A Generalized Framework for Classical Test Theory

Robert C. Foster

Los Alamos National Laboratory

ACCEPTED FOR PUBLICATION IN *THE JOURNAL OF MATHEMATICAL PSYCHOLOGY*

Abstract

This paper develops a generalized framework which allows for the use of parametric classical test theory inference with non-normal models. Using the theory of natural exponential families and Bayesian theory of their conjugate priors, theoretical properties of test scores under the framework are derived, including a formula for parallel-test reliability in terms of the test length and a parameter of the underlying population distribution of abilities. This framework is shown to satisfy the general properties of classical test theory several common classical test theory results are shown to reduce to parallel-test reliability in this framework. An empirical Bayes method for estimating reliability, both with point estimates and with intervals, is described using maximum likelihood. This method is applied to an example set of data and compared to classical test theory estimators of reliability, and a simulation study is performed to show the coverage of the interval estimates of reliability derived from the framework.

Keywords— Classical Test Theory, Reliability, Generalized Models, Bayesian Statistics, Natural Exponential Families, Empirical Bayes.

Robert C. Foster may be contacted at rcfoster@gmail.com.

1 Introduction

This paper presents a generalized framework for classical test theory. The key idea is this: a hierarchical model induces a correlation structure on the resulting data. Given certain assumptions, properties of the correlation structure can be written as a function of a parameter or parameters of the underlying hierarchical model. Rather than treat the the correlation structure itself as the target of inference, the goal becomes to choose an appropriate hierarchical model and estimate its parameters. The correlation structure can then be estimated directly from these parameters. This opens up new avenues for parametric inference, as the hierarchical structure allows for the rich inferential theory of generalized linear or hierarchical Bayesian models to be applied to classical test theory problems. This paper uses the theory of natural exponential families and Bayesian theory of their conjugate priors to develop a simple but flexible hierarchical model for unidimensional test data with equal item difficulty. With the assumption of conditional independence, the mean-variance relationship for natural exponential families can be exploited to derive the parallel-test reliability and Cronbach's alpha as a function of the test length and a single parameter of the underlying distribution of abilities.

Section 2 provides a brief review of early connections between classical testing theory and Bayesian inference in the statistical literature, most of which have made the (overly restrictive, as will be shown) assumption of normality for test scores. Beyond the review, this paper can be divided into two major parts, one theoretical and one applied. The theoretical first part, given in Sections 3 and 4, builds the generalized framework for classical test theory from first principles using natural exponential families and conjugate priors. Section 3 states the assumptions of this framework and derives a simple formula for the parallel-test reliability in terms of the test length and a parameter of the underlying population distribution of abilities. Section 4 shows that the framework satisfies the basic principles of classical test theory and that the parallel-test reliability of the generalized framework is equivalent to commonly used formulas in classical test theory when assumptions are met.

The applied second part, given by Sections 5 and 6, describes how the theoretical framework may be exploited to conduct analyses on real sets of data. Section 5 gives an empirical Bayes method of inference for the test reliability in this framework using maximum likelihood estimation. Section 6 uses an example set of dichotomously-scored test results to directly compare estimation in the framework with common classical test theory formulas, and performs a simulation study to observe that the coverage of interval estimates derived from the framework is nearly nominal and competitive with traditional methods of interval estimation from classical test theory. An appendix is included to demonstrate how the framework may be applied using various natural exponential family distributions, either to simulate data with a certain population value of Cronbach's alpha or estimate alpha for a given set of data.

2 Review

Because the framework relies in part on the use of Bayesian theory for conjugate priors and because the use of Bayesian inference forces specification and consideration of the properties of parametric models, Bayesian inference is of particular interest to this paper. As classical test theory developed in the early and mid twentieth century, however, its origins are unsurprisingly not rooted in Bayesian inference, though early connections were made between classical test theory formulas and Bayesian analyses. In particular, Kelley's formula of Kelley (1923) was identified as matching the mean of a posterior distribution in Novick (1969a). Direct attempts at Bayesian modeling were made in Novick (1969b) and Novick and Thayer (1969), both of which explore the results of Bayesian analyses under the assumption of normality and the former of which presents a model for true score which follows a Poisson distribution rather than a normal. The most complete connection between classical testing theory and Bayesian statistics was made two years later in Novick, Jackson, and Thayer (1971), which assumes normality for test scores and derives a number of posterior quantities for specific priors. Attempts to solve classical testing problems using Bayesian methods were also made in Lindley (1969a), Lindley (1969b), and other works by the same author. Empirical Bayes methods were also developed, as in Lord (1965), Lord (1969), and chapters 22 and 23 of the seminal text on classical test theory Lord, Novick, and Birnbaum (1968). All of these efforts were limited by the same fundamental problem, however: because this research was performed before the advent of Markov chain Monte Carlo (MCMC) techniques and the computational power necessary to perform them, they were forced to deal with models, prior distributions, and techniques for which the posterior distribution could be derived analytically or with only simple numerical methods, often involving an assumption of normality. As computational methods matured Bayesian statistics did find extensive use in psychometric applications, but typically in more complicated item-based models such as item response theory. The history of such efforts is too expansive to describe here but examples may be found in Levy and Mislevy (2016). There have been sporadic attempts to determine how, for example, a Bayesian might estimate quantities such as Cronbach's alpha, as in Li and Woodruff (2002), Padilla and Zhang (2011), and (Najafabadi and Najafabadi, 2016), all of which make assumptions of multivariate normality.

3 Generalized Framework for Reliability

Suppose that a test is performed on a number of subjects, and let the random variable X_i represent the score of the test for subject i with observed value x_i. In this framework, X_i represents a count or a sum of test "items," each of which yields a response which may be either discrete or continuous. Assume that X_i has a probability distribution with parameter θ_i, which is unique to each subject.

$$X_i \sim p(x_i|\theta_i)$$

Following Lord, Novick, and Birnbaum (1968), the parameter θ_i will be referred to as the true "ability" of the subject.

If nothing is assumed about the form of $p(x_i|\theta_i)$, then non-parametric methods may be used to estimate the reliability of the test. This is the logical choice for instances when it is unclear which, if any, assumptions may be made regarding the data. If, however, it can be assumed that $p(x_i|\theta_i)$ is a member of the natural exponential family, then further theory can be derived.

3.1 NEF Distributions and Conjugate Priors

The basic building blocks of the generalized classical test theory framework in this paper are natural exponential family (NEF) distributions, also commonly called the one-parameter exponential family. NEF distributions are described extensively in Morris (1982) and Morris (1983), which serve as a reference for theory and properties of the NEF family. These papers focus on a specific subset of NEF distributions having quadratic variance function; however, this restriction will be avoided in order to remain as general as possible. Many of the most commonly used discrete and continuous distributions belong to the natural exponential family, including the normal, the binomial, and the Poisson.

Since X_i is assumed to be a count or sum, define Y_{ij} as the response for item $j = 1, 2, ..., n_i$ of subject i so that n_i is the test length. It is not necessary for the test length n_i to be equal between subjects. Then X_i can be written as the convolution, in this instance meaning sum, of Y_{ij}.

$$X_i = \sum_{j=1}^{n_i} Y_{ij}$$

Each response Y_{ij} is assumed to be independent conditional on ability θ_i and to identically follow the same NEF distribution. It follows that the X_i are conditionally independent as well. Furthermore, convolutions of NEF distributions are also NEF (Morris, 1982), so

X_i can be assumed to follow an NEF distribution. For example, the normal density can be written as the sum of other normal densities, the binomial as the sum of Bernoulli distributions, the Poisson as the sum of other Poisson distributions, the negative binomial as the sum of geometric distributions, and the gamma density as the sum of exponential distributions. Note that the exponential family is not, in general, closed under convolution. The restriction to the natural exponential family is made to guarantee this closure, and thus, to guarantee the results derived in this paper.

Conditional on having ability θ_i, the expectations of the Y_{ij} are defined as

$$E[Y_{ij}|\theta_i] = \theta_i$$

That the expectation exists, is finite, and may be written in this way is guaranteed by the NEF assumption, though defining θ_i as the expectation of the Y_{ij} may require using a non-standard parameterization of the density or mass function of the random variable Y_{ij}. The particular role of θ_i depends on the NEF distribution chosen for the Y_{ij}. In the case of a Bernoulli distribution, θ_i is a proportion which takes on any real value between 0 and 1. In the case of a Poisson distribution, θ_i is a rate parameter which takes on any positive real number. In the case of a normal distribution, θ_i is a mean which takes on any real number.

Conditional on having mean given by ability θ_i, the expected value of X_i is

$$E[X_i|\theta_i, n_i] = E\left[\sum_{j=1}^{n_i} Y_{ij}\middle|\theta_i\right] = \sum_{j=1}^{n_i} E\left[Y_{ij}\middle|\theta_i\right] = n_i E[Y_{ij}|\theta_i] = n_i\theta_i \tag{1}$$

In testing terms, this means that if a subject is expected to obtain score θ_i on an individual item, then the expected score of the entire test is $n_i\theta_i$. This expectation does not have to be a whole number.

Similarly, the NEF assumption guarantees that the variance exists and is finite. Conditional on ability θ_i, the independence assumption allows the variance of the X_i to be written as

$$Var(X_i|\theta_i, n_i) = Var\left(\sum_{j=1}^{n_i} Y_{ij}\middle|\theta_i\right) = \sum_{j=1}^{n_i} Var\left(Y_{ij}\middle|\theta_i\right) = n_i Var(Y_{ij}|\theta_i)$$

$Var(Y_{ij}|\theta_i)$ is the variance of outcome of each test item conditional on having ability θ_i. For NEF distributions, the variance can be written as a polynomial function of the mean given by the ability θ_i.

$$Var(Y_{ij}|\theta_i) = c_0 + c_1\theta_i + c_2\theta_i^2 + \ldots = V(\theta_i) \tag{2}$$

The function $V(\theta_i)$ is known as the variance function of the NEF, and characterizes the NEF distribution uniquely within the natural exponential family given a sample space (Morris, 1982). For example, defining $Y_{ij} \sim Bern(\theta_i)$ with success probability θ_i gives a variance of $Var(Y_{ij}|\theta_i) = \theta_i(1-\theta_i) = \theta_i - \theta_i^2 = V(\theta_i)$, so it fits Equation (2) with $c_1 = 1, c_2 = -1$ and all other terms 0. The Poisson distribution with mean θ_i has $Var(Y_{ij}|\theta_i) = \theta_i = V(\theta_i)$, so it fits Equation (2) with $c_1 = 1$ and all other terms 0. Defining $Y_{ij} \sim N(\theta_i, \sigma^2)$ with known σ^2 has $Var(Y_{ij}|\theta_i) = \sigma^2 = V(\theta_i)$, so it fits Equation (2) with $c_0 = \sigma^2$ and all other terms 0.

The variance function of the NEF distribution assumed for the Y_{ij} is carried through the convolution process and becomes the variance function of X_i (Morris, 1982). Taking this into account, the variance of the test score X_i is

$$Var(X_i|\theta_i, n_i) = n_i Var(Y_{ij}|\theta_i) = n_i V(\theta_i)$$

Assume that the abilities θ_i themselves follow some distribution $g(\theta_i|\mu, M)$, in a similar manner to the strong true-score theory of Lord (1965). The two-stage model is then

$$\begin{aligned} X_i &\sim p(x_i|\theta_i, n_i) \\ \theta_i &\sim g(\theta_i|\mu, M) \end{aligned}$$

All members of the exponential family are guaranteed to have prior conjugate distributions, and NEF distributions in particular have priors which exist in closed form (Morris, 1983). Assume that $g(\theta_i|\mu, M)$ is conjugate to $p(x_i|\theta_i)$. For example, if $p(x_i|\theta_i, n_i)$ is a normal distribution, then $g(\theta_i|\mu, M)$ is a normal as well. If $p(x_i|\theta_i, n_i)$ is a Binomial distribution, then $g(\theta_i|\mu, M)$ is a beta distribution. If $p(x_i|\theta_i, n_i)$ is a Poisson distribution, then $g(\theta_i|\mu, M)$ is a gamma distribution. The priors do not have to be NEF. A flowchart of the model is shown in Figure 1.

The parameters μ and M of the conjugate distribution of abilities are defined by matching particular moments of the distribution.

$$\begin{aligned} \mu &= E[\theta_i] \\ M &= \frac{E[V(\theta_i)]}{Var(\theta_i)} \end{aligned} \tag{3}$$

In this parameterization, $V(\theta_i)$ is the variance function of the original Y_{ij}, as in Equation (2). The parameter μ is the expected value of the θ_i, and it represents the population mean ability. The parameter M controls, but is not equal to, the variance of θ_i, how spread out the abilities are. Both are assumed, for now, to be known.

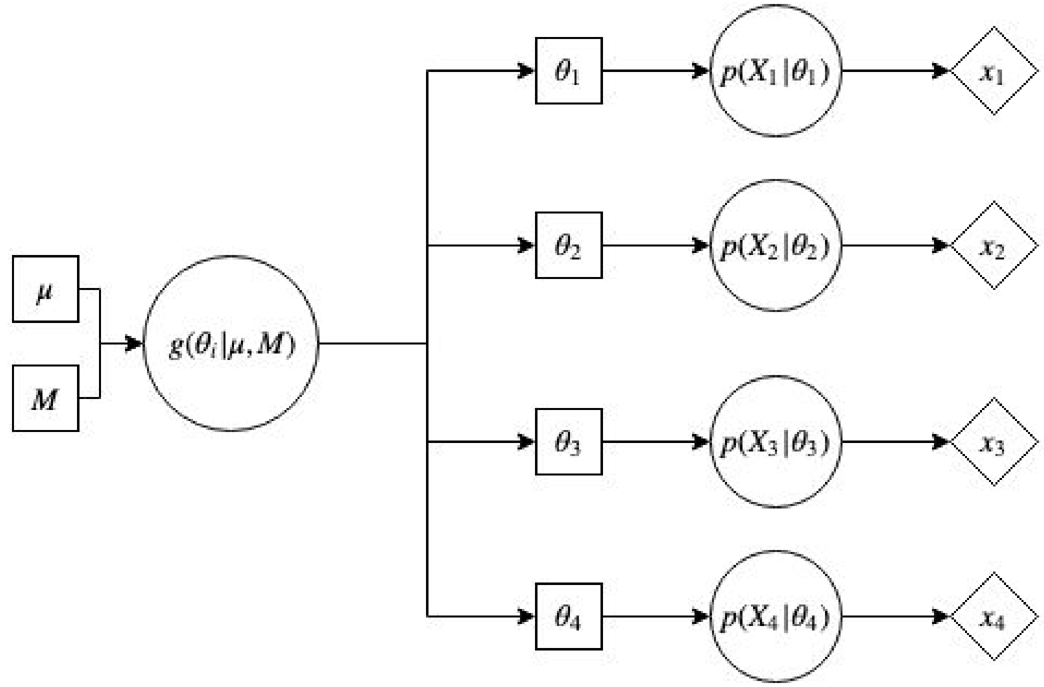

Figure 1: A flowchart illustrating the model considered for classical test theory. Squares indicate parameters, circles indicated probability distributions, and diamonds indicate observed quantities. The hyperparameters μ and M control the distribution $g(\theta_i|\mu, M)$ of true abilities θ_i. For a given subject i, this ability serves as the mean response for each individual item in the random distribution of potential test scores $p(X_i = x_i|\theta_i)$. The observed test score x_i represents one realization of this distribution. The distributions of test scores $p(X_i = x_i|\theta_i)$ are assumed to be a NEF distribution formed as a convolution of "generating" NEF distributions of the response for each item Y_{ij}, with the underlying population distribution of abilities $g(\theta_i|\mu, M)$ being conjugate to the NEF distribution.

Using the law of total expectation, the unconditional expectation of the X_i is

$$E[X_i] = E[E[X_i|\theta_i, n_i]] = E[n_i\theta_i] = n_i\mu$$

and using the law of total variance, the unconditional variance of X_i is

$$Var(X_i) = E[Var(X_i|\theta_i, n_i)] + Var(E[X_i|\theta_i, n_i]) = n_iE[V(\theta_i)] + n_i^2Var(\theta_i) \qquad (4)$$

In Equation (4), the quantity $E[V(\theta_i)]$ is the average variance of the outcome of the individual test items, averaging over all possible abilities θ_i. The quantity $Var(\theta_i)$ is the variance of the abilities themselves, measuring the spread of ability in the population.

Since μ and M are assumed known, Bayes' rule can be used with conjugate prior $g(\theta_i|\mu, M)$ to calculate the posterior distribution for θ_i. NEF families have closed-form posterior densities, which under conjugacy is the same density as that of the prior. For example, the result of placing a beta prior on a binomial success probability θ_i yields a beta posterior, while placing a normal prior on a normal mean θ_i yields a normal posterior. For NEF distributions with conjugate priors, the expected value of the posterior distribution for θ_i is

$$E[\theta_i|x_i, n_i, \mu, M] = \mu + (1 - B)(\bar{x_i} - \mu) = (1 - B)\bar{x_i} + B\mu \qquad (5)$$

where $\bar{x_i}$ is the averaged score for subject i, obtained by dividing x_i by n_i, and B is known as the shrinkage coefficient.For NEF distributions, the form of B is

$$B = \frac{E[n_iV(\theta_i)]}{Var(X_i)} = \frac{n_iE[V(\theta_i)]}{n_iE[V(\theta_i)] + n_i^2Var(\theta_i)} = \frac{E[V(\theta_i)]}{E[V(\theta_i)] + n_iVar(\theta_i)} = \frac{M}{M + n_i} \qquad (6)$$

where $M = E[V(\theta_i)]/Var(\theta_i)$, exactly the parameterization given in Equation (3) (Morris, 1983).

3.2 Parallel-Test Reliability

Suppose that $X_{i,1}$ and $X_{i,2}$ are the resulting scores of subject i with true ability θ_i on each of two parallel tests of length n_i. Each of the two tests here is assumed to be the full length of the original test. Parallel tests are defined as consisting of test items $Y_{ij,1}$ and $Y_{ij,2}$ which have the same NEF distribution with ability θ_i so that each identically and independently follows the model described in Section 3.1.

The goal is to find the correlation coefficient ρ between the scores of the two tests, defined as

$$\rho = \frac{Cov(X_{i,1}, X_{i,2})}{\sqrt{Var(X_{i,1})Var(X_{i,2})}} \tag{7}$$

The numerator of Equation (7) is calculated first. The law of total covariance states that

$$Cov(X_{i,1}, X_{i,2}) = E[Cov(X_{i,1}, X_{i,2}|\theta_i, n_i)] + Cov(E[X_{i,1}|\theta_i, n_i], E[X_{i,2}|\theta_i, n_i]) \tag{8}$$

As previously mentioned, the assumption of conditional independence for items Y_{ij} given ability θ_i yields conditional independence for $X_{i,1}$ and $X_{i,2}$. The first term in Equation (8) is then

$$E[Cov(X_{i,1}, X_{i,2}|\theta_i, n_i)] = E[0] = 0$$

Functionally, this means that given the same test subject i the score on the first parallel test is uncorrelated with the score on the second parallel test.

From Equation (1), the conditional expectation is $E[X_{i,1}|\theta_i, n_i] = E[X_{i,2}|\theta_i, n_i] = n_i\theta_i$. The second term of Equation (8) then becomes

$$Cov(E[X_{i,1}|\theta_i, n_i], E[X_{i,2}|\theta_i, n_i]) = Cov\left(n_i\theta_i, n_i\theta_i\right) = Var\left(n_i\theta_i\right) = n_i^2 Var(\theta_i)$$

For the denominator of Equation (7), each of $Var(X_{i,1})$ and $Var(X_{i,2})$ can be calculated using the law of total variance as shown in Equation (4). Since $X_{i,1}$ and $X_{i,2}$ are assumed to have the same distributional form, they will have the same variance.

$$Var(X_{i,1}) = Var(X_{i,2}) = E[Var(X_{i,1}|\theta_i, n_i)] + Var(E[X_{i,1}|\theta_i, n_i]) = n_i E[V(\theta_i)] + n_i^2 Var(\theta_i)$$

where $E[V(\theta_i)]$ is the average variance of score at the level of test item. Hence, the correlation between them is

$$\rho = \frac{n_i^2 Var(\theta_i)}{n_i E[V(\theta_i)] + n_i^2 Var(\theta_i)} = 1 - \frac{E[V(\theta_i)]}{E[V(\theta_i)] + n_i Var(\theta_i)} = 1 - \frac{M}{M + n_i} = 1 - B \tag{9}$$

where B is the shrinkage coefficient in Equation (6) and $M = E[V(\theta_i)]/Var(\theta_i)$ as in Equation (3). The important result, then, is that for NEF distributions, the parallel-test correlation is equal to one minus the shrinkage coefficient when a conjugate prior is used. The general formula for the parallel-test reliability of NEF distributions is

$$\rho = \frac{n_i}{M + n_i} \tag{10}$$

The parallel-test reliability can be calculated simply from the parameter M of the underlying distribution of abilities θ_i and the test length n_i. In fact, if the assumptions of Section 3.1 are met, the only property of the test which affects the parallel-test reliability is the test length n_i. It should be emphasized that Equation (10) occurs only when the conjugate prior distribution is used for abilities θ_i.

Equation (10) has been previously derived under specific modeling assumptions. Keats and Lord (1962) investigated the negative hypergeometric distribution (a form of the beta-binomial model, further described in Section 6) and obtained Equation (10) exactly – coincidentally, with the same notation. This equation is far more general, however. It applies to the broad and flexible natural exponential family of distributions, and many commonly used formulas in classical test theory can be seen as reducing to or deriving from it. It should also be noted that the hierarchical framework of this section and derivations using laws of total expectation, variance, and covariance have also been explored for specific models, for example with the beta-binomial model in Bechger, Maris, Verstralen, and Béguin (2003), which draws connections between the classical test theory framework and item response theory.

4 Connection to Classical Testing Theory

The basic assumption in classical test theory that the observed score X_i may be decomposed into a "true" component and an "error" component. Within the generalized framework of Section 3, this may be accomplished as

$$X_i = n_i\theta_i + \epsilon_i \tag{11}$$

where $n_i\theta_i$ is the "true" component and ϵ_i is the "error" component. The ϵ_i will have a distribution, but this distribution will depend on the particular NEF distribution of the X_i. The true score is defined as the conditional expectation of the X_i given ability θ_i and test length n_i, as shown in Equation (1). Then from theorem 2.1 of Novick (1966), the errors ϵ_i will have expectation 0 and the true and observed components will be uncorrelated. Since the X_i are conditionally independent by assumption, the errors will also be conditionally independent, and thus uncorrelated. Thus, all general assumptions of classical test theory are met by the framework.

Novick (1966) notes that while a linear regression of the observed score on the true score exists because of the decomposition in Equation (11), a linear regression of the true score on the observed score is not true in general. It is true within the generalized classical test theory framework, however. From Equation (5), the relationship between the true and observed scores is given by

$$E[n_i\theta_i|x_i, n_i, \mu, M] = Bn_i\mu + (1 - B)X_i = \beta_0 + \beta_1 X_i$$

When X_i follows an exponential family distribution, this linear relationship exists if and only if θ_i follows the corresponding conjugate distribution (Diaconis and Ylvisaker, 1979). Hence, the use of the conjugate prior may be seen as justification for the use of a linear regression of true score on observed score. Note that the slope of the relationship is given by the reliability ρ. Note also that when the posterior mean is taken as the estimate of true score, Equation (5) and this formula are essentially Kelley's formula of Kelley (1923), which here applies to the entire natural exponential family. The application of this formula is dependent upon a linear relationship between the posterior mean and the observed score, which occurs if and only if the conjugate prior is used for abilities.

The traditional definition of reliability as the ratio of true-score variance to observed-score variance holds. Let $\sigma_T^2 = Var(n_i\theta_i) = n_i^2 Var(\theta_i)$ be the true-score variance and σ_X^2 be the observed score variance, given by Equation (4). The ratio of true-score variance to observed-score variance is then

$$\frac{\sigma_T^2}{\sigma_X^2} = \frac{n_i^2 Var(\theta_i)}{n_i E[V(\theta_i)] + n_i^2 Var(\theta_i)} = \frac{n_i}{E[V(\theta_i)]/Var(\theta_i) + n_i} = \frac{n_i}{M + n_i} = \rho$$

Other methods of estimating reliability also reduce to simpler forms within this framework. Perhaps the most commonly used formula in classical test theory is Cronbach's alpha (Cronbach, 1951). Let $\sigma_x^2 = Var(X_i)$ and $\sigma_y^2 = Var(Y_{ij})$. The quantity $Var(X_i)$ is given by Equation (4). Similarly, Y_{ij} may be seen as a test of length $n_i = 1$ so that Equation (4) is used to obtain $Var(Y_{ij}) = E[V(\theta_i)] + Var(\theta_i)$. Plugging these into the formula for Cronbach's alpha, it becomes

$$\begin{aligned}\alpha = \frac{n_i}{n_i - 1}\left(1 - \frac{\sum_{i=1}^{n_i} \sigma_Y^2}{\sigma_X^2}\right) &= \frac{n_i}{n_i - 1}\left(1 - \frac{n_i E[V(\theta_i)] + n_i Var(\theta_i)}{n_i E[V(\theta_i)] + n_i^2 Var(\theta_i)}\right) = \frac{n_i}{n_i - 1}\left(1 - \frac{n_i M + n_i}{n_i M + n_i^2}\right)\\ &= \frac{n_i}{n_i - 1}\left(\frac{n_i M + n_i^2 - n_i M - n_i}{n_i M + n_i^2}\right) = \frac{n_i}{n_i - 1}\left(\frac{n_i(n_i - 1)}{n_i(M + n_i)}\right) = \frac{n_i}{M + n_i} = \rho\end{aligned}$$

Hence, Cronbach's alpha can be seen as an estimate of the parallel-test reliability in Equation (10). It can also be shown that formula 21 of Kuder and Richardson (1937) is equivalent to Equation (10) using the framework of Section 3 and assuming a binomial distribution for test scores X_i, which has $V(\theta_i) = \theta_i(1 - \theta_i)$, and a beta prior distribution for θ_i (Keats and Lord, 1962).

The Spearman-Brown prediction formula is also easily obtained. Rearranging Equation (10) gives

$$M = \left(\frac{1-\rho}{\rho}\right) n_i$$

Suppose a reliability of ρ is obtained for test length n_i, but the test length $\tilde{n}_i$ which yields reliability $\tilde{\rho}$ is desired. Equating M for each formula gives

$$\left(\frac{1-\rho}{\rho}\right) n_i = \left(\frac{1-\tilde{\rho}}{\tilde{\rho}}\right) \tilde{n}_i$$

and rearranging terms gives

$$\frac{\tilde{n}_i}{n_i} = \frac{\tilde{\rho}(1-\rho)}{\rho(1-\tilde{\rho})}$$

Lastly, though not a focus of this paper, the framework of Section 3.1 is easily expanded to deal with violated assumptions. A very strong assumption made is that the response Y_{ij} to each item has mean θ_i, assuming in essence that each item is the same difficulty. In practice, this is commonly not the case. This can be addressed by modeling the transformed θ_i values as a function of a subject effect and an item effect. For example, suppose that the Y_{ij} are Bernoulli random variables with $P(Y_{ij} = 1) = \theta_{ij}$. Then modeling

$$\log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) = \alpha_i + \beta_j$$

where α_i is a subject effect and β_i is an item effect yields a basic item response theory (IRT) model (Gelman and Hill, 2007). This derivation, using convolutions of independent exponential family distributions as in Section 3.1, is similar to work found in Rasch (1960). The use of non-conjugate distributions for θ_i induces a non-linear regression of ability (estimated as the posterior mean) on observed score. Conversely, the linear regression induced by the conjugate distribution might be seen as mirroring an IRT procedure with no item parameters and no guessing, where trait or ability is simply randomly distributed within the population.

5 Inference

The formulation of reliability in Section 3 and derivations of classical test theory properties in Section 4 have thus far assumed that the parameters μ and M of the underlying distribution of abilities θ_i are known. In rare cases, there may exist strong prior information which can be used to determine appropriate values to plug in. In most cases, these parameters must be estimated.

5.1 Estimation Methods for μ and M

Maximum likelihood estimation is natural choice for estimation. The act of estimating the parameters of the underlying conjugate distribution of abilities by marginal maximum likelihood and then using them as plug-in estimates to obtain quantities of interest from the posterior distribution makes this an empirical Bayes approach.

The marginal distribution of observed score X_i unconditional on the ability θ_i is obtained by integrating the product of the score distribution and the conjugate prior distribution over θ_i.

$$f(x_i|\mu, M, n_i) = \int_{\theta_i} p(x_i|\theta_i, n_i)g(\theta_i|\mu, M)d\theta_i \tag{12}$$

In the case of a normal distribution for both X_i and θ_i, this marginal density is also normally distributed, simplifying calculations. Other scenarios are more complicated. A complete description of the relationships between the six NEF distributions with quadratic variance function, their conjugate priors, and their marginal distributions can be found in Morris and Lock (2009), and a thorough description of their application in this framework can be found in the Appendix. The marginal density in Equation (12) gives a direct formula for the observed test scores in terms of the desired parameters, and the maximum likelihood estimates $\hat{\mu}$ and $\hat{M}$ are obtained by maximizing the log-likelihood.

$$(\hat{\mu}, \hat{M}) = \arg\max_{(\mu, M)} \sum_{i=1}^{K} \log[f(x_i|\mu, M, n_i)]$$

When neither the distribution of test scores or abilities are normally distributed, this maximization is typically performed with an iterative optimization routine using a computer. An example using the binomial model with a beta prior is given in Section 6.

The advantage of maximum likelihood using an iterative optimization routine is that it presents a direct method of obtaining variance estimates of the parameters as

$$Var(\hat{M}) \approx -J_n^{-1}(\hat{M}) \tag{13}$$

where $\hat{M}$ is the maximum likelihood estimate and $-J_n^{-1}(\hat{M})$ is the observed Fisher information at $\hat{M}$ (Casella and Berger, 2002). This is obtained by calculating the matrix of second partial derivatives of the log-likelihood, called the Hessian matrix, and inverting it at the maximum likelihood estimates $\hat{\mu}$ and $\hat{M}$. As many iterative optimization routines rely on this matrix for determining the direction of the step at each iteration, it may be available. If unavailable, it can calculated numerically.

Another option is to place a hyperprior or hyperpriors on μ and M and perform a full hierarchical Bayesian analysis, which will almost certainly involve Markov chain Monte Carlo sampling. This method is of particular interest where moderately informative prior information may exist about the values of μ and M. Prior sensitivity of the analysis will likely be an issue, especially for small sample sizes. A thorough description of MCMC techniques in psychometrics is given in Levy and Mislevy (2016).

5.2 Empirical Bayesian Point and Interval Estimates for Reliability and Test Length

Given maximum likelihood estimates $\hat{\mu}$ and $\hat{M}$, Equation (10) gives the estimated parallel-test reliability as

$$\hat{\rho} = \frac{n_i}{n_i + \hat{M}} \tag{14}$$

By the invariance of the maximum likelihood estimator, $\hat{\rho}$ is the maximum likelihood estimate of the parallel-test reliability. Note that in Equation (14) the test length n_i is both constant and indexed by subject i. This has two consequences: first, that the test reliability may vary between subjects if the test length n_i varies as well. Second, the reliability may be calculated for any test length, not just the test length n_i used to obtain $\hat{M}$. Supposing one is considering extending the test to new length $\tilde{n_i}$, the new test length may simply be inserted into Equation (14) in order to obtain the maximum likelihood estimate of the parallel-test reliability for that test length.

Assuming the variance for $\hat{M}$ is available, likely from the observed Fisher information obtained by taking the appropriate element of the negative inverted Hessian matrix as in Equation (13) , it can be converted to a variance for the estimated reliability $\hat{\rho}$ by the delta method.

$$Var(\hat{\rho}) \approx \left[\frac{d}{d\hat{M}}\left(\frac{n_i}{n_i + \hat{M}}\right)\right]^2 Var(\hat{M}) = \left(\frac{n_i^2}{(n_i + \hat{M})^4}\right) Var(\hat{M})$$

A $(1-\alpha) \times 100\%$ interval for $\hat{\rho}$ is then given by

$$\left(\frac{n_i}{n_i + \hat{M}}\right) \pm z^* \left(\frac{n_i}{(n_i + \hat{M})^2}\right) \sqrt{Var(\hat{M})} \tag{15}$$

where z^* is the $(1 - \frac{\alpha}{2})$ quantile of the standard normal density.

An interval estimate of the sample size required for a desired reliability is also easily obtained. Given the desired reliability ρ, the estimated test length n_i which will provide the reliability is

$$\hat{n_i} = \left(\frac{\rho}{1-\rho}\right)\hat{M} \tag{16}$$

The variance of the estimated test length for reliability ρ is

$$Var(\hat{n_i}) = Var\left(\left(\frac{\rho}{1-\rho}\right)\hat{M}\right) = \left(\frac{\rho}{1-\rho}\right)^2 Var(\hat{M})$$

and so a $(1-\alpha) \times 100\%$ interval for $\hat{n_i}$ is given as

$$\left(\frac{p}{1-p}\right)\hat{M} \pm z^* \left(\frac{p}{1-p}\right)\sqrt{Var(\hat{M})} \tag{17}$$

where z^* is once again the $(1-\frac{\alpha}{2})$ quantile of the standard normal density.

Lastly, suppose the goal is inference for the estimated reliability if the test length is increased to $\tilde{n_i}$. A point estimate is given by

$$\hat{\rho} = \frac{\tilde{n_i}}{\tilde{n_i} + \hat{M}} \tag{18}$$

with the delta method yielding a corresponding $(1-\alpha) \times 100\%$ interval given by

$$\frac{\tilde{n_i}}{\tilde{n_i} + \hat{M}} \pm z^* \left(\frac{\tilde{n_i}}{(\tilde{n_i} + \hat{M})^2}\right)\sqrt{Var(\hat{M})} \tag{19}$$

Each of the standard errors for the interval estimates is derived from the Fisher information, which depends on asymptotic normality of the maximum likelihood estimator for accuracy. This asymptotic normality in turn depends on both the test lengths n_i and the number of subjects being sufficiently large. When either of these is small, the standard error may be incorrect.

The delta method is, of course, not the only way to obtain variance estimates of the reliability and derived quantities. If a hierarchical Bayesian analysis is performed with MCMC, draws from the posterior distribution for M may be inserted directly into Equation (14) in order to obtain a posterior distribution for ρ or into Equation (16) to obtain a posterior distribution for $\hat{n_i}$. The bootstrap is another option. Both of these will remain sensitive to small sample sizes, a problem for which there is unfortunately no magical cure.

Score	Freq.	Score	Freq.
1	1	11	1
2	1	12	10
3	2	13	8
4	7	14	2
5	13	15	6
6	6	16	1
7	14	17	4
8	5	18	3
9	7	19	2
10	6	20	1

Table 1: Observed score and corresponding frequency from the example data provided with the R package 'CTT' package Willse (2018). The responses are available in the package on a per-item basis, but only the total scores are shown here.

6 Example and Simulation Study

The presentation thus far has been purely theoretical. An example may show how the theory is applied in practice. Suppose that the results of a multiple-choice test are available. This example uses the multiple-choice data included with the R package 'CTT' by Willse (2018). The data consists of the per-item responses on a $n_i = 20$ item multiple choice test conducted on $K = 100$ subjects. Each response is treated as dichotomous, with 0 for an incorrect response and 1 for a correct response. Each subject is assumed to have true ability θ_i, which is unique to each subject and remains constant over all items. The observed scores and number of subjects with the associated scores are shown in Table 1.

6.1 Beta-Binomial Formulation

Under this formulation, the observed score x_i for subject i should be modeled as following a binomial random variable with success probability θ_i. The probability mass function is

$$p(x_i|\theta_i, n_i) = \binom{n_i}{x_i}\theta^{x_i}(1-\theta_i)^{x_i}$$

The distribution of abilities θ_i is assumed to follow the conjugate prior for the binomial mass function, commonly known as the beta distribution. Traditionally, the parameters of a beta distribution are written as α and β. For this paper, the parameterization given in Equation (3) is used. In terms of α and β, this is

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$M = \alpha + \beta$$

This yields $\alpha = \mu M$ and $\beta = (1 - \mu)M$. The density of the prior is then

$$g(\theta_i|\mu, M) = \frac{\theta_i^{\mu M-1}(1-\theta_i)^{(1-\mu)M-1}}{\beta(\mu M, (1-\mu)M)}$$

Integrating out the ability θ_i, the unconditional density of the observed test score x_i in terms of the population parameters is given by

$$f(x_i|\mu, M) = \int_{\theta_i} p(x_i|\theta_i)g(\theta_i|\mu, M)d\theta_i = \binom{n_i}{x_i}\frac{\beta(x_i + \mu M, n_i - x_i + (1-\mu)M}{\beta(\mu M, (1-\mu)M)} \tag{20}$$

The distribution in Equation (20) is known as the beta-binomial distribution and represents the potential distribution of the test score x_i in n_i observations if the subject i is unknown. More importantly, this distribution will serve as the basis for estimation and inference. Given a set of test scores for K subjects, the log-likelihood for parameters μ and M is given by

$$\ell(\mu, M) = \left(\sum_{i=1}^{K} \log[\beta(x_i + \mu M, n_i - x_i + (1-\mu)M)]\right) - K\log[\beta(\mu M, (1-\mu)M)] \tag{21}$$

Maximizing Equation (21) using observed scores x_i and test lengths n_i yields the maximum likelihood estimates $\hat{\mu}$ and $\hat{M}$. As Equation (21) is not analytically tractable, this is typically done with an iterative numerical solver. Similar derivations of the properties of NEF models other than the beta-binomial are given in the Appendix.

6.2 Analysis

For the example data of the ‘CTT’ package, the maximum likelihood estimates obtained by maximizing Equation (21) are $\hat{\mu} = 0.480$ and $\hat{M} = 5.334$. The optimization was performed using the default “optim” command in the R programming language. This routine can be specified to return the Hessian matrix. The estimated variance of $\hat{M}$, given by the negative inverse of the Hessian matrix as in Equation (13), is $Var(\hat{M}) \approx 0.874$. The estimated parallel-test reliability is then

$$\hat{\rho} = \frac{20}{20 + 5.334} \approx 0.789$$

with corresponding 95% interval estimate

$$0.789 \pm 1.96 \left(\frac{20}{(20 + 5.334)^2} \right) \sqrt{0.874} = (0.732, 0.847)$$

A traditional method of calculating the reliability of dichotomously-scored test data is formula 20 of Kuder and Richardson (1937). The KR-20 estimate of reliability for this data set is 0.815 with a 95% interval of $(0.758, 0.864)$ using the method of Feldt (1965) and Feldt, Woodruff, and Salih (1987). Similarly, the KR-21 estimate of reliability for this data set is 0.796 with 95% interval $(0.732, 0.850)$. Unsurprisingly, the empirical Bayesian estimates of reliability and the KR-21 estimates of reliability are close in value, having made the same assumption of equality of test item difficulty.

Suppose that a reliability of $\rho = 0.9$ is desired for the test as a whole. Then following Equation (16), the estimated test length $\hat{n}_i$ is

$$\hat{n}_i = \left(\frac{0.9}{1 - 0.9} \right) 5.334 \approx 48$$

with corresponding 95% interval, following Equation (17), equal to

$$48 \pm 1.96 \left(\frac{0.9}{1 - 0.9} \right) \sqrt{0.874} = (31.50, 64.50)$$

Likewise, suppose that the test length is to be increased to $\tilde{n_i} = 30$ items. The estimated reliability from Equation (18) is then

$$\hat{\rho} = \frac{30}{30 + 5.334} \approx 0.849$$

with corresponding 95% interval, from Equation (19), equal to

$$0.849 \pm 1.96 \left(\frac{30}{(30 + 5.334)^2} \right) \sqrt{0.874} = (0.805, 0.893)$$

These can be repeated to give interval estimates for any desired reliability ρ or new test length $\tilde{n_i}$. A plot of estimated reliability as a function of the test length with 95% bounds and a plot of the estimated test length as a function of the reliability with 95% bounds are shown in Figure 2.

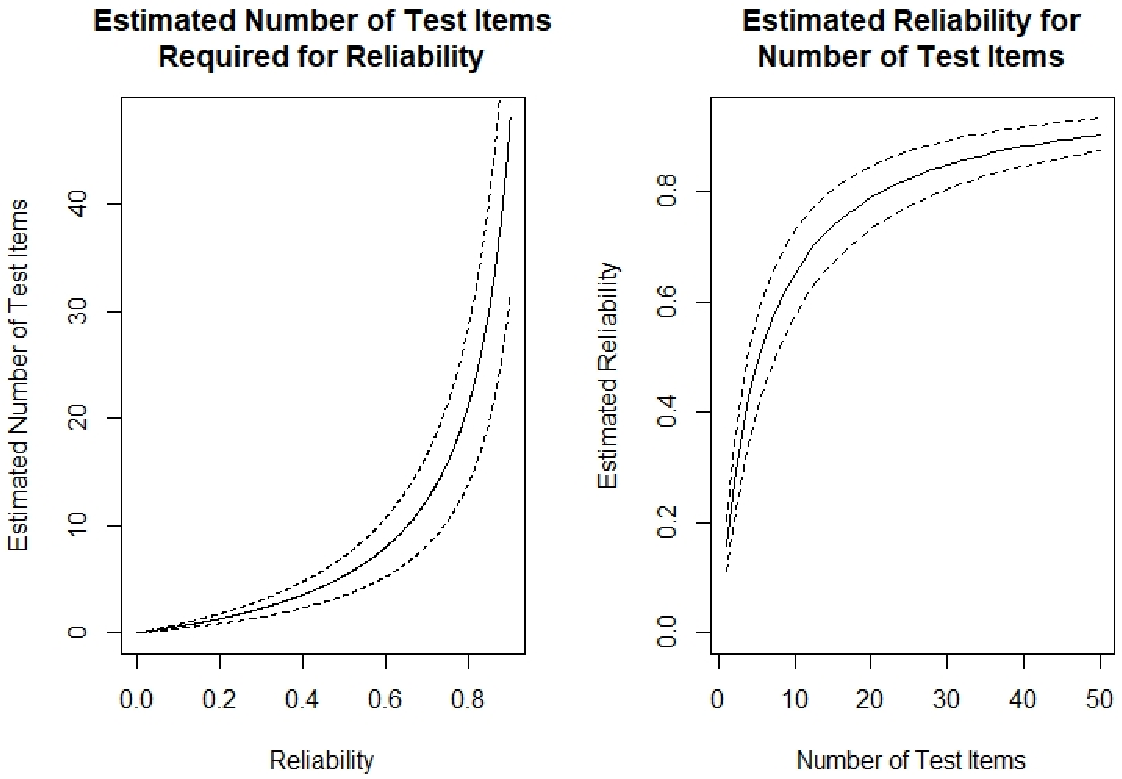

Figure 2: On the left, a plot showing the estimated test length $\hat{n}_i$ required to obtain a test reliability of ρ, with 95% error bounds. The estimate and error bounds are calculated using Equations (16) and (17). On the right, a plot showing the estimated reliability $\hat{\rho}$ for a given test length $\tilde{n}_i$, with 95% error bounds. The estimate and error bounds are calculated using Equations (18) and (19). Both plots are calculated for the example data included with the 'CTT' package of Willse (2018), given in Table 1.

6.3 Simulation Study

The example of Section 6.2 is useful for showing rough equivalence of the empirical Bayesian and classical estimators of reliability for a single data set, but it is unable to show properties of interval estimators. For this, a simulation study is necessary.

The beta-binomial model of Section 6.1 can be used to simulate sets of observed test scores. For this simulation study, three sets of parameters μ and M were chosen. The first set, with $\mu = 0.75$ and $M = 4$, yields a distribution of abilities which is unimodal with mean 0.75, but skewed to the left rather than symmetric. The second set, with $\mu = 0.5$ and $M = 1$, is the classic Jeffrey's prior for a binomial success probability. This distribution of abilities is U-shaped with a mean of 0.5, symmetric, but with no mode. The last set, with $\mu = 0.5$ and $M = 20$, is unimodal, symmetric, and nearly bell-shaped, with only a slight divergence at the tails. For each set of parameters, five sets of test length n_i and number of subjects K were used. These were $n_i = 10$ and $K = 10$, which is small in both and so should pose difficulties in estimation, and each combination of $n_i = 30$ or 100 and $K = 30$ or 100, presenting a moderate or large sample size in each category.

The target of estimation is the true parallel-test reliability in Equation (10). The accuracy of the estimator is judged both pointwise and with 95% intervals. For pointwise accuracy, the root mean squared error is used.

$$RMSE = \sqrt{\frac{1}{N_{sims}} \sum_{s=1}^{N_{sims}} (\rho - \hat{\rho}_s)^2}$$

For interval accuracy, the empirical coverage is used. All intervals are calculated as 95% intervals.

A total of $N_{sims} = 100000$ simulated data sets were created for each set of parameters. For each simulated data set, both a point and 95% interval estimate of reliability are calculated using the empirical Bayesian estimators of Section 5 and the classical formula 20 of Kuder and Richardson (1937) with corresponding interval given by Feldt (1965) and Feldt, Woodruff, and Salih (1987), the default interval implemented in SPSS. The results of the simulation study are shown below in Table 2.

μ	M	n_i	K	ρ	EB RMSE	Coverage	KR20 RMSE	Coverage
0.75	4	10	10	0.714	0.244	0.914	0.233	0.885
0.75	4	30	30	0.882	0.040	0.956	0.038	0.917
0.75	4	30	100	0.882	0.019	0.953	0.019	0.925
0.75	4	100	30	0.962	0.013	0.957	0.012	0.919
0.75	4	100	100	0.962	0.006	0.952	0.006	0.928
0.50	1	10	10	0.909	0.078	0.941	0.062	0.800
0.50	1	30	30	0.968	0.009	0.953	0.009	0.906
0.50	1	30	100	0.968	0.005	0.950	0.005	0.934
0.50	1	100	30	0.990	0.003	0.953	0.003	0.922
0.50	1	100	100	0.990	0.001	0.952	0.001	0.938
0.50	20	10	10	0.333	0.245	0.620	0.515	0.926
0.50	20	30	30	0.600	0.132	0.957	0.119	0.944
0.50	20	30	100	0.600	0.061	0.953	0.059	0.952
0.50	20	100	30	0.833	0.055	0.957	0.049	0.946
0.50	20	100	100	0.833	0.025	0.953	0.024	0.953

Table 2: Results of simulation experiment, simulating from the beta-binomial model described in Section 6.1. The target of estimation is the true parallel test reliability given by $n_i/(n_i+M)$ in Equation (10). The first four columns give the true population parameters μ and M used to simulate the data. The parameter n_i is the number of test items and the parameter K is number of subjects. A total of $N_{sims} = 100000$ data sets were simulated for each set of parameters. For each set of parameters, the test reliability and nominal 95% interval coverage were calculated using the empirical Bayes approach of Section 5.1 and implemented in Section 6.2. This is compared to the estimated reliability using Kuder and Richardson (1937) formula 20 and nominal 95% interval using the method from Feldt (1965) and Feldt, Woodruff, and Salih (1987).

In general, the empirical Bayesian estimator gives a slight increase in RMSE as compared to Kuder and Richardson's formula 20, but with the advantage of providing closer to nominal coverage. The scenario was reversed, however, with the set of parameters $\mu = 0.50, M = 20$, and $n_i = K = 10$, in which the empirical Bayesian method struggled to accurately estimate the underlying variance of abilities when observed scores tended to be clustered together due to both the small sample size and the unimodal and symmetric shape of the distribution of abilities, while the KR-20 method produced more accurate interval estimates. Taken as a whole, it is clear that the empirical Bayesian estimates of reliability are close to traditional formulas for reliability in terms of both point and interval estimation.

7 Discussion and Conclusion

It has been shown that, when unidimensionality and equal item difficulty are assumed, many of the results of classical test emerge directly from the theory of natural exponential families and Bayesian theory of their conjugate priors. When this broader theory is taken into account, it allows for a rich inferential framework which is amenable to common techniques for hierarchical models. Estimates from this framework are competitive, in terms of common measures such as root mean squared error and coverage of intervals, with traditional classical test theory formulas.

The framework can and should be extended, as assumptions of classical test theory are often violated in practice. The generalized framework presents direct ways in which these violated assumptions may be incorporated into a model. For example, it has already been noted that modeling abilities as a function of both subject and item leads directly to an item response model, but there is no reason to limit this to dichotomous data. The NEF assumption allows the framework to extend naturally into generalized linear model theory. Poisson count data, geometric counts of the number of trials until success, or negative binomial counts of the number of successes until a given number of failures may all be modeled in this manner as well. The relatively modest framework here may be embedded into more complicated hierarchical models in order to adapt to peculiar aspects of the test data at hand, such as changing abilities over time or multiple subgroups within the population. It may also be possible to expand the framework to multi-parameter test items through the use of multivariate analogues of natural exponential families, for example by modeling polytomous responses using the Dirichlet-multinomial similarly to modeling dichotomous responses using the beta-binomial. The term framework is appropriate - it is a small piece to be built upon, in order to create a larger whole model which may address the problem at hand.

References

Timo M. Bechger, Gunter Maris, Huub H. F. M. Verstralen, and Anton A. Béguin. Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27(5):319 – 334, September 2003.

G Casella and R Berger. *Statistical Inference, Second Edition*. Duxbury Press, 2002.

Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297 – 334, September 1951.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269 – 281, 1979.

Leonard S. Feldt. The approximate sampling distribution of kuder-richardson reliability coefficient twenty. *Psychometrika*, 30(3):357 – 370, September 1965.

Leonard S. Feldt, David J. Woodruff, and Fathi A. Salih. Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11(1):93 – 103, March 1987.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical model*. Cambridge University Press, 2007.

J. A. Keats and Frederic M. Lord. A theoretical distribution for mental test scores. *Psychometrika*, 27(1):59 – 72, March 1962.

T.L. Kelley. *Statistical method*. New York: Macmillan, 1923.

G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151 – 160, 1937.

Roy Levy and Robert J. Mislevy. *Bayesian psychometric modeling*. Chapman & Hall/CRC, 2016.

Jun Corser Li and David J. Woodruff. Bayesian statistical inference for coefficient alpha. Technical report, ACT Research Report Series, 2002.

Dennis V. Lindley. A bayesian solution for some educational prediction problems. Technical report, Educational Testing Service, Princeton, New Jersey, July 1969a.

Dennis V. Lindley. A bayesian estimate of true score that incorporates prior information. Technical report, Educational Testing Service, 1969b.

Frederic M. Lord. A strong true-score theory, with applications. *Psychometrika*, 30 (3):239 – 270, September 1965.

Frederic M. Lord. Estimating true-score distributions in psychological testing (an empirical bayes estimation problem). *Psychometrika*, 34(3):259 – 299, September 1969.

Frederic M. Lord, Melvin R. Novick, and Allen Birnbaum. *Statistical theories of mental test scores*. Addison-Wesley, Oxford, England, 1968.

Carl N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65 – 80, 1982.

Carl N. Morris. Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, 11(2):515 – 529, 1983.

Carl N. Morris and Kari F. Lock. Unifying the named natural exponential families and their relatives. *The American Statistician*, 63(3):247 – 253, August 2009.

Amir T. Payandeh Najafabadi and Maryam Omidi Najafabadi. On the bayesian estimation for cronbach's alpha. *Journal of Applied Statistics*, 43(13):2416 – 2441, 2016.

Melvin R. Novick. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1):1 – 18, 1966.

Melvin R. Novick. Multiparameter bayesian indifference procedures. *Journal of the Royal Statistical Socity. Series B (Methodological)*, 31(1):29–64, 1969a.

Melvin R. Novick. Bayesian methods in psychological testing. Technical report, Educational Testing Service, Princeton, New Jersey, April 1969b.

Melvin R. Novick and Dorothy T. Thayer. A comparison of bayesian estimates of true score. Technical report, Educational Testing Service, Princeton, New Jersey, September 1969.

Melvin R. Novick, Paul H. Jackson, and Dorothy T. Thayer. Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika*, 36 (3):261 – 288, September 1971.

Miguel A. Padilla and Guili Zhang. Estimating internal consistency using bayesian methods. *Journal of Modern Applied Statistical Methods*, 10(1):277 – 286, 2011.

John T. Willse. *CTT: Classical Test Theory Functions*, 2018. URL `https://CRAN.R-project.org/package=CTT`. R package version 2.3.3.

Georg Rasch *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche, Oxford, England, 1960.

A Appendix

The purpose of this appendix is to explicitly show how the framework in the paper may be implemented for several common hierarchical models based on natural exponential families and their conjugate priors. The beta-binomial model is shown in the paper and is restated here for completion. This appendix adds details for the remaining natural exponential families with quadratic variance function excepting the natural exponential family created as a convolution of generalized hyperbolic secant densities, which is both unfamiliar to most and presents difficulties in sampling and estimation deserving of a full paper of its own.

The distributions of item scores Y_{ij}, test scores X_i, and abilities θ_i are explicitly stated. The density for X_i is given using a parameterization such that $E[X_i] = n_i\theta_i$. Because the use of the framework generally requires non-standard parameterization for the distribution of abilities θ_i, a "standard" form of the density or mass function is be given, followed by the transformation of the standard parameters in order to obtain μ and M and the density or mass function written in terms of μ and M. The marginal density of X_i given μ and M, integrating out over θ_i, is given. The log-likelihood for μ and M given observed test scores x_i is also stated. Though a full mathematical derivation is not shown, the most direct way to obtain $M = E[V(\theta_i)]/Var(\theta_i)$ for distributions with quadratic variance function is through manipulation of $Var(\theta_i) = E[\theta_i^2] - (E[\theta_i])^2$.

There are two immediate uses for the models in this appendix: first, to simulate unidimensional test data with equal item difficulty, not necessarily normally distributed, that has a desired population value of Cronbach's alpha or parallel test reliability (which are equal in this framework). Both are

$$\alpha = \rho = \frac{n_i}{M + n_i}$$

So long as n_i and M are chosen to satisfy the above equation, then by simulating θ_i values from the conjugate prior $g(\theta_i|\mu, M)$ and then summing n_i independent and identical draws y_{ij} from the corresponding NEF distribution of Y_i given mean θ_i for as many subjects as desired, the resulting simulated test data will have a population value of the target Cronbach's alpha. In general, a simulation can be conducted by

1. Select the desired α, the number of test items n_i, and the number of subjects K. The necessary value of M for the target value of α is $M = \left(\frac{1-\alpha}{\alpha}\right) n_i$.
2. Choose an appropriate mean μ for the prior distribution of θ_i. Simulate K abilities θ_i from the prior distribution. This will likely require conversion from μ and M to a more traditional parameterization.

3. For each ability θ_i, simulate n_i total observations from the distribution of Y_{ij}

For example, suppose that the beta-binomial is used to generate dichotomous test data with a target Cronbach's alpha of $\alpha = 0.75$ for $n_i = 15$ test items. The required value of M is $M = \left(\frac{1 - 0.75}{0.75}\right) 15 = 5$. Then choosing, for example, $\mu = 0.5$, the prior parameters of the beta distribution are $\alpha = 0.5(5) = 2.5$ and $\beta = 0.5(5) = 2.5$. A total of K abilities θ_i are simulated from a $Beta(2.5, 2.5)$ distribution. Then for each θ_i, a total of $n_i = 15$ draws from a $Bernoulli(\theta_i)$ are taken to generate the test data.

The second use is in estimation. The log-likelihood may be used directly for parametric empirical Bayes estimation, as described in the paper, or incorporated into an MCMC technique. Given a prior $h(\mu, M)$, a general MCMC iteration can be constructed as

1. Perform Metropolis-Hastings steps to draw new values μ^* and M^*, with log posterior equal to the log likelihood plus the log prior.

2. Perform a Gibbs step to draw from the posterior distributions for θ_i given x_i and draws μ^* and M^*. As $g(\theta_i|\mu, M)$ is the conjugate prior to $p(x_i|\theta_i, n_i)$, this will be a draw from the conjugate posterior.

For example, in the beta-binomial model with test scores x_i assumed to follow a binomial distribution, an MCMC might first draw new values μ^* and M^* using a Metropolis-Hastings step, convert these to $\alpha^* = \mu^* M^*$ and $\beta^* = (1 - \mu^*)M^*$, and perform a Gibbs step for each θ_i by simulating values θ_i^* from the beta posterior of $\theta_i|x_i$ for the beta prior with parameters α^* and β^*.

A.1 Normal - Normal

A.1.1 Model Formulation

The simplest and most common model assumes normality for the response of each question. Ironically, this model presents a difficult case, as the independence of the variance and mean introduces another parameter σ^2 which must be accounted for in the analysis. The univariate case of Section (3.1) assumes that this is known, which is not likely to occur in practice. It should be stated that if normality is assumed for both test scores and abilities, many of the traditional tools for classical test theory, which are often based on an assumption of normality, may be far simpler and easier to use than methods based on this framework.

The "generating" NEF distribution for the normal-normal model is the normal density with mean θ_i and variance σ^2.

$$
\begin{aligned}
Y_{ij}|\theta_i &\sim N(\theta_i, \sigma^2) \\
X_i|\theta_i &\sim N(n_i\theta_i, n_i\sigma^2) \\
\theta_i &\sim N(\mu, \tau^2)
\end{aligned}
$$

The normal distribution has variance function $V(\theta_i) = \sigma^2$, a constant which does not depend on θ_i. The expectation is thus $E[V(\theta_i)] = E[\sigma^2] = \sigma^2$. The variance of θ_i is $Var(\theta_i) = \tau^2$. The parameters μ and M are thus

$$
\begin{aligned}
\mu &= \mu \\
M &= \frac{\sigma^2}{\tau^2}
\end{aligned}
$$

The reliability in the original model is thus equivalent to the reliability in the model

$$
\begin{aligned}
Y_{ij}|\theta_i &\sim N(\theta_i, 1) \\
X_i|\theta_i &\sim N(n_i\theta_i, n_i) \\
\theta_i &\sim N\left(\mu, \frac{1}{M} = \frac{\tau^2}{\sigma^2}\right)
\end{aligned}
$$

Note that M is the precision, not the variance, of the distribution of abilities. The variance is $Var(\theta_i) = \frac{1}{M}$. For purposes other than calculating reliability, the above models are clearly not equivalent to each other.

If σ^2 were truly known, then transforming the test data as $y^*_{ij} = \frac{1}{\sigma}y_{ij}$ would induce the model with $\sigma^2 = 1$. As σ^2 is unknown, it must be estimated. There are many potential ways to do this. One method for equal sample sizes $n_i = n$ is given by

$$
\widehat{\sigma^2} = \left(\frac{1}{n}\sum_{j=1}^{n} Var(y_j)\right) - \left(\frac{1}{n(n-1)}\sum_{j_1 \neq j_2} Cov(y_{j_1}, y_{j_2})\right)
$$

The test data may then be transformed as $y^*_{ij} = \frac{1}{\hat{\sigma}}y_{ij}$ to fit the model with $Y_{ij}|\theta_i \sim N(\theta_i, 1)$

The density function for the normal distribution of X_i is then

$$
p(x_i|\theta_i, n_i) = \frac{1}{\sqrt{2\pi n_i}} e^{-\dfrac{(x_i - n_i\theta_i)^2}{2n_i}}
$$

The conjugate prior for the normal distribution with known variance is also a normal distribution. This prior has density.

$$g(\theta_i|\mu, M) = \frac{1}{\sqrt{2\pi\frac{1}{M}}} e^{-\frac{(\theta_i - \mu)^2}{2\frac{1}{M}}}$$

A.1.2 Marginal Density and Likelihood

The marginal density for this normal-normal model is a normal density with mean $n_i\mu$ and variance $\frac{n_i(M + n_i)}{M}$. This distribution has density

$$f(x_i|\mu, M, n_i) = \frac{1}{\sqrt{2\pi\frac{n_i(M + n_i)}{M}}} e^{-\frac{M(x_i - n_i\mu)^2}{2n_i(M + n_i)}}$$

For a sample of size K subjects, the log-likelihood is

$$\ell(\mu, M) = -\frac{1}{2}\left(\sum_{i=1}^{K}\left[\log\left(\frac{n_i(M + n_i)}{M}\right) + \frac{M(x_i - n_i\mu)^2}{n_i(n_i + M)}\right]\right)$$

A.2 Binomial-Beta

A.2.1 Model Formulation

The binomial-beta model may be used for dichotomous data, which is also common in testing. The “generating” NEF distribution for the beta-binomial model is the Bernoulli distribution with success probability θ_i.

$$Y_{ij}|\theta_i \sim Bernoulli(\theta_i)$$
$$X_i|\theta_i \sim Binomial(n_i, \theta_i)$$
$$\theta_i \sim Beta(\alpha, \beta)$$

The variance function for the Bernoulli distribution is $V(\theta_i) = \theta_i(1 - \theta_i) = \theta_i - \theta_i^2$.

The mass function for the binomial distribution is

$$p(x_i|\theta_i, n_i) = \binom{n_i}{x_i}\theta^{x_i}(1 - \theta_i)^{x_i}$$

The conjugate prior for the binomial distribution is the beta distribution. Traditionally, this is written with parameters α and β and density

$$g(\theta_i|\alpha, \beta) = \frac{\theta_i^{\alpha-1}(1 - \theta_i)^{\beta-1}}{\beta(\alpha, \beta)}$$

The transformation defined by $\mu = E[\theta_i]$ and $M = E[V(\theta_i)]/Var(\theta_i)$ is

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$M = \alpha + \beta$$

Equivalently, $\alpha = \mu M$ and $\beta = (1 - \mu)M$. This parameterization has density function

$$g(\theta_i|\mu, M) = \frac{\theta_i^{\mu M - 1}(1 - \theta_i)^{(1-\mu)M-1}}{\beta(\mu M, (1 - \mu)M)}$$

A.2.2 Marginal Density and Log-Likelihood

The marginal distribution of X_i given n_i, μ, and M is known as the beta-binomial distribution. This distribution has mass function

$$f(x_i|\mu, M) = \binom{n_i}{x_i}\frac{\beta(x_i + \mu M, n_i - x_i + (1 - \mu)M)}{\beta(\mu M, (1 - \mu)M)}$$

For a sample of size K subjects, the log-likelihood is

$$\ell(\mu, M) = \left(\sum_{i=1}^{K} \log[\beta(x_i + \mu M, n_i - x_i + (1 - \mu)M)]\right) - K\log[\beta(\mu M, (1 - \mu)M)]$$

A.3 Poisson-Gamma

A.3.1 Model Formulation

The Poisson-gamma model may be used for data which is a count of an event over a given area of measurement. Usually this is time, but not necessarily. The “generating” distribution for the Poisson-gamma model is the Poisson distribution with mean θ_i.

$$Y_{ij}|\theta_i \sim Poisson(\theta_i)$$
$$X_i|\theta_i \sim Poisson(n_i\theta_i)$$
$$\theta_i \sim Gamma(\alpha, \beta)$$

The variance function for the Poisson distribution is $V(\theta_i) = \theta_i$.

The mass function for the Poisson distribution of X_i is

$$p(x_i|\theta_i, n_i) = \frac{e^{-n_i\theta_i}(n_i\theta_i)^{x_i}}{x_i!}$$

The conjugate prior for the Poisson distribution is the gamma distribution. Traditionally, this is written with parameters α and β (such that $E[\theta_i] = \alpha/\beta$) and density

$$g(\theta_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta_i^{\alpha-1}e^{-\beta\theta_i}$$

The transformation defined by $\mu = E[\theta_i]$ and $M = E[V(\theta_i)]/Var(\theta_i)$ is

$$\mu = \frac{\alpha}{\beta}$$
$$M = \beta$$

Equivalently, $\alpha = \mu M$ and $\beta = M$. This parameterization has density function

$$g(\theta_i|\mu, M) = \frac{M^{\mu M}}{\Gamma(\mu M)}\theta_i^{\mu M-1}e^{-M\theta_i}$$

A.3.2 Marginal Density and Log-Likelihood

The marginal distribution of the Poisson-gamma model is the negative binomial distribution, with mass function

$$f(x_i|n_i, \mu, M) = \frac{M^{\mu M}n_i^{x_i}}{\Gamma(\mu M)x_i!}\frac{\Gamma(x_i + \mu M)}{(n_i + M)^{x_i+\mu M}}$$

For a sample of size K subjects, the log-likelihood is

$$\ell(\mu, M) = K\mu M\log(M) - K\log(\Gamma(\mu M)) + \sum_{i=1}^{K}[\log(\Gamma(x_i + \mu M)) - (x_i + \mu M)\log(n_i + M)]$$

A.4 Gamma - Inverse Gamma

A.4.1 Model Formulation

The gamma - inverse gamma model may be used for test data which measures time between events, response times, or right-skewed data in general. The “generating” NEF distribution for the gamma - inverse gamma model is the exponential density with mean θ_i.

$$Y_{ij}|\theta_i \sim Exponential(\theta_i)$$
$$X_i|\theta_i \sim Gamma(n_i, \theta_i)$$
$$\theta_i \sim InverseGamma(\alpha, \beta)$$

The variance function for the exponential distribution is $V(\theta_i) = \theta_i^2$.

The distribution of the X_i is a Gamma distribution with integer shape parameter n_i, sometimes called the Erlang distribution. This distribution has density

$$p(x_i|\alpha, \beta) = \frac{1}{\theta_i^{n_i}(n_i - 1)!} x_i^{n_i - 1} e^{-\frac{x_i}{\theta_i}}$$

The conjugate prior for the gamma density written in this form is the inverse gamma distribution. Traditionally, this is written with parameters α and β (such that $E[\theta_i] = \beta/(\alpha - 1)$) and density

$$g(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{-\alpha-1} e^{-\frac{\beta}{\theta_i}}$$

The transformation defined by $\mu = E[\theta_i]$ and $M = E[V(\theta_i)]/Var(\theta_i)$ is

$$\mu = \frac{\beta}{\alpha - 1}$$
$$M = \alpha - 1$$

Equivalently, $\alpha = M + 1$ and $\beta = \mu M$. This parameterization has density function

$$g(\theta_i|\mu, M) = \frac{(\mu M)^{M+1}}{\Gamma(M + 1)} \theta_i^{-(M+1)-1} e^{-\dfrac{\mu M}{\theta_i}}$$

A.4.2 Marginal Density and Log-Likelihood

The marginal distribution of the gamma-inverse gamma model is an F distribution with density

$$f(x_i|n_i, \mu, M) = \frac{(\mu M)^{M+1} x_i^{n_i-1} \Gamma(M + n_i + 1)}{(\mu M + x_i)^{M+n_i+1} (n_i - 1)! \Gamma(M + 1)}$$

For a sample of size K subjects, the log-likelihood is

$$\ell(\mu, M) = K(M+1) \log(\mu M) - K \log(\Gamma(M+1)) + \left(\sum_{i=1}^{K} [\log(\Gamma(M + n_i + 1)) - (M + n_i + 1) \log(\mu M + x_i)] \right)$$

A.5 Negative Binomial - F

A.5.1 Model Formulation

The negative binomial distribution may be used for test data which measures the number of non-events before an event occurs. The "generating" distribution for the Negative Binomial - F model is the geometric distribution with expectation θ_i.

$$\begin{aligned} Y_{ij}|\theta_i &\sim Geometric(\theta_i) \\ X_i|\theta_i &\sim NegativeBinomial(n_i, \theta_i) \\ \theta_i &\sim F(d_1, d_2) \end{aligned}$$

The support of Y_i is $y_i \in \{0, 1, 2, \ldots\}$. The mass function of this is traditionally written with probability of the event p_i as

$$p(y_i|p_i) = (1 - p_i)^{y_i} p_i$$

The use of this framework requires an unusual parameterization of this distribution. Define $\theta_i = E[Y_i] = (1 - p_i)/p_i$ so that $p_i = 1/(1 + \theta_i)$. Note that while p_i is constrained to between 0 and 1, θ_i may take any positive real number. The density in terms of θ_i is then

$$p(y_i|\theta_i) = \frac{\theta_i^{y_i}}{(1 + \theta_i)^{y_i+1}}$$

The variance function for the geometric distribution in this parameterization is $V(\theta_i) = \theta_i + \theta_i^2$.

The distribution of the X_i is a negative binomial distribution. It can be thought of as the sum of the number of times a non-event happened before an event occurred over n_i trials. As each individual trial has support $y_i \in \{0, 1, 2, \ldots\}$, the support of x_i is also $x_i \in \{0, 1, 2, \ldots\}$ even though the number of trials n_i is finite. This distribution has density

$$p(x_i|\theta_i, n_i) = \binom{x_i + n_i - 1}{x_i} \frac{\theta_i^{x_i}}{(1 + \theta_i)^{n_i+x_i}}$$

This density, per the framework, has mean $E[X_i|\theta_i] = n_i\theta_i$ and variance $Var(X_i|\theta_i) = n_i(\theta_i + \theta_i^2)$. The conjugate prior for the negative binomial density is the F distribution. Traditionally, this is written with parameters d_1 and d_2 (with $d_2 > 4$ so that the variance is finite), representing degrees of freedom, and density function

$$g(\theta_i|d_1, d_2) = \frac{(\frac{d_1}{d_2})^{\frac{d_1}{2}}}{\beta(\frac{d_1}{2}, \frac{d_2}{2})} \theta_i^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\theta_i\right)^{-\frac{1}{2}d_1+d_2}$$

Thus far, the transformations to μ and M have yielded "nice" densities which are not extremely different from the common parameterizations. The F conjugate prior shows that this does not necessarily have to be the case. The transformation defined by $\mu = E[\theta_i]$ and $M = E[V(\theta_i)]/Var(\theta_i)$ is

$$\mu = \frac{d_2}{d_2 - 2}$$
$$M = \frac{(d_2 - 2)(d_1(d_2 - 2) + d_2)}{d_2(d_1 + d_2 - 2)}$$

Equivalently, this yields the transformations

$$d_1 = -\frac{2\mu(M-1)}{\mu(\mu-1)M-2}$$
$$d_2 = \frac{2\mu}{\mu-1}$$

Note that there are constraints on the potential values of μ and M which yield $d_1 > 0$ and $d_2 > 4$. In general, μ will necessarily be close to 1 for larger values of M.

This parameterization has density function

$$g(\theta_i|\mu, M) = \frac{\left(-\frac{(\mu-1)(M-1)}{\mu(\mu-1)M-2}\right)^{-\frac{\mu(M-1)}{\mu(\mu-1)M-2}}}{\beta\left(-\frac{\mu(M-1)}{\mu(\mu-1)M-2}, \frac{\mu}{\mu-1}\right)} \times \theta_i^{-\frac{\mu(M-1)}{\mu(\mu-1)M-2}}$$
$$\times \left(1 + \left(-\frac{(\mu-1)(M-1)}{\mu(\mu-1)M-2}\right)\theta_i\right)^{\frac{(\mu-1)(M-1)}{\mu(\mu-1)M-2} - \frac{\mu}{\mu-1}}$$

A.5.2 Marginal Density and Log-Likelihood

The marginal distribution of the negative binomial-F is Polya type II distribution. Because the parameterization in μ and M offers no advantages and substantially increases complexity, the marginal density is written in terms of d_1 and d_2. If optimization in terms of μ and M is required, conversion between the two parameterizations can be performed within the iterative optimization routine.

$$f(x_i|n_i,\mu,M) = \binom{x_i+n_i-1}{x_i}\frac{(d_1/d_2)^{d_1/d_2}}{\beta(\frac{d_1}{2},\frac{d_1}{2})}\int_0^\infty \frac{\theta^{x_i+\frac{d_1}{2}-1}}{(1+\theta)^{x_i+n_i}(1+\frac{d_1}{d_2}\theta)^{\frac{1}{2}(d_1+d_2)}}d\theta$$

For a sample of size K subjects, the log-likelihood is

$$\begin{aligned}\ell(d_1,d_2) = K\frac{d_1}{d_2}\left[\log(d_1)-\log(d_2)\right] - K\log\left[\Gamma\left(\frac{1}{2}d_1\right)\right] - K\log\left[\Gamma\left(\frac{1}{2}d_2\right)\right] \\ + K\log\left[\Gamma\left(\frac{1}{2}(d_1+d_2)\right)\right] + \sum_{i=0}^{K}\log\left[\int_0^\infty \frac{\theta^{x_i+\frac{d_1}{2}-1}}{(1+\theta)^{x_i+n_i}(1+\frac{d_1}{d_2}\theta)^{\frac{1}{2}(d_1+d_2)}}d\theta\right]\end{aligned}$$

The integral in the marginal likelihood is not analytically tractable in general, and but may be approximated through any number of numerical integration techniques.