Published in final edited form as:

J Abnorm Psychol. 2010 May ; 119(2): 300–306. doi:10.1037/a0018400.

Limitations of True Score Variance to Measure Discriminating Power: Psychometric Simulation Study

Seung Suk Kang^{1,2,3} and Angus W. MacDonald III^{1,3}

¹Department of Psychology, University of Minnesota

²Veterans Affairs Medical Center, Minneapolis

³Department of Psychiatry, University of Minnesota

Abstract

Demonstrating a specific cognitive deficit usually involves comparing patients' performance on two or more tests. The psychometric confound occurs if the psychometric properties of these tests lead patients to show greater cognitive deficits in one domain. One way to avoid the psychometric confound is to use tests with a similar level of discriminating power, which is a test's ability to index true individual differences in classic psychometric theory. One suggested way to measure discriminating power is to calculate true score variance (Chapman & Chapman, 1978). Despite the centrality of these formulations, there is no systematic examination of the relationship between the observable property of true score variance and the latent property of discriminating power. We simulated administrations of free response tests and forced choice tests by creating different replicable ability scores for two groups, across a wide ranges of various psychometric properties (i.e., difficulty, reliability, observed variance, and number of items), and computing an ideal index of discriminating power. Simulation results indicated that true score variance had only limited ability to predict discriminating power (explained about 10 % of variance in replicable ability scores). Furthermore, the ability varied across tests with wide ranges of psychometric variables, such as difficulty, observed variance, reliability, and number of items. Discriminating power depends upon a complicated interaction of psychometric properties that is not well estimated solely by a test's true score variance.

INTRODUCTION

"...[T]o measure differential deficit in ability, one must match tasks on true-score variance."

Chapman & Chapman, 1978, p. 305

Reliable and valid measures of individual differences provide the basis for describing the cognitive features of psychopathology. For example, executive control deficits in schizophrenia, inhibitory deficits in attention deficit hyper activity disorder, and episodic memory loss in Alzheimer's disease have been identified by psychological tests as primary deficits associated with these disorders. A process-specific deficit, however, is more difficult

Corresponding Author: Angus W MacDonald, Ph.D. University of Minnesota, Department of Psychology, N218 Elliott Hall, 75 E. River Rd., Minneapolis, MN, 55455, angus@umn.edu.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/abn.

to demonstrate. Although cognitive tests are intended to be sensitive to a particular cognitive ability, their measurement inevitably includes variance from other common cognitive and non-cognitive factors (Silverstein, 2008). In addition, patients tend to perform increasingly worse as disease severity increases on almost any test that requires a voluntary response. In other words, patients show generalized performance deficits. Therefore, demonstrating a specific deficit involves showing that test performance is impaired relative to performance on another test, preferably one as similar as possible to the test of interest. However, finding such a group by test interaction is not sufficient to be indicative of an interpretable specific deficit (Strauss, 2001). The reason is that tests may vary in their discriminating power. Discriminating power is a concept from classical psychometrics that represents the sensitivity of a test to individual differences (Lord, 1952). Tests with greater discriminating power are better at differentiating the more competent from the less competent subjects compared to tests with less discriminating power, regardless of the cognitive domain they measure. If a group by test interaction is observed between tests with different degrees of discriminating power, interpretation of the result as a specific deficit is misleading. That is, such an effect may be driven simply by that test's greater discriminating power. This issue is the essence of the psychometric confound (Chapman & Chapman, 1973).

As the epigraph above illustrates, the Chapmans (1978) suggested that true score variance might be a good way to measure discriminating power and that matching tests on this metric would equate them for their sensitivity to individual differences. In classical psychometric theory, an observed test score for a given individual is conceptualized as having two components: a true score and measurement error. True score refers to the portion of the score which is replicable or reliable. Thus, the variance of observed scores was assumed to be the sum of the variances of true scores and of (uncorrelated) measurement errors. It followed that the true score variance could be calculated as the product of observed score variance and reliability of the test, which is considered an estimate of the variance of reliable ability scores measured by a test. Therefore such an estimate of true score variance might be considered a metric of discriminating power.

Reliability and observed variance, however, have been suggested to have a more complicated relationship to discriminating power (Neufeld, 2007; Silverstein, 2008). For example, two components of observed variance, between- and within-group variances, can have different effects on discriminating power. The power of a test decreases as within-group variance increases, and increases as between-group variance increases. That is, high observed variance can increase group separation only if the source of the variance is related to group membership rather than within group variance. For this reason the relationship between true score variance and discriminating power merits further scrutiny.

Based on the test properties influencing reliability and observed score variance, Chapman and Chapman (1978) also suggested that true score variance was influenced by other test parameters, such as the item difficulty and the number of items. However, there has been no study testing the ability of the conventional measure of true score variance to predict discriminating power. The main obstacle to such an investigation is the lack of a direct measure of discriminating power in common experimental psychopathology practices. Although estimation of discriminating power of a test might be inferred with item-response theory (IRT) which allows for the simultaneous estimation of individual's latent ability and the accuracy with which that is measured (Emberson & Reise, 2000; Lord, 1980), IRT has not been able to be widely used in experimental psychopathology research. The reason for this may be that the practice of experimental psychopathology is not optimal for the requirements of IRT. IRT models are most sensitive when a range of items are presented, whereas experimental tasks often (but not always) have a fixed level of difficulty in each condition. Furthermore, most experimental studies have few subjects, whereas the recommended minimum sample size for

IRT parameter estimation is in the 100's. For these reasons, the current study evaluated statistical procedures consistent with classical psychometric practice prevalent in experimental psychopathology today.

The current study examined the relationship between true score variance and discriminating power using computer simulations of large samples of patients and controls taking over 2000 variants of a task that measured the same "cognitive process." The simulation allowed the computation of an ideal estimate of discriminating power to investigate the influence of test parameters on the relationship between true score variance and discriminating power. This allowed us to test the psychometric hypothesis that true score variance accurately measured discriminating power. Should this not be the case for the measurement of a single process, a fortiori it cannot be the basis for comparing tasks of different processes to examine the presence of a specific or differential deficit in a pathological group as suggested by Chapman and Chapman (1978).

METHODS

All simulations were conducted using MATLAB. These simulations were intended to estimate the discriminating powers across a range of tests with various psychometric properties, taken by two groups of participants with different mean ability levels. Based on the classical test theory assuming that test scores consist of true score and measurement error (Lord & Novick, 1968), the general procedure of the simulations consisted of several steps as shown in Figure 1. These steps, described below, included creation of the subjects' replicable ability scores (true scores) and item ability scores (true scores + measurement errors), dichotomization of item ability scores, and creation of test scores as described below. With the simulated test results, reliability (internal consistency), true score variance and discriminating power were estimated.

Creation of Replicable Ability Scores

The replicable ability scores of 10,000 normal subjects and the same number of patients were created to be randomly assigned from a standard normal distribution¹. A mean group difference of 1.0 was imputed by adding 0.5 to the mean of the control group and subtracting 0.5 from the mean of the patient group. This imputed deficit of 1.0 standard deviation in the patient group approximates the overall mean performance difference of 0.98 between schizophrenic patients and controls calculated in meta-analyses of cognitive deficit studies of schizophrenia patients (Dickinson, Ramsey, & Gold, 2007).

Creation of Item Ability Scores

The simulated tests consisted of various numbers of items, from 10 to 100. Each item scored subjects' performances as "correct" or "incorrect", and observed test scores were calculated as a total number of "correct" items. To simulate measurement error, each test item for each individual was assigned an error value. Measurement error values that varied across test items and subjects were created from normal distributions with as many random scores as the number of subjects, which were generated for as many number of test items. Thus, the total number of independent error scores generated was equal to the product of the number of items and the number of subjects. The means of the error score distributions were 0 and the standard deviations were manipulated to be fixed from 0.5 to 5.5 to simulate 11 levels of measurement

¹We ran 10 simulations with the same simulation parameters and computed intra-class correlation coefficients across the 10 simulations for the discriminating power estimate of each test case. The coefficient was .9995 in free response test simulations and .9996 in forced choice test simulations, indicating that the simulations results are almost the same across different runs. Thus we reported only one simulation results here.

error. An item ability score was then calculated for each item by simply adding each individual's replicable ability score to each individual's item error scores. This combined item ability score represented the continuous ability score as measured by that item. Thus item error with a higher standard deviation simulated the underlying instability of an unreliable test.

Dichotomization of Item Ability Scores

The performance on each item was transformed into a dichotomous accuracy score by applying a threshold to each item score. For example, one accuracy threshold score separated the lowest 10% of the distribution from the other 90% of item scores. Subjects with item ability scores below the threshold were determined to have made errors, whereas subjects above the cut point were considered to have responded correctly. The threshold score τ for each difficulty and each measurement error level was determined as

$$\tau = z_{threshold} \times \sqrt{SD_{z-distribution}^2 + SD_{ErrorScore-distribution}^2}$$

where $z_{threshold}$ stands for the score of z -distribution by which the percent correct is determined (e.g., .5244 is the z -score to divide upper 30% from lower 70% in the z -distribution). The SD of z -distribution is always 1, while the SD of error score distribution varied across different levels of measurement error. The threshold reflected difficulty, thus a low threshold score was more likely to lead to an accuracy score of 1 ("true") even for subjects with low ability scores while a high threshold score was more likely to lead to an accuracy score of 0 ("false") even for subjects with high ability scores. The former simulates the ceiling effect of an easy test and the latter simulates the floor effect of a difficult test. This dichotomization was conducted separately for each item in the test for each of 19 thresholds of test difficulty from 5 to 95% correct.

Creation of Test Scores

Each subjects' observed test score was then calculated as the sum of accuracy scores across all items for each levels of difficulty, measurement error, and number of items. Thus the possible perfect observed test scores ranged from 10 to 100, depending on the number of items on each test. In this procedure, we simulated two types of tests, "free response" test (FRT) and "forced choice" test with 2 alternatives (2-choice test; 2CT). The FRT does not allow guesses to improve test performance, such as vocabulary test in Wechsler Intelligence Test Battery, while forced choice test where subjects choose their responses among several alternatives allows guesses to improve test scores. Forced choice tests are common in timed experimental psychopathology studies, and in other circumstances. In the FRT, the test score was determined from only the item ability scores (true score + measurement error). However, in the forced 2CT, we simulated the guessing effect by adding a "guessing score" to the test score after the dichotomization. The guessing score was added when a subject's item score was below threshold, so that the subject still obtained correct scores by chance half the time.

Computation of reliability and indices of discriminating power

In order to factorially examine test conditions with various test parameters including 19 levels of difficulty, 11 levels of measurement error, and 10 levels of number of test items, we simulated a total of 2090 tests with combinations of these variables. Measurement error was the test parameter used to simulate various levels of reliability in our simulation study. Since measurement error can only be inferred in real testing, we estimated the reliabilities at each test condition as the mean of the item intercorrelations using Kuder and Richardson's formula 21 (or KR-21, Kuder & Richardson, 1937)². The KR-21 formula is formally equivalent to alpha (Cronbach, 1951) for dichotomously scored data. Reliabilities were computed from only

the performance of the control group. While the simulation environment allowed direct calculation of reliability, we used KR-21 to follow the standard practice with empirical data. The reliability was computed to be negative value in some tests, especially in 2CTs with high measurement error values, and those cases with negative reliabilities were removed from further analyses. True score variance was then computed as a product of reliability estimate and observed variance (Chapman & Chapman, 1978).

Although effect size has been used to indicate power of a test for group separation, this is not an ideal measure of discriminating power. Effect size mainly focuses on between group separation, while discriminating power should reflect intersubject separation on a test measure. Because we know the true ability scores of subjects in our simulations, we computed the index of discriminating power as a Pearson correlation coefficient between the replicable ability scores (true scores) and the observed test scores of all subjects. A test with high discriminating power should preserve the distribution of true ability scores in its observed score distribution, thus the scatter plot of the two scores should show linear relationship between them. That is, the better a test separates more able from less able subject on its observed test scores, the higher correlation (more linearity) between true replicable score and observed score should be obtained on a fixed scale, ranging from 0 to 1. For example, if the correlation was .99, it could be said that the test had almost perfect discriminating power, whereas lower correlation values corresponded to low discriminating power.

Ability of True Score Variance to Measure Discriminating Power

The ability of true score variance to measure discriminating power was quantified by calculating Pearson correlation coefficients between true score variances and discriminating power indices. The higher positive correlation between them indicates the higher ability of true score variance to predict discriminating power. The influences of test parameters on the ability of true score variance as a measure of discriminating power were examined by comparing the correlation coefficients computed across various levels of test parameters, such as difficulty (percent correct), observed variance, reliability, and number of items. The levels of test parameters were determined to result in similar number of test cases across the levels, thus the increments of the levels were not the same.

RESULTS

Differences in the Test Results between Free Response and Forced Choice Tests

Both types of test showed a similar relationship between true score variance and discriminating power indices (see Figure 2a and 2d). However, there were several main differences between them, including 1) mean % correct (including both normal and patient groups; FRT: 50.08%; 2CT: 79.47%³), 2) mean observed variance (FRT: 264.58; 2CT: 88.15), and 3) mean reliability (FRT: .694; 2CT: .539). As a result, true score variances were generally higher in FRT than in 2CT. However, the mean discriminating power was similar for the two types of tests: .81 and .87 in FRT and 2CT, respectively.

Relationship between True Score Variance and Discriminating Power

As shown in Figure 2a and 2d, true score variance had a weak linear relationship with the discriminating power estimate for both types of tests (R^2 was .148 in the FRT and .072 in the 2CT, respectively)⁴. The linear relationship between them was much improved when the true score variances were log transformed (see scatter plots in lower panel of Figure 2), showing

²The mean % correct was 75.69% in the all 2CT, including all tests with negative reliabilities.

³In this simulation study, no inferential statistical analysis was conducted given the nature of the simulation study with samples of population level. All the statistics reported in this study are descriptive, not inferential ones.

that R^2 was .381 in FRT and .327 in 2CT. Observed variance had weak positive association with discriminating power (see Figure 2b and 2e: R^2 was .101 in FRT and .072 in 2CT), while reliability had a strong positive association in FRT ($R^2 = .749$) but a weaker association in 2CT ($R^2 = .172$; see Figure 2c and 2f).

Influence of Test Parameters on Sensitivity of True Score Variance to Discriminating Power

Given that the true score variance has log-linear, not linear, relationship with discriminating power, the sensitivity of true score variance to discriminating power was indexed by the correlation between discriminating power estimates and true score variances in log scale, not in original scale. The correlation varied across levels of test parameters in both types of tests. Although the correlation was almost the same across different levels of difficulty in FRT (see Figure 2g and 2i), this varied a lot across difficulty levels in 2CT (see Figure 2h and 2i). The correlation was highest in 2CTs with medium level difficulty (i.e., 70~80% correct), while this correlation became negative in the most difficult 2CTs (i.e., < 60 % correct). The correlation was also influenced by observed variance in both types of tests, where the correlation tended to be low in tests with high levels of observed variances (see Figure 2j, 2k, and 2l). The correlation was highest at moderate levels of observed variance (i.e., 200~400) in FRT, while highest at lowest levels of observed variance (i.e., ≤ 20) in 2CT. The correlation was even negative in FRTs with observed variance higher than 800. In comparisons of the correlations across reliability levels, it was found that the correlation also varied across reliability levels, but the effect was complicated. The correlation tended to decrease as reliability increased in 2CT, while this tended to increase as reliability increased in FRT. Interestingly, when the reliability was highest (i.e., >.9 in FRT and >.8 in 2CT, respectively), the correlation was low in FRT, but high in 2CT (see Figure 2m, 2n, and 2o). Finally, the correlation increased as number of items decreased in both types of tests (see Figure 2e, 2j, and 2n). That is, the tests with smallest number of items (i.e., 10~20) showed the highest correlations between true score variance and discriminating power.

DISCUSSION

To test ability of true score variance to measure discriminating power, we simulated more than 2000 test administrations where 10,000 normal controls and 10,000 patients were tested using free response (FRT) and 2-choice (2CT) tests across a variety of psychometric variables. Results indicated substantial differences between FRT and 2CT in terms of the observed psychometric variables, such as mean percent correct, observed variance, and reliability. Despite these differences, the two types of tests had similar levels of discriminating power as well as a similar relationship of true score variance with discriminating power. That is, true score variance predicted about 10% of variance of discriminating power in its original scale, and predicted about 35% of variance in its log scale for both types of tests. Furthermore, while the ability of true score variance to measure discriminating power varied across levels of the psychometric variables of the tests, at no point was its predictive level very adequate. Since true score variance was inadequate in characterizing discriminating power for a single cognitive process measured with many different tasks, per force, it cannot be thought of as adequate for characterizing discriminating power across different cognitive abilities as is required for interpreting a specific or differential deficit.

The simulations showed that the 2CT which modeled even odds of being correct by chance generally had less observed variance, low reliability, and lower true score variance than the FRT. It is likely that the guessing effect in the 2CT reduced not only variation across subjects but also internal consistency of responses within the tests. Although true score variances of the 2CT were generally much smaller than those of the FRT, surprisingly there was no difference

in mean discriminating power estimates between the two types of tests. This discrepancy was consistent with the observation that true score variance did not index discriminating power.

Investigation of the linear associations of true score variance and its two components with discriminating power further indicated the limited ability of true score variance to measure discriminating power. As noted, the FRT and 2CT showed similar but disappointing levels of sensitivity of true score variance to discriminating power in both original and log scales, although there were some differences. This limited predictive ability of true score variance might be related to the modest relationships of observed variance and reliability with discriminating power. Observed variance only explained about 10 % of discriminating power variance in both FRT and 2CT. This should be due to the fact that observed variance had multiple sources, including variance of pre-assigned true scores associated with group difference and error variance. The sources of observed variance are unknowable (Neufeld, 2007), thus observed variance had only a limited relationship to discriminating power. On the other hand, reliability explained much more variance in discriminating power. However, this high predictive ability of reliability was found only in the FRT (which explained 74.9% of variance), not in the 2CT (which explained 17.2% of variance). In the FRT with no probability of being correct by chance, reliability might be very indicative of the sensitivity of a test to individual differences in the ability scores measured by the test. This was not the case in forced choice tests with a high probability of being correct by chance. Consistent with this observation, discriminating power varied according to the levels of test reliability in the FRT, but less so in the 2CT (see Figure 2m and 2n). These findings might be related to the fact that reliability can be increased by either reducing measurement error or increasing true score variance (Neufeld, 1984). Reliability in 2CT might be increased by the increased variance caused by guessing, which is replicable but not related to group separation. As discussed in Silverstein (2008), a test can have good power to differentiate groups, although the test has a low reliability because of reduction in irrelevant true score variance (i.e., true score variance not associated with group membership). That is, high reliability per se cannot guarantee high discriminating power, especially in tests with multiple choices.

In addition, the ability of true score variance to measure discriminating power was found to be influenced by other psychometric variables. The test difficulty was found to be important for the ability of true score variance in 2CT, but not in FRT, where the correlations between true score variance and discriminating power was constant across difficulty levels. In 2CT, when a test was too difficult (e.g., a test with less than 60% correct), the observed variance, the main component of true score variance, was largely driven by guessing. Therefore, not only the observed variance, but also the true score variance of such a difficult 2CT might not be able to reflect true individual variations in the replicable ability scores. The sensitivity of true score variance was also influenced by the levels of observed variance. In both types of tests, there was a tendency for tests with higher observed variances to have less sensitivity of true score variance to discriminating power. Interestingly, although there was a tendency that the correlation between true score variance and discriminating power increased as reliability increased in FRT, the FRT with the highest levels of reliability (i.e., $>.9$) showed a low relationship between true score variance and discriminating power, as high reliability (i.e., $>.9$) corresponded to wide range of discriminating power (see Figure 2c). This might be related to a ceiling or floor effect, where too easy or too difficult tests might have high reliability but relatively low discriminating power. Lastly, the association between true score variance and discriminating power increased as number of items decreased in both types of tests. That is, true score variance could predict discriminating power better when the tests had fewer items.

These results suggested that there were psychometric circumstances where true score variance was a better predictor of discriminating power. Generally speaking, both the FRT and the 2CT with small number of items (e.g., about 20) tended to have true score variance that is more

predictive of discriminating power. On the other hand, true score variance was least predictive of discriminating power in following circumstances; the FRT with very high true score variance due to too big observed variance (i.e., more than 800), low or very high reliability (i.e., lower than .5 or higher than .9), and the 2CT with very high difficulty (i.e., % correct less than 60%) and big observed variance (i.e., more than 150). However, these guidelines are based on the preliminary analyses of the effects of psychometric variables considered in isolation. Given that those psychometric variables appear to interact with each other in complicated ways, it is difficult to identify the psychometric circumstances where true score variance should be relied upon as a measure of discriminating power.

The current simulation study found that, despite a strong theoretical foundation, true score variance even in its log scale had only limited ability to measure discriminating power. Furthermore, psychometric properties of a test influenced the ability of true score variance to measure discriminating power. That is, true score variance could not be a stable measure of discriminating power across tests with various combinations of psychometric variables. Our findings of limited association of the psychometric variables such as observed variance and reliability with discriminating power are consistent with recent discussions of the limitations of task matching strategy based on psychometric variables (Knight & Silverstein, 2001; Neufeld, 2007; Silverstein, 2008). Without consideration of complex interactions amongst psychometric variables and the sources of observed variance and reliability (e.g., pathognomonic vs. non-pathognomonic variances; Silverstein, 2008), matching tests simply on one or two observed psychometric variables may lead to other confounds or misinterpretations of specific or differential deficits. In this case, although we calculated internal consistency as our measure of reliability, the use of retest or other forms of reliability would not alter this central finding.

Given the limitations of true score variance, it may be more useful to focus on developing an alternative stable and sensitive measure of discriminating power that can be calculated from the observed properties of a task in a control population. Such a metric might more fully consider the effects of observed psychometric variables and the interactions amongst them on discriminating power.

Acknowledgments

The authors are grateful for the comments of Michael B. Miller. This work was supported in part by NIH grants MH066629, MH084861 and MH079262 and support from the Sidney R. Baer, Jr. Foundation.

References

- Chapman LJ, Chapman JP. Problems in the measurement of cognitive deficit. *Psychol Bull* 1973;79(6): 380–385. [PubMed: 4707457]
- Chapman LJ, Chapman JP. The measurement of differential deficit. *J Psychiatr Res* 1978;14(1-4):303–311. [PubMed: 722633]
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
- Dickinson D, Ramsey ME, Gold JM. Overlooking the obvious: a meta-analytic comparison of digit symbol coding tasks and other cognitive measures in schizophrenia. *Arch Gen Psychiatry* 2007;64(5): 532–542. [PubMed: 17485605]
- Emberson, SE.; Reise, SP. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
- Knight RA, Silverstein SM. A process-oriented approach for averting confounds resulting from general performance deficiencies in schizophrenia. *J Abnorm Psychol* 2001;110(1):15–30. [PubMed: 11261389]
- Kuder GF, Richardson RM. The theory of the estimation of test reliability. *Psychometrika* 1937;2:151–160.

- Lord, FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.
- Lord, FM.; Novick, MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
- Neufeld RWJ. Re: The incorrect application of traditional test discriminating power formulations to diagnostic-group studies. *J Nerv Ment Dis* 1984;172(6):373–374. [PubMed: 6544811]
- Neufeld, RWJ. Composition and uses of formal clinical cognitive science. In: Shuart, B.; Spaulding, W.; Poland, J., editors. Modeling complex systems: Nebraska symposium on motivation. Vol. 52. Lincoln, NE: University of Nebraska Press; 2007.
- Silverstein SM. Measuring specific, rather than generalized, cognitive deficits and maximizing between-group effect size in studies of cognition and cognitive change. *Schizophr Bull* 2008;34(4):645–655. [PubMed: 18468987]
- Strauss ME. Demonstrating specific cognitive deficits: a psychometric perspective. *J Abnorm Psychol* 2001;110(1):6–14. [PubMed: 11261401]

Figure 1.

Test simulation procedure. For each participant, a replicable ability score was randomly generated from a z-normal distribution. This score was combined with item error scores to generate item ability scores. The item error scores were also randomly generated from a normal distribution with a standard deviation equal to the measurement error level. The item ability scores were thresholded with a criterion threshold score to generate 1 or 0 item test scores (correct or incorrect), which were summed to generate observed test scores. In creation of test scores of two choice tests, guessing scores were added to the sum of item test scores to generate observed test score.

Figure 2.

Scatter-plots of discriminating power estimates versus true score variance and its two components of observed variance and reliability in FRT and 2CT simulations (a-f). Scatter plots of log true score variance versus discriminating power and their correlations were depicted according to difficulty (g-i), observed variance (j-l), reliability (KR-21, m-o), and number of items (p-r). These figures illustrate the non-linear and multi-determinant relationship between true score variance and discriminating power. Abbreviations: OV: observed variance, REL: reliability, NITEMS: number of items.