Power of IRT in GWAS: Successful QTL Mapping of Sum Score Phenotypes Depends on Interplay Between Risk Allele Frequency, Variance Explained by the Risk Allele, and Test Characteristics

Stéphanie M. van den Berg[1]* and Susan K. Service[2]

[1] *Department of Research Methodology, Measurement and Data Analysis, University of Twente, The Netherlands*
[2] *Center for Neurobehavioral Genetics, University of California, Los Angeles, California*

As data from sequencing studies in humans accumulate, rare genetic variants influencing liability to disease and disorders are expected to be identified. Three simulation studies show that characteristics and properties of diagnostic instruments interact with risk allele frequency to affect the power to detect a quantitative trait locus (QTL) based on a test score derived from symptom counts or questionnaire items. Clinical tests, that is, tests that show a positively skewed phenotypic sum score distribution in the general population, are optimal to find rare risk alleles of large effect. Tests that show a negatively skewed sum score distribution are optimal to find rare protective alleles of large effect. For alleles of small effect, tests with normally distributed item parameters give best power for a wide range of allele frequencies. The item-response theory framework can help understand why an existing measurement instrument has more power to detect risk alleles with either low or high frequency, or both kinds. *Genet. Epidemiol.* 36:882–889, 2012. © 2012 Wiley Periodicals, Inc.

Key words: item-response theory (IRT); measurement; statistical power; extreme samples design; case-control design; population sample design

*Correspondence to: Stéphanie M. van den Berg, Department of Research Methodology, Faculty of Behavioral Sciences, University of Twente, Measurement and Data Analysis (OMD), P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: stephanie.vandenberg@utwente.nl

Received 13 July 2012; Revised 13 July 2012; Accepted 3 August 2012
Published online 10 September 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21680

INTRODUCTION

Many diagnostic instruments for a disorder consist of symptom counts. Often the disorder can be seen as the extreme tail of a continuous liability trait: the higher the liability, the more likely a subject shows certain symptoms. High-liability persons will have many symptoms and on the basis of a diagnostic criterion are then labeled as affected with the disorder of interest. In some instances, genome-wide association (GWA) studies are applied on symptom count data, rather than diagnosis, as this allows using more information [Van der Sluis et al., 2012].

Item-response theory (IRT) provides a formal statistical framework for modeling liability and diagnosis, and its application in the medical sciences is increasing [Reise and Waller, 2009]. IRT has also been successfully applied in genetics [Eaves et al., 2005; Van den Berg et al., 2007, 2010; Van Leeuwen et al., 2008]. It provides a useful framework for understanding the relationship between measurement problems and problems in detecting genetic variants. For example, using this IRT framework, Van der Sluis et al. [2010] showed that ignored multidimensionality, measurement bias, and poor reliability can result in poor statistical power in QTL-mapping studies.

Here, we show how power is associated with test characteristics using different study designs, and how this is association is moderated by allele frequency. We link the simulation results to the IRT concept of “test information.” We start out with a brief introduction to IRT and so-called test information functions (TIFs). Next, we describe how this framework makes predictions about statistical power in QTL mapping. Three simulation studies demonstrate the intricate relationship between study design, allele frequency, and the TIF.

ITEM-RESPONSE THEORY

IRT models item data as a function of both item characteristics as well as person characteristics [Embretson and Reise, 2000; Lord, 1980; Lord and Novick, 1968]. An item can be anything from a symptom that is scored in a diagnostic interview as being either present or absent, or an item on a self-report questionnaire that can be answered with yes or no. Items do not have to be dichotomous (i.e., yes/no, or 1/0), but for clarity of exposition, we focus on dichotomous items in our descriptions. In the Discussion, we expand on alternative data types.

The one-parameter logistic IRT model for dichotomous items, or so-called Rasch model, is

$$P\left(X_{ij}=1\,|\theta_i,\beta_j\right)=\frac{1}{1+\exp\left(\beta_j-\theta_i\right)} \tag{1}$$

where $P(X_{ij}=1)$ is the probability of a positive response for person i on item j (or the presence of symptom j). Parameter θ_i is the person parameter for person i and can be thought of as person *i*'s liability for a disorder. Parameter β_j is the item parameter for item *j*. In educational measurement, this β parameter is usually called the difficulty parameter, where it refers to the difficulty of a cognitive test item. A high β value indicates a difficult item, with an overall low probability of making the item correct; only high ability individuals have a reasonable chance of giving a correct answer. Analogously, for questionnaire and clinical data, a high value for β_j indicates a high threshold for a positive response on a questionnaire, or a high threshold for a symptom. A high β value therefore corresponds to low symptom prevalence, as the symptom is not endorsed by many persons. As can be seen in equation (1), when $\theta_i = \beta_j$, the probability of a positive response is 50%. When $\theta_i > \beta_j$, the probability of a positive response is higher than the probability of a negative response, and vice versa for $\theta_i < \beta_j$. In order to make the modeling statistically identifiable, it is usually assumed that the population mean for parameter θ equals 0.

The logistic curve for $P(X_{ij} = 1)$ in equation (1) is sigmoidal (see Figure 1) and has its maximum slope at $\theta_i = \beta_j$: at this point, the change in $P(X_{ij} = 1)$ as a function of θ is at its maximum, rendering maximum discrimination between those individuals with θ values below β (more likely to have a negative response) and those individuals with θ values above β (more likely to have a positive response). Thus, in Figure 1, the slope of the left curve is at its maximum at $\theta = -1.5$, whereas the right curve has its maximum slope at $\theta = 1.5$. Discriminatory power of an item is at its lowest at θ values very far removed from β; for example, in the curve on the left in Figure 1 ($\beta = -1.5$), individuals with high θ values will all have a probability of nearly one for a positive response; but individuals with even higher θ values have about the same probability. All are very likely to have the same positive response. An item *j* therefore yields very little information about individuals and individual differences at θ levels far removed from β_j.

The information that an item *j* with known β_j gives about a trait θ is therefore a function of θ. For the one-parameter model, the Fisher information from item *j* is

$$I_j(\theta) = P_j(\theta)\left(1 - P_j(\theta)\right) \tag{2}$$

where $P_j(\theta)$ is the probability of a positive response for item *j* for a person with liability $= \theta$. The one-parameter model can be extended to a two-parameter model, where the slope of the item-response curve can vary over items, and by consequence the peaked-ness of the item information functions, but for clarity of exposition, we focus on the one-parameter logistic IRT model of equation (1). In the Discussion, we expand to the two-parameter model.

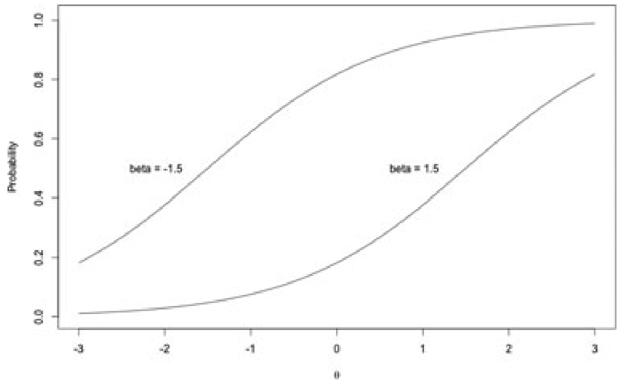

Fig. 1. Item-response curves: The probability of a positive response plotted as a function of θ (see equation (1)) for two different values for β (left –1.5, right 1.5).

The information provided by all test items on trait θ combine additively to the overall shape of the TIF,

$$TIF(\theta) = \sum_{j=1}^{k} I_j(\theta) \tag{3}$$

and with $[TIF(\theta)]^{-1/2}$, we have a formula for the standard error of measurement [SEM; Lord, 1980]. Thus, at θ levels where test information is high, the SEM for the maximum likelihood estimate is small, so that we have high measurement precision. At those levels, it is possible to discriminate between individuals of different levels. With high information content, a small difference in θ level is associated with a difference in response pattern (more or fewer endorsed items, i.e., a different phenotypic sum score). At θ levels with little information, slight differences in θ do not result in different response patterns (sum scores), so there is no information to discriminate among individuals. Figure 2 shows a number of TIFs (top row), with varying distributions of the β values for the items. For instance, for a test with β values uniformly distributed between 1 and 3 (test scenario "Right": all items have low endorsement probabilities, therefore high β values), information reaches its maximum on the right-hand side of the distribution. Persons with high liability will therefore show variation in their sum scores, but not the persons with low liability.

RELATIONSHIP BETWEEN TEST INFORMATION AND DISTRIBUTION OF SUM SCORES

The TIF is directly related to the shape of the distribution of sum scores. If the TIF is symmetric and centered around the average liability value of the population, the resulting distribution of sum scores will be symmetrical, too. If however the TIF has its maximum at the right-hand side of the scale, that is, for above-average levels of θ, the expected distribution of sum scores will be positively skewed. This is observed, for example, when clinical tests are used in population samples: most of the items on clinical tests or diagnostic instruments will have low response rates, as they relate to symptoms that less than half the normal population is likely to exhibit. For such items, the probability of a positive response will exceed 50% only for individuals with high values on the liability trait. The β values for such items will therefore also be highly positively valued (equation (1)), with a resulting TIF that has its maximum at a relatively high θ level. All individuals with average and below-average liability values will have low to very low probability of symptoms, so many will have a sum score of 0 and there is no further distinction possible among them. On the other hand, there will be quite some variation in number of symptoms for above-average trait values. As a result, we see a skewed distribution of symptom counts. So in short, for a given population with the average trait level defined as $\theta = 0$, we expect a positively skewed sum score distribution when most of the β values are positive, a negatively skewed sum score distribution when most of the

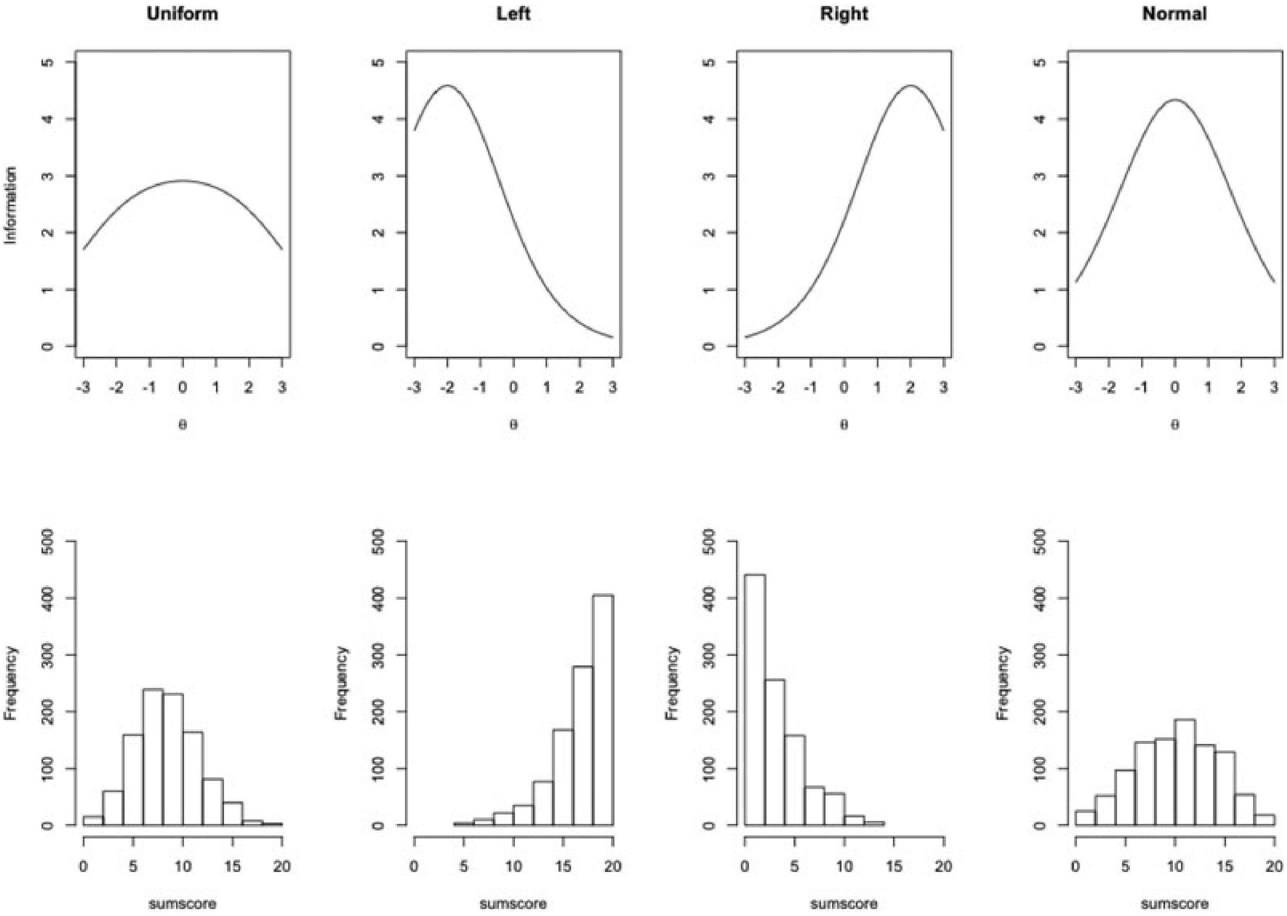

Fig. 2. Test information functions for different types of tests, each consisting of 20 items, but different distributions of the items' β values (top row, see text for explanation of the labels.). Bottom row shows corresponding distributions of sum scores when simulating data from 1,000 individuals with θ values drawn from a standard normal distribution.

β values are negative, and more symmetrical when the β values are scattered across the continuum. Figure 2 shows a number of example TIFs and the typical sum score distributions that result from these (based on simulated data for a 20-item test). The labels "Right," "Left," "Uniform," and "Normal" refer to the distribution of the values of the β parameters in a test: "Right" $\beta \sim U(1, 3)$, "Left" $\beta \sim U(-3, -1)$, "Uniform" $\beta \sim U(-3, 3)$, and "Normal" $\beta \sim N(0, 1)$. Observe in Figure 2, that if the distribution of the β parameters is uniform on the [–3,3] continuum (Uniform), the resulting sum score distribution shows lower variance than if the β parameters are standard normally distributed (Normal).

TEST INFORMATION AND POWER IN DIFFERENT GENE-MAPPING STUDY DESIGNS

As with any mapping study, the power to find a QTL for a liability trait crucially depends on having a data set where there is both genotypic variation and phenotypic variation. More specifically, in the case of sum score phenotypes: there should be sum score variation among those individuals that have different genotypes. If all heterozygotes Aa have the same sum score or the same case-control status as homozygotes AA, there would be lower power to detect a QTL than a situation where heterozygotes have a different average sum score than homozygotes. If a risk allele has a strong phenotypic effect, one would expect that there would be AA, Aa, and aa individuals that have above-average liability, assuming an additive model of action of the A allele. The below-average liability individuals will be mostly aa (see Figure 3). The interesting genotypic variation therefore exists at the right-hand side of the liability scale. It is then important to have good measurement resolution at that part of the scale: one would want the three different genotypes to have different average sum scores. Thus, as a general rule, we expect that the power to detect a QTL for a risk allele with frequency smaller than 0.5 and with a large effect is best for tests that have a lot of test information at high trait levels. Conversely, we expect that the power to detect a QTL for a risk allele with a high frequency (>0.5; that is, rare *protective* alleles) and of large effect is best for tests that have high levels of test information at low trait levels. Following the same logic, studying the expected liability distributions of different genotypes, we expect that alleles of small effect and/or alleles with a frequency of around 0.5 will generally require high levels of test information in the middle of the liability distribution.

This paper presents simulation studies that illustrate the tight relationship between test information, the risk allele frequency of QTLs, the liability variance explained

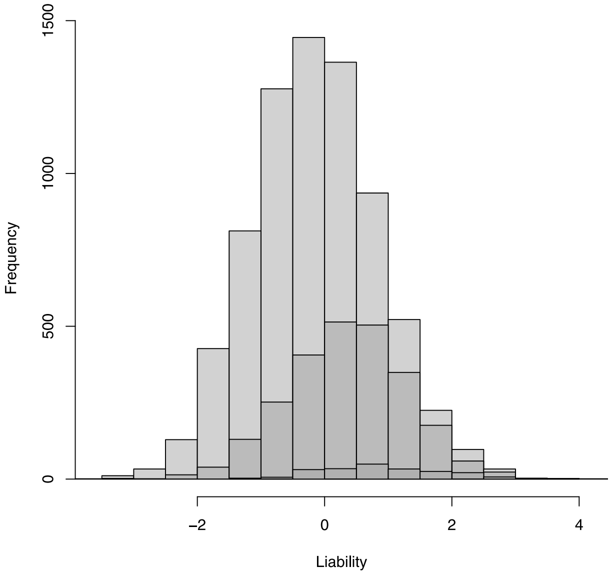

Fig. 3. Superimposed histograms of simulated liability scores of aa genotypes (light gray), Aa genotypes, and AA genotypes (dark gray), with frequency of the A allele equal to 0.15 and variance explained equal to 25%, under Hardy-Weinberg and additive gene action.

by a QTL, and the statistical power to find such QTLs in a gene-mapping study. Three different research designs are studied: a population study where subjects are randomly sampled from the population (Study 1), an extreme samples design, where subjects are sampled from those with only very high or very low sum scores (Study 2), and a case-control design (Study 3). We explore power in situations where QTLs explain little liability variance (1%), and contrast these findings to the situation where QTLs explain a large part of liability variance (25%). This contrast with 25% variance explained serves to illustrate the main principles, rather than being a realistic setting for gene-mapping studies.

METHODS

Number of test items and sample sizes were varied across simulation studies in order to produce a wide range in power estimates to make patterns in the results as clear as possible. Power was defined as the percentage of simulation replicates to be significant at the 0.01 alpha level. The number of replicates for each simulation situation was 1,000 in all studies.

SIMULATION STUDY 1—POPULATION SAMPLE DESIGN

For each test information scenario (“Left,” “Right,” “Uniform,” and “Normal”), data sets were simulated under the assumption of a single-nucleotide polymorphism (SNP) in perfect linkage disequilibrium with a QTL that explains 1% of the variance in a standard normally distributed quantitative liability trait θ under additive gene action. That is, we simulated θ under a linear model where the expectation was equal to an allelic effect of the risk allele times the number of risk alleles, and a normally distributed random term with variance proportional to the variance due to the allelic effect. Theta values were then rescaled to mean 0 and variance 1. The frequency of the risk allele (the allele that increases liability) was varied: 0.001, 0.005, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, 0.995, and 0.999. In each replicated data set, genotypes for 1,000 individuals were simulated assuming Hardy-Weinberg equilibrium. Theta (θ) values were standardized and used for simulating the data for a 20-item test under a one-parameter logistic IRT model (equation (1)). For each individual, the item scores were summed, and the sum score phenotypes were linearly regressed on the number of risk alleles to study power.

Note that the proportion of variance in liability explained by the QTL was fixed to 1%, but not the proportion of variance in the observed phenotypic sum scores. The proportion of variance in the sum scores explained by the QTL is dependent on test information, which in turn depends on the β parameter values for the items in the test. With a limited number of items, the proportion of sum score variance explained is always lower than the proportion of liability variance explained [Van den Berg et al., 2007].

Under the test information scenario referred to as “Left,” the item parameters β were randomly sampled from the uniform distribution $U(-3, -1)$, and under the scenario referred to as “Right” using $U(1, 3)$. Under the scenario “Uniform,” the item parameters were sampled using $U(-3, 3)$, and under the “Normal” scenario, they were sampled from the standard normal distribution, $N(0, 1)$. Figure 2 shows the corresponding TIFs.

Because low allele frequencies lead to very low counts for certain genotypes, the assumptions underlying standard statistical testing might not be met. In practice, if only one or two individuals are homozygous for one allele, one might prefer an empirical *P*-value over the *P*-value resulting from a standard *t*-test for the linear regression on genotype. Moreover, for “Left” and “Right” test scenarios, the sum score distribution is very skewed, also violating the assumptions of standard linear regression. Therefore, all *P*-values were empirically determined, based on 400 permutations per data set.

In a second series of simulations for Study 1, the QTL explained 25% of the liability (θ) variance. This value was chosen as an extreme contrast to the 1% simulations. Power was based on simulating data for 50 individuals on a 10-item test. All other settings were the same as in the first series.

SIMULATION STUDY 2—EXTREME SAMPLES DESIGN

The same test information scenarios were used as in Simulation Study 1. In a first series of simulations, the variance of the liability explained by the QTL was fixed to 1%. In each test information scenario and in each simulation, sum score data were simulated for 10,000 subjects on 20 items. Two hundred subjects were randomly sampled from those subjects with the lowest observed sum score (with replacement), and 200 subjects were sampled from those with the highest observed sum score. In each simulated data set, Fisher’s exact test was performed on the cross-tabulation of SNP genotype and “extreme high”/“extreme low” status. One thousand data sets were simulated for each condition of risk allele frequency: 0.005, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, and 0.995.

In a second series of simulations for Study 2, serving as a contrasting illustration, the variance in liability explained by the QTL was fixed to 25%. All other settings were the same as in the first series, except that now only 10 extremely low scoring and 10 extremely high scoring subjects were sampled from the population of 10,000 subjects that had sum score data based on five items.

SIMULATION STUDY 3—CASE-CONTROL DESIGN

The same test information scenarios were used as in Simulation Studies 1 and 2. In a first series of simulations, the variance of the liability explained by the QTL was fixed to 1%. In each test information scenario and in each simulation, sum score data were simulated for 10,000 subjects on 20 items. Six hundred controls were randomly sampled from those subjects with subthreshold observed sum score (i.e., lower than the cutoff score for diagnosis; sampling with replacement), and 600 cases were sampled from those with the cutoff sum score for diagnosis or higher (with replacement). The cutoff score was defined as the 88th percentile based on the observed sum scores in the simulated samples of 10,000 individuals [cf. Van der Sluis et al., 2012]. In each simulated data set, Fisher's exact test was performed on the cross-tabulation of SNP genotype and case-control status. One thousand data sets were simulated for each condition of risk allele frequency: 0.001, 0.005, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, 0.995 and 0.999.

In a second series of simulations for Study 3, serving as a contrasting illustration, the variance in liability explained by the QTL was fixed to 25%. All other settings were the same as in the first series, except that now 60 cases and 60 controls were sampled from the population of 10,000 subjects that had sum score data based on 20 items.

RESULTS

SIMULATION STUDY 1—POPULATION SAMPLE DESIGN

Power to detect a QTL that accounts for a low percentage of trait variance is best detected when the trait is measured using a test in which the β parameters are normally distributed, except for extreme allele frequencies (Figure 4). Such a test with normally distributed β parameters clearly outperforms a test where item parameters are uniformly distributed. Low power is observed for detecting rare risk alleles for "Left" tests, where item β parameter values are all negative. The same low power is observed for rare protective alleles for "Right" tests with only positive items, that is, a clinical test that discriminates only among high-scoring individuals. For very low frequencies, clinical tests seem optimal, outperforming the power of a test with normally distributed item parameters. For very high frequencies (i.e., very rare protective alleles), "Left" tests outperform "Normal" tests.

When the variance of liability explained is 25%, however, "Left" and "Right" type of tests outperform "Normal" tests, depending on allele frequency (Figure 4). For risk allele frequencies lower than 0.2, "Right" tests give best power; for risk allele frequencies higher than 0.8, "Left" tests give best power. For risk allele frequencies between 0.2 and 0.8, "Normal" tests seem to give best power.

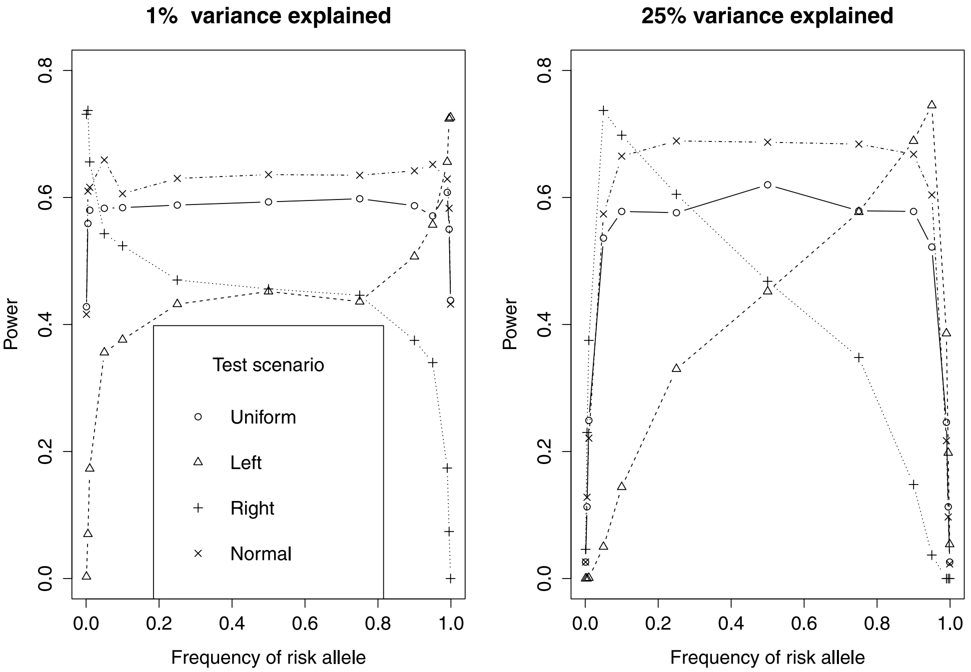

Fig. 4. Power to detect a QTL as a function of risk allele frequency and test characteristics, using a population sample design. Percentage liability variance explained by QTL is 1% (left panel), and 25% (right panel).

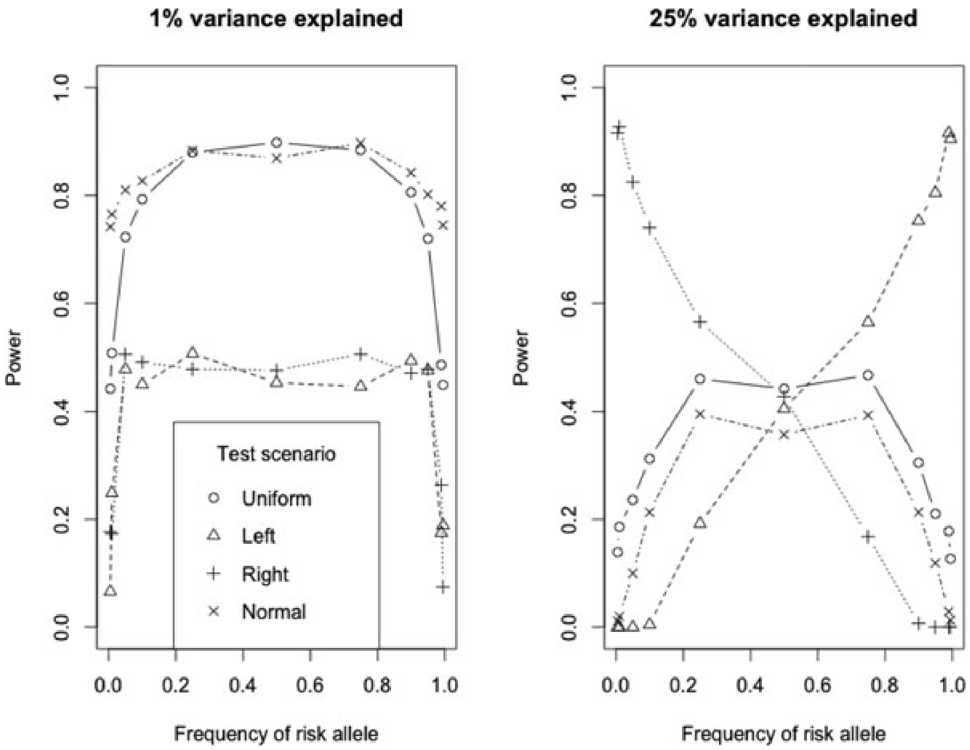

Fig. 5. Power to detect a QTL as a function of risk allele frequency and test characteristics, using an extreme samples design. Percentage liability variance explained by QTL is 1% (left panel) and 25% (right panel).

SIMULATION STUDY 2—EXTREME SAMPLES DESIGN

When the percentage variance explained by the QTL is 1%, lowest power is observed for "Left" and "Right" types of tests, irrespective of risk allele frequency (Figure 5). Overall best power is shown by tests with normally distributed β parameters. In contrast, when percentage variance explained by the QTL is 25%, the "Left" and "Right" types of tests outperform the other two: "Left" tests are optimal for risk allele frequencies above 0.5, and "Right" tests are optimal for risk allele frequencies below 0.5.

SIMULATION STUDY 3—CASE-CONTROL DESIGN

When the percentage variance explained by the QTL is 1%, lowest power is observed for "Left" types of tests, irrespective of risk allele frequency (Figure 6). Overall, "Right" types of tests show highest power, while tests with normally distributed β parameters are a close second. When percentage variance explained by the QTL is 25%, power is again lowest for "Left" types of tests. In addition, there is a decrease in power with increasing frequency of the risk allele for allele frequencies larger than 0.10.

DISCUSSION

In educational measurement, the definition of the "best test" is the test that minimizes measurement error over the target of measurement [Wright and Stone, 1979]. For instance, when constructing a law school admission test, one would want to have maximum test information at the point on the scale where the threshold for sufficient aptitude is located; one would want to have confidence that the decision of pass/fail is reliable. There is less interest in reliably quantifying individual differences at the lower end of the scale. Translating this concept of "optimal test design" to gene-finding studies leads to constructing or searching for a test that discriminates best among those subjects with different genotypes for the type of QTL one hopes to find.

Power to detect a QTL naturally depends both on variation in genotypes and variation in liability, but more crucially, it depends on whether the measurement tool that is used to assess liability discriminates among the genotypes. First, power depends on how much genotypic variance there is and also where it is: as Figure 3 shows, allele frequency and variance explained determine in unison where and how much genetic variance there is: allele frequency determines the relative surface areas of the three histograms, whereas variance explained (actually, allelic effect) determines how far apart the means of the three histograms are. Together, allele frequency and variance explained determine whether most of the genotypic variation is among high-liability individuals, low-liability individuals, or individuals of average liability. Second, the TIF shows how well a test discriminates between individuals, and where on the scale on the scale it does so more clearly. Optimally therefore, the location of the maximum of the TIF on the liability scale should be exactly there where there is the most variation in genotypes. In the case of Figure 3, for example, a test that has a TIF that has its maximum value on the far left-hand side of the distribution, with very low

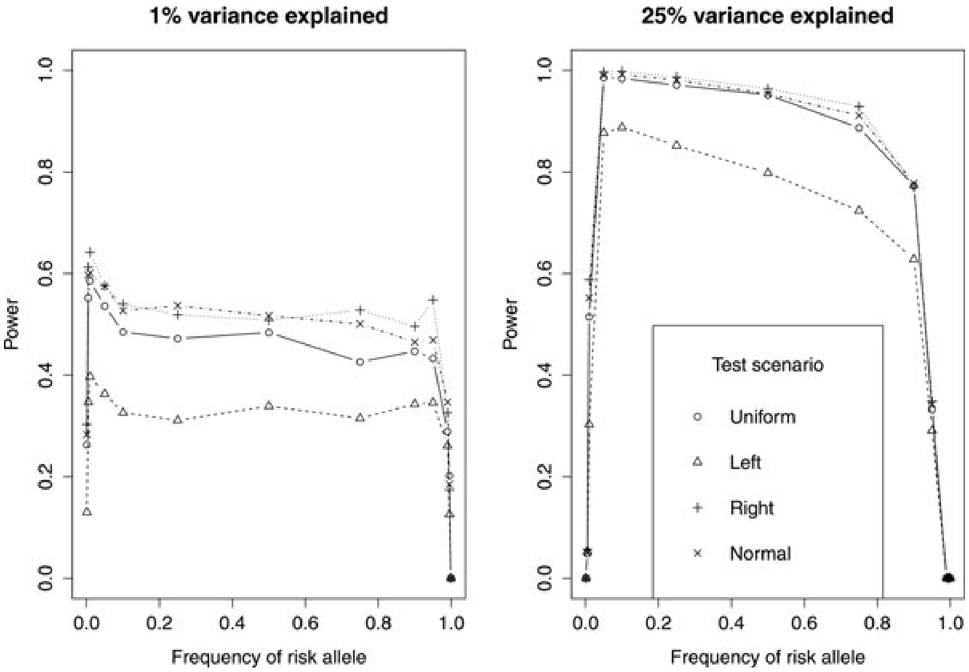

Fig. 6. Power to detect a QTL as a function of risk allele frequency and test characteristics, using a case-control design. Percentage liability variance explained by QTL is 1% (left panel) and 25% (right panel).

information content on the right-hand side of the distribution, will result in extremely low power to find a QTL with risk allele frequency 0.15, even when variance explained is 25%: sum scores would phenotypically only discriminate among individuals that were homozygous for the protective allele (i.e., they would vary in their sum score), whereas most heterozygotes and homozygotes for the risk allele would have the same maximum sum score (cf. ceiling effect).

This is exactly what was seen in the simulations, both for a population design and an extreme samples design: even with highly influential risk alleles (25% variance explained), the power to detect them can be dramatically low when the wrong phenotypic measurement tool is used. In contrast, for case-control designs, this effect was less dramatic.

In the case of QTLs that explain only little of the liability variance, we see all three genotypes throughout the phenotypic range; all genotypes are scattered across the entire scale, with all three genotype means close to 0. For maximum power, the information function should therefore also be more spread out across the continuum, but also be high where most of the individuals are. A test that is highly informative at scale locations where only few of the individuals are located generates little statistical power. Our simulations clearly showed that the optimal TIF for a wide range of allele frequencies and variance explained was where the TIF is the result of item parameters that are standard normally distributed. As can be seen in Figure 2 the corresponding sum score distribution then also shows the largest variance.

Even in a case-control design (Study 3), tests with normally distributed item parameters performed very well, and approached the power of clinical tests. Nevertheless, a test that maximizes the discrimination between cases and controls is the best tool for a case-control study, independent of risk allele frequency or variance explained.

Thus, overall, when looking for alleles of small effect, tests with normally distributed item parameters give best power for a wide range of allele frequencies, and for different study designs. Only in the case of a population sample design and very extreme allele frequencies (below 0.01, above 0.99), such tests are outperformed by “Right” and “Left” tests, respectively. In case-control designs, “Right” types of tests are slightly more powerful than “Normal” types of tests. In the study design phase, therefore, care should be devoted to choosing the optimal diagnostic test or self-report questionnaire. The TIF of the phenotyping instrument can be used as an aid for the identification of potential measurement problems. It can be plotted using standard IRT software packages such as Multilog [Thissen et al., 2003], or the ltm library in R [Rizopoulos, 2006] and compared to the TIFs in Figure 2. When interest is in finding QTLs that explain only a small amount of variance of a wide range of allele frequencies, the information function of the phenotyping instrument should ideally look like the “Normal” information function in Figure 2.

As all results shown here are directly related to the shape of the TIF, the results are easily extended to two-parameter IRT models (where items may have different factor loadings or discrimination parameters), and to IRT models for polytomous items, where instead of yes/no or absent/present, the data consist of scales with multiple categories, for instance, “never,” “sometimes,” “often,” and “always.” Such IRT models can be fitted using standard software after which the TIF can be plotted. These TIFs are directly comparable to those based on the one-parameter IRT model for dichotomous data as used in the present simulations.

Usually in genome-wide association studies, SNPs with minor allele frequencies (MAFs) lower than 1–2% are not included in analyses. In these simulations, we varied allele frequency in the population. Focus on low-frequency alleles is increasing with the increased interest in resequencing studies. In case-control and extreme samples designs, sample frequencies of risk alleles will be higher than their population values. Given the right design and the right sort of phenotypic test, studies can be very powerful to detect allelic effects with low MAF, see, for example, Figures 4 and 5, where high power was observed for MAFs below 1%.

The results reported here may also partly explain why relatively few QTLs of small effect have been identified for disorders measured using clinical tests [e.g., Manolio et al., 2009]. Thus far, the reasons mentioned in the literature for the lack of success include mainly limited sample sizes [Sullivan, 2012], or reasons of genetic nature [e.g., Crow, 2011]. Relatively little attention has been paid to psychometric properties of phenotypic measures [but see Van der Sluis et al., 2010, 2012]. This study clearly shows that genetic studies could be much more successful if researchers selected their phenotypic instruments to suit their study aim. Rather than using the phenotypic measurement that is the most valid clinical tool for diagnosis, one should choose the phenotypic measurement tool that gives highest power to find genetic variants that explain variation in liability. Changing the phenotypic instrument might be much cheaper than ever-increasing sample sizes with the same clinical test. Such a change might be as trivial as rewording an item, such as, for example, changing "I like chocolate very much" into "I like chocolate." Making the item less extreme will increase prevalence of positive responses, resulting in lower item β parameter values and therefore a maximum of the information function lower on the liability scale, with a corresponding change in statistical power. Simulation Study 2 shows that power to find QTLs of small effect in an extreme samples design can be dramatically increased when changing from a clinical test to a test that shows more variability in low-liability individuals, with the same number of items.

In addition to carefully choosing a measurement instrument for the phenotyping, test information, and therefore power, can be further increased by simply adding items to a scale (see equation (3)). One approach is to combine data from different phenotypic measurement instruments to optimize the TIF for a measure of interest. For example, Van den Berg et al. [under revision] optimized the TIF for a clinical measure for schizotypy by adding information from a self-report instrument. This was done through the use of a multidimensional IRT measurement model, which resulted in increased measurement precision, especially at the lower end of the liability continuum. As shown by the present results, this should increase overall power to find a QTL of small effect. In conclusion, IRT is a useful framework to identify potential strengths and limitations of an existing measurement instrument based on symptom counts or questionnaire items, and to remedy potential problems through more sophisticated modeling of multivariate phenotypic data sets.

ACKNOWLEDGMENTS

This work was made possible in part due to a grant from the University of Twente Incentive Fund awarded to SMvdB. The authors thank Sophie van der Sluis and an anonymous reviewer for helpful comments.

REFERENCES

Crow TJ. 2011. The missing genes: what happened to the heritability of psychiatric disorders? Mol Psych 16:362–364.

Eaves L, Erkanli A, Silberg J, Angold A, Maes HH, Foley D. 2005. Application of Bayesian inference using Gibbs sampling to item-response theory modelling of multi-symptom genetic data. Behav Genet 35:765–780.

Embretson ES, Reise SP. 2000. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum.

Lord FM. 1980. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Lord FM, Novick MR. 1968. Statistical Theories of Mental Test Scores. Reading, UK: Addison-Wesley.

Manolio TA, Collins, FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Reise SP, Waller NG. 2009. Item Response Theory and clinical measurement. Ann Rev Clin Psychol 5:27–48.

Rizopoulos D. 2006. ltm: an R package for latent variable modelling and Item Response Theory analyses. J Stat Software 17:1–25.

Sullivan P. 2012. Don't give up on GWAS. Mol Psych 17:2–3.

Thissen D, Chen WH, Bock RD. 2003. Multilog (Version 7) [Computer Software]. Lincolnwood, IL: Scientific Software International.

Van den Berg SM, Glas CAW, Boomsma DI. 2007. Variance decomposition using an IRT measurement model. Behav Genet 37:604–616.

Van den Berg SM, Fikse F, Arvelius P, Glas CAW, Strandberg E. 2010. Integrating phenotypic measurement models with animal models. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, August 1–6, 2010. Giessen, Germany: German Society for Animal Sciences.

Van den Berg SM, Paap MCS, Derks EM, Genetic Risk and Outcome of Psychosis (GROUP) investigators. Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. Psychiatry Research (in press).

Van Leeuwen M, van den Berg SM, Boomsma DI. 2008. A twin-family study of general IQ. Learn Individ Differ 18:76–88.

Van der Sluis S, Verhage M, Posthuma D, Dolan CV. 2010. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. Plos One 5:e13929.

Van der Sluis S, Posthuma D, Nivard MG, Verhage M, Dolan CV. 2012. Power in GWAS: lifting the curse of the clinical cut-off. Mol Psych. doi:10.1038/mp.2012.65. Epub ahead of print.

Wright BD, Stone MH. 1979. Best Test Design-Rasch Measurement. Chicago, IL: MESA Press.