and extends traditional notions of reliability, which reflects how well a test measures an attribute in a group of individuals possessing a given distribution of trait levels.

The concepts of IF and reliability have tremendous utility in psychological measurement. On some occasions, however, one might wish to quantify measurement precision in a way that is not quite captured by either of these concepts. For example, one might want to describe how much psychometric information a measure could potentially provide overall, without reference to a particular sample or distribution of trait values. Alternatively, one might want to express the amount of psychometric information a measure provides given a specific pattern of observed responses from an individual or a set of such patterns in a group of individuals.

The purpose of the current article is to address the question of how to quantify the overall amount of psychometric information provided by a measure. To do so, I introduce the concept of information utility, which reflects the degree to which a measure increases the posterior probability of trait estimates that are in fact most probable, relative to their prior probability. Although the concept of information utility has been recognized for some time in the statistical literature (e.g., Bernardo, 1979a; Lindley, 1956), it has thus far not been applied to psychometric problems. Information utility complements existing notions of psychometric precision, in that it quantifies the precision of a measure for an observed response or set of responses, as opposed to latent trait value or set of trait values. In this regard, information utility may be especially useful in situations where information about latent trait level is vague or otherwise unclear. Importantly, although information utility itself derives from a Bayesian paradigm, indices related to information utility have non-Bayesian interpretations as well. In this way, information utility indices generalize notions of local test information to a global information setting. Information utility has important applied implications, such as for the choice of priors in Bayesian trait estimation, as well as hypothesis testing and adaptive testing.

The article begins with an overview of standard conceptualizations of test information, before introducing and explaining various indices of information utility, which are summarized in Table 1. Information utility is introduced via its application to individual response patterns. Reference priors, which maximize information utility and play a central role in information utility theory, are then discussed. Expected information utility, which summarizes information utility in a sample of response patterns from a group of individuals, is then introduced and compared to reliability, which plays a parallel role in traditional test information theory. Sample-independent quantities that derive from information utility theory, the information utility bounds and the criterion information utility, are then introduced. These two indices are shown to generalize roles of local test information (e.g., in likelihood ratios and confidence intervals) to a global information setting. The article ends with a discussion of practical issues involved in use of information

---

1 A preliminary R library implementing the methods summarized in Table 1 is available from the author.
utility. Examples of the use and properties of information utility-related quantities are provided throughout.

**Item Response Models and Fisher Information**

In order to explain the notion of information utility, its properties, and implications, this article focuses on the situation in which one is interested in estimating an individual’s value on a single latent trait \( \theta \) (e.g., depression, intelligence, or disinhibition) from their pattern of item responses \( x \). Although simplified, much of current psychometric theory and practice is based on this basic situation; extensions to this scenario are discussed later in the article.

In this prototypic scenario, the trait value \( \theta \) is not observed, but must be estimated based on the pattern of observed item responses \( x \). The probability of a particular type of item response is assumed to vary as a function of the latent trait in a way that can be described by an item response model. For example, someone’s level of disinhibition might not be directly observed, but could affect their probability of responding positively to an item on a personality measure in a way that is described by an item response model. Examples of such models are the Rasch model, the two-parameter logistic model, or other models that specify directly or indirectly the probability of response patterns conditional on latent trait value, \( P(x|\theta) \). In most item response models, parameters describing characteristics of the items (e.g., the difficulty or discrimination) are independent of parameters describing individuals (e.g., \( \theta \); Expected response patterns are not person-specific as they are in classical test theory, and are assumed to be the same for two individuals with the same level of the latent trait.

In one general formulation (e.g., Mellenbergh, 1994; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004), an item response model for the responses to an item \( j \) by a person \( i \) can be written as

\[
g(\tau_{ij}) = b_j + a \theta_i, \quad (1)
\]

where \( g \) is a link function relating the latent trait to an expected value of the response variable, and \( \tau_{ij} \) is the expected value of the response variable given a value of the latent trait (i.e., \( E(x_{ij}|\theta_i) \)). The link function \( g \), which derives from generalized linear modeling, depends on the assumed distribution of the response variable and choices regarding the specific model of interest (e.g., normal versus logistic). It can be thought of as a function that transforms the scale of the latent variable (e.g., continuous) into the assumed scale and distribution of the response variable (e.g., an ordinal polytomous variable or a count variable). The parameter \( a \) reflects the discrimination or loading of the item, and the parameter \( b \) is an intercept term reflecting the difficulty or severity of the item.

One popular choice for polytomous discrete response variables is to use a cumulative logit link function, where the expected values \( \tau_{ijk} \) are the probabilities of responding in a category \( k \) of the item:

\[
g(\tau_{ij}) = \ln \left( \frac{\tau_{ijk} + \ldots + \tau_{ijm}}{\tau_{ij1} + \ldots + \tau_{ijk-1}} \right) = b_j + a \theta_i, \quad (2)
\]

For polytomous items, this model is equivalent to Samejima’s (1994) graded response model (Mellenbergh, 1994). In the dichotomous case, this model reduces to the two parameter model logistic model (the examples in this article used these cumulative logit models, as fit in the ltm library for R; Rizopoulos, 2006).

Given an individual’s responses and an item response model, one can estimate that individual’s level of the latent trait. This is often done using a maximum likelihood approach, where the trait estimate that maximizes the likelihood of the data is chosen (e.g., choose an estimate of \( \theta \) such that \( P(x|\theta) \) is maximized). However, a Bayesian approach can also be used, where some prior information about the trait value is incorporated as a probability density, \( P(\theta) \), and the estimate of \( \theta \) is selected based on the posterior probability of the data, \( P(\theta|x) \), which is proportional to \( P(x|\theta)P(\theta) \).

When estimating individuals’ trait values, it is often desirable to have some index of the accuracy of those estimates, or more generally to have some index of how much information the test is providing about the trait values. For example, when choosing between tests, how does one decide which test might be the most informative about trait level? How does one quantify how much information has been provided about the trait level of a person responding to a particular test?

In item response theory, these issues are traditionally addressed by the item or test information function, which describes the amount of information provided by an item or test about the latent trait value. The test information is usually the sum of the item information values. Test information, moreover, is inversely proportional to the standard error of estimation conditional on a value of the latent trait (Mellenbergh, 1996).

The information function of item response theory is more generally known as the Fisher information function, \( I(\theta) \). It is proportional to the second derivative of the log-likelihood function,
which reflects the amount of curvature in the log-likelihood function at a particular value of the latent trait. Its inverse is the variance of the estimate at a particular value of the latent trait, which is also the squared standard error of the estimate at that value:

\[ \hat{I}(\theta) = -L''(\theta) = 1/\text{var}(\hat{\theta}) = \text{se}(\hat{\theta})^{-2}. \]  

(3)

As an example, consider the response patterns in the first column of Table 2, from the Disinhibition scale of the Schedule for Nonadaptive and Adaptive Personality (SNAP-2; Clark, Simms, Wu, & Casillas, in press). The SNAP-2 is a self-report inventory assessing characteristics across the range of normal and abnormal personality, and is intended to measure dimensions underlying personality pathology. The SNAP-2 comprises 15 nonoverlapping scales assessing three superordinate traits (Negative Temperament, Positive Temperament, and Disinhibition) and 12 subordinate traits (Mistrust, Manipulativeness, Aggression, Self-Harm, Dependency, Exhibitionism, Detachment, Impulsivity, Propriety, and Workaholism). The Disinhibition scale in particular assesses personality characteristics related to lack of conscientiousness, disorganization, carelessness, recklessness, and antisocial behavior. Values in the table are based on item parameter estimates calculated using the two-parameter logistic model in the SNAP-2 normative sample (N = 561).

Trait estimates corresponding to each response pattern are given in the second column. The Fisher information function at each of these values is given in the third column (the other columns are returned to later). For example, the second response pattern produces an estimated trait value approximately one standard deviation below the mean, with a Fisher information equal to 1.045, a variance of the estimate of .957, and a standard error of the estimate of .978. Consistent with its role as a measure of pathology, the Disinhibition scale provides greater information—and therefore, more precise trait estimates, with lower standard errors—at greater levels of the trait.

Note that the Fisher information function reflects measurement precision as a function of the latent trait. As the latent trait is generally unknown—uncertainty regarding trait level is the primary reason for administering the test—it might be desirable to incorporate uncertainty regarding trait level into the assessment of measurement precision. For example, the second response pattern in Table 2 has an associated trait estimate of 1.876, but this is only an estimate—the actual trait value for a person producing that pattern is unknown. It follows that the actual Fisher information for that person’s trait value, and the actual amount of measurement error for that person, is also unknown.

Note also that the Fisher information function is a measure of local measurement precision, reflecting the degree of measurement precision for a specific value of the latent trait. By considering quantities that are functions of the Fisher information function, however, one can develop measures of measurement precision that span a larger range of the latent trait. Reliability, for example, indexes measurement precision over a particular distribution of the latent trait. Over an even larger span, information utility can be used to derive indices of global measurement information over the entire range of the latent trait.

**Information Utility**

The information utility (Bernardo, 1979a) or Lindley information (Lindley, 1956) associated with a set of observed responses to a measure is defined as

\[ u(x) = \int P(\theta|x)\ln \left( \frac{P(\theta|x)}{P(\theta)} \right) d\theta, \]

(4)

where \( P(\theta) \) is the prior distribution assumed for the latent trait, and \( P(\theta|x) \) is the posterior distribution of the trait given the set of observed responses. The information utility is the same as the Kullback–Leibler distance or relative entropy between the posterior and the prior, and provides an index of similarity between the posterior and prior trait distributions. Values of \( u(x) \) are near zero when the posterior distribution is similar to the prior, and positive to the extent that the posterior is greater than the prior for probable values of \( \theta \).

In a psychometric context, information utility reflects the extent to which a set of item responses increases knowledge about the latent trait relative to what is known a priori. It can be thought of as a type of average, reflecting how much a set of item responses increases the probabilities assigned to different trait estimates, where the average is weighted by posterior probabilities of the trait values. When a set of item responses provides little information, the posterior will differ little from the prior, and the information utility will be near zero. When a set of item responses provides substantial information, the posterior will be increased relative to the prior for probable trait values, and the information utility will be positive.

Information utility or Lindley information can also be useful for quantifying the psychometric precision of measures (e.g., items,
tests), as they will differ in the amount of information they provide relative to a given prior. Given two tests, all other things being equal, one should select the test that elicits item responses that shift trait estimates toward more probable values than what is already known. Information utility is equivalent to certain classical criteria for optimal choice of experimental design (D-optimality criteria; Bernardo & Smith, 2000); by extension, it could be considered a criterion for optimal choice of psychological measure, if one considers a psychological measure a type of experiment aimed at estimating an individual’s value on a latent trait.

Note that information utility or Lindley information incorporates uncertainty regarding someone’s trait value, as it weights the information provided by the test at a given trait value by the probability of that trait value. This contrasts with Fisher information, which places weight on a single point, the estimated trait value, and implicitly puts zero weight at other points. Relatedly, because it considers the entire range of the trait, information utility or Lindley information can be thought of as a measure of global information, in contrast to Fisher information, which is local in nature.

Reference Priors

One advantage of information utility as an index of psychometric precision is that it makes clear that the value of a measure depends on prior information about the trait. This has been noted many times by others (e.g., Dawes, 1962; Mehl & Rosen, 1955): When prior knowledge about an individual’s standing on the trait is very vague, a relatively weak measure may nevertheless provide useful information; conversely, when prior knowledge about the trait is very accurate, even a relatively powerful measure may provide little useful information. In this way, the utility of psychometric information is always defined as being relative to what is already known.

The dependency of information utility therefore raises the question of how to choose a prior. There are many ways to do so, and the optimal choice of prior likely depends on the purposes of the test and the situation. Global or local base rate information may be used to create a prior, for example—so might background information, such as demographic information or history.

Williamson (2010) reviewed theories regarding selection of priors, and suggested that there are generally two strategies for doing so. First, he argued, are subjective Bayesian approaches, where one defines a prior to reflect one’s preexisting beliefs about the parameter being estimated—in this case, the trait value. These prior beliefs could be based on prior empirical evidence (e.g., base rates or previous test results), or could be completely theoretical in nature. Second, however, is the objective Bayesian approach to selecting a prior, where one chooses a prior so as to be maximally conservative about one’s preexisting beliefs. In this approach, one defines the prior so that preexisting beliefs influence parameter estimation as little as possible. Objective Bayesian approaches are based on the premise that preexisting evidence can be nonexistent, ambiguous, or otherwise difficult to interpret, and that in those cases, the optimal prior is one that makes minimally consequential assumptions about possible values of the estimate.

One interesting and important choice of prior is the prior that maximizes the information utility of the test. This prior, known as the reference prior (Berger, Bernardo, & Sun, 2009; Bernardo, 1979b; Good, 1968), operationalizes the objective Bayesian philosophy of choosing maximally conservative priors. The reference prior can be thought of as the prior that maximizes the empirical usefulness of a test, or maximizes the psychometric “weight” placed on the empirical results of the test in estimating the trait value. Technically, the reference prior is the prior maximizing the information utility of the test, as the number of the items in the test goes to infinity but other aspects of the test are held constant (e.g., as in a hypothetical scenario where the test items are administered an infinite number of times).

This definition initially might seem counterintuitive, as an infinitely long test provides perfect information. However, as test length increases, information will not increase evenly across the latent trait, instead increasing in a manner that is proportional to the Fisher information function. As a hypothetical test is repeated infinitely, the height of the Fisher information function will increase, but its shape will have a certain form. It is this shape that defines the reference prior—as a prior must sum to one, the absolute magnitude of the information provided in this theoretical setting matters less than the shape of the information function per se. The prior optimally weighting the empirical results of the test will place weight unevenly across the trait, in proportion to what the test can empirically distinguish, which is reflected in the Fisher information function.

Computing the Reference Prior: Monte Carlo Methods and the Jeffreys Prior

Detailed methods for calculating the reference prior in general using Monte Carlo methods are provided by Berger et al. (2009; summarized in Appendix A). In cases where the posterior distribution of a parameter estimate is normal, however—as is often the case when estimating the value of a trait (e.g., Chang, 1996; Chang & Stout, 1993)—the reference prior coincides with the Jeffreys prior (Clarke & Barron, 1994; Ghosh, Mergel, & Liu, 2011; Jeffreys, 1946). Given this, in general, the reference prior for a trait estimate is given by

$$P'(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}},$$  \hspace{1cm} (5)

where \(I(\theta)\) is the information function for the measure (e.g., the item or test information function). The integral is taken over possible values of the trait, and represents a normalizing constant, so that the prior sums to one.

The reference prior is therefore generally proportional to the square root of the test information function—that is, the inverse of the standard error of the estimate. This is intuitively reasonable, as it suggests that maximum prior probability should be placed on trait values that the test has optimal power to distinguish. Conversely, minimum prior probability should be placed on trait values that cannot be distinguished well by empirical observations provided by the test. From this perspective, the reference prior can be thought of as an optimally conservative prior, in that it maximizes the information obtained from the test responses relative to the information obtained from the prior.

Through the Jeffreys prior, the reference prior has another interpretation, in terms of distinguishable trait regions or confidence intervals (Balasubramanian, 1997; Grunwald, 2007; Myung,
Jeffreys prior and Monte Carlo estimates of the reference prior

Figure 1. Jeffreys prior and Monte Carlo estimates of the reference prior for the Schedule for Nonadaptive and Adaptive Personality (SNAP-2) Disinhibition scale. The Jeffreys prior is plotted as the dark solid line; Monte Carlo estimates of the reference prior are plotted as dotted lines. Each dotted line corresponds to using either 16, 32, 64, 128, or 256 items in the estimate; estimates using more items are plotted in a darker shade. The test information function, normalized to be on the same scale, is shown for comparison; it is represented by the light solid line.

Non-Bayesian Interpretations of the Reference Prior

Interestingly, in addition to being the information utility-maximizing prior, the reference prior plays a special role in weighted likelihood estimation (WLE; Warm, 1989). In particular, the Jeffreys prior—which is the same as the reference prior for item response theory (IRT) models with normal posterior distributions—is equivalent to the bias-minimizing weight in WLE for one and two-parameter logistic models (Hojitjink & Boomsma, 1995; Magis & Raiche, 2012; Warm, 1989). In this way, if the prior is thought of as a weight for maximum likelihood estimation, the reference prior is the weight that minimizes the bias of the resulting ML estimate, for one and two-parameter logistic models.
Caution is warranted in applying this bias-minimizing interpretation to models other than the one and two-parameter logistic models. Magis and Raiche (2012) show, for example, that the bias-minimizing weight for the three-parameter model is not, in general, the Jeffreys prior, although the two will produce essentially equivalent estimates for extreme trait levels. The Jeffreys prior, however, appears to produce estimates with lower standard errors than the bias-minimizing weight in WLE under many conditions, consistent with its role as an information-maximizing prior. Whether Magis and Raiche’s conclusions extend to the reference prior more broadly, as well as to the Jeffreys prior—if the two are in fact not the same for the three-parameter model—is unclear. In either case, better understanding the relationship between weighted likelihood estimation, Bayesian estimation with the reference prior, and information utility, is an interesting area for future research. It is tempting to speculate, for example, that the reference prior will tend to reduce bias or variance of estimates more depending on where the major source of error in trait estimates is for that particular model.

Information Utility for Groups

Expected Information Utility

Note that the information utility of a test as defined in Equation 4, $u(x)$, is defined with reference to a specific observed response pattern. That is, it is a characteristic of an individual, or of an individual who exhibits that response pattern. As such, it provides an index of the amount of information a test can be expected to provide about an individual with that response pattern.

It is also possible to define the expected information utility, or the average information utility of a test in a group of individuals (Bernardo & Smith, 2000):

$$u(X) = \sum p(x)u(x). \quad (6)$$

Here, $p(x)$ is the probability of response pattern $x$. When using expected information utility to summarize the information utility in a specific observed sample, $p(x)$ will generally be the observed proportion of individuals that have produced a given item response pattern $x$. As will be explained, however, $p(x)$ can also reflect the predicted probability of observing a response pattern (e.g., in the future, or in a population). The expected information utility in general reflects the typical amount of information a test is expected to provide on average for a group of individuals with a given set of observed response patterns.

For example, the expected information utility of the sample of response patterns presented in the first column of Table 2 is simply the average of the information utilities in the fourth column, or 1.316. If a sample consisted of the first three response patterns in equal proportion (i.e., each response pattern was observed in a third of the respondents), the expected information utility would be $(1/3)1.291 + (1/3)1.284 + (1/3)1.264 = 1.280$. A sample where half of the respondents produced the sixth response pattern and a quarter produced each of the last two response patterns would have an expected information utility of $(1/2)1.305 + (1/4)1.335 + (1/4)1.492 = 1.359$.

Relationship Between Expected Information Utility and Reliability

Expected information utility is similar to test reliability in that it provides an index of overall measurement quality in a sample or group of individuals. In this regard, it plays a role in Lindley information or information utility that is somewhat analogous to the role that reliability plays with regard to Fisher information. However, whereas the reliability reflects measurement quality in the form of Fisher information for a sample of latent trait values, the expected information utility reflects measurement quality in the form of information utility for a sample of observed response vectors. Moreover, whereas expected information utility is proportional to the average psychometric information provided by a test above and beyond prior information, test reliability is proportional to the average percent variance attributable to measurement error.

Mellenbergh (1996) noted that reliability can be expressed in terms of the Fisher information function by expressing expected measurement error variance in terms of the latter:

$$\text{Reliability} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\hat{\theta} \mid \theta)]} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\theta^{-1})^{-1}]. \quad (7)$$

where $E$ indicates the expectation (i.e., average). Reliability is therefore the proportion of total variance due to trait variance, where the total variance is the sum of the trait variance and the average measurement error variance, which in turn can be expressed in terms of the information function. Note that, as error variance changes with the value of the latent trait, measurement error enters into reliability through the average error variance, taken across the trait distribution. When the error variance—and therefore, the information function—is constant, Equation 7 reduces to a form analogous to the classic formula for reliability in classical test theory (Equation 7 will actually reduce to a lower bound to classical reliability, as it ignores the distinction between error and reliable unique variance).

Figure 2 illustrates the relationship between expected information utility and reliability for different tests in the same sample of individuals. Plotted are values of $u(X)$ and estimated reliability for the 15 scales of the SNAP-2 in the normative sample, using the Jeffreys prior to calculate information utility. These are the same values presented in Table 3. For example, the uppermost point represents the Positive Temperament scale, with $u(X)$ calculated using Equations 4, 5, and 6, and the reliability calculated using Equation 7, using the trait estimates (obtained using maximum a posteriori estimation with a Jeffreys prior) as plug-in estimates of the latent trait values (i.e., substituting $\hat{\theta}$ for $\theta$ in the rightmost formula of Equation 7). As can be seen, tests that are more reliable in the sample generally provide greater expected information. This relationship is not perfect however, as tests with lower reliabilities may have moderate information utilities. The test with the lowest reliability in Figure 2—for example, the Self-Harm scale—has an information utility comparable to that of the other scales.

The relationship between expected information utility and reliability can also be somewhat counterintuitive, however, with greater reliabilities being associated with less expected information utility. To illustrate this, Figure 3 presents expected information utilities as a function of estimated reliability for a single test.

Relationship Between Expected Information Utility and Reliability

Expected information utility is similar to test reliability in that it provides an index of overall measurement quality in a sample or group of individuals. In this regard, it plays a role in Lindley information or information utility that is somewhat analogous to the role that reliability plays with regard to Fisher information. However, whereas the reliability reflects measurement quality in the form of Fisher information for a sample of latent trait values, the expected information utility reflects measurement quality in the form of information utility for a sample of observed response vectors. Moreover, whereas expected information utility is proportional to the average psychometric information provided by a test above and beyond prior information, test reliability is proportional to the average percent variance attributable to measurement error.

Mellenbergh (1996) noted that reliability can be expressed in terms of the Fisher information function by expressing expected measurement error variance in terms of the latter:

$$\text{Reliability} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\hat{\theta} \mid \theta)]} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\theta^{-1})^{-1}]. \quad (7)$$

where $E$ indicates the expectation (i.e., average). Reliability is therefore the proportion of total variance due to trait variance, where the total variance is the sum of the trait variance and the average measurement error variance, which in turn can be expressed in terms of the information function. Note that, as error variance changes with the value of the latent trait, measurement error enters into reliability through the average error variance, taken across the trait distribution. When the error variance—and therefore, the information function—is constant, Equation 7 reduces to a form analogous to the classic formula for reliability in classical test theory (Equation 7 will actually reduce to a lower bound to classical reliability, as it ignores the distinction between error and reliable unique variance).

Figure 2 illustrates the relationship between expected information utility and reliability for different tests in the same sample of individuals. Plotted are values of $u(X)$ and estimated reliability for the 15 scales of the SNAP-2 in the normative sample, using the Jeffreys prior to calculate information utility. These are the same values presented in Table 3. For example, the uppermost point represents the Positive Temperament scale, with $u(X)$ calculated using Equations 4, 5, and 6, and the reliability calculated using Equation 7, using the trait estimates (obtained using maximum a posteriori estimation with a Jeffreys prior) as plug-in estimates of the latent trait values (i.e., substituting $\hat{\theta}$ for $\theta$ in the rightmost formula of Equation 7). As can be seen, tests that are more reliable in the sample generally provide greater expected information. This relationship is not perfect however, as tests with lower reliabilities may have moderate information utilities. The test with the lowest reliability in Figure 2—for example, the Self-Harm scale—has an information utility comparable to that of the other scales.

The relationship between expected information utility and reliability can also be somewhat counterintuitive, however, with greater reliabilities being associated with less expected information utility. To illustrate this, Figure 3 presents expected information utilities as a function of estimated reliability for a single test.

Relationship Between Expected Information Utility and Reliability

Expected information utility is similar to test reliability in that it provides an index of overall measurement quality in a sample or group of individuals. In this regard, it plays a role in Lindley information or information utility that is somewhat analogous to the role that reliability plays with regard to Fisher information. However, whereas the reliability reflects measurement quality in the form of Fisher information for a sample of latent trait values, the expected information utility reflects measurement quality in the form of information utility for a sample of observed response vectors. Moreover, whereas expected information utility is proportional to the average psychometric information provided by a test above and beyond prior information, test reliability is proportional to the average percent variance attributable to measurement error.

Mellenbergh (1996) noted that reliability can be expressed in terms of the Fisher information function by expressing expected measurement error variance in terms of the latter:

$$\text{Reliability} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\hat{\theta} \mid \theta)]} = \frac{\text{var}(\hat{\theta})}{\text{var}(\hat{\theta}) + E[\text{var}(\theta^{-1})^{-1}]. \quad (7)$$

where $E$ indicates the expectation (i.e., average). Reliability is therefore the proportion of total variance due to trait variance, where the total variance is the sum of the trait variance and the average measurement error variance, which in turn can be expressed in terms of the information function. Note that, as error variance changes with the value of the latent trait, measurement error enters into reliability through the average error variance, taken across the trait distribution. When the error variance—and therefore, the information function—is constant, Equation 7 reduces to a form analogous to the classic formula for reliability in classical test theory (Equation 7 will actually reduce to a lower bound to classical reliability, as it ignores the distinction between error and reliable unique variance).

Figure 2 illustrates the relationship between expected information utility and reliability for different tests in the same sample of individuals. Plotted are values of $u(X)$ and estimated reliability for the 15 scales of the SNAP-2 in the normative sample, using the Jeffreys prior to calculate information utility. These are the same values presented in Table 3. For example, the uppermost point represents the Positive Temperament scale, with $u(X)$ calculated using Equations 4, 5, and 6, and the reliability calculated using Equation 7, using the trait estimates (obtained using maximum a posteriori estimation with a Jeffreys prior) as plug-in estimates of the latent trait values (i.e., substituting $\hat{\theta}$ for $\theta$ in the rightmost formula of Equation 7). As can be seen, tests that are more reliable in the sample generally provide greater expected information. This relationship is not perfect however, as tests with lower reliabilities may have moderate information utilities. The test with the lowest reliability in Figure 2—for example, the Self-Harm scale—has an information utility comparable to that of the other scales.

The relationship between expected information utility and reliability can also be somewhat counterintuitive, however, with greater reliabilities being associated with less expected information utility. To illustrate this, Figure 3 presents expected information utilities as a function of estimated reliability for a single test.
across 24 different, nonoverlapping samples. Plotted values reflect estimated reliabilities and expected information utilities (calculated with a Jeffreys prior) for SNAP-2 Disinhibition in the different samples. Sample sizes ranged from 75 to 875 (Simms, Turkheimer, & Clark, 2007); nine were college samples, eight were clinical samples, three were community samples, and four were military samples; male composition of the samples varied from 20% to 88%, with a mean of 47%. Mean Disinhibition trait score estimates varied significantly across samples ($MSB = 45.60; MSW = 1.13; F(23, 8666) = 40.53, p < .0001; mean trait estimates ranged from .053 to 1.297).

As can be seen, in this case, as reliability increases, expected information utility decreases, albeit only slightly (note that the

Table 3

<table>
<thead>
<tr>
<th>Scale</th>
<th>Reliabilities</th>
<th>$\mu(X)$</th>
<th>$\upsilon(X)$</th>
<th>$\upsilon^*(X)$</th>
<th>$\gamma(X)$</th>
<th>$\eta(X)$</th>
<th>$NMRU$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Temperament</td>
<td>0.878</td>
<td>0.880</td>
<td>0.562</td>
<td>1.509</td>
<td>1.512</td>
<td>1.558</td>
<td>1.533</td>
</tr>
<tr>
<td>Positive Temperament</td>
<td>0.826</td>
<td>0.873</td>
<td>−0.904</td>
<td>1.532</td>
<td>1.523</td>
<td>1.586</td>
<td>1.560</td>
</tr>
<tr>
<td>Disinhibition</td>
<td>0.663</td>
<td>0.752</td>
<td>1.829</td>
<td>1.307</td>
<td>1.292</td>
<td>1.351</td>
<td>1.327</td>
</tr>
<tr>
<td>Mistrust</td>
<td>0.740</td>
<td>0.837</td>
<td>1.222</td>
<td>1.350</td>
<td>1.326</td>
<td>1.394</td>
<td>1.369</td>
</tr>
<tr>
<td>Manipulativeness</td>
<td>0.679</td>
<td>0.764</td>
<td>1.475</td>
<td>1.468</td>
<td>1.458</td>
<td>1.477</td>
<td>1.452</td>
</tr>
<tr>
<td>Aggression</td>
<td>0.684</td>
<td>0.814</td>
<td>1.535</td>
<td>1.442</td>
<td>1.402</td>
<td>1.443</td>
<td>1.418</td>
</tr>
<tr>
<td>Self-Harm</td>
<td>0.520</td>
<td>0.798</td>
<td>1.619</td>
<td>1.395</td>
<td>1.313</td>
<td>1.347</td>
<td>1.320</td>
</tr>
<tr>
<td>Eccentric Perceptions</td>
<td>0.714</td>
<td>0.763</td>
<td>1.300</td>
<td>1.284</td>
<td>1.269</td>
<td>1.328</td>
<td>1.303</td>
</tr>
<tr>
<td>Dependency</td>
<td>0.714</td>
<td>0.811</td>
<td>1.697</td>
<td>1.362</td>
<td>1.348</td>
<td>1.407</td>
<td>1.382</td>
</tr>
<tr>
<td>Exhibitionism</td>
<td>0.810</td>
<td>0.795</td>
<td>0.387</td>
<td>1.320</td>
<td>1.326</td>
<td>1.376</td>
<td>1.350</td>
</tr>
<tr>
<td>Entitlement</td>
<td>0.808</td>
<td>0.815</td>
<td>−0.045</td>
<td>1.386</td>
<td>1.394</td>
<td>1.456</td>
<td>1.432</td>
</tr>
<tr>
<td>Detachment</td>
<td>0.784</td>
<td>0.838</td>
<td>0.778</td>
<td>1.361</td>
<td>1.351</td>
<td>1.429</td>
<td>1.404</td>
</tr>
<tr>
<td>Impulsivity</td>
<td>0.731</td>
<td>0.799</td>
<td>1.216</td>
<td>1.358</td>
<td>1.348</td>
<td>1.397</td>
<td>1.371</td>
</tr>
<tr>
<td>Propriety</td>
<td>0.818</td>
<td>0.788</td>
<td>−0.562</td>
<td>1.364</td>
<td>1.360</td>
<td>1.404</td>
<td>1.378</td>
</tr>
<tr>
<td>Workaholism</td>
<td>0.804</td>
<td>0.830</td>
<td>0.778</td>
<td>1.425</td>
<td>1.428</td>
<td>1.477</td>
<td>1.452</td>
</tr>
</tbody>
</table>

Note. Reliabilities and expected information utilities, $\mu(X)$, are presented for the Schedule for Nonadaptive and Adaptive Personality (SNAP-2) normative sample and combined patient sample as described in the text. Reliabilities were calculated using the definition in Equation 7. $\theta^*$ values are the trait estimates implied by the reference prior, or equivalently, the trait values at the maxima of the test information functions; greater values of $\theta^*$ indicate greater, more characteristic levels of the trait. Criterion information utilities, $\upsilon^*(X)$, were estimated using Monte Carlo methods as described in the text and are presented together with lower and upper limits (LLs and ULs) of their 99% confidence intervals. The information utility bounds, $\gamma(X)$ and $\eta(X)$, and normalized minimum reduction in uncertainty ($NMRU$) were calculated using Equations 8–11.
Using Information Utility to Quantify Measurement Precision Independent of Sample

It is important to emphasize that this inverse relationship between test information and reliability is most clearly observed when discussing differences in measurement across samples rather than measures. As such, it is desirable to have an index of global test information that is independent of persons, in much the same way that the information function provides an index of local test information that is independent of persons. This would allow one to quantify the measurement information provided by a test independent of its application in any particular sample, and to more directly compare the global psychometric characteristics of measures per se.

Bounds on Information Utility

Lower bound. One important approach to deriving a sample-independent index of global measurement information is to note that, for a unidimensional trait, when using the reference prior, the information utility is bounded from below by the following quantity (Clarke & Barron, 1994):

$$u(X) = \frac{1}{2} \ln \left[ \frac{1}{2\pi e} \right] + \ln \int \sqrt{I(\theta)}d\theta,$$

where $I(\theta)$ is again the test information function, and $\pi$ and $e$ are the standard constants. Because the first term is a constant, the bound is primarily determined by the second term, the logarithm of the integral of the square root of the test information function. This term, and therefore the bound, is proportional to the area under the square root of the test information function.

Samejima (1994) has shown that for single dichotomous items, for a wide variety of common IRT models (e.g., generally for two-parameter models), the integral term of Equation 8 is always equal to $\log \pi$ (see also Grunwald, 2007, who presents similar results). However, it is not always equal to this value: it will vary for single items for certain models (e.g., the three-parameter logistic model), and most importantly, for measures comprising more than one item. That the value should often be the same for single items can be explained by noting that for trait estimation using a single item, most of the information about the trait estimate will likely be coming from the prior. As the number of items in the test increases, more of the information in the trait estimate is coming from the pattern of test responses per se.

Using a quantity that is proportional to the area under the test information function as a global index of measurement information has intuitive appeal, and is theoretically important for reasons that are explained in greater detail later in this article. To the extent that the information function is person-independent, the area under its square root, and therefore the bound in Equation 8, will inherit those properties. In this way, the information utility bound itself can be seen as an index of the measurement precision of the test, one that is independent of a particular response pattern or set of response patterns. Given that the square root of the information function is inversely proportional to the standard error of the estimate, the information utility bound in Equation 8 is directly proportional to the integrated inverse standard error of the estimate for the measure. Moreover, given that the test information function generally increases with test length, the bound in Equation 8 will increase with test length as well. The information utility bound can be roughly interpreted as the minimum information utility one might obtain using the reference prior for the test.

Upper bound. The information utility is also bounded from above—that is, there is a maximum amount of information that the test can provide about the trait. This maximum derives from the fact that the information utility can reexpressed as the mutual information between the test responses and the latent trait (e.g., Clarke & Barron, 1994). The mutual information is the relative entropy between the joint probability of the latent and observed variables and the product of their marginal probabilities. It reflects the dependence versus independence of the latent and observed variables, with complete independence when the joint probability equals the product of the marginal probabilities. Values of zero indicating complete independence of the latent and observed variables and a completely uninformative test. Increasing values of the mutual information indicate increasing dependence of the latent and observed variables and an increasingly informative test. The information utility (Equation 4) reexpressed as the mutual information between the test responses and the trait is:

$$u(x) = \int P(\theta|x)\ln \left[ \frac{P(\theta, x)}{P(\theta)P(x)} \right] d\theta.$$

Under this formulation, the information utility can be seen as reflecting how much the test reduces uncertainty about the trait relative to what is known a priori. The maximum amount of uncertainty that the test can reduce is equal to the amount of
uncertainty prior to test administration, which is reflected in the entropy of the prior:

$$\psi(X) = H[P(\theta)] = -\int P(\theta) \ln[P(\theta)] d\theta.$$  \hspace{1cm} (10)

The entropy of the prior is an upper bound to the information utility, a relationship that follows directly from its definition in terms of the mutual information (e.g., Cover & Thomas, 1991; see also Haussler & Opper, 1997). A perfectly informative test would completely reduce the uncertainty reflected in the prior, and would have an information utility equal to the entropy of the prior. Unlike the lower bound, the upper bound is less tight, and might be much larger than the observed information utility. However, the upper bound applies to all priors, whereas the lower bound is a maximum lower bound attained only with use of the reference prior. In this sense, the upper bound is stricter than the lower bound.

The lower bound of Equation 8 can be rescaled by the upper bound to provide an estimate of the normalized minimum reduction in uncertainty (NMRU) provided by the test when using the reference prior:

$$\frac{\psi(X)}{\psi(X)} = \frac{\psi(X)}{H[P(\theta)]}.$$  \hspace{1cm} (11)

This quantity answers the following question, in the form of a proportion: overall, what is the minimum amount one can expect the test to reduce uncertainty about the trait, when the test is providing the maximum amount of information possible? Tests with large values of the NMRU have large “potential,” in the sense that they provide large amounts of information—or conversely, reduce uncertainty—a great deal in some population. In this regard, note that, as all the terms in Equations 8 and 10 are response- and sample-independent, the NMRU will be response- and sample-independent as well. However, variants of the NMRU could be constructed, in which raw information utilities (Equation 4 or 6) are used instead of the lower bound, or where a prior other than the reference prior is used in calculation of the upper bound, to index the percent reduction in uncertainty for specific response patterns or alternative priors.

**Criterion Information Utility**

Another way to define a sample-independent index of global measurement information is to define it as the expected information utility in a criterion population, one that is defined solely by properties of the test and not by any particular observed sample. This would facilitate comparisons across measures by creating expected information utilities that entirely reflect the measures themselves.

One useful and theoretically important choice of criterion population, one that is closely related to the information utility lower bound (e.g., Barron, Rissanen, & Yu, 1998), is one in which response patterns are observed according the probabilities predicted by the test and the reference prior (which itself is a function of the test). In this case, $p(x)$ in Equation 6 becomes the probability of observing response pattern $x$ in a population of individuals whose latent trait values are distributed according to the reference prior, $P^*(x)$. In this case, Equation 6 becomes

$$\psi(x) = \sum p^*(x) \ln[p(x)].$$  \hspace{1cm} (12)

This defines a criterion information utility, $\psi^*(X)$, a variant of the expected information utility in a population that reflects the properties of the test.

In general, the criterion information utility can be interpreted as the average or expected information utility in a population of individuals that are uniquely suited to the test in a certain sense. This is not necessarily to say that the population defining the criterion information utility is the population maximizing the information utility, as the distribution of individuals’ trait values in such a population perfectly conforms to expectations under the prior. The reference prior maximizes information utility; the criterion information utility addresses the question of what information utilities would be expected if the reference prior actually defined the population of examinees. The criterion information utility is also closely related to the information utility lower bound in ways that are explained later, and will produce similar values.

The criterion information utility can be estimated using standard Monte Carlo methods. In this approach, one would simulate responses to a test from a population of individuals whose latent trait values are distributed according to the reference prior. This simulation would consist of randomly sampling trait values from the reference prior, and randomly generating an observed response pattern for each trait value. The criterion information utility could then be estimated by calculating the average log posterior probability of the responses over the prior. Confidence intervals around the estimated criterion information utility could be calculated using standard formulas for the confidence interval of a mean (see Appendix B for a more detailed explanation of the Monte Carlo procedure).

**Information Utility Bounds and the Criterion Information Utility: An Illustration**

Table 3 illustrates the use of the lower bound to information utility, including its interpretation in terms of the NMRU, and the criterion information utility. The table presents the estimates of the criterion information utilities, their 99% CIs, and bounds on the information utility (Equations 8–11) for SNAP-2 scales. Estimates of the $\psi^*(X)$ values were calculated using Monte Carlo methods as described above, sampling 5,000 random response vectors for each scale. Also presented are reliabilities (Equation 7) and expected information utilities of the SNAP-2 scales in the normative sample ($N = 561$) and a combined patient sample ($N = 1,333$ patients; Simms et al., 2007). Trait estimates implied by the reference prior are also presented; as these are equivalent to the trait values at the maximum of each respective test information function, they provide as an index of how severe or difficult each test is.

As is evident in the table, reliability is a sample-specific quantity, and will vary from sample to sample. As such, it may be unclear from the reliability how useful a test might be in a different sample. The Self-Harm scale, for example, has a relatively low reliability in the normative sample, but a higher, more typical reliability in a patient sample. This is partially attributable to the scale’s severity, as reflected in the fact the maximum of the test information occurs at a relatively high trait level. The reliability of the test in the normative sample would thus be misleading as an index of the test’s overall measurement precision, as the test measures relatively severe phenomena.
The criterion information utility and information utility bounds, in contrast, are invariant across samples, and provide estimates of how well the test is measuring globally. The estimated $I^*(X)$ and bound $I(X)$ for Self-Harm, for example, are roughly comparable to that of other SNAP-2 scales, albeit lower than the average, and its NMRU is the second largest among the scales, suggesting it is measuring about as well as, if not better than, the other scales in general over the entire range of the trait. In contrast to the reliabilities in the normative sample, which might provide a misleading idea of the scale's potential, the information utility indices point toward the Self-Harm scale's potential use in other samples.

Non-Bayesian Explanations of the Information Utility Lower Bound and Criterion Information Utility

Although the presentation of information utility here has so far focused on its Bayesian interpretation, the sample-independent information utility quantities—the lower bound $I(X)$ and the criterion information utility $I^*(X)$—can be explained within a frequentist or information-theoretic paradigm as well, without invoking concepts related to a prior. In particular $I(X)$ and $I^*(X)$ can be interpreted as minimax expected log-likelihood ratio values, minimized over alternative trait value hypotheses and maximized over true trait values. $I(X)$, moreover, can be interpreted in terms of the maximum number of distinguishable confidence intervals or trait values that a test can accommodate, or equivalently, in terms of the average width of confidence intervals produced by the test.

Local information, the log-likelihood ratio, and Fisher information. In order to better understand information utility from a non-Bayesian perspective, it is helpful to first consider the log-likelihood ratio as it is typically used, and how it relates to local test information in the form of the Fisher information (Equation 3). In this case, the interest is usually in evaluating the hypothesis that the trait equals some posited true value $\hat{\theta}$—where the posited true value is often the trait estimate—against some alternative hypothesis that the trait equals some other value, $\theta_a$. In the case that the true value hypothesis is actually correct, the expected value of the log-likelihood ratio is (van der Vaart, 1998)

$$
\sum_x P(x|\hat{\theta}) \left( \frac{P(x|\hat{\theta})}{P(x|\theta_a)} \right) = \frac{1}{2}(\hat{\theta} - \theta_a)^2 I(\hat{\theta}),
$$

(13)

where $I(\hat{\theta})$ is the Fisher information at the posited true value, and the sum is taken over all possible response patterns.

Equation 13 is useful because it summarizes the statistical power of a measure at a particular value of the latent trait, and relates this power to local information in the form of the Fisher information. The expected log-likelihood ratio, $E[\lambda]$, relates directly to Type I and Type II error rates in statistical theory, $E[\lambda]$ with inversely related to Type II error at a given significance level (Cover & Thomas, 1991). As the difference between the true value and alternative value increases, and as the value of the Fisher information at that point increases, so does the expected log-likelihood ratio, and therefore, the power of the test at that point.

Although the expected log-likelihood value in Equation 13 is useful, two important points should be emphasized. First, it is local in nature: It reflects the ability of a test to make inferences about a specific trait value. Second, the specific trait value usually of interest is actually unknown: Although in practice, some estimate is substituted for the true value (e.g., the maximum likelihood estimate), the trait value itself—and therefore, the power of the test at that value—is subject to error.

Global information, the log-likelihood ratio, and information utility. Information utility—or more specifically, the lower bound to the information utility and criterion information utility, $I(X)$ and $I^*(X)$—generalize this log-likelihood ratio interpretation of test information globally. In particular, $I(X)$ and $I^*(X)$ can be both be interpreted as expected log-likelihood ratios, involving the alternative hypothesis value that the test has the worst maximal power to test against.

To better understand what is meant by “worst maximal power to test against,” imagine each possible alternative hypothesis trait value $\theta_a$ For each of those alternatives, imagine finding the true trait value that produces the largest log-likelihood ratio statistic, on average. $I(X)$ and $I^*(X)$ reflect the expected log-likelihood ratio for the alternative hypothesis having the worst of these largest average log-likelihood ratio statistics.

The information utility lower bound and criterion information utility, $I(X)$ and $I^*(X)$, answer variants of the question “what is the expected log-likelihood ratio under the worst best-case power scenario?” The information utility lower bound and criterion information utility represent different answers to that question, depending on how the question is interpreted. Specifically, they differ in what type of expected log-likelihood ratio is involved.

Information utility lower bound. The lower bound to the information utility, $I(X)$, addresses the question of a test’s global measurement power in terms of a test of the maximum likelihood estimate against an alternative. Specifically,

$$
\min_{\theta_a} \max_{\hat{\theta}} \sum_x P(x|\hat{\theta}) \left[ \frac{P(x|\hat{\theta})}{P(x|\theta_a)} \right] = I(X) + \frac{1}{2},
$$

(14)

where $\hat{\theta}$ is the maximum likelihood estimate of the trait value given the true value and a response pattern, and the sum is taken over all possible response patterns $x$ (Barron et al., 1998; Drmota & Szpankowski, 2004; Granwald, 2007; Rissanen, 1996, 2001).

Note that the maximum likelihood estimate, $\hat{\theta}$, will change with $x$ from sample to sample, even though the true value, $\theta_a$ remains the same.

From this perspective, the information utility lower bound can be interpreted in terms of an expected log-maximum-likelihood ratio value, for the value of the alternative $\theta_a$ that the test has the worst best-case power to test against. Again imagine each possible alternative hypothesis trait value $\theta_a$, and for each of those alternatives, imagine finding the true value of the trait that produces the largest maximum-likelihood ratio statistic, on average. The information utility lower bound reflects the expected log-maximum-likelihood ratio for the alternative hypothesis having the worst of these largest average maximum-likelihood ratio statistics. $I(X)$ roughly answers the question “given the test, what’s the best expected maximum-likelihood ratio statistic in the worst-case setting?”

For example, $I(X)$ for the SNAP-2 Disinhibition scale is 1.241 (see Table 3). As the information utility lower bound, this number reflects the lower bound to the information utility when using the reference prior with the test. However, this number (or more
accurately, 1.741) also represents another sort of lower bound for the test, a lower bound to the largest expected log-likelihood ratio comparing the maximum likelihood against the likelihood of an alternative. Although this log-likelihood ratio is somewhat of an abstract quantity, it nevertheless reflects the global power of the test.

**Criterion information utility.** The criterion information utility, \( v^*(X) \), addresses the question of a test’s global measurement power slightly differently, in terms of the true trait value, rather than the maximum likelihood estimate. Specifically,

\[
\min_{\theta} \max_{\theta_a} \sum_{x} P(x|\theta) \left[ \frac{P(x|\theta_a)}{P(x|\theta)} \right] = v^*(X) \tag{15}
\]

(Barron et al., 1998; Davisson & Leon-Garcia, 1980; Drmota & Szpankowski, 2004; Feder & Merhav, 1996; Gallager, 1979; Merhav & Feder, 1995; Ryabko, 1981). That is, \( v^*(X) \) is equal to the expected log-likelihood ratio for the value of the alternative that the test has the worst best-case power to test against. In contrast to the information utility lower bound, however, the criterion information utility expresses this power in terms of the probability of the data under the true value, rather than in terms of the probability of the data under the maximum likelihood estimate.

Returning to the SNAP-2 Disinhibition scale, its estimated \( v^*(X) \) is 1.351 (see Table 3). As with the information utility lower bound, this number also represents a lower bound for the test, a lower bound to the largest expected log-likelihood ratio. However, in contrast to \( \psi_l(X) \), which reflects an expected log-likelihood ratio involving the maximum likelihood, \( v^*(X) \) reflects an expected log-likelihood involving the likelihood of the data under the true value. \( v^*(X) \) is less than \( \psi_l(X) + .5 \) (1.351 vs. 1.741) because the largest maximum likelihoods will tend to be slightly larger than the likelihoods under the true values, due to overfitting (e.g., Grunwald, 2007).

Comparing Equations 14 and 15, it can be seen that both sample-independent information utility indices reflect the global power of the test in a log-likelihood ratio sense, under a worst case optimal power setting. \( \psi_l(X) \) relates to the log-likelihood ratio of an alternative against the maximum likelihood estimate, and \( v^*(X) \) relates to the log-likelihood ratio of an alternative against the true value. \( \psi_l(X) \) and \( v^*(X) \) can be thought of as conservative estimates of how much power a test will have, in general, to test a given trait value against an alternative, even though in particular scenarios this might not hold.

Examining Table 3, for example, it can be seen that the SNAP-2 Impulsivity scale has greater \( \psi_l(X) \) and \( v^*(X) \) values (1.307 and 1.397, respectively) than the Disinhibition scale (1.241 and 1.351). This indicates that in general the Impulsivity test, conservatively, under a worst case optimal power scenario, has greater power to test against arbitrary alternatives than the Disinhibition test. In this sense, the Impulsivity test provides greater global information than the Disinhibition test. Locally, however, the converse may be true, in that there may be certain scenarios where the Disinhibition test has greater power (which could be determined using the Fisher information function if one had sufficient information about the trait value of interest; Equation 13).

**Local and global information, confidence intervals, and the number of distinguishable trait values.** In addition to its interpretation in terms of log-likelihood ratio values, the lower bound to the information utility, \( \psi_l(X) \), also has an interpretation in terms of confidence intervals and what can be thought of as the maximum number of trait values a test can distinguish. In this sense, \( \psi_l(X) \) provides a generalization of the confidence interval interpretation of local test information (as is implied by Equation 3) to the global information setting.

Consider, for example, Figure 4, which illustrates a hypothetical test information function (i.e., Fisher information function) and a set of associated confidence intervals at five trait values (note that the confidence intervals are not drawn to scale). The width of a confidence interval at a given trait value (through the variance or standard error of the trait estimate; see Equation 3) is inversely related to the test information function. Greater Fisher information at a trait value is associated with a smaller confidence interval around that value, and a more precise trait estimate. As the Fisher information function generally changes over the range of the trait, the confidence interval at a given point may differ in width from the confidence interval at another point.

If one defines the trait as ranging over a certain interval (e.g., from –10 to 10, or a range defined by computer precision), one can also define a set of adjacent but nonoverlapping confidence intervals covering that range, as illustrated in the bottom of Figure 4. This set of confidence intervals can be thought of in an abstract

![Figure 4](image_url)

**Figure 4.** Illustration of the relationship between confidence intervals and local and global test information. The figure is not drawn to scale. A hypothetical test information function is plotted in the top portion of the figure; confidence intervals around five trait values \((\theta_0, \theta_1, \ldots)\) are plotted below. Local test information in the form of Fisher information is inversely proportional to the width of a confidence interval at a given point; global test information in the form of the information utility lower bound is proportional to the average width of confidence intervals over the range of the trait, or to the number of non-overlapping confidence intervals the test can accommodate.
sense as representing a set of distinguishable trait values, in that each trait value is empirically distinguishable from the others within a certain error tolerance defined by the $\alpha$ level of the confidence interval.

Interestingly, a number of authors have shown that the maximum number of these nonoverlapping confidence intervals is directly related to the information utility lower bound, $\ell(X)$ (Grunwald, 2007; Myung et al., 2000; Qian & Kunsch, 1998; Razavi & Giurcaneanu, 2009; Rissanen, 2006, 2007; see also Ghosal, Ghosh, & Ramamoorthi, 1997). In particular, the maximum number of nonoverlapping confidence intervals is equal to

$$e^{\ell(X)} = c(1 - \alpha)^{-1},$$

(16)

where the confidence intervals are all at the $(1 - \alpha)$ level (e.g., are 99% CIs for an $\alpha$ of .01) and $c$ is a constant to be discussed further that is the same for all tests measuring a single trait (Grunwald, 2007). Therefore, the information utility lower bound, $\ell(X)$, can be thought of as reflecting the maximum number of distinguishable trait values supported by a test within the range of the trait, within a certain confidence level. Note that this is not the number of trait estimates the test can provide, but the maximum number of those trait estimates that are distinguishable in the sense of having nonoverlapping confidence intervals.

From this perspective, the local information provided by a test at a certain trait value is related to the precision of the trait estimate at that point. The global information, in contrast, is related to how many trait values the test can distinguish at a given level of error tolerance. As the local information at each point increases, the widths of the confidence intervals at those points decrease, and more of them can be accommodated within the range of the trait. This conceptualization of $\ell(X)$ and global information is broadly consistent with the log-maximum likelihood ratio conceptualization just discussed (Equation 14), in that as $\ell(X)$ increases, the number of nonoverlapping confidence intervals increases, which increases the probability that any given pair of trait values will be significantly different from one another.

In the literature, the value of the constant $c$ in Equation 16 is usually not explicitly defined (e.g., Ghosal et al., 1997; Grunwald, 2007), which makes it difficult to use the information utility lower bound to directly calculate the number of confidence intervals or distinguishable trait estimates a test can provide in an absolute sense (Qian & Kunsch, 1998, and Rissanen, 2006, do provide means of calculating the absolute number of distinguishable trait estimates, but for a certain optimal confidence interval width, which is outside the scope of this article). However, it is still possible to interpret the difference in information utility lower bounds between two tests as reflecting the relative number of confidence intervals or distinguishable trait estimates the two tests can provide. The relative number of nonoverlapping confidence intervals or distinguishable trait values two tests can accommodate is given by

$$\frac{e^{\ell(X)} - c(1 - \alpha)^{-1}}{e^{\ell(Y)} - c(1 - \alpha)^{-1}} = e^{\Delta\ell(X)},$$

(17)

where $\Delta\ell(X)$ is the difference in information utility lower bounds between the two tests. Note that because they will be the same for any two tests, both the constant $c$ and the term related to the significance level cancel and drop out, suggesting that the difference in information utility lower bounds is related to the relative number of trait values the tests can distinguish independent of significance level.

An important corollary of the interpretation of $\ell(X)$ in terms of the number of distinguishable trait values or confidence intervals is that if the range of the trait is considered fixed across tests, it can also be interpreted in terms of the average confidence interval width. That is, if more confidence intervals are being fit into the same range of the trait, their widths on average will be smaller. This provides a more direct analogy between local and global test information and confidence intervals: local information in the form of Fisher information is related to the width of a particular confidence interval at a particular point; global information in the form of Lindley information (through its lower bound) is related to the average width of confidence intervals across the range of the trait.

As an example, consider the Positive Temperament and Exhibitionism scales of the SNAP-2 (see Table 3). Both scales are considered to measure traits in the domain of positive emotionality or extroversion, the Positive Temperament scale being a direct measure of positive emotionality, and the Exhibitionism scale measuring possibly pathological positive emotionality as manifest in socially prominent attention-seeking behavior (Clark et al., in press). Applying Equation 17 to the information utility lower bounds in Table 3, one has $\exp(1.496 - 1.276) = 1.246$, suggesting that the Positive Temperament scale can distinguish between 25% more trait values than the Exhibitionism scale, in the sense that it can accommodate 25% more nonoverlapping confidence intervals. This relationship holds, moreover, regardless of the significance level of the confidence intervals being considered. Equivalently, it can be said that confidence intervals on average will be narrower for trait estimates produced by the Positive Temperament scale than for trait estimates produced by the Exhibitionism scales.

**Relationships Between $v^*(X)$ and $\ell(X)$**

As is suggested by Equations 14 and 15, the criterion information utility and lower bound to the information utility are closely related. In fact, as Drmota and Szpankowski (2004) note, $v^*(X) \leq \ell(X) + .5$, implying that

$$\ell(X) \leq v^*(X) \leq \ell(X) + \frac{1}{2},$$

(18)

because $\ell(X)$ is, by definition, a lower bound to $v^*(X)$. Moreover, as test length increases, the two rightmost terms will become increasingly similar (Barron et al., 1998; Drmota & Szpankowski, 2004; Grunwald, 2007). The two will generally be within a certain amount of one another, suggesting that they might be interchangeable for many practical purposes.

Given that the two quantities are quite similar, it is reasonable to ask how one might choose between them. This choice can be approached from two perspectives, a theoretical one and a practical one. From a theoretical perspective, $\ell(X)$ might be of more interest, as it is more closely related to the expected maximum likelihood ratio, which involves observable, knowable quantities, rather than the true likelihood ratio, which involves unobservable, unknowable quantities. On the other hand, one might argue that $v^*(X)$ generalizes the Fisher information more directly in some sense, as the global relationship in Equation 15 more directly generalizes the local relationship in Equation 12.
From a practical perspective \( \psi(X) \) is probably easier to calculate than \( \psi^*(X) \), if the test information function is known explicitly, which is probably often the case. However, \( \psi^*(X) \) might be easier to calculate when the test information is difficult to evaluate analytically, or when it is otherwise easier to sample from the reference prior than calculate the integral of the information function. Such situations might be encountered, for example, with nonparametric IRT. From yet another perspective, as \( \psi^*(X) \) is estimated using Monte Carlo methods, it will be subject to Monte Carlo estimation error in addition to errors in estimation of test parameters, whereas \( \psi(X) \) will only be subject to estimation errors through the latter. Ultimately, the decision to express global test information using the criterion information utility or lower bound probably depends on the setting and test.

**Practical Interpretation of Information Utility and Its Use in Test Selection**

The Use of Information Utility in Test Selection: An Example

The use of information utility in test selection—in terms of how it compares to the use of reliability, the impact of priors, and sample characteristics—is perhaps best illustrated with a specific example. Samuel, Simms, Clark, Livesley, and Widiger (2010) examined the IRT properties of various normal and abnormal personality scales as measures of the Big Four traits—emotional instability, extroversion, antagonism, and introversion. Among these scales, they included the Dimensional Assessment of Personality Pathology (DAPP; Livesley & Jackson, 2009) Anxiety and Suicidal Ideation scales as alternative measures of emotional instability (i.e., negative emotionality or neuroticism). In their work, they demonstrated that items of the Anxiety and Suicidal Ideation scales could be scaled on a single dimension of emotional instability, consistent with other psychometric (Markon, Krueger, & Watson, 2005), longitudinal (Fergusson, Horwood, & Boden, 2006), genetic (Kendler, Neale, Kessler, Heath, & Eaves, 1992), and family history (Seeley, Kosty, Farmer, & Lewinsohn, 2011) literature on anxiety, depression, and related personality traits. Like the SNAP-2, the DAPP is a self-report measure of personality pathology. Unlike the SNAP-2, which comprises dichotomous items, the DAPP comprises polytomous items. The DAPP Anxiety scale includes 16 items and the Suicidal Ideation scale includes 12 items.

Figure 5 presents test information functions for DAPP Anxiety and Suicidal Ideation scored on a common latent trait metric, as reported in Samuel et al. (2010). Consistent with what one might expect given its item content, the Suicidal Ideation scale is the most informative at relatively high levels of emotional instability, providing more information than the other scale for much of the extreme range of the trait. However, the Anxiety scale is the most informative overall, even though it provides maximum information at lower levels of the trait.

How might one go about choosing between the two DAPP emotional instability tests in this situation? One possibility is to choose the test with maximum test information—or, similarly, reliability—for a sample of interest. For example, in a hypothetical inpatient sample where the mean level of emotional instability is 2.5 standard deviations above the population mean, the Suicidal Ideation scale might be expected to have maximum test information and reliability. Another possibility, however, is to choose the test with maximum capacity for information utility, which might suggest the Anxiety scale, given that the total area under its information function is largest (cf. Equation 8).

Tables 4 and 5 illustrate the psychometric properties of the tests and the consequences of choosing one test over another. These tables present psychometric properties of the tests and summaries of estimation error for three samples of 100 simulated individuals, drawn from each of three normal distributions. As can be seen in Table 4, the Suicidal Ideation scale does have the maximum test information at the mean of the most severe sample, and is most reliable in that sample. The Anxiety scale, in contrast, has the greatest capacity for information utility, as reflected in the information utility indices based on the reference prior (information utilities using a normal prior are presented for comparison). The criterion information utilities and information utility lower bounds, \( \psi^*(X) \) and \( \psi(X) \), are greater for the Anxiety (2.024 and 1.793) than Suicidal Ideation scales (1.798 and 1.430). These values indicate that even though the Suicidal Ideation scale might locally provide a more powerful test of a trait value against an alternative (i.e., for extreme levels of emotional instability), the Anxiety scale can conservatively be relied on to globally provide more powerful tests of a trait value against an alternative (i.e., when the entire range of emotional instability is considered). The difference between the information utility lower bounds, moreover, suggests that the Anxiety scale can accommodate estimation of 43% more distinguishable trait values, in the sense of nonoverlapping confidence intervals, than the Suicidal Ideation scale. Equivalently, the confidence intervals around trait estimates from the Anxiety scale will be narrower on average than confidence intervals for trait estimates from the Suicidal Ideation scale. Finally, note that as the
samples become less and less severe, the test information at the sample means—and by extension, reliability—decrease substantially for the Suicidal Ideation scale, but less so for the Anxiety scale, which is relatively stable in its reliabilities.

Table 5 illustrates the consequences of choosing one test over the other, in terms of measurement error. Even though the Suicidal Ideation scale is more reliable in the severe sample, the Anxiety scale resulted in less measurement error in that sample in terms of RMSE. In fact, the Anxiety scale had the lowest RMSE in all three samples. Moreover, as the sample became less and less severe, the bias, variance, and RMSE for the Suicidal Ideation scale increased dramatically, reflecting its relatively peaked test information function.

This example illustrates the relative roles of reliability and information utility in quantifying measurement precision. When prior information about trait level is very accurate, tests with large reliabilities can provide relatively accurate assessments. However, given that the trait values are not known for any individual—and are in fact the primary reason for administering the test—a putatively reliable test may nevertheless provide less accurate estimates than another test with lower reliability but greater information utility. Tests with greater information utility tend to provide more total information over the range of the trait, and are likely to provide more accurate estimates when prior information is very vague or inaccurate. Ultimately, of course, tests with large information utilities as well as reliabilities are likely to provide the most accurate estimates of all. The two indices of measurement precision likely complement one another, with one index of measurement precision likely to be preferable over the other depending on the accuracy of prior information.

**Priors and Information Utility**

Certain points regarding the role of priors in information utility and testing more broadly should be emphasized. Depending on the setting, it may be more or less clear what prior to use for any specific individual, or it might be deemed inappropriate to use any prior at all. Assessments differ, for example, in whether their goal is comparative evaluation of interindividual differences in a population of respondents, or evaluation of intraindividual change (or similarly, state) in a single respondent, and whether the assessment is cooperative or competitive in nature (Carver, 1974; Matarazzo, 1990). It is reasonable to assume that priors might be used differently across these settings, or might not be used at all. These choices regarding priors will impact trait estimates, and therefore, the criteria used to choose and interpret tests and the information obtained from them. In different settings with different priors, different information utility indices, or different forms of those indices, might be differentially useful.

**Strong priors.** For example, although the examples discussed here have assumed the same prior for every individual taking a particular test in a particular sample, this need not be the case. In some settings (e.g., clinical or diagnostic assessments), it might be seen as desirable to make use of strong prior information about a particular examinee. The prior assumed for one individual might be different from those assumed for other individuals, and it is even conceivable that the prior might differ for the same individual on different occasions. In such scenarios, indices such as $\psi(X)$ or $\psi(X)$ might be more useful, as they focus on specific observed

<table>
<thead>
<tr>
<th>Scale</th>
<th>Bias</th>
<th>Variance</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicidal Ideation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(2.5, 1)</td>
<td>.017</td>
<td>.155</td>
<td>.392</td>
</tr>
<tr>
<td>N(0, 1)</td>
<td>.088</td>
<td>.150</td>
<td>.395</td>
</tr>
<tr>
<td>N(−1, 1)</td>
<td>.395</td>
<td>.318</td>
<td>.684</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(2.5, 1)</td>
<td>−.006</td>
<td>.151</td>
<td>.387</td>
</tr>
<tr>
<td>N(0, 1)</td>
<td>−.024</td>
<td>.093</td>
<td>.304</td>
</tr>
<tr>
<td>N(−1, 1)</td>
<td>.042</td>
<td>.108</td>
<td>.329</td>
</tr>
</tbody>
</table>

*Note.* Values in the table are the bias, variance, and root-mean-square error (RMSE) of the maximum likelihood trait estimates using the Dimensional Assessment of Personality Pathology (DAPP) Suicidal Ideation and Anxiety scales, for a sample of 100 simulated individuals drawn from one of three normal distributions.
response patterns or sets of patterns, in the context of the priors used with each pattern. If situation-specific priors are deemed optimal, \(u(x)\) or \(u(X)\) could be calculated with regard to the priors used for those individual assessments to summarize or predict information obtained from tests (the latter section on computerized adaptive testing provides an example of prediction). In contrast, indices such as \(v^*(X)\), \(\zeta(X)\), or \(NMRU\), which are focused on properties of measures in the context of general reference priors, might be less useful.

Note that, when interpreting information utility indices such as \(u(x)\) or \(u(X)\) with strong priors, the indices may reflect the quality of the prior as much as the test. That is, in the presence of poor prior information, even a poor test may provide substantial information and \(u(x)\) or \(u(X)\) may have relatively large values; conversely, in the presence of good prior information, a good test may provide relatively little information, and \(u(x)\) or \(u(X)\) may have relatively small values. In this regard, \(u(x)\) and \(u(X)\) may be thought of as indexing the information provided by a particular measurement (e.g., as produced by an administration of a test on a particular occasion, using a particular prior), rather than a measure per se (e.g., a test or item). This contrasts with indices such as \(v^*(X)\), \(\zeta(X)\), or \(NMRU\), which are independent of any observed response patterns, and index the potential information provided by a measure, rather than any given empirical measurement that measure may have produced.

**Reference priors.** Often, the prior may be extremely vague, or it may be unclear what prior to assign, for empirical, practical, or ethical reasons (e.g., legal or equity considerations). In such cases, the reference prior is an attractive choice, as it maximizes the “weight” being placed on the empirical results of the test relative to any prior information. Use of the reference prior also results in the same prior being assigned to all individuals taking the same test, which might provide a desirable compromise between considerations of fairness or equity and a desire to make use of some prior assumptions to improve estimates. As the reference prior is a property of a test, and maximizes the information obtained from the test, use of the reference prior implicitly transforms a decision about priors into a decision about tests. It is also worth noting again that the reference prior acts as a bias minimizing prior for some models, providing further justification for its use as a fair or equitable prior in some settings (Hojitink & Boomsma, 1995; Magis & Raiche, 2012; Warm, 1989).

In cases where the reference prior is deemed useful as a general prior, \(u(x)\) or \(u(X)\) could again be useful in summarizing the information provided by individual measurements or test administrations. In addition, however, indices such as \(v^*(X)\), \(\zeta(X)\), or \(NMRU\) become useful summaries of tests as used in those contexts. The criterion information utility, information utility bounds, and \(NMRU\) all derive from the assumed use of a reference prior and would apply to all examinee responses in those scenarios.

**No priors.** Finally, it should be recognized that in certain cases, it might be deemed inappropriate to use any prior at all. For example, for some tests, response models, and settings, the use of any prior information in trait estimation might be seen as biasing, inequitable, or unfair. In such cases, the sample-independent information utility indices—the criterion information utility and information utility lower bound, \(v^*(X)\) and \(u(X)\)—still provide useful metrics for summarizing the global informativeness of the test. These indices can be interpreted without regard to any prior, in terms of expected global log-likelihood ratio bounds, or in terms of the relative number of distinguishable trait estimates (or equivalently, nonoverlapping confidence intervals) the test can accommodate.

**The Response and Scale Dependence and Independence of Information Utility Indices**

It is important to emphasize that information utility per se (as reflected in Equations 4 and 6) is response-dependent—that is, it indexes the amount of information provided by a specific response pattern or sample of response patterns. The criterion information utility, information utility bounds, and \(NMRU\), in contrast, are sample independent, depending on the characteristics of the test alone. As such, these different quantities will likely have different uses, depending on the application (some of these are described next, in the following section).

Although they are response- and sample-independent, the criterion information utility and information utility bounds are scale-dependent. For example, they will both increase as test length increases, consistent with intuition, but in a manner that is somewhat complex (consider the last term of Equation 8, for example). For this reason, it is recommended for the time being that their use be restricted to cases where the tests are linked on a common scale, or for other applications (such as hypothesis testing) that are scale-independent. In contrast, as the \(NMRU\) is scaled as a proportion (of information units—e.g., percent logits or bits), it should be scale- as well as response-independent. Further research is needed to explore how well this holds in practice.

**Other Applications of Information Utility**

**Hypothesis Tests**

One particularly important application of the information utility is in hypothesis testing. Note that the information utility can be interpreted as the average increase in posterior probability over the prior for different values of the trait. As such, it represents a sort of benchmark increase in posterior relative to the prior, one that can be used to test hypotheses about particular values of a trait. Clarke and Barron (1990; Clarke & Barron, 1994) have shown that the value

\[
-2 \left[ \log \left( \frac{P(\theta_0|x)}{P(\theta_0)} \right) - \psi(x) \right],
\]

(19)

where \(\theta_0\) is the hypothesized trait value, is distributed as \(\chi^2 - 1\) with 1 degree of freedom for a unidimensional trait. Therefore, in addition to being interpretable as an index of global psychometric information, the information utility can also be interpreted as a baseline value in a sort of Bayesian analogue to the likelihood ratio test. Rather than taking twice the log-likelihood, one is taking twice the log posterior over the prior, which is equivalent to a log normalized likelihood.

The hypothesized trait value, \(\theta_0\), could be any value of interest, such as a criterion value for selection or treatment. In clinical settings, for example, there are commonly adopted thresholds for what constitutes clinical significance, which becomes critically relevant to diagnostic interpretations, treatment decisions, or legal
status. In such settings, it would often be important to test whether the estimated trait value differs significantly from the threshold (e.g., a score significantly greater than two standard deviations above the mean might qualify the examinee for social services). Another possible choice, relevant to the issue of test selection, is motivated by whether the observed pattern of responses differs significantly from what would have been hypothesized based on the selection of the test alone. The trait value with the maximum prior probability represents a prior hypothesis implied by the selection of test. By setting $\theta_0$ equal to this trait value, one can formally evaluate whether the measure or test significantly added to what was hypothesized implicitly by administering that particular test at all.

Table 2 illustrates some of these concepts using examples. The last four columns present the results of tests that the estimated trait values differ significantly from the trait value implied by the test prior ($H_o: \theta_0 = 0$; i.e., $H_o: \theta_0 = 1.829$, from Table 2) and the population average ($H_o: \theta_0 = 0$).

As expected, response patterns with greater frequencies of endorsement result in larger trait estimates. The pattern consisting of no endorsements, for example, resulted in a trait estimate of $-2.099$; the first pattern in the table, in which most of the items were endorsed, resulted in a trait estimate of 2.713. Information utilities varied across the response patterns, reflecting varying amounts of information each pattern provided about trait level. Differences in the magnitudes of utilities varied across the response patterns, reflecting varying amounts of information each pattern provided about trait level. Differences in the magnitudes of utilities reflect a combination of how far the trait estimate was from the prior, and how much information the test provides at each trait level. The largest $u(x)$ values were for response patterns associated with $\theta$ estimates far from the value implied by the prior; the smallest $u(x)$ value was for the response pattern associated with a $\theta$ estimate of 1.002.

Hypothesis tests were calculated using Equation 19, using values of $u(x)$ shown in the table. $p$-values were calculated by adding 1 to each test statistic, and comparing them to a $\chi^2$ distribution with 1 df. Results of the hypothesis tests are generally consistent with expectations, with test statistics larger, and $p$-values smaller, for response patterns whose trait estimates were closer to the trait values under the null hypotheses. The second response pattern, for example, has an associated trait estimate very close to the value implied by the reference prior ($\hat{\theta} = 1.876$). The trait estimate for that pattern, correspondingly, does not significantly differ from the prior estimate ($\chi^2 = -1 = -0.953; p = .829$), but does significantly differ from zero ($\chi^2 = 1 = 16.237; p < .001$). In this sense, the test did not significantly change prior estimates of the trait value based on the test that was selected. Conversely, the fifth response pattern has a trait estimate of zero ($\hat{\theta} = 0.000$), which did differ significantly from the prior estimate ($\chi^2 = -1 = 13.026; p = .000$), but not from average ($\chi^2 = 1 = -0.907; p = .760$).

**Computerized Adaptive Testing**

Another important application of information utility is in computerized adaptive testing (CAT). Two central issues in CAT are which criteria to use for item selection, and when to stop administering items. Information utility could be used for both purposes, by selecting items that maximize the expected information utility of the resulting test, and stopping when the change in information utility decreases below some threshold.

Note that in the CAT context, after the initial item administration, the posterior distribution for a trait estimate becomes the prior for the next item administration. In other words, in Equation 4, $P(\theta|x)$ of one administration becomes $P(\theta)$ for the next administration. One can then choose the next item by selecting the item that maximizes the criterion

$$\sum P(x_\theta)u(x).$$

where $P(x)$ is the marginal probability of a possible response pattern $x$, including possible responses to the next item as well as previous responses (i.e., the posterior predictive probability). This maximum expected information utility (MEIU) criterion resembles the expected information utility in Equation 6, but instead of summing over observed response patterns, one is summing over possible response patterns, weighting by the probability of that response, given previous observed responses and possible responses to the next item.

In each administration, the total information provided by the CAT relative to the initial prior should increase, but the rate of increase should decline over administrations. Moreover, $u(x)$ should decrease with each item administration, with diminishing returns in each administration, as the prior for each succeeding trial becomes more precise. As such, the change in $u(x)$ could be used as a criterion for test termination: when the change in $u(x)$ decreases beyond a certain threshold, additional items are not adding sufficient information, and the test can terminate.

It is useful to conceptually compare the MEIU criterion with other CAT item selection criteria. In contrast to commonly used criteria—for example, the maximum information criterion (MI), the maximum likelihood weighted information (MLWI; Veerkamp & Berger, 1997), or the maximum posterior weighted information (MPWI; van der Linden, 1998)—which tend to minimize the expected standard error of the estimate, or some similar function of the standard error—the MEIU maximizes a form of expected log-likelihood. Noting that the posterior divided by the prior (Equation 4) is a form of normalized likelihood—that is, the likelihood normalized by the marginal probability of the data—the MEIU can be seen as maximizing the expected log normalized likelihood.

The MEIU criterion possibly most resembles the global information criterion of Chang and Ying (1996), which can be written as

$$\int \sum P(x_\theta) \ln \left( \frac{P(x_\theta)}{P(x_\theta|\hat{\theta})} \right) d\hat{\theta},$$

where $\hat{\theta}$ is the current estimate of the trait on a given trial. Comparing Equations 21 and 4, it is apparent that the MEIU and Chang–Ying criteria are very similar. For example, the likelihood of a response vector in Equation 21 is replaced with the posterior of the trait estimate in Equation 4 (recall that in Bayesian CAT, the posterior of one trial becomes the prior for the next). Also, the expectation over the data is taken inside the integral rather than outside. In contrast to other criteria, which are based on local information concepts, the MEIU and the Chang–Ying criterion are both global information criteria, and both maximize a function of a log-likelihood. In contrast to the MEIU, however, the Chang–Ying criterion can be seen as maximizing an expected log likeli-
Figure 6 illustrates the application of information utility to CAT. The figure presents the information utility for a respondent to a series of simulated CAT item trials for the SNAP-2 Negative Temperament subscales. In this case, responses were taken from an actual respondent in the normative data set; items were administered using the MEIU as explained above, with a Jeffreys prior, continuing until all 28 items were exhausted. In the figure, the open circles represent the information utility at each item trial. The solid line represents an estimate of the information utility based on Equation 8, using the cumulative test at each trial to calculate the test information, and using the difference in cumulative tests at each trial to estimate the contribution of each new item. Overall, the information utility declines as the trials proceed, reflecting the increasing accuracy of the prior at each trial. The information utility does not monotonically decrease, because occasionally the examinee will respond differently than expected, shifting the estimate from prior expectations. Finally, the bound on the information utility given by Equation 8 seems reasonably accurate, as it tracks the decline in information utility fairly closely.

Recent work suggests that the MEIU criterion performs better relative to other CAT item selection criteria, both in terms of the accuracy of the estimates as well as efficiency of administration (Wang & Chang, 2010). Interestingly, this may be especially true for individuals whose trait values are at the extreme ranges of the trait distribution and are otherwise measured poorly according to maximum Fisher information criteria. Estimates using the MEIU criterion tended to be more uniformly accurate across the range of the trait, with less sharp declines along the extremes (Wang & Chang, 2010). These findings are generally consistent with the examples discussed above, where selection of tests with larger information utilities tended to result in accuracy rates that were more stable across the range of the trait.

**Interpreting the Absolute Scale of Information Utility and Related Indices**

One particularly important issue requiring further research is how to interpret the absolute scale of information utility. Reliability, for example, can be interpreted on an absolute scale, as it ranges from 0 to 1, and its widespread use has set implicit, if not explicit, standards for what is acceptable. Additional research is needed to establish similar standards and better understand what to expect with regard to the absolute scale information utility indices.

Various information utility indices discussed here do have an interpretable scale, in an absolute as well as relative sense. The information utility bounds, scaled as the NMRU, for example, also varies from 0 to 1, and can be interpreted in a percent entropy or uncertainty metric. Similarly, the criterion information utility and information utility lower bounds themselves can be interpreted in terms of expected log-likelihood ratio values. Importantly, information utility values can be interpreted on a relative scale: differences in information utility lower bounds, for example, can be interpreted in terms of the relative number of distinguishable trait estimates two tests can produce, or the relative average width of confidence intervals produced by the two tests.

Nevertheless, further research is needed to better establish what is typical for information utility indices and how to interpret them on an absolute scale. The instruments used here as examples, the SNAP-2 and DAPP, are extremely well-validated, reliable measures that have been used in numerous contexts. As such, they may provide some rough guidelines about what to expect for similarly reliable, well-validated measures. For example, all $\psi(X)$ were greater than 1.2, $\psi_u(X)$ were greater than 1.3, and the NMRU values were greater than .45, suggesting that these values might provide heuristic guidelines for tests with similar characteristics. However, it is somewhat unclear how well these might generalize to other settings. Relatedly, future research should clarify how to interpret the scale of various indices in the context of one another—for example, would a given value of NMRU be interpreted differently depending on the absolute value of $\psi(X)$ or $\psi_u(X)$?

**Impact of Model Characteristics and Assumptions**

Another important issue requiring further inquiry is how characteristics of the IRT models used—in terms of their properties, assumptions, and appropriateness—impact characteristics of information utility and its estimation. It is conceivable, for example, that estimates of information utility might be misleading in some cases if assumptions about the IRT model being used are incorrect for any given individual or sample. These concerns might be somewhat less relevant to sample-independent indices of information utility (e.g., the criterion information utility or NMRU), but even in these cases it is important characterize how estimation and
model assumption errors impact the indices. For example, the
distribution used to estimate item parameters or the form of the
item response function may be misspecified, and there will always
be errors in item parameter estimates due to sampling variation. Of
course, the importance of correct model assumptions is important
in any assessment scenario, but further research is needed to clarify
how it impacts use of information utility in particular.

Inferences on Information Utility

Yet another issue, related to both the scale of information utility
and the impact of model misspecification, is how to make and
evaluate inferences about information utility statistics. It is cur-
cently unclear, for example, how to most appropriately determine
whether two information utility indices significantly differ,
whether one frames the question in frequentist or Bayesian terms.
Monte Carlo estimation of the criterion information utility pro-
vides a framework for partially addressing this question (e.g., one
could test the difference between two estimates using the same
framework for calculating confidence intervals around a single
estimate, as explained in Appendix B), but still does not account
for sampling error in the item parameter estimates. Sampling error
in the item parameter estimates is probably the primary source of
error in making inferences about the information utility bounds,
and is probably a primary source of error in making inferences
about information utility as well. Understanding how stochastic
errors in the estimation of item parameters affects the process of
estimating and making inferences about information utility statistics
will be an important question to address in future research.

Multidimensional Tests

Another important issue is how to extend concepts of informa-
tion utility to multidimensional tests. In general, evaluating infor-
mation utility becomes more complex in the presence of multiple
traits, as an estimate of one trait potentially provides information
about other traits.

Extending information utility itself to multidimensional tests
may be relatively straightforward, in that one can reexpress Equa-
tion 4 in terms of a vector of trait estimates instead of a single
scalar trait estimate—for example, \( P(\theta) \) instead of \( P(0) \). However,
related concepts, such as the reference prior, have proved more
challenging to extend to multidimensional settings. Many re-
searchers have suggested using the determinant of the information
matrix in place of the information function in the Jeffreys prior in
multidimensional settings (e.g., Berger et al., 2009; Clarke &
Barron, 1994; Good, 1968; Jeffreys, 1946). Although this shares
many of the properties of the unidimensional Jeffreys prior (e.g.,
Clarke & Barron, 1994), questions remain about its general suit-
ability as a reference prior in multidimensional settings, as well as
how to estimate the reference prior in general in the multidimen-
sional case (Berger et al., 2009). Further research is needed to
address these issues.

Summary

Overall, information utility serves a number of important func-
tions. First, it provides indices of the global information provided
by a measurement, for an individual response pattern or a sample
of response patterns, relative to what is known about trait level a
priori. Second, information utility forms the basis for a particularly
useful prior for trait estimation, the reference prior, the prior that
maximizes the information provided by the test, and in this sense
is the least possible informative prior. Finally, indices derived
from information utility—the criterion information utility and
lower bound to information utility—provide sample-independent
measures of global information that generalize the role Fisher
information plays in test selection, hypothesis testing, and confi-
dence intervals.

As noted earlier, the global nature of information utility suggests
it is a useful complement to more local indices of measurement
precision and information, such as Fisher information and reliabil-
ity. Information utility might be more useful in scenarios where
trait level is relatively uncertain, or where a global index of
measurement precision or information is required; Fisher informa-
tion and reliability might be more useful in scenarios where trait
level is more certain, or where a more localized index of measure-
ment precision or information is required. Use of the two types of
information, Lindley or Fisher information, will depend on the
circumstances and questions at hand, as they reflect different
aspects of measurement quality.

Relatively, information utility indices might prove useful in
situations where use of reliability or Fisher information has been
deemed inappropriate. Clinical research, for example, often fo-
cuses on identifying deficits specific to a given form of psycho-
pathology. In doing so, it is important to avoid psychometric
confounds, where a measure discriminates between groups not
because it reflects a pathology specific to one group, but because
it is more informative in general. Noting problems with reliability
and related indices of measurement precision in this scenario (due
to dependence on latent and observed variance), Kang and Mc-
Donald (2010) concluded that it might be “useful to focus on
developing an alternative stable and sensitive measure of discrim-
inating power that can be calculated from the observed properties
of a task in a control population” (p. 306). Information utility
indices might be appropriate in such situations.

Overall, use of information utility and quantities that derive
from it may be particularly useful for quantifying measurement
precision in situations where knowledge about trait level is inac-
curate or vague. Such indices are also useful in quantifying global,
as opposed to local, information. Information utility has a number
of important applications, such as in hypothesis testing and com-
puter adaptive testing. Further research is needed to better under-
stand how it can be implemented in more complex measurement
scenarios.

References

Balasubramanian, V. (1997). Statistical inference, Occam’s Razor, and
statistical mechanics on the space of probability distributions. Neural
length principle in coding and modeling. IEEE Transactions on Infor-
mation Theory, 44, 2743–2760. doi:10.1109/18.720554
reference priors. Annals of Statistics, 37, 905–938. doi:10.1214/07-
AOS597
of Statistics, 7, 686–690. doi:10.1214/aos/1176344689
Appendix A

Monte Carlo Estimation of the Reference Prior

As noted in the text, in cases where the posterior of the trait estimate is asymptotically normally distributed, the Jeffreys prior is the reference prior. Although this covers many, if not most, cases in item response theory (IRT), there may be some cases where it is necessary to estimate the reference prior using Monte Carlo simulations. Berger, Bernardo, and Sun (2009) have provided a very general algorithm for estimating the reference prior in this way. In the current context, their algorithm is as follows:

1. Choose a suitably large value of $J$, the number of items in the simulation test. In most situations, this would be some multiple of the number of items in the actual test, as the reference prior is defined as a limit where the number of items goes to infinity. Although larger $J$ increases the accuracy of the simulations, practical constraints will limit the size of $J$. The simulation test can be thought of an imaginary test where the items of the actual test are duplicated an arbitrary large number of times, to a total of $J$ items.

2. Choose a number of Monte Carlo replications, denoted here by $M$.

3. For any given latent trait value $\theta$, repeat, for $m = 1, ... M$:
   a. Generate a set of $M$ response vectors $\{x_1, \ldots, x_m, \ldots x_M\}$, each response comprising $J$ items, according to the IRT model of interest, denoted here by $P(x|\theta)$.
   b. Calculate a normalizing constant for each response vector, $c_m = \int P(x_m|\theta)\,d\theta$. Note that this integral is taken over the entire range of the trait.
   c. Calculate the log normalized likelihood for each response vector, $r_m = \ln P(x_m|\theta)/c_m$. 
   d. Compute $P^*(\theta) = \exp[M^{-1}\sum_m r_m]$ and store the pair $(0, P^*(0))$. These points define the prior.

4. Repeat routine (3) for all latent trait values $\theta$ for which the prior is required. Berger et al. (2009) have noted that interpolation or smoothing techniques may be used for continuous latent traits if desired.

(Appendices continue)
Appendix B

Monte Carlo Estimation of the Criterion Information Utility

A. The criterion information utility can be reexpressed from its form in Equation 12, into a form that facilitates Monte Carlo estimation (cf. Equation 9). Note that the sum is taken over all possible response patterns.

\[ \psi(x) = \sum P'(x) \psi(x) \]

\[ = \sum \left[ \int P(\theta|0) P'(\theta) \right] \psi(x) \left[ \frac{P'(0|x)}{P'(\theta)} \right] d\theta \]

\[ = \sum \left[ \int P'(0|x) P'(\theta) \right] \psi(x) \left[ \frac{P'(0|x)}{P'(\theta)} \right] d\theta \]

\[ = \int \int P(\theta|x) P'(\theta) \left[ \psi(x) \left[ \frac{P'(0|x)}{P'(\theta)} \right] d\theta dx \]

\[ = \int \int P'(0|x) \left[ \psi(x) \left[ \frac{P'(0|x)}{P'(\theta)} \right] d\theta dx \]

B. Given this form, the criterion information utility can be estimated using the following Monte Carlo procedure:

1. Choose a number of Monte Carlo replications, denoted here by \( M \).

2. Generate \( M \) values of the latent trait, \( \theta_m \), for \( m = 1, \ldots, M \), from the distribution implied by the reference prior \( P*(\theta) \). That is, sample \( M \) values of \( \theta \), with \( \theta \sim P*(\theta) \).

3. For each simulated trait value, \( \theta_m \), randomly generate a corresponding response pattern \( x_m \) according to the test model. That is, generate \( x_m \) for \( m = 1, \ldots, M \), with \( x_m \sim P(x|\theta_m) \).

4. The criterion information utility is estimated by

\[ \hat{\psi}(X) = \frac{1}{M} \sum \left[ \frac{P'(\theta_m|x_m)}{P'(\theta_m)} \right] . \]

Confidence intervals can be calculated around the estimated criterion information utility using standard formulas for the confidence interval around a mean. That is, defining the logarithm term above as

\[ \delta_m = \ln \left[ \frac{P'(\theta_m|x_m)}{P'(\theta_m)} \right] . \]

confidence intervals can be calculated by setting the standard error equal to the standard deviation of the sampled values \( \delta_m \) divided by the square root of the size of the Monte Carlo sample—that is,

\[ sd(\delta_m)/\sqrt{M} \]

Received February 10, 2010
Revision received June 19, 2012
Accepted June 25, 2012