JOURNAL OF APPLIED MEASUREMENT, *18*(4),

Approximate Functional Relationship between IRT and CTT Item Discrimination Indices: A Simulation, Validation, and Practical Extension of Lord's (1980) Formula

John T. Kulas
Corporate Mr. Fixit

Jeffrey A. Smith
State Farm Mutual Automobile Insurance Company

Hui Xu
U.S. Bancorp

Lord (1980) presented a purely conceptual equation to approximate the nonlinear functional relationship between classical test theory (CTT; aka true score theory) and item response theory (IRT) item discrimination indices. The current project proposes a modification to his equation that makes it useful in practice. The suggested modification acknowledges the more common contemporary CTT discrimination index of a corrected item-total correlation and incorporates item difficulty. We simulated slightly over 768 trillion individual item responses to uncover a best-fitting empirical function relating the IRT and CTT discrimination indices. To evaluate the effectiveness of the function, we applied it to real-world test data from 16 workforce and educational tests. Our modification results in shifted functional asymptotes, slopes, and points of inflection across item difficulties. Validation with the workforce and educational tests suggests good prediction under common assumption testing conditions (approximately normal distribution of abilities and moderate item difficulties) and greater precision than Lord's (1980) formula.

Requests for reprints should be sent to

The approximate functional relationship between classical test theory (CTT) and item response theory (IRT) estimates of item discrimination as described by Lord (1980):

$$a_i \cong \frac{r_i}{\sqrt{1-r_i^2}}\ , \tag{1}$$

$$r_i \cong \frac{a_i}{\sqrt{1+a_i^2}}, \tag{2}$$

was not intended as an accurate or practical description of the relationship between IRT parameters and CTT statistics. Rather it was presented as a rough approximation of a general relationship that may assist the conceptual comprehension of the IRT discrimination parameter for assessment specialists who were more familiar with CTT procedures and item indices. Lord (1980) states very clearly that the equations, "…are given here not for practical use" (p. 33). Yet the formulas (or more generally the association between *a* and *r*) are being applied in practice, and have appeared in several research articles (see, for example, Dawber, Rogers, and Carbonaro, 2009; Hambleton and Jones, 1993; Weitzman, 2009).

The formulas have also been notably absent from other investigations of IRT parameters and CTT indices. MacDonald and Paunonen (2002), for example, cite strong associations between IRT and CTT estimates of ability and difficulty, but occasionally poor (linear) associations between IRT and CTT estimates of discrimination. In fact, the authors state that, "With the exception of a test containing items with both a wide range of discrimination values and a narrow range of difficulty values, any expectations of high item discrimination comparability between IRT and CTT may be unfounded" (p. 932). We believe the authors came to this conclusion in part due to a reliance upon the Pearson's product-moment correlation as an index of relationship (e.g., the true form of the relationship is exponential, as specified by Lord [1980]; see Figure 1).

Our primary interest in conducting the present studies lies in the CTT asymptote of Lord's function (which occurs at a value of 1.0). Although Lord's CTT discrimination index was the item-test biserial correlation, the more common contemporary index of discrimination is

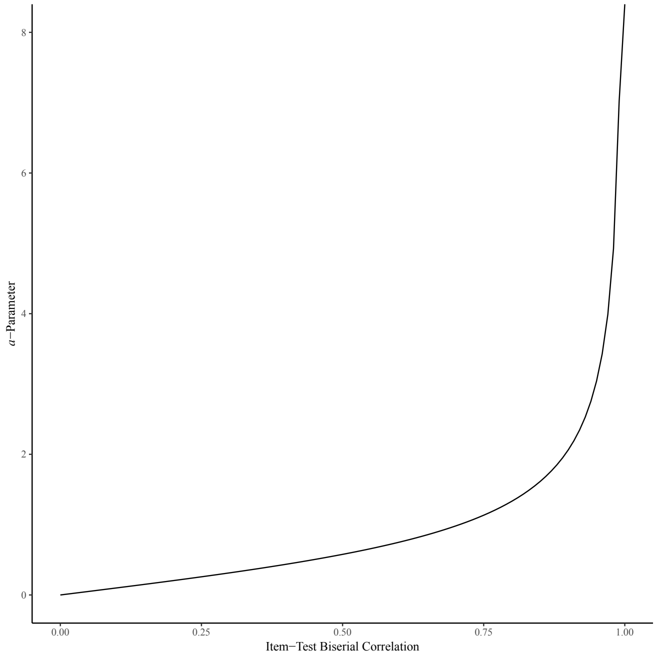

Figure 1. Lord's (1980) functional relationship between the IRT *a*-parameter and CTT item-test biserial correlation.

the corrected item-total point-biserial correlation (explained further below). In practical testing applications, this corrected item total correlation does not approach such an extreme asymptote (regardless of test or sample characteristics). Our goals, then, were to: 1) determine the extent to which Lord's function provides a reasonable (shape) approximation of the IRT-CTT relationship via simulated data, 2) provide an adjustment (location) to this function if the empirical association evidences a lower vertical asymptote, and 3) evaluate the predictive utility of the revised and original functions via direct application to suitable real-world test data (test data resulting from judicious person and item sampling procedures).

Classical Test Theory Discrimination Indices

Item discrimination (regardless of psychometric lens) refers to the extent to which a given item's response provides informative differentiation across individuals who possess varying levels of a target construct. Within the scope of CTT (and knowledge or ability *testing*) discrimination is most commonly estimated via a determination of the association between respondents' total test score and binary actions (correct [1] or incorrect [0] response) on a focal item. Particularly with shorter tests the inclusion of any individual item response into computation of the total test score will influence the item-total association and it is for this reason that an adjustment is commonly made to the total score—removing the item's influence via computing separate *corrected* total scores for each focal item.

The *correlation* between a dichotomous and (effectively) continuous variable—regardless of whether the continuous variable has a correction or not—is commonly estimated via one of two alternative indices: the biserial or point-biserial correlation coefficient. These terms are not typically encountered in contemporary psychometric works but their technical differences are important to note. The point-biserial correlation can be estimated via simple application of Pearson's product-moment formula to a situation where one variable is dichotomous and the other continuous. The common corrected-item total correlation—calculated via application of Pearson's product-moment correlation - can therefore be classified as a point-biserial correlation.

The point-biserial correlation assumes the categorical variable represents a true dichotomy (Brogden, 1949; Lord, 1963; Ludlow, 2001). The biserial correlation, in contrast, treats the dichotomy as contrived. Kemery et al. (1989) present the distinction nicely, with proper *choice of index* dependent on whether the binary variable represents "either a true dichotomy or a dichotomized representation of an underlying continuous variable" (p. 418). The biserial coefficient provides an upward adjustment to the point-biserial—this adjustment is smallest when category membership is evenly split (50% of cases occur in one category and 50% occur in the complement). The biserial coefficient is therefore always greater than the (absolute) magnitude of the point-biserial (unless the point-biserial value is zero). The point-biserial is observed whereas the biserial is estimated.

The CTT index referenced in Lord's (1980) formula is the item-test biserial correlation whereas the common contemporary discrimination index is the item-corrected total point biserial correlation. Lord's (1980) specification therefore deviates from the currently dominant CTT discrimination index in at least two important ways. One, the correction of focal-item deletion from the total test score is not specified (e.g., his "total" is a true total test score). Two, Lord's CTT index of association yields relatively greater magnitude values than does the typical contemporary index. The corrected item total (point-biserial) correlation is the more common CTT index of discrimination that is used in contemporary psychometric applications, and it is *primarily* for this reason that it is the CTT index we retain and incorporate into our suggested revisions to Lord's (1980) formula.[1]

[1] The (debatable) characterization of correct versus incorrect item response as a true dichotomy and (undebatable) possibility of out-of-range estimated values for the biserial coefficient are additional considerations here. Also note that although our focus is the corrected item-total correlation, the biserial coefficient is incorporated as a secondary index into the Study 2 method and further acknowledged in the discussion.

Item Response Theory—A Brief Overview

For binary responses, item response functions (IRF's) specifying the form of relationship between person parameter (θ), item parameters (*b* [difficulty/extremity], *a* [discrimination], and *c* [pseudo-guessing]) and the probability of a (1) response may be expressed in either normal or logistic form. The function describing the normal ogive IRF requires integral specification:

$$P(\theta) = c + (1-c)\int_{-\infty}^{a(\theta-b)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt, \qquad (3)$$

Whereas the logistic function is mathematically elegant:

$$P(\theta) = c + (1-c)\frac{1}{1+e^{-1.7a(\theta-b)}}, \qquad (4)$$

The logistic model described in equation 2 is preferred in practical work, and our current study focus is placed on (one variant of) this 3 parameter logistic (3PL) model as well. Both the normal and logistic three-parameter models presented in equations 1 and 2 state that the probability of an individual, *j*, responding "correctly" to an item, *i*, is only attributable to four things: the three item parameters and person ability (θ). Knowing ability level as well as item characteristics is sufficient for estimating probability of correct item response—knowing something about how the examinee performed on other items is uninformative (this constitutes the assumption of *local independence*). Substituting a value of zero for the *c*-parameter in equation 2 yields the two parameter model (2PL); additionally constraining *a*-parameter estimates to a value of one yields the 1PL.

Item response functions of different location, slope, and lower asymptote can be specified by substituting different values for *b, a,* and *c* in either of the above formulas. The 3PL was applied in this capacity to plot the functions presented in Figure 2. Here, for example, one IRF is defined by *b* = .5, *a* = 1, and *c* = 0 (solid line function—this slope and intercept are equivalent to fixed parameter values imposed in a 1PL specification). The freed-parameter *estimation* of item discrimination can be seen visually via slope differentiation in the dashed line (*b* = 0, *a* = .5, *c* = .15) and dotted line (*b* = –1, *a* = 2, *c* = 0) functions. With the 1PL

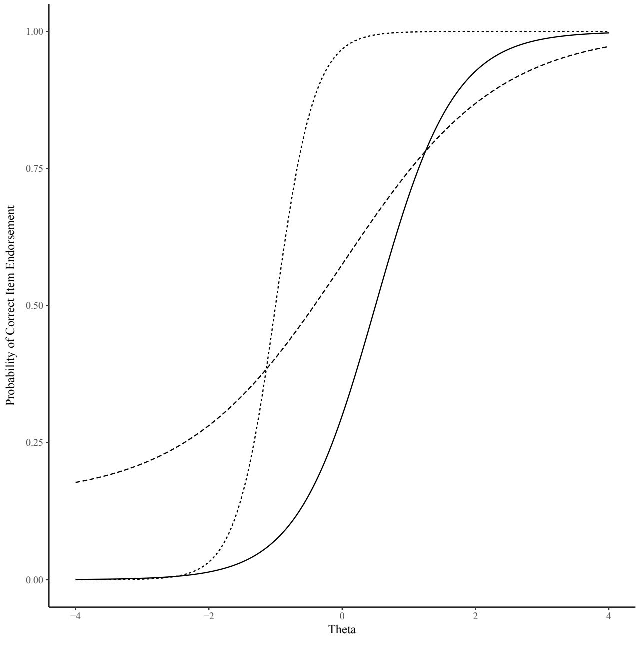

Figure 2. Example item response functions (IRFs).

(*b* freed, *a* and *c* fixed), discrimination as well as pseudo-guessing are constrained to be equal across items. The 2PL (*b* and *a* freed, *c* fixed) allows for direct estimation of the function slope (reflecting potentially differential item discrimination). The full 3PL allows for estimation of differential slope variation but also accommodates the lower asymptote (*c*-parameter; e.g., to be non-zero—the dashed-line function reflects such a specification).

The multitude of IRFs that can be specified via equations 1 or 2 accommodate a wide range of possibilities regarding the nature of the relationship between underlying construct standing and probability of correct response to an item (e.g., the formulas accommodate many different functional forms—it is reasonable to assume that one such function may provide a good fit to a given pattern of test data). In addition to *c*- and/or *a*-parameter constrainment (e.g., 2PL, 1PL), these three item parameter models of assessment response can also be expanded. Prentice (1976), for example, modeled skew and other item characteristics, with Lord also stating that a 5-parameter model (incorporating some of Prentice's parameters) is a fairly straightforward extension of the three item parameter models. The focus of our investigation lies in the *a*-parameter, which reflects an item's discrimination (IRF slope at the point of inflection), and our IRT model of focus therefore held the *c*-parameter constant (at a value of zero)—this is consistent with Lord's original set of restrictions.

Comparison of IRT and CTT Item Indices

The concept of invariance undoubtably represents the primary advocated psychometric advantage of a focus on IRT parameters over CTT item statistics. It also represents the primary obstacle to and ceiling for the utility of any function relating CTT and IRT item indices (because of the CTT index sample-dependency/lack of invariance [LOI]). The invariance property, within the context of the 2PL, refers to stability of difficulty and discrimination indices regardless of the nature of the underlying ability/trait distribution.[2] This means that the same item parameters will be obtained regardless of the distribution of theta within a population (note however that procedures are required to place parameters on equivalent scales—these linear transformation scaling procedures are fully algebraic and certainly not as elaborate as full equating *methodologies* that are generally pursued in CTT applications). It is important to note that IRT parameter invariance is only achieved if the chosen IRT model exactly fits population data (see, for example, Hambleton, Swaminathan, and Rogers, 1991).

Although parameter invariance is an ideal categorical population property, with "degrees" of invariance being a technically inappropriate conceptualization, sample (parameter *estimate*) evidence of the *degree to which a LOI is likely* has, with some controversy, been conceptualized as a graded continuum (see, for example, Hambleton et al., 1991, pp. 23-24; Rupp and Zumbo, 2004, 2006). When viewed thusly in degree, sample estimates of IRT parameters across different calibration runs can be investigated for evidence of a possible lack of population parameter invariance. There is no such debate with CTT item indices: they will always be sample-dependent. However (within this general consideration of index "invariance"), sampling strategy can greatly limit the impact of this dependency. We therefore propose that the general approach taken in large-scale high-stakes testing is in fact likely to result in comparable CTT item and person statistics across testing populations and occasions (e.g., moderate to high levels of index stability). With this orientation in mind, we note that an adjustment to Lord's (1980) functional relationship between the "non-invariant" CTT statistic and "invariant"

[2] Invariance also refers to stability of theta scores regardless of administered test. In this presentation of invariance we take careful note of an important distinction between the population parameter property of invariance and the common sample use of the term (applied to parameter *estimates*) that implies coefficient stability. To acknowledge this important distinction, when the common sample use of the term is referenced within the current paper, quotations are used around the term "invariance."

IRT parameter may prove useful beyond mere conceptual application.

Consistent with the above perspective, several empirical investigations note high levels of "invariance" of CTT estimates (e.g., MacDonald and Paunonen, 2002). The cross-sample stability of CTT item estimates can, in fact, surpass that of estimated IRT parameters (see, for example, Fan, 1998 who documented consistently higher levels of index "invariance" for CTT vs. IRT difficulty estimates; Fan states in summary that the IRT and CTT "…frameworks produced very similar item and person statistics…in terms of the degree of invariance" [p. 379]). The primary obstacles to CTT discrimination index stability are sampling bias (particularly resulting in sample homogeneity) and item difficulty (which can attenuate discrimination estimates). We therefore acknowledge here that true parameter invariance is a categorical IRT property, but pursue our functional modification noting that careful item and person sampling can also at least partially mitigate the most severe threats to CTT index "invariance" (e.g., high degree of index similarity obtained across samples).

Purpose and Summary

This manuscript investigates the extent to which Lord's (1980) formula provides a reasonable representation of the relationship between IRT and CTT indices of discrimination. This has been investigated by others (e.g., Dawber, et al., 2009) but with a focus on *robustness* of Lord's (1980) discrimination and difficulty relationships across assumptions violations. The primary reason for our exploration of this relationship is a bit different and primarily concerns the liberal item-test asymptote of Lord's original formula. Lord's (1980) function is characterized by a vertical asymptote at a biserial (item-total) correlation of 1.0 (see Figure 1). This asymptote is thought to be liberal particularly because contemporary CTT discrimination estimates do not approach such an extreme value in practical test applications, where the more common point-biserial discrimination index also applies a correction to the test estimate (by omitting the influence of the investigated item in the test score definition; the so-called *corrected* item-total correlation). Our question therefore is one of practical versus conceptual application of this functional specification: is there a practical function that provides a more accurate relationship between the IRT *a*-parameter and the most common contemporary CTT (corrected item-total) discrimination index?

There are many good references (conceptual, theoretical, and empirical) relating CTT concepts, procedures, and indices to IRT applications. We do not fully address these comparisons in the current manuscript beyond our brief acknowledgement of the concept of "invariance," but rather direct the interested reader to these sources (such as Bechger, Maris, Verstralen, and Béguin [2003], Fan [1998], and Reise and Henson [2003]). It is important here to acknowledge that there will never be an absolute association between indices from the different psychometric models because the models and underlying theories differ in important qualitative ways. The focus of the current paper is therefore placed simply on a contemporary refinement of the "crude" (p. 33) discrimination indices formula presented in Lord (1980).

Study 1—Simulation and Function Specification

Materials and Methods

Study 1's purpose was to modify Lord's (1980) formula:

$$a_i \cong \frac{r_i}{\sqrt{1-r_i^2}}\ , \tag{5}$$

in such a manner that the average (actual- minus predicted-*a*) residual would be minimized. Our simulations therefore had a least-squares orientation. We generated several datasets, eventually resulting in the retention of two revised formulas based on three idealized functions at constant *b*-parameter values of 0, |±1|, and |±2|. Each function was defined by either CTT or IRT modifying coefficients for the variable *r* and the exponential ratio:

$$\left(\frac{e^r - e^{-r}}{e - e^r}\right).$$

To integrate these functions we fit quadratic regressions to the three separate sets of coefficients—our revised functions could therefore be characterized as smoothed estimates.

Procedure and Materials

The goal of the simulations was to obtain data of controlled characteristics that could be considered a reasonable facsimile of knowledge/ability test information. The initial simulations altered: 1) number of test items (k), 2) a-parameter mean, and 3) a-parameter standard deviation. k-specifications began at $k = 4$ and increased by 2 item increments until $k = 50$. a-parameter means started at $a = 0.5$ and increased by 0.5 unit increments until $a = 7.0$. Such a large a-parameter would not typically be found in testing applications, but we extended our simulations to this high magnitude value in order to gain item indices across as wide a range of values as possible (e.g., to investigate the nature of the relationship across many [extreme and common] values along the specified equation-defined function). a-parameter standard deviations ranged from $SD = 0.5$ to $SD = 1$. Parameters which were initially held constant in each simulation include: 1) n-size (10,000 examinees), 2) b-parameter (mean = 0, standard deviation = 0.25), and 3) an approximately normal (mean = 0, standard deviation = 1) theta distribution.

To run these simulations, we built a series of nested loops via custom Ruby script and permuted the variable data specifications (within k, a-parameter mean, and a-parameter standard deviation) to create syntax and cue files for WinGen 3.0 (Han, 2007).[3] WinGen read the cue files to create binary (correct or incorrect) datafiles given our desired k, a, b, Θ, and n specifications. We constructed Ruby code to compute corrected item-total correlations within each of these datafiles and this information was then merged back with corresponding WinGen-estimated a- and b-parameters. In total, we performed 12,600 simulations, with 6,096,384 items and 126,000,000 examinees.

Results

Upon initial inspection of our simulations (via a-r scatter plots), it was evident that Lord's (1980) functional shape was present (with the anticipated shifted asymptote), but there was also substantial skew to distributions of corrected item-total correlations within each set of a-parameter values. Targeted constrainment of simulation variables revealed that this scatter could be largely accounted for by b-parameter magnitude. Simulations that provided the most clarity regarding the nature of the a-r relationship therefore constrained b standard deviations to a value of zero (while varying the other previously specified characteristics). The strong relationships yielded by this constraint did begin to weaken between b's of $|2|$ and $|3|$ and our resulting *simulation*-generated formulas are therefore only recommended for b's between – 2 and +2 (or corresponding p-values between .1 and .9 [discussed further below]).

The exponential form

$$\frac{e^r - e^{-r}}{e - e^r}$$

fit the desired general function properties of the a and r relationship, but the simulations identified systematic slope and inflection differences across items with differing b values. We therefore fit three different sets of modifiers of r and

$$\frac{e^r - e^{-r}}{e - e^r}$$

for b specifications of 0, 1, and 2 (we did not *directly* model any negative b's when holding the parameter constant). Assuming a smoothed quadratic relationship between these parameter specifications (e.g., different values of b—for instance an intermediate value of 1.2) allowed the estimation of an inelegant but residual-minimizing function:

$$\hat{a} \cong \left[\left(0.51 + 0.04|b| + 0.11b^2\right)r\right] + \left[\left(0.57 + 0.01|b| + 0.07b^2\right)\frac{e^r - e^{-r}}{e - e^r}\right]. \quad (6)$$

[3] Information regarding the use of cue files can be found in Han and Hambleton (2007; p. 43). More detailed information regarding the specifics of our simulations is available from the authors upon request.

The above formula can be used in situations where b has been estimated (e.g., the 1PL). Because there may also be interest in $\hat{a}$ estimates in situations where b has not been estimated, we generated a formula based on an index reflective of p-values (generally, the results of b and p-value investigations support strong associations between the two indices; see, for example, Fan, 1998; MacDonald and Paunonen, 2002; our current Study 2 samples also returned a b-p-value correlation of – .93). In order to scale the p-values in a manner similar to b's (e.g., interval-level), we specified an absolute normalized deviation from a focal p–q ratio (for notational purposes we refer to this index as z_g).

The calculation of this index is straightforward: let $g = |p - 0.5|$. This "g" is simply the absolute deviation from 50% answering an item correctly and 50% answering an item incorrectly. The normalization of this index is accomplished via identifying the standard normal deviation score (e.g., z) associated with g. For example, referencing the standard normal distribution will reveal that a g of zero corresponds to a z_g of zero, a g of .25 corresponds to a z_g of .674, and a g of .4 corresponds to a z_g of 1.282 (these z_g's are in fact the values we specified to generate our equation—these g's are CTT estimated approximations for b's of 0, |±1|, and |±2|):

$$\hat{a} \cong \left[\left(0.51 + 0.02z_g + 0.301z_g{}^2\right)r\right] + \left[\left(0.57 - 0.009z_g + 0.19z_g{}^2\right)\frac{e^r - e^{-r}}{e - e^r}\right]. \quad (7)$$

The *visual* result of either re-specification (z_g or b) can be seen in Figure 3, where both Lord's (1980) formula and the re-specified formulas are presented (at b values of 0, |±1|, and |±2| [or corresponding z_g values of 0, .674, and 1.282]). Note that: 1) our predicted a parameter is scaled to the normal metric, and 2) Lord's function assumes a biserial coefficient abscissa.

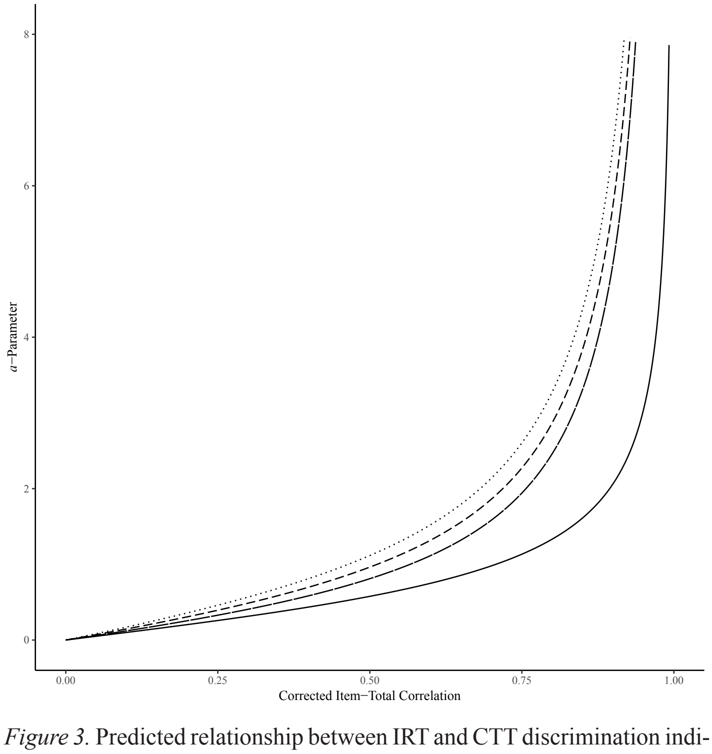

Figure 3. Predicted relationship between IRT and CTT discrimination indices—Lord's (1980) formula and the current proposed formulas (at $b = 0/ z_g = 0$ [long dash], $b = |\pm1|/ z_g = .674$ [dashed], and $b = |\pm2|/ z_g = 1.282$ [dotted]).

Study 2—Revised Function Validation

Materials and Methods

We applied both the revised and original formulas to real-world test data from two large testing organizations. The practical utility of any such function is dependent on acceptable levels of index "invariance"—from both IRT and CTT calibrations. Here we note that although CTT item indices are unquestionably sample dependent, with large-scale high-stakes standardized tests (of the kind most likely to also be estimated via IRT), we consider this dependency to be of less concern, as heterogeneously representative respondent samples are the norm rather than exception. We also note that parameter invariance in the IRT application is not a given, either, and is only achieved in situations with perfect model-data fit. Again we note here that the threats to "invariance" across both applications are at least partially mitigated by careful person and item sampling (of the type commonly realized in high-stakes testing applications).

Procedure, Materials, and Participants

Test data was obtained from one large national workforce assessment organization and one regional educational testing organization (see Table 1 for test characteristics). The two proctored workforce readiness assessments were administered to high-school students enrolled in career-oriented programs, unemployed individuals using job services centers, and job-seekers applying for positions in client organizations that require test results for personnel selection. One test measures work-embedded reading and comprehension skills and one assesses mathematical reasoning. Each test administration is capped at 45 minutes, although accommodations are made for individuals who request extended time. The educational assessments were administered in the spring of 2007 to students in public school grades 3 through 11 and are used for accountability purposes regarding the United States' Elementary and Secondary Education Act of 1965. Students are given two hours to complete these assessments.

Results

All IRT parameters were estimated via marginal maximum likelihood with BILOG-MG 3.0 (Zimowski et al., 2003). Fit of the 2PL to all test data was estimated via chi-square singles, doubles, and triples (see, for example, Drasgow et al., 1995). These fit indices reflect levels of convergence between expected (IRT calibrated) and observed (hold-out validation sample) response probabilities. We computed chi-squares via MODFIT (Stark, 2001) by simple random sampling within each of our test datasets to create hold-out validation samples (each of n = 3,000). Chi-square/df ratios did support the 2PL fitting the data across tests:

(singles $\overline{X^2/df} = 1.21, SD = 1.77,$

doubles $\overline{X^2/df} = 1.61, SD = 1.60,$

triples $\overline{X^2/df} = 1.79, SD = 1.29$).

Across our comparisons, 11.2% of singles, 14.2% of doubles, and 12.6% of triples had 2PL chi-square to degree of freedom ratios greater than three.

Table 2 presents bivariate associations between item indices. Note in this table that 1PL fit was estimated in addition to 2PL fit because one eventual end-purpose of our formulas could be a rough $\hat{a}$ approximation for assessment specialists who have 1PL and corrected item-total information but no 2PL specification (e.g., application of equation #3). The a-parameter—corrected item total association was quite high across our items (r = .79), although note that this *overall* index reported in Table 2 also reflects differences across tests (e.g., variability in average item discrimination across the 16 tests would result in a larger magnitude Table 2 correlation). Note also that the majority of investigated items can be represented in the 'horizontally flat' region of the Figure 3 function(s)—prior to function inflection toward a vertical asymptote.

Residuals

Our residuals are the differences between the formula-predicted a-parameter value and the BILOG calibrated a-parameter value (residual =

Table 1

Study 2 Test Characteristics

						a-Parameter	*b*-Parameter	Corrected Item-Total (*r*)	*p*-value	Residual (*b*-generated)	Residual (z_g-generated)
Test Name			*n*	*k*	α	2PL Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
Work Force		Reading	11,187	33	.85	.823 (.373)	–0.298 (1.55)	.355 (.106)	.595 (.263)	.055 (.167)	.065 (.087)
		Math	11,739	33	.82	.664 (.295)	–0.814 (1.13)	.328 (.062)	.664 (.204)	.156 (.286)	.163 (.196)
Educational - ESEA	Math	Grade 3	53,353	42	.90	.783 (.236)	–1.384 (.819)	.402 (.091)	.768 (.126)	–.046 (.109)	–.015 (.045)
		Grade 4	53,782	41	.89	.687 (.189)	–1.351 (.871)	.375 (.088)	.752 (.131)	–.074 (.089)	–.021 (.034)
		Grade 5	54,207	41	.89	.713 (.234)	–1.210 (.838)	.386 (.093)	.735 (.128)	–.046 (.092)	–.009 (.033)
		Grade 6	55,720	40	.89	.675 (.203)	–1.139 (.846)	.381 (.089)	.712 (.112)	–.040 (.098)	.008 (.027)
		Grade 7	57,991	39	.89	.734 (.261)	–0.837 (.792)	.395 (.104)	.677 (.144)	.010 (.081)	.031 (.030)
		Grade 8	59,682	40	.90	.731 (.210)	–0.812 (.617)	.404 (.065)	.675 (.132)	.016 (.071)	.027 (.035)
		Grade 11	61,810	39	.88	.657 (.257)	–0.403 (1.03)	.378 (.104)	.605 (.156)	–.002 (.078)	.032 (.040)
	Reading	Grade 3	57,740	38	.91	.877 (.298)	–1.081 (.525)	.445 (.073)	.748 (.128)	–.010 (.087)	–.034 (.057)
		Grade 4	57,892	40	.90	.806 (.283)	–1.010 (.566)	.425 (.090)	.720 (.125)	–.007 (.079)	–.013 (.035)
		Grade 5	57,899	41	.90	.792 (.252)	–1.189 (.672)	.411 (.087)	.751 (.134)	–.033 (.082)	–.031 (.052)
		Grade 6	59,272	42	.89	.771 (.353)	–1.089 (.670)	.390 (.088)	.729 (.143)	.011 (.134)	.007 (.059)
		Grade 7	61,191	42	.89	.745 (.243)	–1.336 (.636)	.392 (.076)	.765 (.117)	–.049 (.088)	–.030 (.048)
		Grade 8	62,931	41	.89	.729 (.227)	–1.059 (.621)	.397 (.073)	.721 (.130)	–.024 (.058)	–.016 (.037)
		Grade 10	65,421	42	.91	.824 (.283)	–1.126 (.638)	.421 (.083)	.748 (.130)	–.045 (.088)	–.018 (.046)

BILOG a – predicted a). Positive values reflect under-prediction of a whereas negative values reflect over-prediction of a. Figure 4 presents frequency distributions of four sets of these residuals—one for the residuals computed from b estimates (equation 3; $M = -0.009$; $SD = 0.123$), one for the residuals computed from z_g estimates (equation 4; $M = 0.006$; $SD = 0.077$), one for Lord's (1980) function applied to corrected item-total point-biserial coefficients ($M = 0.315$; $SD = 0.191$), and one for Lord's (1980) function applied to item-total biserial coefficients (e.g., Lord's intended CTT index; $M = -.479$; $SD = .829$).[4]

Table 2

Bivariate Correlations among Study 2 Item Characteristics

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Residual (*b*-generated)	—								
2. Residual (z_g-generated)	.82*	—							
3. 2PL Chi-square	.03	.02	—						
4. 1PL Chi-square	.21*	.20*	.12*	—					
5. *a*-parameter	.60*	.21*	.05	.33*	—				
6. corrected item-total	.31*	–.08	.06	–.11*	.79*	—			
7. *p*-value	.02	–.17*	.01	.55*	.56*	.25*	—		
8. 2PL *b*-parameter	.15*	.18*	.00	–.51*	–.32*	–.02	–.93*	—	
9. 1PL *b*-parameter	–.10*	.06	–.01	–.67*	–.56*	–.16*	–.98*	.92*	—

Note. * $p < .05$.

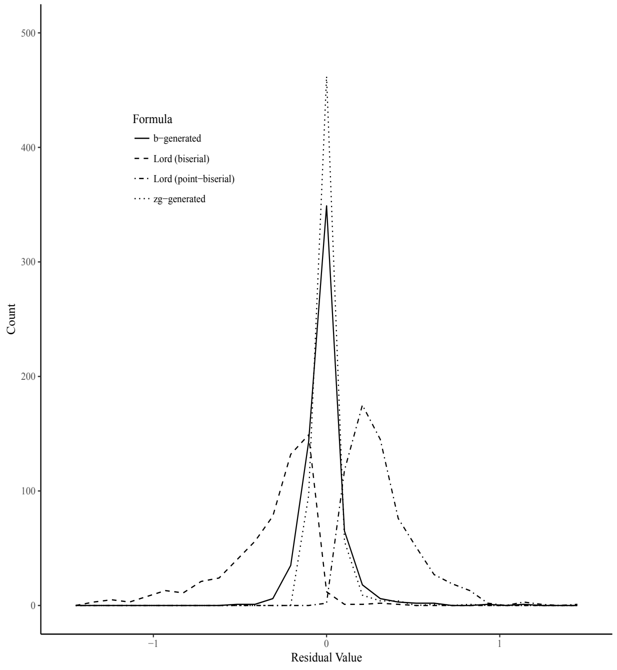

Figure 4. Normal-metric a-parameter residuals from the four investigated formulas.

[4] There were only 594 computed residuals from Lord's (1980) item-test biserial correlation due to 10 biserial coefficients exceeding values of 1.0 (resulting in a requested square root of a negative number—see Lord's formula).

The distributions of residuals from the modified equations 3 and 4 exhibit good prediction (small standard deviations) with desired locations. However, Table 2 also documents relationships between residuals and item discrimination indices (possibly reflecting under [over] prediction with highly [less] discriminating items) as well as discrimination and difficulty index associations (such that items of greater difficulty tended to exhibit less discrimination). Because of these Table 2 associations, we searched for systematic patterns of poor, fair, good or excellent fit within the sets of residuals. Figures 5 and 6 present, respectively, the *b*- and z_g-generated residuals grouped by both *a*-parameter and difficulty index magnitude (difficulty indices are categorized [at IRT-CTT rough equivalent values] to facilitate graphical presentation). As can be seen in these residual plots, systematic patterns do occur such that *extreme* item difficulties result in *a*-parameter under-prediction at high actual *a*-parameter values and over-prediction at low *a*-values when difficulties are modeled via the IRT *b*-parameter (e.g., Figure 5). This effect is less prominent with z_g-generated residuals (e.g., Figure 6), where the most evident pattern is greater prediction error with more highly discriminating items.

In addition to the above-noted index associations and residual patterns, 1PL chi-square associations also exhibited moderate association with residual values (see Table 2). The residuals reported in Table 2, however, reflect both magnitude and valence. Chi-square fit indices were therefore also correlated with the absolute value of residuals, with similar results (to those reported in Table 2) for 2PL chi-square association but much stronger 1PL chi-square associations (*r*'s = .49 [*b*-generated residual] and .37 [z_g-generated

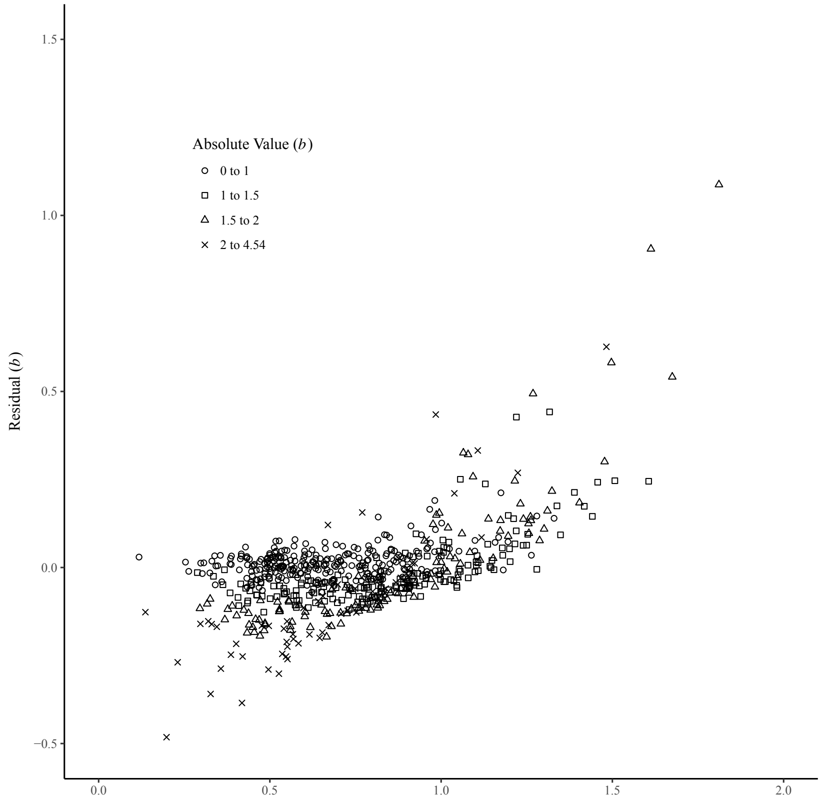

Figure 5. *b*-generated residuals (formula 3) and *a*-parameter magnitude (organized by *b* magnitude).

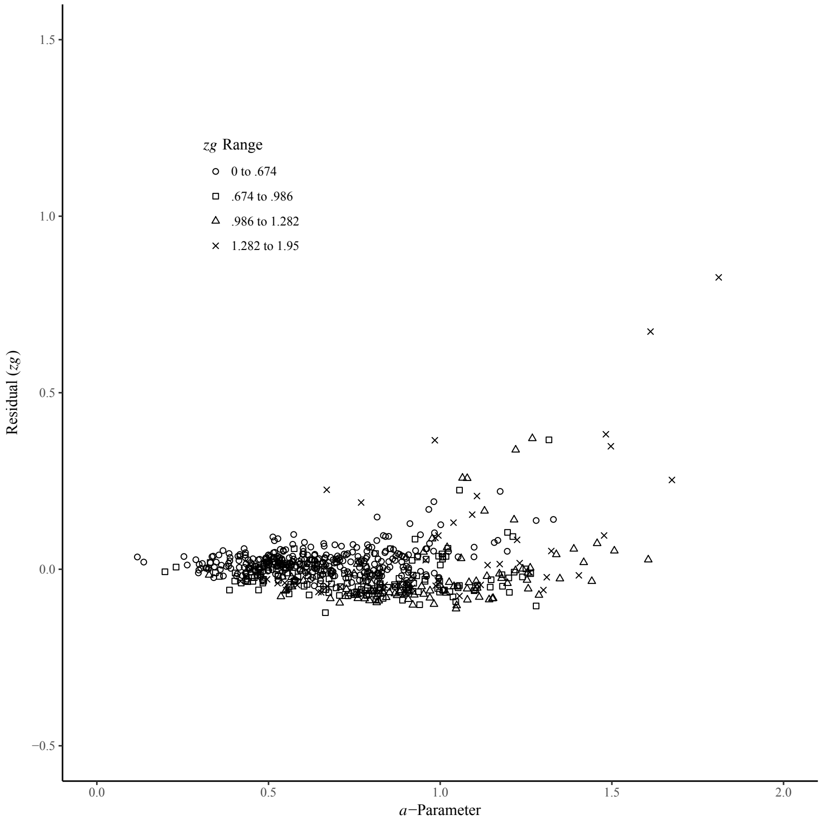

Figure 6. z_g-generated residuals (formula 4) and *a*-parameter magnitude (organized by z_g magnitude).

residual]). Note that 1PL misfit likely captures differential item discrimination (which is not modeled via 1PL specification) and it is therefore difficult to tease apart the 1PL (mis)fit residual associations from the *a*-parameter residual associations (e.g., these associations may be at least partially reflecting the same source—deviations from the fixed *a*-parameter value).

In spite of these noted systematic patterns, because the applied data has peculiarities (as would any set of applied data), we do not advocate a revision to our formulas. Rather than pursue further function refinement we therefore acknowledge a limitation – predicted *a*-parameters should be interpreted with more caution if they are associated with extreme item difficulties (beyond $b = |\pm 1.5|$ or corresponding *p*-values greater than .85 or less than .15). Also note that the purpose of functions that locate residuals on a value *closer* to zero is achieved here (Figure 4; Table 1). The adequacy of these functions across different contexts (particularly regarding *bias* of estimates at different levels of discrimination and/or difficulty) certainly warrants further attention – although again note that the simple goal (to obtain a contemporary refinement to the “crude” approximation) was attained.

Discussion

Our formulas are not as elegant as Lord’s (1980), but they are computationally simple, accommodate the item difficulty influence, and minimize residuals. The same caveats also exist—they are (somewhat refined) crude approximations and CTT estimates will always be sample-dependent. We do not, however, consider the sample dependency to be an *acute* problem because CTT indices can exhibit adequate levels of stability when item and person sampling is carefully practiced (as is common in high-stakes testing applications). Our suggested modifications pull the CTT asymptote off of the 1.0 value, but also adjust functional slopes and points of inflection. These adjustments are fully dependent on item difficulty. These functions can be contrasted with the original formula, which across both simulation and validation datasets was demonstrated to provide an *under*-estimate of the IRT index given a normal-metric *a*-parameter (when the CTT index is the *corrected* item-total correlation) and an *over*-estimate when the CTT index is the item-test biserial correlation.

We note here that Lord did also effectively take the difficulty-discrimination association into account via specification of a *biserial* correlation. It is quite possible that Lord's (1980) specification of the biserial coefficient was made in consideration of comparative index "invariance." The choice of estimated biserial as opposed to observed point-biserial as an index of discrimination was historically in agreement with the above noted limitation of CTT—the biserial is "more invariant over examinee samples" (Hambleton and Jones, 1993; p. 253). However, primarily because this index is not commonly encountered in contemporary psychological assessment applications, our recommended revision takes a different route to incorporate item difficulty information (e.g., our difficulty index is accommodated in the function [relationship between IRT and CTT discrimination indices] rather than the correlation [CTT discrimination index]).

Our Study 1 simulations suggested that the accuracy of our functions is strongest within a range of *b*-parameter values of –2 to +2 (approximate corresponding *p*-values of .1 to .9). The validation data generally supported this finding (current samples evidence functional degradation at *b*-values closer to |±1.5|). We do not, however, consider either constraint to represent a major limitation. Reise and Henson (2003) in fact state, "typical [*b*-parameter] item difficulties range from –2 to 2" (p. 94).

We believe that our primary contribution lies in the simple acknowledgement of the function presented in Figure 1 and its (likely) systematic under-estimation of IRT *a*-parameter values (particularly when the CTT index is a corrected point-biserial item-total correlation). Our initial validation of the revised functions across 16 educational and workforce tests do support their potential utility (e.g., good *a*-parameter prediction under reasonable testing conditions). We envision the formulas in their current form to be most useful to: 1) assessment specialists who wish to compare test characteristics when one test or sample has (only) CTT and the other has IRT item information, 2) individuals in need of starting values for iterative IRT parameter estimation procedures, and 3) researchers interested in pursuing future comparative IRT-CTT investigations.

Author Note

State Farm Mutual Automobile Insurance Company, its subsidiaries, and its affiliates were not involved with the creation of this paper, and no State Farm information was used in its development. The conclusions and opinions expressed in this document are solely those of the authors, and State Farm neither approves of nor endorses the conclusion and opinions expressed in the document.

This paper presents the views of the authors, and these views are neither approved nor endorsed by U.S. Bancorp. The opinions presented in this paper are solely those of the authors.

References

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., and Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*, 319-334.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. *Psychometrika*. *14*, 169-182.

Cook, L. L., and Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37-45.

Dawber, T., Rogers, W. T., and Carbonaro, M. (2009). Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models. *The Alberta Journal of Educational Research*, *55*, 512-533.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., and Mead, A. D. (1995). Fitting polytomous IRT models to multiple choice tests. *Applied Psychological Measurement*, *19*, 143-165.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357-381.

Hambleton, R. K., and Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practices*, *12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, *31*, 457-459.

Han, K. T., and Hambleton, R. K. (2007). User's manual: WinGen (*Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.

Kemery, E. R., Dunlap, W. P., and Bedeian, A. G. (1989). The employee separation process: Criterion-related issues associated with tenure and turnover. *Journal of Management*, *15*, 417-424.

Lord, F. M. (1963). Biserial estimates of correlation. *Psychometrika. 28*, 81-85.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Ludlow, L. H. (2001). Teacher test accountability: From Alabama to Massachusetts. *Education Policy Analysis Archives*, *9*, 1-22.

MacDonald, P., and Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 921-943.

Reise, S. P., and Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*, 93-103.

Rupp, A. A., and Zumbo, B. D. (2004). A note on how to quantify and report whether item parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*, 588-599.

Rupp, A. A., and Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*, 63-84.

Stark, S. (2001). MODFIT [Computer software]. Urbana-Champain, IL: University of Illinois.

Weitzman, R. A. (2009). Fitting the Rasch model to account for variation in item discrimination. *Educational and Psychological Measurement*, *69*, 216-231.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (2003). BILOG-MG 3.0 [Computer software]. Chicago, IL: Scientific Software.