6

Genetic Modeling of Tobacco Use Behavior and Trajectories

Hermine H. Maes and Michael C. Neale

Genetic studies have provided strong evidence that heritable factors generate individual differences in smoking behavior. Shared environmental factors appear to play a larger role in tobacco use at earlier ages. Improved modeling techniques hold the potential to better differentiate between genetic and environmental factors in tobacco use. This chapter examines genetic modeling issues in the study of smoking trajectories and behavior, including

- *Methodological and conceptual issues such as inferring potential dependence and trajectories in nonsmokers, issues in measurement invariance, and the use of epidemiological methods in genetically informative studies*
- *Statistical modeling considerations such as the use of structural equation modeling (SEM) to assess whether covariation between traits is due to genetic or environmental causes, the identification of genetic latent classes, and the analysis of molecular genetic data from linkage and association studies*
- *A review of prior genetic studies of smoking behavior, including twin and extended twin family studies, multivariate genetic studies, and molecular genetic studies*
- *A study applying an item response theory (IRT) approach to an analysis of smoking trajectories with data from the Virginia Twin Registry, examining tobacco initiation, regular tobacco use, and items on the modified version of the Fagerström Tolerance Questionnaire (FTQ)*

The IRT study in this chapter underscores the importance of assessing measurement invariance in establishing the heritability of nicotine dependence and its variation with gender.

The analyses described herein were supported by Public Health Service grant RR008123 and National Institute of Health grants CA085739, CA093423, DA011287, DA016977, DA018673, MH001458, MH049492, MH065322, MH068521, and a grant from the Virginia Tobacco Settlement Foundation. Mx development was previously supported by Public Health Service grants RR008123 and National Institute of Health grant MH001458. Data were kindly provided by Dr. Kenneth S. Kendler and the Mid-Atlantic Twin Registry.

Introduction

This chapter examines issues in studying the heritability of tobacco use behavior and trajectories and their methodological implications for future genetic research in nicotine dependence. Its goal is to follow the discussion of tobacco use trajectories in chapter 5 to examine what can be learned about these trajectories through studies using genetically informed data. Areas discussed include methodological and conceptual issues, a review of existing genetic studies of smoking, and the results from a multivariate genetic analysis of nicotine dependence using the Virginia Twin Registry.

A key question in many epidemiological studies is the extent to which parents influence their children. For example, one may ask whether parental cigarette smoking in and of itself increases the chance that their children will smoke. At the simplest level, one might compare the proportion of smokers in parents whose children smoke to the parents whose children do not. A higher rate of smoking in the parents of smokers may be taken as evidence that behavioral modeling is operating; that is, children have learned their behavior from their parents. Alternatively, it might be thought that the secondhand smoke ingested by the child of a smoker kindles the smoking habit. In practice, however, such conclusions may be unwarranted, because—except in the case of adoption—parents share genetic factors with their children. Genetically informative research designs, such as data collected from monozygotic and dizygotic twins, or from adopted and biological relatives, permit a closer inspection of the nature of parent-child resemblance, and indeed, of any association between a putative risk factor and an outcome. In principle, any random effect, such as variation in level or slope in a growth curve model, or membership in a particular latent trajectory class, may be partitioned into genetic and environmental components. However, the value of data collected from family members does not end here. In addition to the potential to resolve genetic and environmental components of variance, it is possible to measure covariance between variables that cannot be measured with data from unrelated individuals. For example, one can test whether liability to initiate smoking is related to quantity smoked or propensity to become nicotine dependent. Such information is of particular value when one considers whether to expend efforts on the prevention of tobacco initiation or on the alteration of trajectories of tobacco consumption once initiation has occurred. Therefore, this chapter provides a review of these methods, with a view to integrating both molecular and nonmolecular approaches into the same framework.

First, some of the methodological and conceptual issues in tobacco use research are considered. A statistical framework is then discussed within which these issues may be tackled. The approach is general enough to encompass both latent trait and latent class models and is suited to a wide variety of both genetic and nongenetic analyses. This methodological review is followed by a substantive one, considering the findings of genetic studies of smoking initiation on nicotine dependence. The final section applies multivariate genetic analysis of tobacco use and nicotine-dependence symptoms to data collected from relatives in the Virginia Twin Registry. The results are described in more detail than those of published studies because these results integrate a focus on assessing the phenotype with the traditional partitioning of the variance of that phenotype into genetic and environmental sources. The analyses also take into account that nicotine dependence is contingent on smoking initiation and progression to regular smoking.

Methodological and Conceptual Issues

One of the problems with studying tobacco use is that many of the symptoms and signs of abuse or dependence are *contingent*. Thus, it is not possible to observe the rate of increase in use of cigarettes in those who have never smoked. Whether it is correct to regard the nonsmoker's increase in cigarette consumption as zero is an empirical question. There is an assumption that a nonsmoker does not experience symptoms of nicotine dependence. However, in trying to understand the population from an epidemiological perspective, it is often better to ask the question of whether a nonsmoker would have experienced symptoms of dependence if he or she had initiated cigarette use. Certain research designs permit such inferences. For example, data from pairs of siblings might show that nicotine-dependence symptoms are more common in individuals with a sibling who has also become a tobacco user than in those with a sibling who has not. Such data imply a relationship between initiation and dependence. Ordinarily, with data collected from unrelated individuals, it is typically not possible to assess the relationship between initiation and dependence symptoms because dependence data are missing in those who have not initiated. Therefore, the modeling of this contingent type of data is described.

A similar issue arises with the analysis of the relationship between age at onset of tobacco use and its sequelae, such as trajectory. While it is possible to compare trajectories of those who initiate at a young age to those who initiated at a later age, it remains impossible to examine the trajectories of those who have not initiated use. Again, a research design that includes data collected from relatives provides a framework within which the relationship between age at onset and liability to use may be estimated. In this context, it becomes possible to tease apart factors that influence initiation, which, in turn, influences trajectory, from those that influence trajectory only.[1] In addition, it may prove useful to examine substance use as a function of time onset rather than of chronological age.[2]

One of the impediments to research on behavioral and psychological traits, such as tobacco use, is that behavior is intrinsically difficult to measure. For the most part, the quantification of daily tobacco use is limited to an ordinal scale (0, 1–5, 6–10, 11–20, 20+), and the assessment of symptoms of dependence is typically only binary (e.g., do you find it difficult to cut down?). Many of the more modern models for the analysis of growth or change have been developed on the assumption that measurement has been at the *interval* level. For the most part, it is not wise to simply pretend that the data have been measured on a continuous scale and proceed with data analysis as usual. However, it is often possible to extract continuous-level information from ordinal data by modeling it appropriately,[3] although at the cost of additional computer time. Yet, even given an appropriate analytical framework for ordinal data, things can go wrong at the measurement level. For example, a questionnaire item—do you find it difficult to wait for your first cigarette of the day—may provide a good indicator of dependence for those attending high school if smoking at home is not permitted. Those who no longer live at home may never have to wait, and therefore, the question loses its relevance as a measure of dependence. Such failures of *measurement invariance* are important to detect and should be controlled wherever possible.[4–7] That is, it is important to distinguish change in behavior or symptoms over time from change in the way that the measurement instrument works. This chapter examines this issue of measurement invariance with data from twins assessed with the FTQ.[8]

Many statistical frameworks are constructed around the assumption that the population is homogeneous in some respect. Thus, a simple regression equation, $y = \beta_0 + \beta_1 x$, implies that the effect of the independent variable x on the dependent variable y is the same for all subjects in the sample. In practice, however, it is possible that the strength of the regression—for example, between liability to initiate tobacco use and liability to progress to nicotine dependence—varies as a function of other variables. Such *moderation* of relationships may occur as a function of either variables that have been measured, such as age or gender, or of variables that have not been measured, such as an unidentified polymorphism at a particular region of the genome or the quantity of secondhand smoke experienced as a child.

Much progress has been made in tying together statistical methods used in epidemiological studies of unrelated individuals with those in use with genetically informative studies. For example, analyses of growth curves, measurement invariance, factor analysis, and latent class analysis all have been adapted and extended for use with data collected from relatives. Multilevel analysis might be considered to be almost ubiquitous in the study of relatives in that the family provides a level. However, it is clear that several areas have yet to be implemented for use in family data. For example, factor mixture modeling and growth curve transition modeling are in need of further development. While technical challenges remain (e.g., the likelihood of longitudinal ordinal data collected on a large pedigree may involve numerical integration over a very large number of dimensions), this is an area of active research. The future, with improvements in computer architecture and software that exploit it, holds much promise for furthering the understanding of genetic and environmental factors in the etiology, development, and interaction of complex traits.

Statistical Framework

Structural Equation Modeling

The majority of statistical modeling of genetically informative data is carried out within the framework of SEM. In its basic form, SEM involves the specification of two types of variables: (1) observed variables that have been directly measured and (2) latent variables that have not been directly measured. Two types of relationship between these variables may be specified: linear regression and covariance. This type of model may be represented as a path diagram[9–11] in which observed variables are shown as boxes, latent variables are shown as circles, regression paths are drawn as single-headed arrows from the independent variable to the dependent variable, and covariance paths are shown as double-headed arrows. Any description of the model, be it a simple list of the paths involved, or matrices thereof, or a correctly drawn path diagram, is mathematically complete and can be used to derive predicted covariances between variables. Three extensions of this framework are becoming popular. One is the depiction of means,[12] usually drawn as a triangle that has a constant value of one, which enables specification of mean structure as well as of covariances.[13] The second is the specification of "definition variables," which are values attached to specific paths in the diagram. These may specify a different predicted covariance structure for every subject in the sample.[14] They are thus of value in the specification of models for data that were collected at different sets of ages, as opposed to the unlikely scenario that, for example, all subjects were assessed precisely on their 10th, 12th, and 15th birthdays.[15] The third extension is that the population may be described as a mixture of two or more subpopulations in which different mean and/or covariance structures exist. This third addition subsumes latent class and latent profile analyses as special cases; growth curve

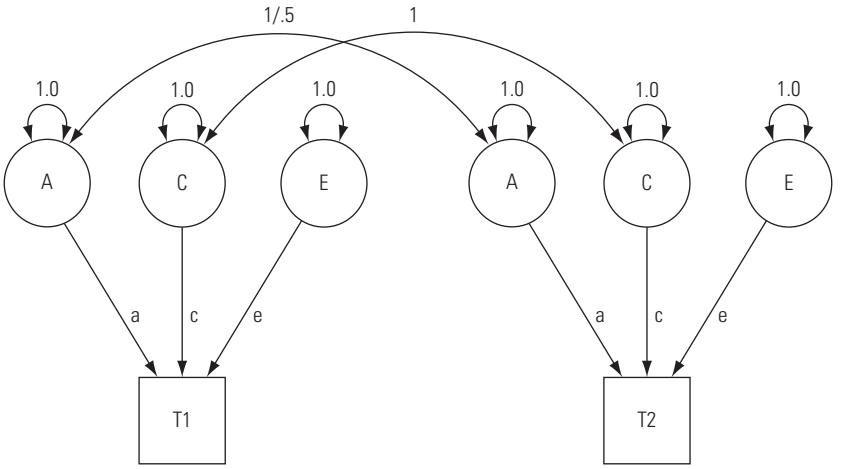
mixture modeling is a popular example.[16,17] This framework is referred to as "extended structural equation modeling" (i.e., XSEM).

Originally, SEM was devised for the analysis of data that were distributed according to the multivariate normal distribution, and it is still used in this way in many applications. The addition of mean structures, which may differ according to group or definition variables, makes the method appropriate for the analysis of data that are distributed according to a conditional multivariate normal distribution (the data from the sample as a whole will not be normally distributed if there are group mean differences). Applications to binary or ordinal data have become popular, because such data are commonly encountered in behavioral and other research. For the most part, these methods shift the distributional assumption to a level above the actual measurement, such that it is assumed that there is an underlying normal distribution of liability in the population, but that it is only possible to discriminate whether a given subject falls in a particular range or band of this distribution. For example, if a subject indicates that he or she has tried smoking cigarettes, as do some 60% of subjects, then the subject's liability is assumed to be in the top 60% of the distribution, or above a threshold of –0.253 measured in *z*-score units. It turns out that with ordinal data with at least three categories, it is possible to estimate the same parameters as in the continuous case by fixing the first threshold at zero, the second threshold at one, and estimating the mean and variance of the measure instead.[3] It is also possible to fit growth curves to binary item data,[18,19] although item-specific variances are confounded with item means in this case.

Structural Equation Model for Twin Data

The basic path diagram for the analysis of data collected from pairs of monozygotic and dizygotic twins—the most widely used genetically informative design[20]—is shown in figure 6.1. The diagram includes three variance components: additive genetic factors (A), which correlate perfectly between monozygotic twins and .5 between dizygotic twins; common or shared environmental factors (C), which correlate perfectly between twins regardless of their zygosity; and random or specific environmental factors (E), which are those influences unique to each member of a twin pair (including measurement error and genotype by specific environment interaction). The key to identifying the parameters of this model (the regression paths a, c, and e) is the availability of three statistics: the variance, the covariance between monozygotic twins, and the covariance between dizygotic twins. These data, together with the equal environment assumption (for more details, consult, for example, Loehlin and Nichols,[21] Rose and colleagues,[22] and Kendler and colleagues[23] for theoretical and empirical reasons that the equal environment assumption is unlikely to be violated), allow unique estimates of the parameters to be obtained. Alternatively, one could include a dominance parameter (D) instead of shared environment; the effects of both are confounded in the classical twin design. It is important to note that the classical twin study is really just a starting point for the genetic epidemiological investigation of a trait. Extending the design to include, for example, adoptees, parents and offspring, half siblings, or more distant relatives allows for resolving a greater variety of genetic and environmental parameters, such as genetic nonadditivity and assortative mating,[24] which are assumed to be zero when fitting the ACE model. Other assumptions include no genotype by environment correlation or interaction. The power of the classical twin study has been described in detail for the continuous case[25] and the ordinal case.[26] Of note is that for ordinal data, three times the sample size is needed for equivalent

Figure 6.1 Basic Path Diagram for the Analysis of Data Collected from Pairs of Monozygotic and Dizygotic Twins



Note. The correlation between additive genetic factors is fixed at either 1.0 or 0.5, according to whether the twins are monozygotic or dizygotic. A = additive genetic; C = common or shared environment; E = specific or unique environment; a, c, e = regression paths; T1 = twin 1; T2 = twin 2.

power to the continuous case when the threshold is at the optimal 50%, and this ratio increases rapidly for more extreme thresholds. In general, the twin study has more power to reject false models when the true world involved shared environmental effects than when familial aggregation was genetic. While false models that involve no familial aggregation are easy to reject, models including incorrectly specified sources of resemblance (e.g., AE instead of CE) are difficult to reject.

The basic ACE model can be straightforwardly extended to multivariate or longitudinal data. In the multivariate context, it becomes possible to partition *covariation* into the same components as is variation. Thus, one can establish whether two traits covary primarily because the same genetic factors influence both or because the same environmental factors do so. In addition, it is possible to detect relationships between variables that do not covary within an individual but, in fact, share genetic and environmental factors whose influences counterbalance—for example, a correlation of +.7 due to environmental factors but –.7 because of genetic factors. This same partitioning of covariation between traits may be applied to the same trait measured on repeated occasions to address whether development and change have primarily genetic or environmental origins. Four specific extensions to this model are considered below.

Extensions of the Basic Twin Model

Extended Twin Family Studies

While twin studies provide an excellent design to disentangle genetic and shared environmental influences, several assumptions are made, and only a limited number of sources of variance can be estimated simultaneously (A, C, and E or A, D, and E, with C and D being confounded). Three statistics provide the information for the partition: the total phenotypic variance,

the monozygotic covariance, and the dizygotic covariance. Data from other types of relatives provide additional, qualitatively different statistics, which (subject to identification of the model) permit estimation of additional sources of variance. Conceptually, this approach is similar to that used in plant and animal breeding experiments in which different types of cross provide information about different types of genetic effect.[27] Early contributions to developing methods for the analysis of data from human populations were provided by Jencks,[28] Eaves and colleagues,[29,30] and Fulker.[31]

Extending the twin design to include siblings allows a test of whether twins resemble each other more than do regular siblings. The usual way to model the addition of siblings is as a special twin environment variance component, *T,* for which twins (monozygotic or dizygotic) are specified to correlate perfectly, while siblings are specified to correlate with zero. Several potential contributors to a variance component are specified in this way. The most obvious source is twins who share trait-influencing environmental factors to a greater extent than do siblings. A second possibility is that twins influence each other, although typically this would result in different total variances of monozygotic, dizygotic, and siblings. A third potential contributor is interaction between age or cohort and the variable under study. Nontwin siblings are commonly measured at different ages and may therefore have reduced similarity compared with siblings measured at the same age and time. The addition of half siblings or adoptive siblings also permits estimation of genetic dominance as well as shared environmental influences—two sources that are confounded in the classical twin study.

Further extensions, such as including parents of twins, provide a test for the presence of assortative mating (process of mate selection based on the phenotype) and cultural transmission or whether parents influence their children's behavior through environmental pathways in addition to passing on their genes. Different mechanisms could account for environmental transmission: (1) parents can influence the environment of their children directly; this is referred to as phenotypic cultural transmission (or P [phenotype] to C [shared environment] transmission); and (2) the parental environment directly influences the children's environment, which is known as social homogamy (C to C transmission). Similarly, assortment, evidenced through significant marital correlations, can be a function of the phenotypes of the spouses (phenotypic assortative mating). Alternatively, social homogamy may result in spousal concordance, or direct influence between the spouses may lead to increased similarity over time. The extended twin kinship model, which extends the classical twin study with not only siblings and parents but also spouses and children of twins, was developed[32] for simultaneous estimation of additive and dominance genetic as well as unique and shared environmental (cultural transmission, nonparental, special twin) factors in the presence of assortment. The specification includes phenotypic cultural transmission and phenotypic assortative mating.[33] It is important to note that the correlation between parents and their children alone provides information to sort out whether parents directly influence their children's smoking behavior in that they also share genes with one another. However, a design that includes additional types of relatives (with differing degrees of genetic similarity), such as monozygotic and dizygotic twins, allows one to disentangle genetic from environmental transmission and "controls for" the genetic relatedness of parents and offspring.

Another design that also allows for disentangling genetic from environmental transmission is the children of twins (COT) design, which collects data from adult twin

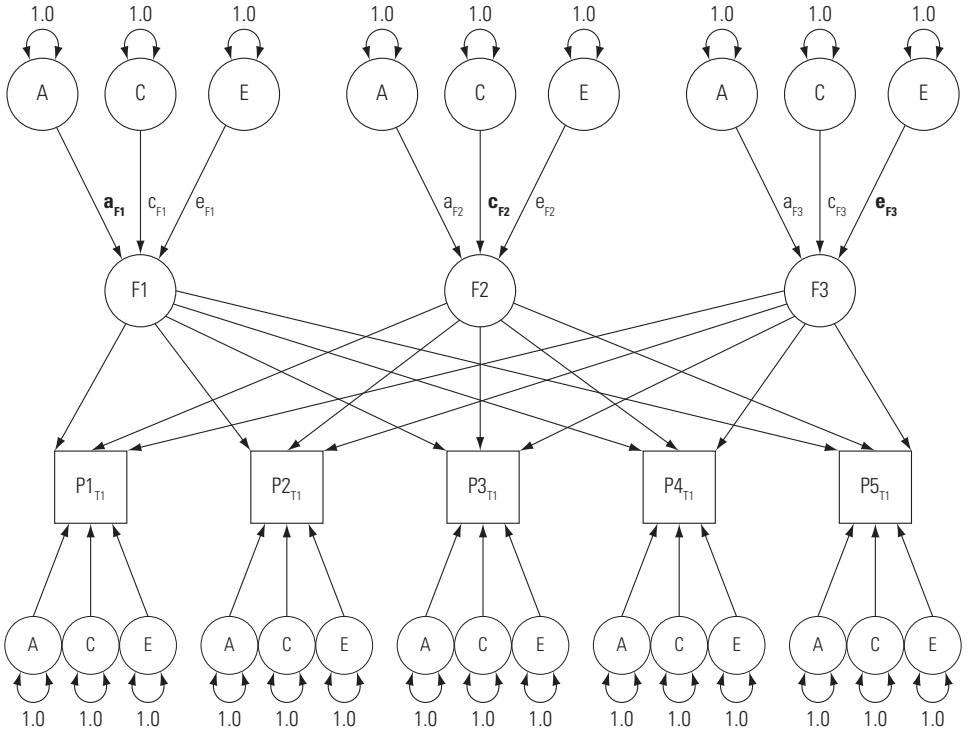
pairs and their children (and possibly their spouses). One specific application relevant to tobacco use research is the comparison of the prevalence of smoking initiation in children and the parent-offspring correlation as a function of the smoking status of the parents: either nonsmoker, former smoker, or current smoker.

Multivariate Factor Model

A basic model for multivariate data collected from twins is shown for one member of a twin pair in figure 6.2. This model, known as a "latent phenotype" or "common pathway" model,[34,35] includes three latent phenotypes (factors) that influence all the observed measures (shown in squares). It is a natural extension of a psychometric common factor model to twin data. All the covariation between twins' items occurs through correlations between the additive genetic (A) and common environment (C) latent variables in twin 1 and their counterparts in twin 2. These correlations are fixed, in accordance with genetic theory, at 1.0 for monozygotic twins for both A and C, and at .5 and 1.0 for A and C, respectively, in dizygotic twins. Note that residual or "measure-specific" covariation between an observed measure and that of the co-twin may occur through the A and C paths shown at the bottom of the figure. Also note that the variance components A, C, and E for factor 1 may correlate with their counterparts for factors 2 and 3. Thus, there is an analog of the oblique factor model in psychometrics.

An important submodel of this three-factor model is one in which the path coefficients

Figure 6.2 Three-Factor Latent Phenotype Model



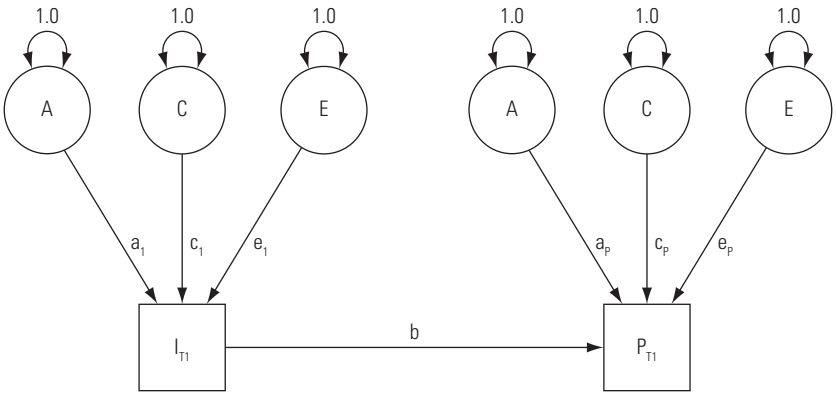
Note. A = additive genetic; C = common or shared environment; E = specific or unique environment; a, c, e = regression path coefficients; F1–F3 = latent phenotypes; P1–P5 = observed phenotypes; T1 = twin 1.

a_{F2}, a_{F3}, c_{F1}, c_{F3}, e_{F1}, and e_{F2} are fixed to zero, and the coefficients (a_{F1}, c_{F2}, and e_{F3}) are fixed to unity. This submodel, known as the "independent pathway" or "biometric factor" model,[34,35] estimates loadings from variance components that are specified as having only one source of variation. Another important submodel is one in which only one latent phenotype is specified; that is, *F2* and *F3* are omitted. This model is often called the "common pathway" model, because the genetic and environmental components (at the top of the diagram) are combined into a latent common factor before they affect the measured variables.

Causal Contingent Common Pathway Model

To analyze contingent data, such as the presence of symptoms of nicotine dependence, for which nicotine use is prerequisite, a simplified bivariate model is used. The model is necessarily simplified because not all the data that would ordinarily identify a multivariate model for twin data are available. There is no information on the within-person covariance between initiation and progression because there is no variation in initiation when progression is observed. However, it is possible to estimate the strength of this relationship in twin data because the co-twin data provide a proxy form of information about the relationship between progression and initiation. Thus, this information comes from discordant pairs; that is, one twin progresses from initiation but the co-twin does not progress. The diagram in figure 6.3 shows progression regressed onto initiation. Each variable has its own A, C, and E components, which are specific to either initiation or progression. All covariance between these two variables is assumed to arise via the regression path. This model has a number of extensions, including the multivariate case and more than two-stage phenomena.[36,37] The key here is the use of twins to overcome the problem of systematically missing data, which is exploited in the application below in the section "Item Response Theory Approach: Application to Virginia Twin Registry Data." Ordinarily, it is not possible to identify the loading of a binary initiation variable on a common factor when the remaining items that load on the factor (e.g., measures of dependence) are contingent on it. However, when data are collected from twins, and when the factors correlate, the model is identified.[38]

Figure 6.3 Causal Contingent Common Pathway Model



Note. A = additive genetic; C = common or shared environment; E = specific or unique environment; a, c, e, and b = regression path coefficients; I = initiation; P = progression; T1 = twin 1.

Item Response Theory

In both clinical practice and research, it is common practice to collect data on the presence or absence of multiple symptom criteria for a given disorder or trait. The item-level data are often collapsed into either a single affected versus unaffected classification or summarized into a score by summing the endorsed symptoms. For example, the Fagerström criteria are widely used to provide an nicotine-dependence score (either FTQ or Fagerström Test for Nicotine Dependence [FTND]) or a binary dependence diagnosis. Both types of summary statistic have problems. In the binary classification case, much of the available information is not utilized. The sum score approach assumes that the criteria are equally important. These assumptions can be tested in an item response framework. In addition, one can evaluate the role of potential covariates, such as gender and age, on the measurement of the phenotype of interest. For genetic studies, failure of these assumptions can lead to erroneous conclusions about the genetic architecture of the trait of interest. Lubke and Neale[39] noted that studies of genotype by environment interaction, including genotype by age and genotype by gender interactions, are subject to potential confounding of measurement artifacts when sum scores or diagnoses are analyzed. Since changes in heritability over time or across groups are fundamental to the genetic analysis of trajectories, it is a crucial first step to assess whether one is measuring the same construct at different times and with the same accuracy. Therefore, these methods are applied in the section below, "Item Response Theory Approach: Application to Virginia Twin Registry Data."

Genetic Latent Growth Curve Models

Of particular interest to those studying trajectories of tobacco use is the information provided by structured latent growth curve (LGC) models. These models are described in detail in chapter 5. Their treatment here is brief and focuses on their extension to genetically informative data,[40] nonlinear models, and switching. By way of preamble, the authors of this chapter support "putting the individual back in growth curves" as proposed by Mehta and West.[15] That is, arbitrary categorization of subjects into age bands (e.g., 10 years old, 11 years old) should be avoided if possible, and the analysis should proceed at the raw data level with subjects' actual ages at testing. This method eliminates the biases that can accrue when there is variation between subjects' ages at a particular occasion of measurement. The LGC model is essentially a factor model with some specific restrictions. In the linear case, it is hypothesized that there is random variation in two factors: initial level and slope. These factors may be correlated. A natural extension of this model for genetically informative data is, therefore, to apply the variance and covariance partitioning to these latent variables. Thus, with twin data, one expects the variation in the level and slope factors to arise from the action of genetic and environmental (C or E) factors, and the covariance between level and slope can be partitioned in the same way. In addition, one can partition the residual occasion-specific variance into the three usual A, C, and E sources. Note, however, that this departs from the idea that the residual variance is purely random measurement error; such a model would eliminate the A and C components and may be fitted to explicitly test this hypothesis. Initial attempts to model growth data collected from relatives[41] used a two-stage approach in which individual growth curves were estimated (to obtain person scores for level and slope), followed by biometric analysis of the scores themselves. This is a practical approach that is suitable when all subjects are measured at equal intervals. However, when data are missing or there is variation in the intervals between measurements, the individual growth curves will vary in their accuracy. The initial summary step does not

capture these differences in precision and therefore may yield biased or inaccurate estimates of the biometrical parameters. It is for this reason that one should prefer, whenever possible, single-step analysis of what has been measured.[42]

A critical issue in LGC modeling is the assumption that growth is linear. While it is likely that the majority of variation in many traits will be captured by this component, it is unlikely to be the case for all traits, especially those measured over a wide range of ages. This was recognized as early as the eighteenth century by Malthus,[43] who developed mathematical equations for alternative growth curves. Fundamental work by Browne[44] provided methods to fit such nonlinear growth curves to data. Perhaps due, in part, to limitations of some of the popular software packages, such nonlinear growth curves have not proved popular.[44] Fitting such models is not technically difficult, even for the case of data from relatives[45] with very long time series; these are typically handled by using time series analysis.[46] Ecological momentary assessments, which may contain thousands of repeated measures for each individual, present an obvious technical challenge. Research in this area typically extracts summary statistics in a two-stage approach. While practical, there may be unwarranted or undesirable assumptions in such an approach, which a more direct analytic method could avoid if it became practical.

The assessment of tobacco use patterns is no different from most other behavioral and psychological domains in that it typically begins with a collection of binary or ordinal items. The analysis of such measures represents a serious challenge for growth curve modeling because the methods were developed for continuous data. Two main challenges present themselves. One is that computing the likelihood of ordinal data is typically done by integrating the multivariate normal distribution. In a growth curve model with m occasions of measurement, multiple integrals must be computed over as many dimensions as there are occasions, which becomes computationally intensive with more than 10 dimensions. This problem is more acute with data from relatives; pairs of twins doubles the number of dimensions of integration, and larger pedigrees (e.g., size f) further exacerbate the problem to mf integration. Worse still, when measures of dependence are being derived from a set of p items, mpf dimensional integration is needed. It is nonetheless possible to apply models for both mean and covariance structure (of which LGCs are an example) to ordinal data, as described by Mehta and colleagues[3] and Wirth and Edwards.[47] The second key issue in the analysis of multivariate data (such as a measure derived from a number of questionnaire items) is that it is very important to assess measurement invariance.[5] Analysis of sum scores could provide misleading results influenced by variance specific to any of the items rather than by the factor itself. Conversely, analysis of individual items subsumes factor variance and item-specific variance for which patterns of familial resemblance (and relative magnitude of variance components) may differ.

An addition to modeling of growth curve mixture models is the notion of switching[48] in which individuals may belong to different trajectory groups at different times. The specification of these models is not straightforward, because it is necessary to consider all possible latent states in which an individual might be at each occasion of measurement. With r trajectory classes and s occasions of measurement, there are r^s possible states for each individual and, thus, r^s components to the mixture distribution. The situation is exacerbated when the model is extended to data collected from pairs of relatives in that r^{2s} components are required. One may therefore envisage analysis of relatively few occasions of measurement

with this approach. Nevertheless, this approach has some attraction for the analysis of data on nicotine use. Transitions between user and nonuser classes are of key importance in the study of the uptake and cessation of tobacco use. In future work, it is hoped to extend the model to the genetic epidemiology of the probability of transitions between different latent states.

Genetic Latent Class Models

Historically, latent class models and factor models developed separately. Factor models can be traced to the work of Spearman.[49] Latent class analysis was developed in the mid-twentieth century.[50,51] Although its use has been less widespread than that of latent trait models (which have been very popular for the last 20 years), it is still a popular method.[52] Under certain circumstances, latent class models and factor models are equally able to account for mean and covariance structure;[53] they have distinct conceptual frameworks and can be distinguished by analysis of raw data.[39] In the latent class model, the population is regarded as a mixture of subgroups, whose item response probabilities (or item means and variances in the continuous case, known as the latent profile model) vary between the groups. Within each subgroup, the items are specified to be uncorrelated (the assumption of *conditional independence*). The model is one example of a finite mixture model;[54] along with other such models, it is becoming popular in many areas.

Eventually, structural equation models and latent class models were combined in a single comprehensive model.[55–60] Slightly different combined models have been proposed with names including "finite mixture structural equation model," "mixtures of conditional mean- and covariance-structure models,"[55] and "finite mixture confirmatory factor models."[58] In what follows, the combined model is referred to as the "factor mixture model" (FMM). The FMM features two types of latent variables—namely, a latent class variable and one or more continuous factors within each class. The continuous factors have several observed indicators (e.g., items of a questionnaire), which can be binary, ordinal, or continuous. The FMM is therefore a model for multivariate data. Muthén and Asparouhov[40] describe application of an FMM to data collected from twins. These models may be fitted with either Mx or Mplus.

Several genetic latent class models were described by Eaves and colleagues.[61] In these models, the conditional independence assumption is retained, both within individuals and across relatives. Complexity arises in the modeling of familial resemblance for class membership. Several choices are possible. A simple Mendelian model of a diallelic major locus that controls class membership (*AA* versus *Aa* versus *aa* genotypes corresponding to three latent classes) would yield a pattern of identical class membership for monozygotic twin pairs with frequencies p^2, $2pq$, and q^2, where $p = 1 - q$ is the frequency of allele *A* in the population. The dizygotic proportions of class membership are more complex, involving pairs discordant for class membership, but are straightforward to derive. It is also possible to construct a two-class concatenation of this single locus model, where genotypes *AA* and *Aa* are both associated with class 1, while *aa* is associated with class 2. Eaves and colleagues also describe more complex models that specify a binary environmental factor that interacts with the major locus to generate four classes. The environmental factor is allowed any degree of association between relatives, according to the pattern

$$\begin{matrix} \alpha^2 + \delta & \alpha\beta - \delta \\ \alpha\beta - \delta & \beta^2 + \delta \end{matrix}$$

where $\alpha = 1 - \beta$ is the frequency of the first environmental condition, and δ is

the association parameter for familial resemblance, which has to satisfy the range constraint

$$0 < (\alpha\beta - \delta) / [(\alpha^2 + \delta)\ (\beta^2 + \delta)]^{½} < 1$$

Such nonlinear inequality constraints are easily specified in Mx or MPlus, although no implementation of the model was found. Other specifications of familial resemblance for class membership are possible. Gillespie and Neale[62] described a finite mixture distribution model for genotype by environment interaction in which a major locus, a continuous threshold model, a shared environment, or a nonshared environment factor controlled group membership. This area is underdeveloped in genetic modeling, particularly in view of developments such as growth curve mixture modeling.[48,56,63,64]

Molecular Genetic Analysis

Linkage Analysis

The focus of structural equation modeling of data has largely been on testing the significance and quantifying the contributions of genetic and environmental latent sources of variance to individual traits or the comorbidity of traits. This is referred to as either "basic" or "advanced genetic epidemiology." The 1990s saw a huge upswing in the analysis of data collected from molecular genetic studies, which continues to increase to this day. These studies attempt to establish whether measured specific genetic variants contribute to variation in the trait of interest and thus identify the actual genes involved. There are two main types of molecular genetic study: linkage and association. Linkage analysis uses related individuals to evaluate the correlation between similarity at a genetic locus with similarity of the trait value. Association studies mostly employ unrelated individuals and compare the frequency of genetic variants at a locus in cases and controls. Typically, these analyses are repeated for a range of locations across the genome, either using a candidate gene approach or by scanning the genome. While traditionally a limited set of markers across the genome was included, genome-wide association studies now employ chips with a million loci. This section describes in brief the connection between structural equation modeling and linkage analysis, setting the stage for the integration of models for gene action with growth curves or stages of tobacco use trajectories.

Linkage analysis is closely analogous to the analysis of twin data. In practice, the molecular biologist assays several markers along the genome. Originally, these markers were chosen to be highly polymorphic, such that there were some 15–20 different alleles at a "microsatellite" locus, and some 300–400 loci were placed at approximately equal intervals along the genome. Today, a larger number of two-allele loci are usually assayed, using single nucleotide polymorphism (SNP) genotyping technologies. In either case, the idea in linkage analysis is to assess how many alleles a pair of siblings (for example) share at a particular location along the genome. Sib pairs can then be classified into those sharing zero, one, or two alleles identical by descent (IBD) at the locus. The possible IBD configurations for sib pairs can be tabulated by labeling parents' alleles as *AB* for the father and *CD* for the mother.[65,66] Their possible offspring are *AC, AD, BC,* and *BD;* the possible pairwise combinations of these offspring are shown in table 6.1. The cells of this table indicate the number of alleles shared IBD by each of the 16 possible sib pair types. Since each combination is expected to be equally frequent, the expectation is that one-fourth of the pairs will be IBD 2, one-half will be IBD 1, and one-fourth will be IBD 0.

Detection of linkage occurs when IBD 2 pairs are more similar than IBD 1 pairs, who in

Table 6.1 Number of Alleles Shared Identical by Descent for a Pair of Full Siblings

	AC	*AD*	*BC*	*BD*
AC	2	1	1	0
AD	1	2	0	1
BC	1	0	2	1
BD	0	1	1	2

Note. Parental genotypes are *AB* and *CD*.

turn are more similar than IBD 0 pairs. This is very much the same as the twin study apart from three important exceptions. First, rather than fitting an ACE model, an estimate is made of the contributions to the variance of the genetic variants at a specific locus, the quantitative trait locus (QTL), the residual familial factors (F), and unique environmental factors (E), sometimes referred to as the QFE model. Second, the IBD 0 pairs correspond to unrelated pairs, such as adopted children reared in the same family. Third, the information about IBD sharing is imperfect because the markers do not unambiguously classify sib pairs into those sharing 0, 1, or 2 alleles IBD. There are two main approaches to overcoming this limitation. One is to use an estimate of π, the proportion of alleles shared IBD, and π is specified as the covariance between the variance component that represents the effect of the QTL. Alternatively, and mathematically more consistent, the imperfect classification can be represented as a mixture distribution.[67,68] The likelihood of a sibling pair's phenotypes can be written as the weighted sum of three likelihoods: the IBD status is zero, one, or two. In either approach, one uses the definition variable approach described above in the subsection on "Structural Equation Modeling" to specify the model. The significance for linkage is evaluated by the logarithm of odds (LOD) score, a statistic that represents the likelihood of the odds of linkage over the odds of no linkage. Criteria have been established to classify results as suggestive, significant, or confirmed evidence for linkage.[69]

In practice, most linkage analysis is conducted with specialized software such as Merlin[70] or GENEHUNTER.[71] However, such programs are designed for the analysis of a single trait. Fortunately, they permit export of IBD probabilities that can be used in other software for modeling multivariate or longitudinal data or simply modeling traits that are assessed by using a collection of binary or ordinal items. For example, it is possible to conduct a linkage scan for quantitative trait loci that cause variation in level or slope of a growth curve model.

Association Analysis

Association analysis is in principle simpler than linkage analysis in that it can be conducted with groups of cases and controls. Conceptually, the idea is to compare between groups the allele frequency at a particular locus. From a statistical point of view, this is a simple comparison that can be conducted using a χ^2 test. However, certain pitfalls have the potential to generate false positives or false negatives. One is population admixture in which there exist two or more subpopulations whose allele frequencies differ and whose trait mean values differ for entirely different reasons. Several approaches exist to control for such admixture (or stratification). One is to obtain a set of alleles in noncoding regions of the genome to assess whether there is stratification.[72] A second is to use data collected from relatives.[73] Since families come from the same stratum of the population, any allele-phenotype association observed within families cannot be due to population stratification. An additional advantage of the family-based research design is that it permits joint analysis of linkage and association information, which in turn assists with fine mapping of quantitative trait loci.[74]

There is much focus on genome-wide association studies, which have become practical to conduct with the advent of inexpensive SNP chips. These microarray

chips permit the assaying of a very large number (500,000, for example) of SNPs across the genome. Such density permits exploitation of linkage disequilibrium in which short strands of DNA are transmitted intact with low chance of recombination. Thus, it becomes possible to identify very small regions likely to contain a polymorphism that accounts for variation in a trait. Much of the software development in this area is targeted at the rapid analysis of this large number of data points and with handling the high type 1 statistical error rates that ensue. Redden and Allison[75] note that assortative mating can increase the risk of type 1 error in association studies. Again, the focus is single trait oriented rather than multivariate. However, it has been noted that association data have considerable potential to resolve alternative pathways between phenotypes.[76] The integration of association data into a more sophisticated modeling framework is straightforward in principle, but much remains in the way of opportunities to develop and test the models. For example, in a latent growth curve mixture model, one might specify that alleles at a locus affect the mean of the level or growth factors. Alternatively, one might specify that an individual's class membership probabilities are a function of genotype. Thus, one would explicitly model the allele effects as another model parameter. A simpler two-step approach would be to assess allele frequencies between those classified as belonging to one or another class. This latter method would have the advantage of analytic simplicity at the cost of losing the information about the precision of the class membership classification.

Review of Genetic Studies of Smoking

The role of genes and environment in initiating smoking has been the subject of a growing number of twin and family studies and several reviews.[77,78] Evidence from these studies generally points to an important role of genetic factors in explaining individual differences in starting to smoke. In addition to additive genetic factors, however, shared environmental factors also contribute significantly to the variation, especially in adolescent samples. The literature is reviewed here from a range of perspectives, with the aim of providing a better understanding of the process of developing the smoking habit and subsequent dependence. As shown in epidemiological studies, approximately 50% of the individuals who start to smoke continue to do so and go on to become dependent on nicotine. Prevalence rates for adults in 2006 suggest that 42% have ever smoked in their lifetime and 24% of men and 18% of women still were smoking.[79] One obvious question is whether the factors that lead to individuals starting to smoke also contribute to whether they persist in their smoking behavior. First, this section reviews the most prominent twin studies on adolescent smoking. Second, additional information is considered that can be obtained from extending the classical twin design to other relatives—for example, parents, siblings, and spouses. Third, the focus is on studies that have included measures of smoking initiation and progression to discern the role of genes and environment to the different stages of the smoking process. Finally, molecular studies of adolescent smoking are reviewed to show how the direct assessment of molecular genetic polymorphisms can enhance understanding of the trajectory from initiating the smoking habit to nicotine dependence.

Twin Studies of Adolescent Smoking

Eight published papers were identified that report results from twin studies on smoking behavior in adolescence. The first, by Boomsma and colleagues,[80] reported data on 1,600 Dutch adolescents aged 13–22 years, concluding that the majority

of interindividual variation in smoking behavior was due to shared environmental factors (59%), with 31% attributed to genetic factors. Results, however, were not consistent across age groups, with heritability estimates decreasing with age in males but increasing in females. When age was included in the analyses, 9% of the variance could be accounted for by age, reducing the proportion explained by shared environmental factors to 50%. Furthermore, the shared environmental factors differed between males and females (correlation between shared environmental factors, $r_C = .65$). A follow-up study including more than 2,600 pairs of Dutch adolescents[81] showed a more consistent trend of an increasing role of genetic factors in smoking behavior from ages 12 to 22 years, with a corresponding decline in the contribution of shared environmental factors. Up to age 17, heritability was not significantly different from zero. However, 33% (95% confidence interval [CI], 31%–54%) of the variance was attributed to genetic factors in young adult females, and 66% (95% CI, 43%–86%) in males. Again, shared environmental factors, which accounted for the majority of the variance in adolescence, appeared partially different for males and females. Similar results were obtained in a sample of 1,419 16-year-old Finnish twin pairs.[82] Shared environmental factors accounted for the majority of the variance in smoking behavior (having smoked 50 cigarettes or more)—75% in males and 63% in females. Heritability was estimated at 17% and 30%, respectively. In analyzing FinnTwin12 smoking data from twins and their classmate controls, heritability (h^2) was 11% and shared environment could be split into familial influences (49%) and school-based neighborhood effects (24%).[83,84]

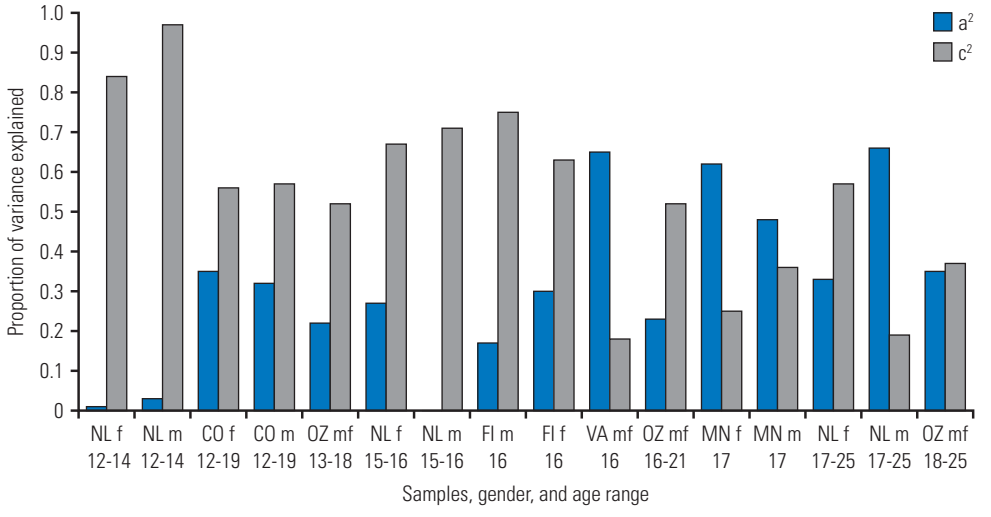
Data from 16-year-old twins ($N = 159$) studied in the first wave of the Virginia Twin Study of Adolescent Behavioral Development[85] suggested that additive genetic factors accounted for 65% (95% CI, 10%–93%) of the variance in liability to lifetime smoking and 60% (95% CI, 0%–93%) for current tobacco use, with nonsignificant contributions of shared environmental factors (18% and 21%, respectively). While the prevalence of smoking was statistically different for males and females, the contribution of genetic and environmental factors did not differ by gender. Gender differences were also not statistically significant in analyses of 500 17- to 18-year-old twin pairs from the Minnesota Twin Family Study,[86] resulting in estimates of 36% for the heritability of tobacco use and 44% for shared environmental factors. When analyzed separately by gender, the predominant source of variance was genetic (59%) for males and shared environmental (71%) for females. An updated report[87] on a slightly larger sample ($N = 626$) with primarily additional female twins showed a heritability of 56% for ever having used tobacco and a smaller contribution of shared environmental factors (30%). Genetic factors were the predominant source of variance in males (48%) and females (62%). The contributions of both genes (38%) and shared environment (52%) were significant in data from 682 twin pairs (306 biological siblings and 74 adoptive sibling pairs) aged 12 to 19 years assessed by the Center for Antisocial Drug Dependence in Colorado.[88] Again, gender differences were not statistically significant. The slightly larger role of the shared environment is consistent with the inclusion of younger adolescents.

Shared environmental factors were also the predominant (52%) source of variation in smoking initiation in a sample of 414 same-gender twin pairs aged 13–18 years from the Australian Twin Registry (ATR).[89] Heritability reduced from 22% to zero when the model was adjusted for smoking by peers and parents. While shared environmental factors were more important in males, and genetic factors accounted for the largest proportion of variance in females, the gender difference

was not significant. Data from two follow-up waves showed a gradual shift from shared environmental to genetic influences, with each accounting for about 35% of the variance in 18- to 25-year-olds, consistent with results from other studies. A 2005 study, using a genetically informative subsample of 2,142 sibling pairs, aged 11–20 years, participating in two waves of the National Longitudinal Study of Adolescent Health,[90] presented heritability estimates for smoking frequency of 52% and between 28% and 35% for high levels of smoking frequency. The role of shared environmental factors was greater for high levels of smoking frequency (25%–38%) than for overall smoking frequency (7%). In contrast to previous studies, smoking frequency was used rather than a measure of smoking initiation.

As discussed in a review of twin and adoption studies of adolescent substance use,[91] shared environmental influences appear stronger in younger adolescents, whereas genetic influences are more substantial in older adolescents and young adults (figure 6.4). However, it is possible that the earliest stage of cigarette smoking (i.e., first experimentation) is mostly due to environmental factors, whereas later stages (conditioned on previous exposure to nicotine) are more likely to be due to genetic factors. The issue with many studies is that the presence of smoking initiation is determined by a somewhat vague item, such as, "Have you ever smoked?" It is possible that this question is more likely to be interpreted by a younger adolescent (i.e., one closer to first exposure to cigarettes) as "whether they've ever tried even a single cigarette," while by older adolescents or young adults as meaning onset of regular smoking. Studies that use such a vague measure of initiation with a broad age range may inadvertently be measuring different behaviors (i.e., different stages of the smoking habit) in early adolescents compared with those subjects

Figure 6.4 Estimates of the Contributions of Additive Genetic (a^2) and Shared Environmental (c^2) Factors to Smoking Initiation by Sample, Age, and Gender in Published Studies of Adolescent Twins



Note. NL = Netherlands Twin Register; f = female; CO = Center for Antisocial Drug Dependence in Colorado; m = male; OZ = Australian Twin Registry; FI = Finnish Twin Registry; VA = Virginia Twin Study of Adolescent Behavioral Development; MN = Minnesota Twin Family Study.

in young adulthood. It should also be noted that these comparisons are based on estimates from studies with varying sample size and thus varying precision. Ideally, a meta-analysis should be undertaken that appropriately accounts for these differences. Alternatively, a mega-analysis that combines the raw data of several related studies would provide more accurate estimates of the potentially changing role of genes and environment across adolescence.

Extended Genetic Epidemiology

Extended Twin Family Studies of Transmission of Smoking

Relatively few studies have either included or analyzed data collected from other relatives. Rhee and colleagues reported results of fitting a model to data on twins, nontwin siblings, and adoptees.[88,92] The major finding was that the proportions of variance associated with the special twin environment and with genetic dominance were small. Data from the Netherlands Twin Register suggested significant assortment between spouses, with the correlation between husband and wife for "currently smoking" larger than for "ever smoking."[80] Furthermore, there was no evidence that parental smoking encouraged smoking in their offspring, as resemblance between parents and offspring was significant, but rather low, and could be completely accounted for by genetic relatedness. If included, cultural transmission estimates were negative. Similar results were obtained for data on twins and their parents from the Finnish Twin Registry,[82] showing significant assortment (husband-wife correlation = .42), and low but significant parent-offspring correlations.

Nongenetic analyses of data on 3,906 twins confirmed significant associations between the smoking behavior of the twin with that of the co-twin. Odds ratios, ranked highest to lowest, were given when an individual had a smoking monozygotic co-twin, a smoking same-gender dizygotic co-twin, or a smoking opposite-gender dizygotic co-twin—suggesting a role for both genetic factors and gender. In addition, associations were also significant for smoking behavior of parents, siblings, and friends, and were gender dependent (stronger associations for same-gender smoking family members).[93,94] In fact, the risk to initiate smoking when having friends who smoke was similar to that of having a smoking co-twin and greatly exceeded that of having a parent who smokes.

Similarly, data from the Virginia 30,000 Study, including about 15,000 twins and their first degree relatives (parents, siblings, spouses, children), showed little evidence for the role of parents in influencing the smoking behavior of their children through other than genetic pathways.[95,96] Analyses of these extended twin kinship data supported the role of additive genetic factors, accounting for more than one-half of the variance in smoking initiation, partly due to the consequences of assortative mating, which was highly significant. About 20% of the variance was accounted for by specific environmental factors. Furthermore, the contributions of shared environment and special twin environment were both significant. The environmental paths from the parents to their children were estimated to be negative, but this was not significant. Note that these analyses were based on data from different generations of adults and should ideally be performed on data sets of adolescent twins augmented with parents.

Multivariate Genetic Studies

Only a few studies have investigated whether the same genetic or the same environmental factors account for the co-occurrence of several smoking behaviors. Genetic analyses of data from young adult Australian twins[97] reporting any cigarette use were undertaken to examine whether there are genetic

factors specific to nicotine withdrawal after controlling for factors for smoking progression and quantity smoked. Significant genetic overlap was found for smoking progression, quantity smoked, and nicotine withdrawal, but evidence for specific genetic influence to nicotine withdrawal remained. An extension of the causal contingent common (CCC) pathway models (see also the subsection below, "Progression from Smoking Initiation to Nicotine Dependence") was used to explore the interrelationship of smoking age at onset, cigarette consumption, and smoking persistence.[98] Smoking initiation was operationalized as an ordinal variable with three categories—nonsmokers, late-onset smokers, and early-onset smokers—assuming a single underlying distribution and thus referred to as age at onset. This allows the authors to fit a full multivariate model, rather than the CCC pathway model, according to Heath and colleagues,[99] and partition both the variation and covariation into genetic and environmental contributions. The authors found significant heritability for all three phenotypes in males and females and slightly higher genetic correlations in males than in females. The relationship of smoking age of onset, cigarette consumption, and smoking persistence was also mostly due to shared genetic influences. A similar analysis of age at initiation, amount of smoking, and smoking cessation was done on data from adult Finnish twins.[100] The study found that genetic factors were important in amount of smoking and smoking cessation, but these were largely independent of genetic influences on age at initiation.

Progression from Smoking Initiation to Nicotine Dependence

Most individuals who initiate smoking progress to regular smoking, and many become dependent on nicotine.[79] It is, therefore, important to evaluate whether the same factors influence whether someone starts to smoke and whether one continues to smoke. Reports that analyze measures of persistence or dependence without taking initiation into account assume that the dimensions underlying initiation and progression are independent (if only smokers are included) or assume that persistence is an extreme version of initiation on the same single liability dimension (if nonsmokers are included but score zero on the progression measures). Heath and colleagues[101] recognized this and developed alternative models to test these assumptions. First, studies are reviewed that estimated the role of genes and environment on the measure of dependence without taking initiation into account. McGue and colleagues[87] reported no gender differences in the role of genetic and environmental factors for nicotine dependence in a sample of 626 17-year-old twin pairs, with genes accounting for 44% and shared environment for 37% of the variance. Although Rhee and colleagues[88] found no significant gender differences for initiation, shared environmental factors were significant for tobacco use and problem use in males but not in females, explaining 45%–48% of the variance in a sample of more than 1,000 twins and siblings. Heritability estimates were 24%–26% in males and 95% in females, respectively.

As far as known, only one study has simultaneously analyzed data on smoking initiation and persistence in a juvenile sample. Koopmans and colleagues[102] published analyses from 1,676 Dutch adolescents. They found separate smoking initiation and quantity dimensions, which were not completely independent. The total heritability of quantity smoked was estimated at 86%. Five studies were found of smoking initiation and progression in adults. Data from 4,000 male twin pairs from the Vietnam Era Twin (VET) Registry[103] found that genetic and shared environmental factors accounted for 50% and 30%, respectively, of the variance in liability to initiate smoking. However, no evidence for shared environment was found for factors specific to persistence,

for which variation was estimated to be 70% additive genetic. Significant heritability for nicotine dependence (60%) was also found in a follow-up study of 3,356 male VET Registry pairs.[104] Using nonmetric multidimensional scaling, Heath and colleagues[105] found that the etiologic factors that determined which individuals were at risk of becoming smokers differed from those that influenced age of smoking initiation. The role of genes and shared environment in the onset of smoking differed by cohort and gender, and only genetic factors accounted for twin resemblance in the age at which smoking onset occurred.

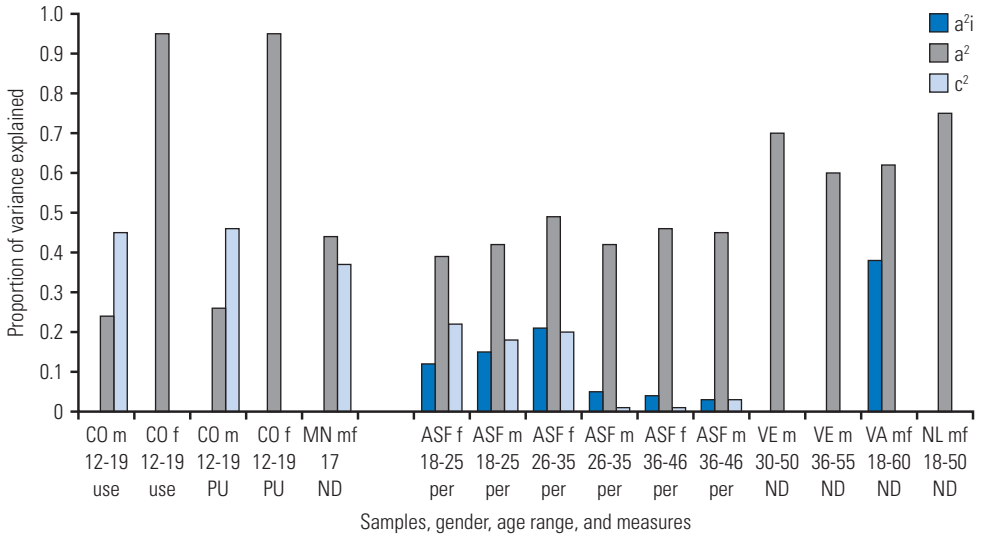
As described above, Kendler and colleagues[1] developed a model that estimates the correlation between liability to smoking initiation and liability to nicotine dependence and applied it to data on 1,898 female twins from the Virginia Twin Registry. Results indicated that etiological factors that influence initiation and dependence, while overlapping, are not perfectly correlated. Thus, genetic factors contributed 72% to variance in liability to nicotine dependence, of which 69% also influence initiation and 31% are unique to nicotine dependence. Madden and colleagues[106] fitted a similar correlated liability dimensions model to data from large samples of male and female same-gender twins from three countries—Australia (1,535 pairs), Sweden (5,916 pairs), and Finland (4,438 pairs)—further subdivided by age bands. The authors also found that familial influence on risk for persistence in smoking cannot be entirely explained by the same factors responsible for risk of smoking initiation. Total genetic variance for smoking persistence ranged between 39% and 48% in women and 42% and 45% in men, of which only 7%–35% was accounted for by factors in common with initiation. Although shared environmental factors contributed significantly to smoking initiation, there were no significant additional shared environmental contributions to smoking persistence.

Maes and colleagues[37] extended the liability models to include smoking initiation, regular tobacco use, and nicotine dependence, and applied them to data on both female and male twins from the Virginia Twin Registry. Results showed that the liabilities to all three stages of smoking behavior were correlated, with 80% of the variance in liability shared between initiation and regular use, and 50% between regular use and nicotine dependence.[37] The heritability of nicotine dependence was estimated at 62%, of which 24% was specific to nicotine dependence, 10% shared with regular tobacco use, and the remaining 28% shared with smoking initiation as well. Data on 1,572 Dutch adult twins also showed that the smoking initiation dimension is not independent from the nicotine-dependence dimension.[107] As shown in other data sets, shared environmental factors contributed significantly to the variance in liability to initiation but not to nicotine dependence, which was strongly (75%) influenced by genetic factors (figure 6.5).

Assessment of Nicotine Dependence

Almost all studies that measure nicotine dependence use either a sum score or a binary diagnosis. The most widely used measures are the FTQ[8] items and the criteria based on the *Diagnostic and Statistical Manual of Mental Disorders (DSM),*[108] either of which may be dichotomized by imposing a threshold for affection status. The latter approach reduces the information available, which, in genetic studies, would typically result in reduced statistical power.[26,109] Using sum scores assumes that the scale of measurement is invariant and that the underlying liability is unidimensional. The FTQ correlates with other proposed measures of nicotine dependence such as carbon monoxide, nicotine, and cotinine levels.[8] However, the nicotine rating and inhalation items were found to be unrelated to biochemical measures, and a revised scoring was proposed, the FTND.[110] Both

Figure 6.5 Estimates of the Contributions of Additive Genetic Factors in Common with Initiation (a^2i), Total Additive Genetic (a^2), and Shared Environmental (c^2) Factors to Smoking Persistence by Sample, Gender, Age, and Measure of Persistence



Note. The first five studies do not take initiation into account. CO = Center for Antisocial Drug Dependence in Colorado; m = male; MN = Minnesota Twin Family Study; f = female; ASF = Australian, Swedish, and Finnish Twin Registries; VE = Vietnam Era Twin Registry; VA = Virginia Twin Registry; use = tobacco use; PU = problem tobacco use; per = persistent tobacco use; ND = nicotine dependence.

the FTQ and the FTND were highly reliable, and internal consistency was greater for the FTND than for the FTQ.[111] Retrospectively assessed FTQ-FTND scale scores also have acceptable reliability.[112] Furthermore, the six FTQ items were positively correlated with cotinine values in adolescent smokers.[113]

Several studies have attempted to evaluate the dimensionality of nicotine dependence using factor analysis of either the FTQ or the FTND items. An exploratory factor analysis of FTND in young adult smokers resulted in two factors.[114] The first factor, labeled "smoking pattern," included items assessing the number of cigarettes smoked per day, time to first cigarette, difficulty refraining from smoking, and smoking when ill. The second factor, labeled "morning smoking," consisted of two items measuring whether one smokes more in the morning and whether the first cigarette is most satisfying, Confirmatory factor analysis, however, only confirmed the first factor. Similar factors resulted from an exploratory factor analysis of FTND in an adult sample, with a third factor related to the brand of cigarettes when all eight FTQ items were included.[115] These analyses were repeated in the drug abuse patient sample, with similar results, except that time to first cigarette loaded on both factors.[116] Factors were named "persistence in maintaining nicotine levels during waking hours" and "urgency in restoring nicotine levels after nighttime abstinence." A confirmatory analysis in hospital patients confirmed that the items of the FTND were best modeled as two correlated factors with a cross-loading.[117] Furthermore, a four-item single factor ("daytime smoking factor") fitted the data reasonably well. This confirms previous studies showing that both the four-item and the Heavy Smoking Index

(based on two FTND items: the number of cigarettes smoked per day and the time to first cigarette) represent the FTND well.[118,119] Few studies have compared the questionnaire-based FTQ-FTND measures with those based on structured interviews (i.e., *DSM, the International Statistical Classification of Diseases and Related Health Problems*), and have found only moderate concordance,[120,121] which may indicate that they tap into different aspects of nicotine dependence. Only one analysis was found that was based on factor analysis/item response of nicotine-dependence measures using a genetically informative sample. In a genetic factor analysis of nicotine-dependence items measured in adult Australian twins, item covariation was best captured by two genetic but one shared environmental factor for both women and men; however, item factor loadings differed by gender.[122] None of these studies included initiation as an item. Later in this chapter, results are presented from a genetic item analysis of adult Virginia twin data that include both initiation and regular smoking in the analysis. As nicotine-dependence symptoms were only assessed in individuals who had initiated smoking and become regular smokers, it is shown here how including these conditional items affects the estimates of factor loadings and thresholds.

Genetic Latent Growth Curves and Latent Class Analysis

Although the epidemiological literature on growth curve and latent class analysis of smoking behavior is rapidly expanding (chapter 5), genetically informative applications of this type of analysis were not identified.

Molecular Genetic Studies of Smoking

Besides the extensive literature on genetic epidemiological studies of smoking behavior and nicotine dependence, the literature on gene-finding approaches for nicotine dependence is increasing rapidly, reflecting the general trend in the genetic analysis of complex traits. The major results from linkage and association studies on smoking behavior and nicotine dependence are briefly summarized below. Given that very few molecular genetic studies of smoking behavior have included adolescent subjects, results are provided from adult samples.

Linkage Studies

In 2003, three linkage scans of smoking-related measures had been completed. Since then, at least seven more have been published and others are under way. The first genome scan was conducted using a sample of 130 sibling pairs concordant for nicotine dependence from Christchurch, New Zealand (CNZ) and a replication sample from Richmond, Virginia.[123] Several publications have resulted from data made available to investigators participating in Genetic Analysis Workshops (GAW). As part of GAW11, data on 105 families from the Collaborative Studies on Genetics of Alcoholism (COGA) were examined for linkage for smoking-related traits, including smoking initiation, and habitual smoking, defined as ever smoking at least one pack (20 cigarettes) daily for six months or more.[124] Data from a genome scan with 330 extended families participating in the Framingham Heart Study (FHS) were made available to investigators participating in GAW13, resulting in several reports on maximum cigarettes per day (maxcig)[125] on a typical day.

Several genome scans have been performed with samples initially selected for phenotypes other than smoking. As part of a Netherlands Twin Study of Anxious Depression (NETSAD) collaborative project, a genome scan was performed on 646 sibling pairs in 212 families for the three smoking phenotypes: smoking initiation, maxcig, and

age of first cigarette.[126] A scan for regular and persistent tobacco use was performed with data collected from a community sample of Mission Indians as part of a larger study exploring risk factors for substance dependence.[127] Another scan was conducted using a Yale University sample originally collected for linkage analyses of anxiety disorders, which also included a measure of cigarette smoking.[128] Similarly, data on tobacco use and nicotine dependence were available for a sample ascertained for affected sibling pair linkage studies of cocaine or opioid dependence.[129] In the latter study, analyses were conducted separately for subjects with European American versus African American descent.

A number of studies have been published on samples specifically ascertained for smoking behavior. A linkage study focused entirely on a sample of African American origin from the Mid-South Tobacco Family (MSTF)[130] cohort, with assessments of tobacco use and nicotine dependence. Swan and colleagues[131] performed a genome-wide screen for nicotine dependence susceptibility loci on tobacco use data collected from families obtained through participants in the Smoking in Families Study (SMOFAM). Saccone and colleagues[132] analyzed a smoking quantitative trait in Australian and Finnish families with at least one heavy smoker. In 2006, the first study in linkage analysis for smoking initiation and cigarette consumption was published that incorporates gender differences by using Australian twin families (ATR).[133] None of the reported linkage scans of smoking-related phenotypes have included data from adolescents.

Published linkage scans have resulted in only a few regions that have exceeded levels of genome-wide significance.[134] Saccone and colleagues[132] reported the largest LOD score (5.98) for nicotine use on chromosome 22q12. The second highest LOD score (4.22) was found for maxcig on chromosome 20 at 72 centimorgans (cM),[132] which replicates an earlier result in the FHS sample.[135,136] A similarly high LOD (4.17) was reported for chromosome 10 between 92 and 94 cM for quantity smoked[130] in the MSTF sample. Furthermore, this result was supported by suggestive linkage in the same location for three other nicotine dependence measures. This region was part of a broader region initially reported by Straub and colleagues[123] for which the highest LOD score (1.28) for nicotine dependence was obtained in the CNZ sample. A modest signal (LOD 2.16) was also found for a location close to this region (80 cM) for FTND in a European American sample.[129] An LOD score of 3.71 was found for smoking rate in the FHS sample on chromosome 11 at 70 cM.[135] This result has not been replicated so far, although a modest LOD score of 1.64 was found at 87 cM for heavy smoking.[124] Suggestive evidence for linkage (LOD = 3.04) was also reported at 95 cM on chromosome 5 for FTND.[129]

At least 12 other 10-cM chromosomal regions contain positive findings from at least two different samples; however, neither reach criteria for significant linkage. For chromosome 5, Vink and colleagues[137] reported an LOD of 2.09 at 205 cM for age at first cigarette in NETSAD, and Saccone and colleagues[125] obtained an LOD of 1.02 at 100 cM for maxcig in FHS. Five reports converged on locations between 50 and 65 cM on chromosome 6 with LOD scores ranging from 1.1 to 3 for different tobacco use phenotypes.[126,127,133,137,138] Four reports converge on an area on chromosome 7 between 140 and 164 cM.[127,131–133] Two regions on chromosome 8 showed some evidence for linkage: one between 24 and 31 cM for maxcig and nicotine dependence in the FHS and SMOFAM, the other between 110 and 115 cM for regular tobacco use in the Mission Indian and FHS samples. The largest region identified, with seven "hits," is in a 25-cM region (91–116 cM) on chromosome 9 for phenotypes ranging from lifetime smoking to nicotine dependence. An additional

region on chromosome 9 (between 165 and 172 cM) also showed modest to suggestive evidence for linkage for ever smoking (COGA sample) and maxcig (in the FHS). Two positive reports were found for an area between 38 and 43 cM on chromosome 11 in the MSTF and the ATR. On chromosome 13 (41–42 cM), two positive linkage signals were found for quantity smoked, one in FHS and the other in MSTF. LOD scores between 1.29 and 3 were reported for the exact same location on chromosome 14 (88 cM) in three independent samples (COGA, NETSAD, and the FHS). Another region with support from at least two samples includes locations 127 and 135 cM on chromosome 15 for smoking rate (FHS) or ever smoking (COGA).[129] Given the range of phenotypes, methods, selection criteria, and sample sizes, the accumulated data have at least identified regions of interest for susceptibility loci for nicotine use phenotypes. Collaborations and meta-analyses might assist in resolving some of these findings.[139]

Association Studies

The number of association studies of candidate genes for smoking initiation and nicotine dependence has grown steadily. A search identified only 10 studies published before 2000 and five or less papers per year from 2001 to 2003. Yet, in 2004, 13 papers were published on the subject, a trend that has continued with 17 papers in 2005 and 15 in 2006, bringing the total to more than 70 articles.

Several reviews have summarized the findings.[140–146] They can be broadly divided into four categories: (1) metabolism of nicotine, (2) nicotine receptors, (3) the dopaminergic reward system, and (4) the serotonergic reward system. Obvious candidate genes are those that influence the metabolism of nicotine, such as the cytochrome P-450 (CYP) system. Interest has focused on *CYP2A6,* which is involved in the metabolism of nicotine to cotinine. At least 10 out of 14 studies show significant associations to smoking behavior, primarily smoking status and quantity, with 3 reporting positive associations with nicotine dependence. The second group of candidates are genes involved in sensitivity to nicotine, the major addictive substance in tobacco. Evidence from mouse knockouts suggests that the gene coding for the nicotinic acetylcholine receptor beta2-subunit *(CHRNB2)* is necessary for the full reinforcing properties of nicotine. Four studies of humans did not find association between *CHRNB2* and smoking initiation or nicotine dependence. However, a nominally significant allelic and genotypic association was found for *CHRNB2* and three other nicotinic cholinergic receptors and smoking initiation.[147] Furthermore, some evidence suggests variation in the *CHRNA4* gene may be associated with reduced risk for nicotine dependence. Two other receptors (*CHRNB1* and *CHRM1*) have also been implicated in the risk for nicotine dependence.

A third group of studies has examined the association of smoking with variations in genes involved in the dopamine system, motivated by findings that the mesolimbic dopaminergic system appears to play a significant role in the reinforcing effects of addictive drugs, including nicotine. A number of studies have examined the association between several aspects of smoking behavior and variants in the dopamine receptors and a repeat polymorphism in the dopamine transporter protein (DAT/SLC6A3). About two-thirds of the findings for *DRD2* suggested an association with smoking status. Evidence for an association of DAT with smoking behavior was even stronger: five out of six reports presented significant positive findings. Analyses of other dopamine receptors *(DRD4, DRD5)* have largely produced nonsignificant results. A number of studies have examined genes related to dopamine synthesis or degradation.

Mostly significant associations have been reported for DOPA decarboxylase (DDC) and dopamine β-hydroxylase (DβH) with, respectively, three out of three and three out of four studies showing significant results. Several studies have examined polymorphisms in the monoamine oxidase *(MAOA, MAOB),* catechol-*O*-methyl transferase *(COMT),* and tyrosine hydroxylase *(TH)* genes with mixed results.

The fourth group of genes examined in association studies of smoking involves the serotonin system on the basis of evidence that nicotine withdrawal may be modulated by serotonergic transmission. The most studied gene in this system is the serotonin transporter *5-HTT,* particularly the functional polymorphism *5-HTTLPR,* which is implicated in alcoholism and major depression. These studies have produced conflicting results, with one-half of the reports indicating a significant association. Variation in another serotonin system gene, *TPH,* has been associated with smoking behavior in three out of five reports. Finally, other genes have been tested for associations with smoking behavior, such as the phosphatase and tensin homolog gene *(PTEN),* and the cholecystokinin gene *(CCK),* but so far these results have not been replicated.

In addition, two genome-wide association studies of nicotine dependence have nominated several novel genes while also identifying known candidate genes.[148,149]

In summary, although this research area is in an early stage, and may be limited by several methodological weaknesses, various trends are starting to emerge. Most studies did not examine nicotine dependence directly; they used smoking status as the outcome. Sample sizes have tended to be relatively modest, the statistical criteria have been liberal, and multiple testing has been common. Therefore, the chance that these findings contain false positive results is high.

Item Response Theory Approach: Application to Virginia Twin Registry Data

This section applies the psychometric factor model to data on nicotine initiation and dependence collected from twins. The model is described above in the subsection "Item Response Theory." These analyses are novel in that initiation and dependence are being analyzed together, exploiting the information on co-twins' dependence as a function of a twin's initiation status. The model tests for measurement noninvariance of nicotine dependence as a function of age and gender and their interaction.

Subjects

Participants in the present investigation were drawn from two longitudinal studies of adult twins, conducted in parallel;[1] the first consisted of female-female twin pairs (FF) and the second of male-male and male-female twin pairs (MMMF). Each sample was obtained from the population-based Virginia Twin Registry, which is now part of the Mid-Atlantic Twin Registry. The first study was of zygosity determination and was based on questionnaire responses and DNA polymorphisms when required.[150] Telephone interviews were collected from 1,846 individuals in the FF study and from 4,959 individuals in the MMMF study. The final sample includes 1,503 monozygotic males, 1,085 dizygotic males, 1,078 monozygotic females, 768 dizygotic females, and 2,371 dizygotic opposite-gender twin pairs.

Measures

Interviews for both the FF and MMMF studies were highly homologous. In the MMMF study, all common forms of tobacco

self-administration (cigarettes, cigars, pipe tobacco, chewing tobacco, snuff) were assessed, whereas FF study participants were asked only about cigarettes. The focus was on tobacco initiation (TI), regular tobacco use (RTU), and items on the modified version of the FTQ.[8] TI was defined according to the responses to the questions "have you ever smoked cigarettes?" and the follow-up query "not even once?" RTU was defined as the use of an average of at least seven cigarettes per week for a minimum of four weeks. Individuals who met criteria for RTU were given the FTQ. This scale consists of eight items; three are scored on a two-point scale (number of cigarettes per day, inhale, nicotine level of cigarette brand) and five on a one-point scale (first cigarette soon after waking, difficulty refraining when forbidden, smoking when ill in bed, smoking most in morning, first cigarette most satisfying). The revised FTND[110] scale includes only six items (inhale and nicotine level were dropped), with two other items scored on a three-point scale (number of cigarettes per day, first cigarette soon after waking).[110] It should be noted that the FTQ and FTND scales are not universally agreed-upon definitions of nicotine dependence, and results obtained with other measures—that is, *DSM* criteria[108]—could vary.

Methods

IRT models were used to estimate parameters that represent the "locations" of items on a latent continuum. The model describes the probability of a discrete response to an item as a function of a person parameter (their location on the latent trait) and one or more item parameters. In the two-parameter case, one parameter represents the location, and in the case of attainment testing, is referred to as the "item difficulty." The second parameter estimates the discrimination of the item—that is, the degree to which the item distinguishes between persons who have different scores on the latent trait. This second parameter characterizes the slope of the item characteristic curve. The difficulty parameter relates to the location of the curve on the continuum. These models can be extended to data on pairs of twins, and the trait variance can be partitioned into sources due to additive genetic, shared environmental, and specific environmental factors. The parameterization of these models is similar to that of the common pathway/latent phenotype model,[20] which allows for variance partitioning at two levels: (1) the latent trait—that is, nicotine dependence, and (2) the residual item variances. The parameters of the genetic IRT model thus include item discrimination parameters (which correspond to factor loadings), item difficulties (which correspond to thresholds), and genetic and environmental parameters of the items and construct. As is typical for twin analyses, factor loadings are constrained to be the same for monozygotic and dizygotic twins. This assumption seems reasonable because it is unlikely that zygosity has a main effect on the measurement of the latent trait; however, it could be evaluated empirically by testing for measurement invariance of factor loadings as a function of zygosity. In the present analysis, item thresholds were also constrained to be equal across zygosity. Again, this assumption could be relaxed to test for possible sibling interaction, which results in differences in thresholds by zygosity.[151] All analyses were performed using the Mx statistical modeling package;[14] Mx scripts are available on the Mx website.[152] Note that for identification purposes, the variance of the factor was fixed to one (but allowed to differ as a function of the covariates) and an estimate was made of all factor loadings rather than arbitrarily fixing one factor loading to one. This has implications for the choice of model testing for measurement invariance.

Results

The twin sample contained 6,805 individuals; 55% were male and 44% were female.

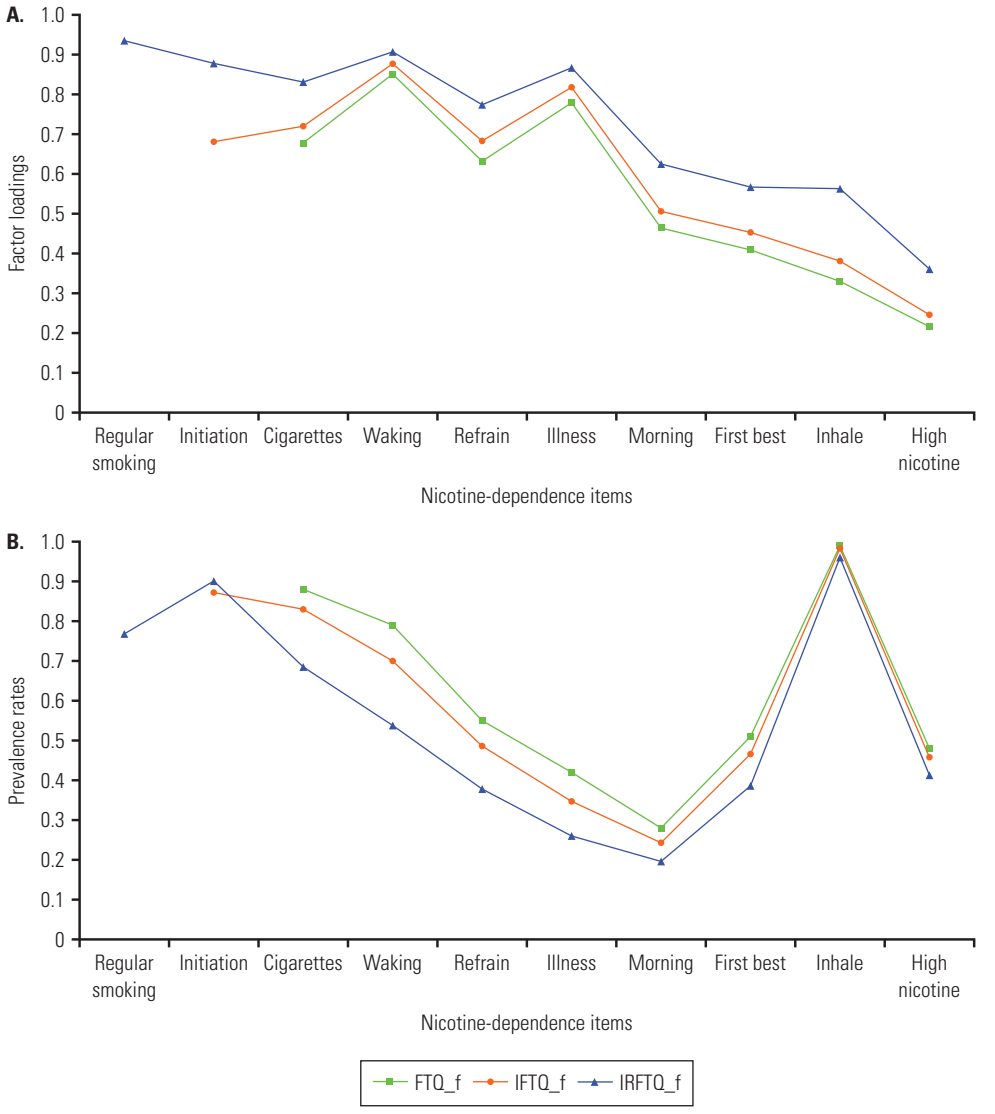
The mean age was 36.2 (standard deviation 8.6) years with a range of 20.4–59.5 years. Overall, 78% reported lifetime TI, and 54% had smoked regularly and thus completed the FTQ. Consistent with previous analyses, which included TI and RTU when estimating the contributions of genetic and environmental factors to nicotine dependence in the CCC pathway model, TI and RTU were included together with the eight FTQ items, assuming neither independence nor unidimensionality of TI and nicotine dependence. Results were compared by using the traditional eight FTQ items, allowing for multiple thresholds as necessary, and the revised six FTND items. Results showed that factor loadings were consistently higher when including TI and RTU compared to those from analyses that (1) included TI alone and (2) included neither of the conditional variables (figure 6.6A). Similarly, prevalences were consistently lower when including TI and RTU, properly adjusting these parameters for the fact that only a selected sample was given the FTQ (figure 6.6B). Of note is that the genetic and environmental parameters were biased when not taking TI and RTU into account. These findings were observed for males and females.

When comparing results from analyses including the FTQ scoring with those including only the six FTND items, the two items not included in the FTND showed a different pattern of factor loadings and prevalences than the other items (figures 6.7A and 6.7B). The inhale item showed very high prevalence, resulting in little variance; the high-nicotine-level item exhibited the lowest factor loadings and clear difference in prevalence by gender. Therefore, the presentation of the results of the genetic analyses with IT, RTU, and the six FTND items is limited here, although the results using the full FTQ items did not differ substantially.

A strict order of model testing was followed and measurement properties were evaluated before testing alternative genetic models. One of the common hypotheses to test with twin data for females and males is whether the contributions of genes and environment are the same (in magnitude and nature) between both genders. However, if differences exist in the assessment of the phenotype in the two genders, then false conclusions may be drawn from genetic analyses if these measurement differences are not taken into account.[4] For example, one might conclude that the heritability for the latent phenotype of interest is significantly greater in females than in males, when in fact there are significant gender differences in the factor loadings and/or thresholds, but not in the sources of individual differences. Accordingly, a series of homogeneity and heterogeneity models were fitted to evaluate the degree of measurement invariance (see Neale and Cardon[20] for a detailed description of heterogeneity models).

Homogeneity models assume that the contributions of genes and environment to the variance (both at the level of the factor, and at the item level, that is, residual variances) are equal for both genders. First, a measurement invariant model was used in which factor mean and variance, factor loadings, and thresholds were the same by gender and age. Then the factor mean and/or factor variance were tested for difference by gender and age (given the large age distribution of the sample) and their interaction. Further testing was conducted to determine significant effects of the covariates on the factor loadings in addition to the factor mean or on the item thresholds in addition to the factor variance. Given that one estimates all factor loadings and fixes the factor variance, one cannot at this stage estimate all factor loadings in addition to the factor variance. Finally, the most saturated measurement model was fitted allowing for covariate effects on both the item thresholds and factor loadings. This series of tests was then repeated for heterogeneity models,

Figure 6.6 Estimates of Factor Loadings (A) and Thresholds (B) of Nicotine-Dependence Items in Female Twins from the Virginia Twin Registry

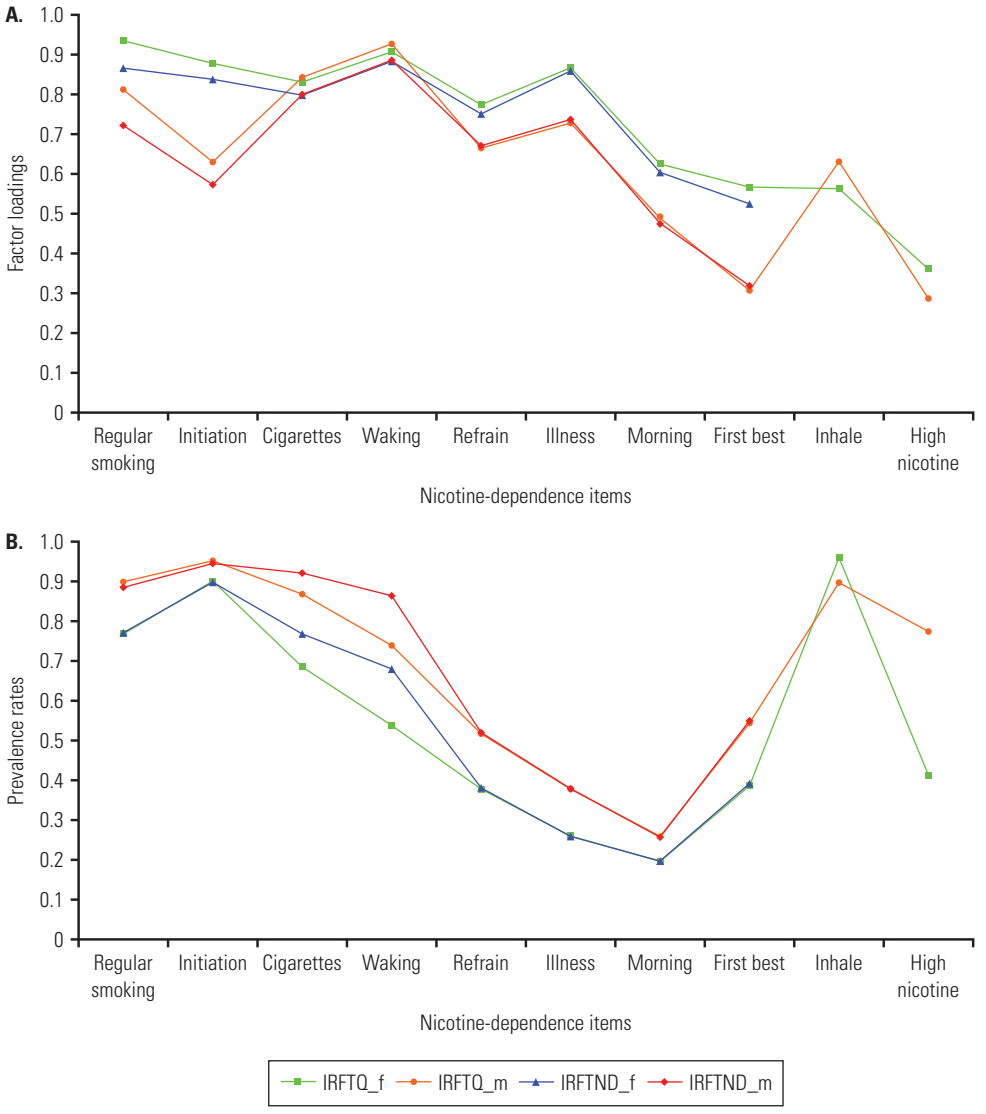


Note. Separate lines depict the effect of including tobacco initiation (TI) and regular tobacco use (RTU) items. FTQ = Fagerström Tolerance Questionnaire; f = female; IFTQ = FTQ items plus initiation; IRFTQ = FTQ items plus initiation and regular tobacco use.

which allow the magnitude of the genetic and environmental contributions to differ by gender. A comparison of the two series of models provides a gender heterogeneity test for the role of genes and environment. Table 6.2 presents selected results from these genetic analyses of nicotine dependence.

When fitting the homogeneity models (columns 2–4) to the adult nicotine-dependence data, significant effects were found of gender and age on both the factor mean and factor variance (models 2 and 3). Differences by gender and age were then tested at the item level—that is, differences

Figure 6.7 Estimates of Factor Loadings (A) and Thresholds (B) of Nicotine-Dependence Items Plotted by Gender and Measurement Instrument (FTQ or FTND Scale)



Note. FTQ = Fagerström Tolerance Questionnaire; FTND = Fagerström Test for Nicotine Dependence; IRFTQ = FTQ items plus initiation and regular tobacco use; f = female; IRFTND = FTND items plus initiation and regular tobacco use; m = male.

in thresholds and factor loadings. Results indicated that thresholds were significantly different by age (model 4) and gender (model 5). Furthermore, factor loadings differed significantly by age (model 6) and between males and females (model 7). When testing was conducted for measurement invariance of the factor loadings allowing for differences in thresholds by gender, age and their interaction, only gender differences in factor loadings were found to be significant (model 9). Similar results were obtained when heterogeneity models were fitted (columns 5–7). The gender heterogeneity

Table 6.2 Results from Fitting Measurement Noninvariance and Gender Heterogeneity Models to Nicotine Initiation and Dependence Data Collected from Twins

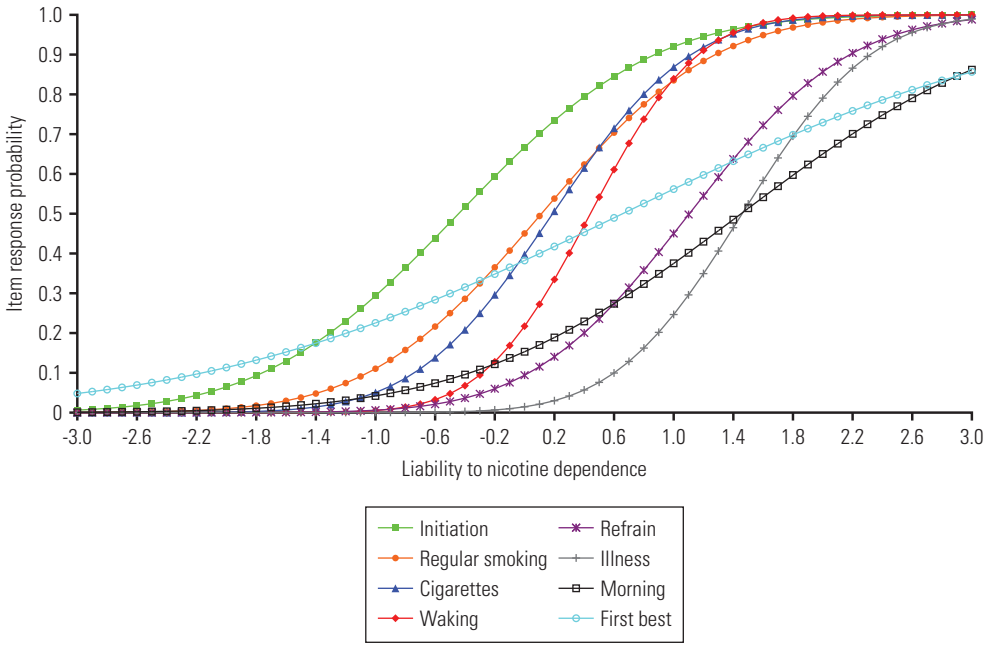
	Homogeneity models			Heterogeneity models			Gender heterogeneity test		
	–2LL	ep	AIC	–2LL	ep	AIC	$\Delta\chi^2$	df	*p*
1. Invariance	52474.92	38	-24729.1	52436.76	57	-24729.2	38.16	18	0.00
Factor mean and factor variance									
2. Age	52075.79	42	-25120.2	52039.78	61	-25118.2	36.01	18	0.01
3. Gender	52075.19	42	-25120.8	52044.87	61	-25113.1	30.33	18	0.05
Item thresholds and factor variance									
4. Age	52259.44	49	-24922.6	52222.72	68	-24921.3	36.72	18	0.01
5. Gender	51976.54	49	-25205.5	51952.14	68	-25191.9	24.40	18	0.14
Factor mean and factor loadings									
6. Age	52046.74	49	-25135.3	52014.43	68	-25129.6	32.31	18	0.02
7. Gender	52017.26	49	-25164.7	51989.19	68	-25154.8	28.07	18	0.06
Item thresholds and factor loadings									
8. Age	51765.04	70	-25375.0	51747.36	89	-25354.6	17.68	18	0.48
9. Gender	51742.11	70	-25397.9	51724.14	89	-25377.9	17.97	18	0.46
Submodels of model 9									
10. GE variance of factor	51732.84	73	-25401.2				9.27	2	0.01
11. GE variance of items	51733.48	86	-25374.5				8.63	16	0.93

Note. –2LL = minus twice the log-likelihood of the data; ep = number of estimated parameters; AIC = Akaike Information Criterion; $\Delta\chi^2$ = difference chi-square statistic; df = degrees of freedom; *p* = probability; GE = genetic and environmental.

tests (last three columns, 8–10) compare the corresponding homogeneity and heterogeneity models. For the measurement invariant models (model 1), as well as for models with limited measurement variance—that is, age variant and gender invariant (models 2, 4, and 6)—the gender heterogeneity tests are significant, suggesting that the contributions of genes and environment to the factor and the items differ significantly. However, when allowing for gender differences at the measurement level (thresholds and factor loadings), the combined genetic and environmental parameters—that is, at the factor and item level—did not differ significantly between males and females, resulting in homogeneity model 9 as the best fitting model (by the Akaike Information Criterion [AIC]).

A further exploration was made of whether the difference in fit between homogeneity and heterogeneity models 9, although not significant, was explained by differences in the genetic and environmental contributions to the factor variance or to the residual item variances. A model allowing for different magnitudes of genetic and environmental contributions to the latent construct (but equating genetic and environmental parameters at the item level between the genders) (model 10) further significantly improved the overall fit of the model over model 9 and resulted in a lower AIC. The converse—different variance components at the item level but not at the factor level (model 11)—did not result in improvement of fit over model 9. Thus, the overall conclusion is that significant gender differences exist at the measurement level (both thresholds and factor loadings). If one is prepared to assume that the same factors are operating in males and females of different ages, and that the measurement

Figure 6.8 Estimates of Nicotine-Dependence Item Characteristic Curves for 20-Year-Old Females



Note. The initiation and regular smoking items are farthest on the left, reflecting that more people initiate smoking and become regular smokers than exhibit dependence symptoms. The least endorsed items are "smoking when ill" and "smoking most in the morning." "Smoking when ill" is also one of the most discriminating items, together with "first cigarette soon after waking" and "number of cigarettes per day."

noninvariance is due to differential sensitivity of certain items, then it would appear that the genetic and environmental factors have different magnitudes of effect at the factor level, but not at the item level. Age also has a significant effect on the thresholds but not on the factor loadings. If these differences in measurement had been ignored, it would have been wrongly concluded that the genetic and environmental contributions were different for males and females not only at the factor level but also at the item level.

The information about the contributions of the individual items to the latent construct of nicotine dependence is best viewed using ICCs in which the slope of the curve reflects the factor loading or indicates how well the item discriminates people who have nicotine dependence from those who have not. The threshold corresponds to the point of inflection of the curve that marks the level at which individuals have a 50% chance of endorsing the item and relates to the endorsement frequency of the items. Thus, the higher the factor loading, the steeper the curve; the higher the threshold, the more to the right of the underlying liability distribution is the curve. As these measurement parameters may be moderated by gender and age, the curves will depend on the particular values of the covariates. Figure 6.8 shows the ICCs for 20-year-old females.

The curves have fairly good coverage in that from –2 to 3 SDs there is likely to be variation in the response patterns. Below –2 SDs, almost all respondents would likely

respond in the lowest category on all items. At +3 SDs, responses would be almost all in the highest response category, although some 15% may be expected not to do so for the "first best" and "morning" items. Because of its relatively flat slope, the "first best" item would also be most likely to be responded to positively at the low end of the scale. In a sum score approach, this item would therefore perform inconsistently and might be considered for deletion.

When curves were compared by the level of the covariates, the ICCs for males are shifted to the left compared to those for females, reflecting the more frequent endorsement of most of the items by males than by females. Similarly, curves for older individuals are shifted to the left of those of younger individuals. The slopes of the curves differ only by gender, and all but one (first cigarette soon after waking) is steeper for males than females. The same information is gleaned from figures 6.9A and 6.9B, which depict the factor loadings and thresholds, respectively.

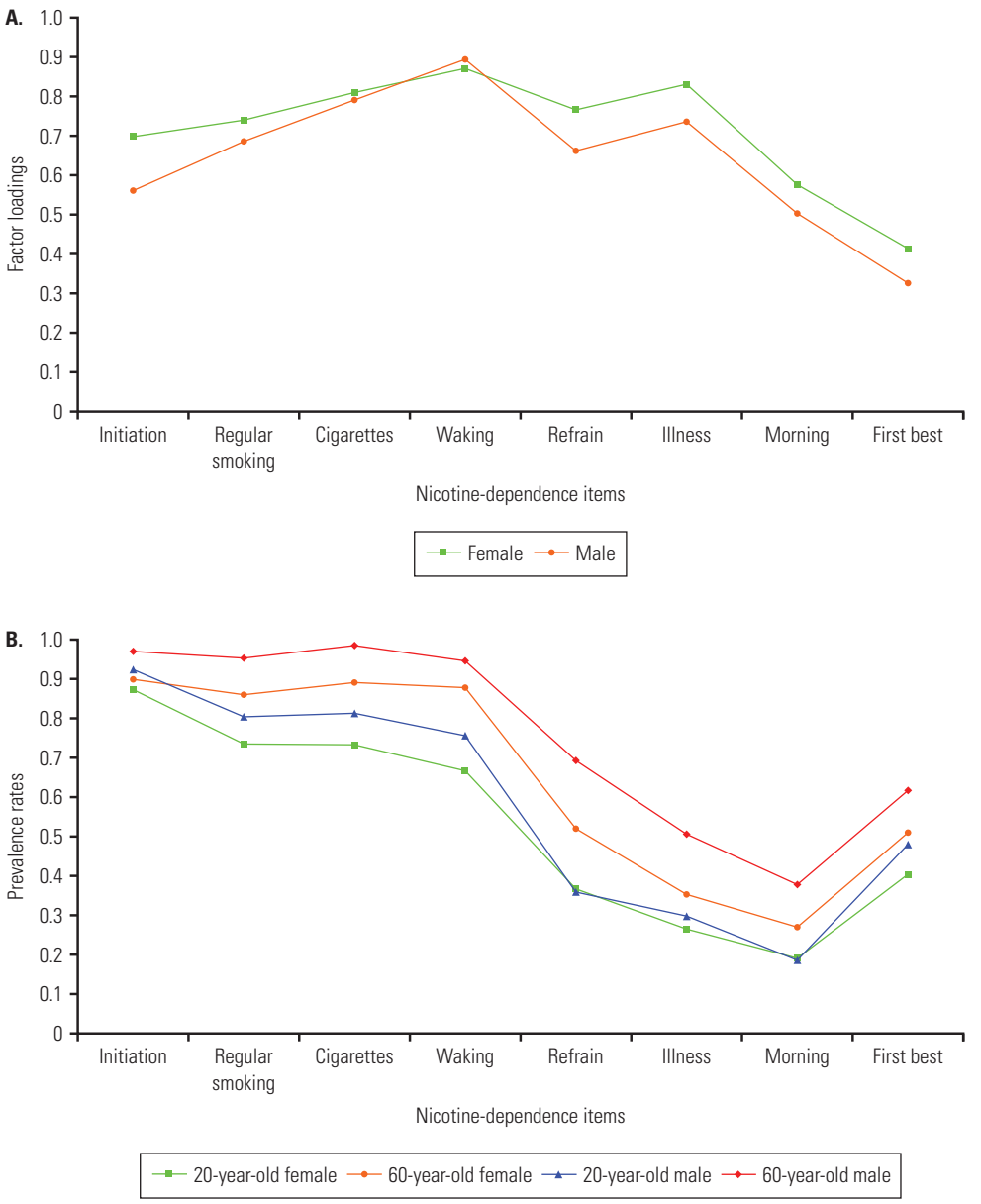
Separate lines represent the different levels of the categorical covariates. For continuous covariates, such as age, estimates are shown for minimum and maximum of the range of the covariate in the sample. The remaining two panels of figure 6.9 present the estimates of the genetic variance (heritability) of each of the items separately for the heritability through the common factor and the residual heritability. The heritability of the latent factor is also shown. Note that the latter was significantly different in males and females, explaining, respectively, 80% and 58% of the variance. Thus, the heritability of the items resulting from the latent factor also differed by gender and reflects the factor loadings of the items.

Interestingly, the pattern of genetic contributions to the residual variance of the items, independent of the latent nicotine-dependence factor, is quite distinct, with initiation and regular smoking exhibiting the largest genetic variance specific to them. This is consistent with previous results from fitting CCC pathway models to these data, which suggested that the genetic factors for TI, RTU, and nicotine dependence were correlated, but not identical dimensions, and that specific genetic factors influence each stage of the smoking behavior continuum. Shared environmental factors contributed about 20% to the latent nicotine-dependence factor in females but were negligible in males (not shown). They also accounted for zero to 8% of the residual item variances. Specific environmental factors explained about 20% of the factor variance in males and females, and between 18% and 32% of the residual item variances, except for "smoking most in morning" and "first cigarette most satisfying," which accounted for about 65% of the variance.

Study Conclusions

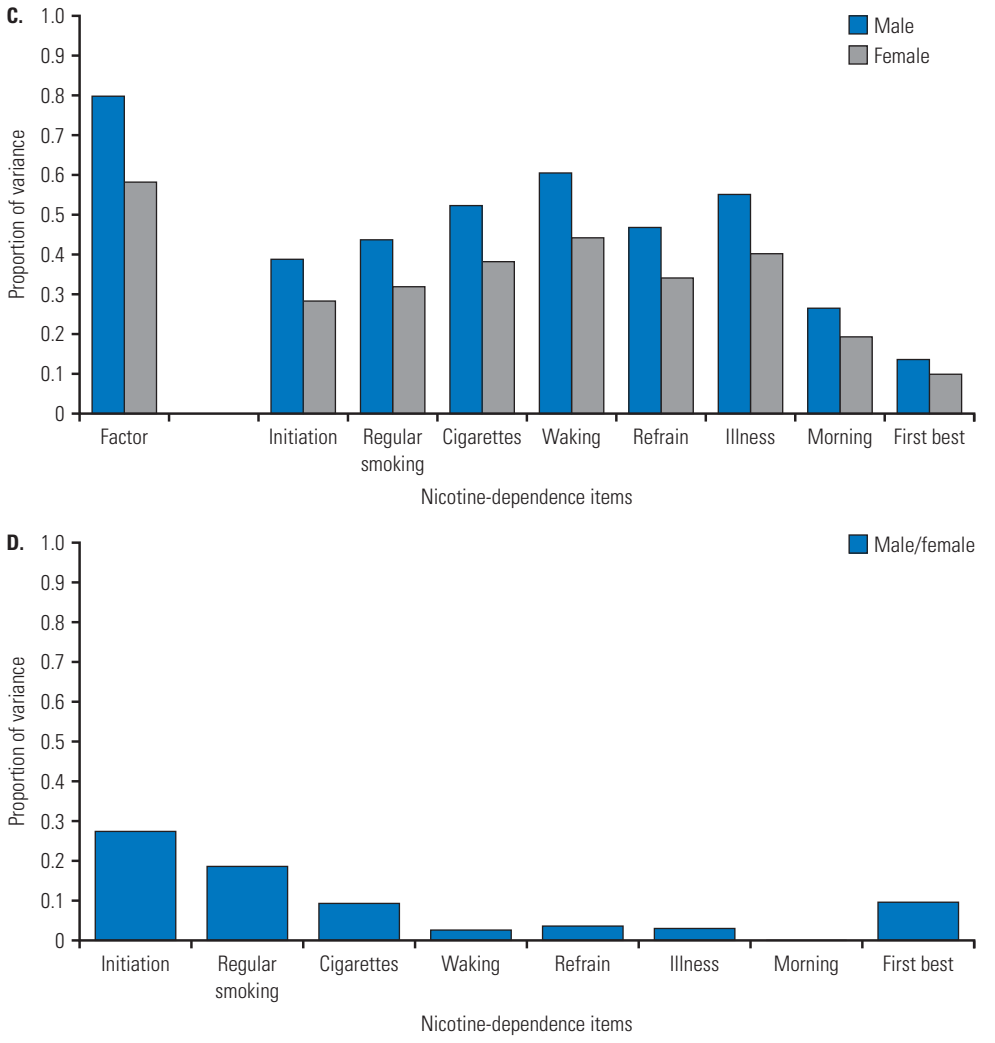
This analysis has shown the importance of taking the assessment of nicotine dependence into account when estimating the role of genetic and environmental factors in the liability to nicotine dependence. When measurement invariance of nicotine initiation and dependence by age and gender was assumed, significant gender heterogeneity was found in the contributions of genes and environment to both age and gender at the factor level and the item level. However, when measurement invariance was accounted for, the overall gender heterogeneity test was not significant, suggesting no differences in the magnitude of genetic and environmental influences in males and females. Model fit further improved when these influences were allowed to differ at the factor level but not the item level. One could argue that when measurement is not invariant by gender, the common factor is measuring something different, or at least the latent factor is measured on a different metric in males and females, which makes it difficult to

Figure 6.9 Estimates of Factor Loadings (A), Thresholds (B), Genetic Variance Components Due to the Factor (C), and Residual Item-Specific Genetic Variance Components (D) in Virginia Twin Registry Males



interpret results of differences in heritability of the common factor. Although caution in interpretation is needed, it is argued that by allowing for limited measurement differences—due to differential sensitivity of items by gender—inferences about heterogeneity of heritability by gender at the latent factor become more meaningful.

Figure 6.9 Estimates of Factor Loadings (A), Thresholds (B), Genetic Variance Components Due to the Factor (C), and Residual Item-Specific Genetic Variance Components (D) in Virginia Twin Registry Males *(continued)*



In the current application, factor loadings are slightly shifted upward in males versus females and prevalences are consistently higher in males and older individuals. In situations in which some items have substantially higher factor loadings and/or thresholds in one gender and other items in the other gender, comparing gender heterogeneity at the factor level may become problematic.

Second, it appears that an item response framework is to be preferred over a sum score approach in which differential item functioning would be obscured and each item weighted equally, regardless of their correlation with the latent construct to be measured. Finally, as was shown previously in using the CCC pathway model to estimate the heritability of nicotine dependence, and repeated in the current

analysis, it is important to include smoking initiation and regular smoking to obtain unbiased estimates of the factor loadings and thresholds. Factor scores from such an analysis should provide a more accurate quantitative phenotype that will improve the ability to find and replicate susceptibility genes for nicotine dependence.

Previous studies of the heritability of nicotine dependence reported estimates between .4 and .7, with little evidence for gender differences. The current analysis estimated heritability of nicotine dependence to be .8 in males and .6 in females. These estimates are significantly different. Estimates of factor loadings and thresholds increased or decreased up to .2 units when measurement variance was allowed and conditional variables such as initiation and regular smoking were included. At this point, one can only speculate about whether this difference proves to be relevant in the search for specific genes or environments that influence smoking behavior and whether it might guide prevention efforts.

Limitations

Although the twin study is one of the most powerful designs to estimate the contributions of genetic and environmental factors to a phenotype of interest, several assumptions are made. One of the most often voiced criticisms of the classical twin study is the equal environments assumption, which states that the degree to which trait-relevant environments are shared is the same for monozygotic and dizygotic twins. In one of the few formal examinations of the validity of the equal environment assumption in twin studies, it appeared not to be violated for regular smoking.[153] Some assumptions can be tested, that is, the random mating can be evaluated when data on spouses are available. A common way to include such data with the twin design is by extending it with their parents. The twin-parent design also allows one to disentangle genetic transmission from environmental transmission, which are confounded in nuclear family parent-offspring correlations. Other designs, however, such as adoption studies or COT designs may be more powerful to sort out transmission from parents to offspring.

Another limitation of studies is the vague assessment of the phenotype. Typically, smoking initiation is assessed with simple questions such as, "Have you ever smoked?" However, how initiation is operationalized for data analysis varies considerably, or the same question may be interpreted differently by younger and older individuals or vary by the status/stage of an individual's smoking behavior. With regard to nicotine dependence, the FTND is widely used, although it is not considered the gold standard. Typically, a sum score is derived, although factor analyses have suggested that factor loadings vary considerably and that two factors might better account for the item correlations. There also appears to be limited overlap between the FTND-based and *DSM*-based assessment of nicotine dependence.

Except for six studies of smoking behavior in adolescents, the majority of the research has focused on adult samples. It is likely incorrect to assume that the same results would be obtained with adolescent samples. Additional complications arise, however, when using data on adolescents in that adolescents have not passed through the main period of uptake of smoking behavior. Thus, the data are left censored and may require approaches based on survival analysis.

While most studies have included both males and females, this is not true for ethnicity, and no adolescent studies on non-Caucasian populations were found. Furthermore, the vast majority of studies on adults are based on Caucasian samples.

Redden and Allison[75] note that assortative mating can increase the risk of type 1 error in association studies; accordingly, this is a risk for association studies of nicotine initiation and dependence.

Summary

The main goal of this chapter is to provide a review of methods for, and applied analyses of, the genetic epidemiological study of nicotine dependence. A secondary aim is to demonstrate that data collected from relatives provide qualitatively different information, which can be used to overcome certain limitations of data collected from unrelated individuals. In the process, the ability to assess the relationship between initiation and dependence by estimating parameters of a factor model applied to data on nicotine initiation as well as the FTND, was exploited. In addition to providing the reader with a general impression of the genetic epidemiology of nicotine dependence, this review has identified a number of further opportunities for model development and data analysis.

Data from relatives do not merely provide a way to partition variation into genetic and environmental components. Especially important for the study of tobacco use, abuse, and dependence is the potential to examine the association between initiation and subsequent progression. The proxy information gleaned from comparing the rate of progression in pairs of relatives who are concordant for initiation to the rate in those who are discordant for initiation allows a number of hypotheses and assumptions to be tested. At the most basic level, one can assess measurement invariance assumptions, which address whether dependence items perform equally well at measuring the latent trait of nicotine dependence in males and females and at different ages. To some extent, this is a sine qua non of epidemiological research into complex behavioral traits. In the absence of measurement invariance tests, one cannot draw unambiguous conclusions about development, trajectories, or even the efficacy of treatments.

Conclusions

1. Data from twin studies suggest that shared environmental factors are the predominant source of familial resemblance in liability to smoking initiation in young adolescents, while additive genetic factors appear more important in older adolescents.
2. Results from extended twin designs show that significant assortative mating exists for smoking initiation and that the parent-child correlations can be almost entirely accounted for by genetic factors. This implies a limited environmental influence of parental smoking initiation on smoking initiation in their children.
3. In contrast to the significant role of shared environmental factors in smoking initiation, the liability to smoking persistence and nicotine dependence appears to be primarily accounted for by additive genetic factors. Furthermore, the liabilities to initiation and progression appear to be substantially correlated. Molecular genetic studies may be expected to find some genetic variants that contribute specifically to initiation—some that are specific to dependence and some that contribute to both.
4. Future development and applications of genetic latent growth curve models and genetic latent class models promise to improve the understanding of the role of genes and environment in smoking trajectories and transitions from nonsmoker to smoking dependence.
5. The search for susceptibility loci for smoking-related traits, either through

linkage or association studies, has not identified any convincing replicated findings. However, several genomic regions and several candidate genes have been found to be associated with smoking behavior in more than one study.

6. Improving the assessment of nicotine initiation and dependence by allowing for differences in measurement by age and gender and taking conditionality into account might provide more accurate estimates of the contributions of genes and environment to different stages of smoking.

7. Meta-analyses or mega-analyses of studies of smoking phenotypes—both genetic epidemiological and molecular genetic—should prove useful in summarizing the available data and results. Possibly, certain data sets may produce results that are outliers, and controlling for their effects would permit finer resolution between hypotheses and more accurate parameter estimates.

References

1. Kendler, K. S., M. C. Neale, P. Sullivan, L. A. Corey, C. O. Gardner, and C. A. Prescott. 1999. A population-based twin study in women of smoking initiation and nicotine dependence. *Psychological Medicine* 29 (2): 299–308.
2. Sung, M., A. Erkanli, A. Angold, and E. J. Costello. 2004. Effects of age at first substance use and psychiatric comorbidity on the development of substance use disorders. *Drug and Alcohol Dependence* 75 (3): 287–99.
3. Mehta, P. D., M. C. Neale, and B. R. Flay. 2004. Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods* 9 (3): 301–33.
4. Lubke, G. H., C. V. Dolan, and M. C. Neale. 2004. Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Research* 7 (3): 292–98.
5. Meredith, W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58 (4): 525–43.
6. Vandenberg, R. J., and C. E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods* 3 (1): 4–69.
7. Vandenberg, R. J. 2002. Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods* 5 (2): 139–58.
8. Fagerström, K. O., and N. G. Schneider. 1989. Measuring nicotine dependence: A review of the Fagerström Tolerance Questionnaire. *Journal of Behavioral Medicine* 12 (2): 159–82.
9. Wright, S. 1934. The method of path coefficients. *Annals of Mathematical Statistics* 5: 161–215.
10. Bollen, K. A. 1989. *Structural equations with latent variables.* Oxford, UK: John Wiley & Sons.
11. Wright, S. 1921. Correlation and causation. *Journal of Agricultural Research* 20:557–85.
12. Sörbom, D. 1974. A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology* 27: 229–39.
13. McArdle, J. J., and S. M. Boker. 1986. *RAMpath - path diagram software.* Denver: Data Transforms.
14. Neale, M. C., S. M. Boker, G. Xie, and H. H. Maes. 1997. *Mx: Statistical modeling.* 4th ed. Richmond, VA: Virginia Commonwealth Univ.
15. Mehta, P. D., and S. G. West. 2000. Putting the individual back into individual growth curves. *Psychological Methods* 5 (1): 23–43.
16. Muthén, B. 2001. Latent variable mixture modeling. In *New developments and techniques in structural equation modeling,* ed. G. A. Marcoulides and R. E. Schumacker, 1–33. Mahwah, NJ: Lawrence Erlbaum.
17. Bauer, D. J., and P. J. Curran. 2003. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods* 8 (3): 338–63.
18. Raudenbush, S. W. 2001. Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology* 52:501–25.
19. Stanek, E. J. 3rd, and S. R. Diehl. 1988. Growth curve models of repeated binary response. *Biometrics* 44 (4): 973–83.
20. Neale, M. C., and L. R. Cardon. 1992. *Methodology for genetic studies of twins and families.* New York: Kluwer Academic/ Plenum Publishers.
21. Loehlin, J. C., and R. C. Nichols. 1976. *Heredity, environment, and personality: A study of 850 sets of twins.* Austin, TX: Univ. of Texas Press.
22. Rose, R. J., J. Kaprio, C. J. Williams, R. Viken, and K. Obremski. 1990. Social contact and sibling similarity: Facts, issues, and red herrings. *Behavior Genetics* 20 (6): 763–78.
23. Kendler, K. S., M. C. Neale, R. C. Kessler, A. C. Heath, and L. J. Eaves. 1994. Parental treatment and the equal environment assumption in twin studies of psychiatric illness. *Psychological Medicine* 24 (3): 579–90.
24. Truett, K. R., L. J. Eaves, E. E. Walters, A. C. Heath, J. K. Hewitt, J. M. Meyer, J. Silberg, M. C. Neale, N. G. Martin, and K. S. Kendler. 1994. A model system for analysis of family resemblance in extended kinships of twins. *Behavior Genetics* 24 (1): 35–49.

25. Martin, N. G., L. J. Eaves, M. J. Kearsey, and P. Davies. 1978. The power of the classical twin study. *Heredity* 40 (1): 97–116.
26. Neale, M. C., L. J. Eaves, and K. S. Kendler. 1994. The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics* 24 (3): 239–58.
27. Mather, K., and J. L. Jinks. 1982. *Biometrical genetics 3rd ed.* New York: Chapman and Hall.
28. Jencks, C. 1972. *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.
29. Eaves, L. J., K. A. Last, P. A. Young, and N. G. Martin. 1978. Model-fitting approaches to the analysis of human behaviour. *Heredity* 41 (3): 249–320.
30. Young, P. A., L. J. Eaves, and H. J. Eysenck. 1980. Intergenerational stability and change in the causes of variation in personality. *Personality and Individual Differences* 1 (1): 35–55.
31. Fulker, D. W. 1982. Extensions of the classical twin method. *Progress in Clinical and Biological Research* 103 Pt A: 395-406.
32. Eaves, L., A. Heath, N. Martin, H. Maes, M. Neale, K. Kendler, K. Kirk, and L. Corey. 1999. Comparing the biological and cultural inheritance of personality and social attitudes in the Virginia 30,000 study of twins and their relatives. *Twin Research* 2 (2): 62–80.
33. Maes, H. H., M. C. Neale, N. G. Martin, A. C. Heath, and L. J. Eaves. 1999. Religious attendance and frequency of alcohol use. Same genes or same environments: A bivariate extended twin kinship model. *Twin Research* 2 (2): 169–79.
34. Kendler, K. S., A. C. Heath, N. G. Martin, and L. J. Eaves. 1987. Symptoms of anxiety and symptoms of depression. Same genes, different environments? *Archives of General Psychiatry* 44 (5): 451–57.
35. McArdle, J. J., and H. H. Goldsmith. 1990. Alternative common factor models for multivariate biometric analyses. *Behavior Genetics* 20 (5): 569–608.
36. Neale, M. C., E. Harvey, H. H. Maes, P. F. Sullivan, and K. S. Kendler. 2006. Extensions to the modeling of initiation and progression: Applications to substance use and abuse. *Behavior Genetics* 36 (4): 507–24.
37. Maes, H. H., P. F. Sullivan, C. M. Bulik, M. C. Neale, C. A. Prescott, L. J. Eaves, and K. S. Kendler. 2004. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychological Medicine* 34 (7): 1251–61.
38. Neale, M. C., S. H. Aggen, H. H. Maes, T. S. Kubarych, and J. E. Schmitt. 2006. Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors* 31 (6): 1010–34.
39. Lubke, G., and M. C. Neale. 2006. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research* 41 (4): 499–532.
40. Muthén, B., and T. Asparouhov. 2006. Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors* 31 (6): 1050–66.
41. Baker, L. A., C. Reynolds, and E. Phelps. 1992. Biometrical analysis of individual growth curves. *Behavior Genetics* 22 (2): 253–64.
42. Neale, M. C., E. Roysamb, and K. Jacobson. 2006. Multivariate genetic analysis of sex limitation and G x E interaction. *Twin Research and Human Genetics* 9 (4): 481–89.
43. Malthus, T. R. 1789. *An essay on the principle of population as it affects the future improvement of society, with remarks on the speculations of Mr. Godwin, M. Condorcet, and other writers.* London: J. Johnson.
44. Browne, M. W. 1984. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37 (Pt 1): 62–83.
45. Neale, M. C., and J. J. McArdle. 2000. Structured latent growth curves for twin data. *Twin Research* 3 (3): 165–77.
46. Box, G., G. M. Jenkins, and G. Reinsel. 1994. *Time series analysis: Forecasting and control.* 3rd ed. Englewood Cliffs, NJ: Prentice Hall.
47. Wirth, R. J., and M. C. Edwards. 2007. Item factor analysis: Current approaches and future directions. *Psychological Methods* 12 (1): 58–79.
48. Dolan, C. V., V. D. Schmittmann, G. H. Lubke, and M. C. Neale. 2005. Regime switching in the latent growth curve mixture model. *Structural Equation Modeling* 12 (1): 94–119.

49. Spearman, C. 1904. 'General intelligence,' objectively determined and measured. *American Journal of Psychology* 15 (2): 201–93.
50. Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent structure analysis.* Boston, MA: Houghton Mifflin.
51. Lazarsfeld, P. F. 1950. Some latent structures. In *The American soldier: Studies in social psychology in World War II,* vol. 4, ed. S. A. Stouffer. Princeton, NJ: Princeton Univ. Press.
52. Vermunt, J. K., and J. Magidson. 2002. Latent class cluster analysis. In *Advances in latent class analysis,* ed. J. Hagenaars and A. McCutcheon, 89–106. Cambridge, UK: Cambridge Univ. Press.
53. Bartholomew, D. J. 1987. *Latent variable models and factor analysis.* New York: Oxford Univ. Press.
54. McLachlan, G., and D. Peel. 2000. *Finite mixture models.* New York: John Wiley and Sons.
55. Arminger, G., P. Stein, and J. Wittenberg. 1999. Mixtures of conditional mean- and covariance-structure models. *Psychometrika* 64 (4): 475–94.
56. Dolan, C. V., and H. L. J. van der Maas. 1998. Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika* 63 (3): 227–53.
57. Jedidi, K., H. S. Jagpal, and W. S. DeSarbo. 1997. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science* 16 (1): 39–59.
58. Yung, Y.-F. 1997. Finite mixtures in confirmatory factor-analysis models. *Psychometrika* 62 (3): 297–330.
59. Muthén, B. O. 2001. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In *New methods for the analysis of change,* ed. L. M. Collins and A. Sayer, 291–322. Washington, DC: American Psychological Association.
60. Muthén, B. O., and K. Shedden. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55 (2): 463–69.
61. Eaves, L. J., J. L. Silberg, J. K. Hewitt, M. Rutter, J. M. Meyer, M. C. Neale, and A. Pickles. 1993. Analyzing twin resemblance in multisymptom data: Genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys. *Behavior Genetics* 23 (1): 5–19.
62. Gillespie, N. A., and M. C. Neale. 2006. A finite mixture model for genotype and environment interactions: Detecting latent population heterogeneity. *Twin Research and Human Genetics* 9 (3): 412–23.
63. Muthén, B., T. Asparouhov, and I. Rebollo. 2006. Advances in behavioral genetics modeling using Mplus: Applications of factor mixture modeling to twin data. *Twin Research and Human Genetics* 9 (3): 313–24.
64. Li, F., T. E. Duncan, and H. Hops. 2001. Examining developmental trajectories in adolescent alcohol use using piecewise growth mixture modeling analysis. *Journal of Studies on Alcohol* 62 (2): 199–210.
65. Li, M. D. 1976. *First course in population genetics.* Pacific Grove, CA: Boxwood Press.
66. Morton, N. E. 1982. *Outline of genetic epidemiology.* New York: Karger.
67. Fulker, D. W., and S. S. Cherny. 1996. An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics* 26 (5): 527–32.
68. Eaves, L. J., M. C. Neale, and H. Maes. 1996. Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior Genetics* 26 (5): 519–25.
69. Lander, E., and L. Kruglyak. 1995. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics* 11 (3): 241–47.
70. Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30 (1): 97–101.
71. Kruglyak, L., M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics* 58 (6): 1347–63.
72. Pritchard, J. K., and N. A. Rosenberg. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 65 (1): 220–28.
73. Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* 64 (1): 259–67.

74. Cardon, L. R., and G. R. Abecasis. 2000. Some properties of a variance components model for fine-mapping quantitative trait loci. *Behavior Genetics* 30 (3): 235–43.

75. Redden, D. T., and D. B. Allison. 2006. The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. *Behavior Genetics* 36 (5): 678–86.

76. van den Oord, E. J., and H. Snieder. 2002. Including measured genotypes in statistical models to study the interplay of multiple factors affecting complex traits. *Behavior Genetics* 32 (1): 1–22.

77. Sullivan, P. F., and K. S. Kendler. 1999. The genetic epidemiology of smoking. *Nicotine & Tobacco Research* 1 Suppl. 2: S51–S57, S69–S70.

78. Li, M. D. 2003. The genetics of smoking related behavior: A brief review. *American Journal of the Medical Sciences* 326 (4): 168–73.

79. Centers for Disease Control and Prevention. 2007. Cigarette smoking among adults—United States, 2006. *Morbidity and Mortality Weekly Report* 56 (44): 1157–61.

80. Boomsma, D. I., J. R. Koopmans, L. J. Van Doornen, and J. F. Orlebeke. 1994. Genetic and social influences on starting to smoke: A study of Dutch adolescent twins and their parents. *Addiction* 89 (2): 219–26.

81. Koopmans, J. R., L. J. van Doornen, and D. I. Boomsma. 1997. Association between alcohol use and smoking in adolescent and young adult twins: A bivariate genetic analysis. *Alcoholism, Clinical and Experimental Research* 21 (3): 537–46.

82. Kaprio, J., D. I. Boomsma, K. Heikkilä, M. Koskenvuo, K. Romanov, R. J. Rose, R. J. Viken, and T. Winter. 1995. Genetic variation in behavioral risk factors to risk for atherosclerosis: A twin family study of smoking and cynical hostility. In *Atherosclerosis X: Proceedings of the Xth International Symposium on Atherosclerosis,* ed. F. P. Woodford, J. Davignon, and A. Sniderman, 634–37. Amsterdam: Elsevier Science.

83. Dick, D. M., S. Barman, and T. Pitkänen. 2006. Genetic and environmental influences on the initiation and continuation of smoking and drinking. *Socioemotional development and health from adolescence to adulthood,* ed. L. Pulkkinen, J. Kaprio, and R. J. Rose, 126–45. New York: Cambridge Univ. Press.

84. Rose, R. J., R. J. Viken, D. M. Dick, J. E. Bates, L. Pulkkinen, and J. Kaprio. 2003. It does take a village: Nonfamilial environments and children's behavior. *Psychological Science* 14 (3): 273–77.

85. Maes, H. H., C. E. Woodard, L. Murrelle, J. M. Meyer, J. L. Silberg, J. K. Hewitt, M. Rutter, et al. 1999. Tobacco, alcohol and drug use in eight- to sixteen-year-old twins: The Virginia Twin Study of Adolescent Behavioral Development. *Journal of Studies on Alcohol* 60 (3): 293–305.

86. Han, C., M. K. McGue, and W. G. Iacono. 1999. Lifetime tobacco, alcohol and other substance use in adolescent Minnesota twins: Univariate and multivariate behavioral genetic analyses. *Addiction* 94 (7): 981–93.

87. McGue, M., I. Elkins, and W. G. Iacono. 2000. Genetic and environmental influences on adolescent substance use and abuse. *American Journal of Medical Genetics* 96 (5): 671–77.

88. Rhee, S. H., J. K. Hewitt, S. E. Young, R. P. Corley, T. J. Crowley, and M. C. Stallings. 2003. Genetic and environmental influences on substance initiation, use, and problem use in adolescents. *Archives of General Psychiatry* 60 (12): 1256–64.

89. White, V. M., J. L. Hopper, A. J. Wearing, and D. J. Hill. 2003. The role of genes in tobacco smoking during adolescence and young adulthood: A multivariate behaviour genetic investigation. *Addiction* 98 (8): 1087–1100.

90. Rende, R., C. Slomkowski, J. McCaffery, E. E. Lloyd-Richardson, and R. Niaura. 2005. A twin-sibling study of tobacco use in adolescence: Etiology of individual differences and extreme scores. *Nicotine & Tobacco Research* 7 (3): 413–19.

91. Hopfer, C. J., T. J. Crowley, and J. K. Hewitt. 2003. Review of twin and adoption studies of adolescent substance use. *Journal of the American Academy of Child & Adolescent Psychiatry* 42 (6): 710–19.

92. Jedidi, K., H. S. Jagpal, and W. S. DeSarbo. 1997. STEMM: A general finite mixture structural equation model. *Journal of Classification* 14 (1): 23–50.

93. Vink, J. M., G. Willemsen, R. C. Engels, and D. I. Boomsma. 2003. Smoking status of parents, siblings and friends: Predictors of regular smoking? Findings from a longitudinal twin-family study. *Twin Research* 6 (3): 209–17.

94. Vink, J. M., G. Willemsen, and D. I. Boomsma. 2003. The association of current smoking behavior with the smoking behavior of parents, siblings, friends and spouses. *Addiction* 98 (7): 923–31.

95. Maes, H. H., M. C. Neale, K. S. Kendler, N. G. Martin, A. C. Heath, and L. J. Eaves. 2006. Genetic and cultural transmission of smoking initiation: An extended twin kinship model. *Behavior Genetics* 36 (6): 795–808.

96. Maes, H. H., M. C. Neale, and K. S. Kendler. 2006. Testing for measurement invariance in genetic analyses of smoking and nicotine dependence. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 141B (7): 700.

97. Pergadia, M. L., A. C. Heath, N. G. Martin, and P. A. Madden. 2006. Genetic analyses of DSM-IV nicotine withdrawal in adult twins. *Psychological Medicine* 36 (7): 963–72.

98. Morley, K. I., M. T. Lynskey, P. A. Madden, S. A. Treloar, A. C. Heath, and N. G. Martin. 2007. Exploring the inter-relationship of smoking age-at-onset, cigarette consumption and smoking persistence: Genes or environment? *Psychological Medicine* 37 (9): 1357–67.

99. Heath, A. C., N. G. Martin, M. T. Lynskey, A. A. Todorov, and P. A. Madden. 2002. Estimating two-stage models for genetic influences on alcohol, tobacco or drug use initiation and dependence vulnerability in twin and family data. *Twin Research* 5 (2): 113–24.

100. Broms, U., K. Silventoinen, P. A. Madden, A. C. Heath, and J. Kaprio. 2006. Genetic architecture of smoking behavior: A study of Finnish adult twins. *Twin Research and Human Genetics* 9 (1): 64–72.

101. Heath, A. C., P. A. Madden, and N. G. Martin. 1998. Statistical methods in genetic research on smoking. *Statistical Methods in Medical Research* 7 (2): 165–86.

102. Koopmans, J. R., W. S. Slutske, A. C. Heath, M. C. Neale, and D. I. Boomsma. 1999. The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. *Behavior Genetics* 29 (6): 383–93.

103. True, W. R., A. C. Heath, J. F. Scherrer, B. Waterman, J. Goldberg, N. Lin, S. A. Eisen, M. J. Lyons, and M. T. Tsuang. 1997. Genetic and environmental contributions to smoking. *Addiction* 92 (10): 1277–87.

104. True, W. R., H. Xian, J. F. Scherrer, P. A. Madden, K. K. Bucholz, A. C. Heath, S. A. Eisen, M. J. Lyons, J. Goldberg, and M. Tsuang. 1999. Common genetic vulnerability for nicotine and alcohol dependence in men. *Archives of General Psychiatry* 56 (7): 655–61.

105. Heath, A. C., K. M. Kirk, J. M. Meyer, and N. G. Martin. 1999. Genetic and social determinants of initiation and age at onset of smoking in Australian twins. *Behavior Genetics* 29 (6): 395–407.

106. Madden, P. A., A. C. Heath, N. L. Pedersen, J. Kaprio, M. J. Koskenvuo, and N. G. Martin. 1999. The genetics of smoking persistence in men and women: A multicultural study. *Behavior Genetics* 29 (6): 423–31.

107. Vink, J. M., G. Willemsen, and D. I. Boomsma. 2005. Heritability of smoking initiation and nicotine dependence. *Behavior Genetics* 35 (4): 397–406.

108. American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed., text rev. (DSM–IV–TR). Arlington, VA: American Psychiatric Publishing.

109. Aggen, S. H., M. C. Neale, and K. S. Kendler. 2005. DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine* 35 (4): 475–87.

110. Heatherton, T. F., L. T. Kozlowski, R. C. Frecker, and K. O. Fagerström. 1991. The Fagerström Test for Nicotine Dependence: A revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction* 86 (9): 1119–27.

111. Pomerleau, C. S., S. M. Carton, M. L. Lutzke, K. A. Flessland, and O. F. Pomerleau. 1994. Reliability of the Fagerström Tolerance Questionnaire and the Fagerström Test for Nicotine Dependence. *Addictive Behaviors* 19 (1): 33–39.

112. Hudmon, K. S., C. S. Pomerleau, J. Brigham, H. Javitz, and G. E. Swan. 2005. Validity of retrospective assessments of nicotine dependence: A preliminary report. *Addictive Behaviors* 30 (3): 613–37.

113. Prokhorov, A. V., C. De Moor, U. E. Pallonen, K. S. Hudmon, L. Koehly, and S. Hu. 2000. Validation of the modified Fagerström Tolerance Questionnaire with salivary cotinine among adolescents. *Addictive Behaviors* 25 (3): 429–33.

114. Haddock, C. K., H. Lando, R. C. Klesges, G. W. Talcott, and E. A. Renaud. 1999. A study of the psychometric and predictive properties of the Fagerström Test for

Nicotine Dependence in a population of young smokers. *Nicotine & Tobacco Research* 1 (1): 59–66.

115. Radzius, A., E. T. Moolchan, J. E. Henningfield, S. J. Heishman, and J. J. Gallo. 2001. A factor analysis of the Fagerström Tolerance Questionnaire. *Addictive Behaviors* 26 (2): 303–10.
116. Radzius, A., J. J. Gallo, D. H. Epstein, D. A. Gorelick, J. L. Cadet, G. E. Uhl, and E. T. Moolchan. 2003. A factor analysis of the Fagerström Test for Nicotine Dependence (FTND). *Nicotine &Tobacco Research* 5 (2): 255–40.
117. Richardson, C. G., and P. A. Ratner. 2005. A confirmatory factor analysis of the Fagerström Test for Nicotine Dependence. *Addictive Behaviors* 30 (4): 697–709.
118. de Leon, J., F. J. Diaz, E. Becona, M. Gurpegui, D. Jurado, and A. Gonzalez-Pinto. 2003. Exploring brief measures of nicotine dependence for epidemiological surveys. *Addictive Behaviors* 28 (8): 1481–6.
119. John, U., C. Meyer, A. Schumann, U. Hapke, H. J. Rumpf, C. Adam, D. Alte, and J. Ludemann. 2004. A short form of the Fagerström Test for Nicotine Dependence and the Heaviness of Smoking Index in two adult population samples. *Addictive Behaviors* 29 (6): 1207–12.
120. Kandel, D., C. Schaffran, P. Griesler, J. Samuolis, M. Davies, and R. Galanti. 2005. On the measurement of nicotine dependence in adolescence: Comparisons of the mFTQ and a DSM-IV-based scale. *Journal of Pediatric Psychology* 30 (4): 319–32.
121. Hughes, J. R., A. H. Oliveto, R. Riggs, M. Kenny, A. Liguori, J. L. Pillitteri, and M. A. MacLaughlin. 2004. Concordance of different measures of nicotine dependence: Two pilot studies. *Addictive Behaviors* 29 (8): 1527–39.
122. Lessov, C. N., N. G. Martin, D. J. Statham, A. A. Todorov, W. S. Slutske, K. K. Bucholz, A. C. Heath, and P. A. Madden. 2004. Defining nicotine dependence for genetic research: Evidence from Australian twins. *Psychological Medicine* 34 (5): 865–79.
123. Straub, R. E., P. F. Sullivan, Y. Ma, M. V. Myakishev, C. Harris-Kerr, B. Wormley, B. Kadambi, et al. 1999. Susceptibility genes for nicotine dependence: A genome scan and followup in an independent sample suggest that regions on chromosomes 2, 4, 10, 16, 17 and 18 merit further study. *Molecular Psychiatry* 4 (2): 129–44.
124. Bierut, L. J., J. P. Rice, A. Goate, A. L. Hinrichs, N. L. Saccone, T. Foroud, H. J. Edenberg, et al. 2004. A genomic scan for habitual smoking in families of alcoholics: Common and specific genetic factors in substance dependence. *American Journal of Medical Genetics A* 124 (1): 19–27.
125. Saccone, N. L., E. L. Goode, and A. W. Bergen. 2003. Genetic analysis workshop 13: Summary of analyses of alcohol and cigarette use phenotypes in the Framingham Heart Study. *Genetic Epidemiology* 25 Suppl. 1: S90–S97.
126. Vink, J. M., A. L. Beem, D. Posthuma, M. C. Neale, G. Willemsen, K. S. Kendler, P. E. Slagboom, and D. I. Boomsma. 2004. Linkage analysis of smoking initiation and quantity in Dutch sibling pairs. *Pharmacogenomics Journal* 4 (4): 274–82.
127. Ehlers, C. L., and K. C. Wilhelmsen. 2006. Genomic screen for loci associated with tobacco usage in Mission Indians. *BMC Medical Genetics* 7:9.
128. Gelernter, J., X. Liu, V. Hesselbrock, G. P. Page, A. Goddard, and H. Zhang. 2004. Results of a genomewide linkage scan: Support for chromosomes 9 and 11 loci increasing risk for cigarette smoking. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 128 (1): 94–101.
129. Gelernter, J., C. Panhuysen, R. Weiss, K. Brady, J. Poling, M. Krauthammer, L. Farrer, and H. R. Kranzler. 2007. Genomewide linkage scan for nicotine dependence: Identification of a chromosome 5 risk locus. *Biological Psychiatry* 61 (1): 119–26.
130. Li, M. D., T. J. Payne, J. Z. Ma, X. Y. Lou, D. Zhang, R. T. Dupont, K. M. Crews, G. Somes, N. J. Williams, and R. C. Elston. 2006. A genomewide search finds major susceptibility loci for nicotine dependence on chromosome 10 in African Americans. *American Journal of Human Genetics* 79 (4): 745–51.
131. Swan, G. E., H. Hops, K. C. Wilhelmsen, C. N. Lessov-Schlaggar, L. S. Cheng, K. S. Hudmon, C. I. Amos, et al. 2006. A genome-wide screen for nicotine dependence susceptibility loci. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 141 (4): 354–60.
132. Saccone, S. F., M. L. Pergadia, A. Loukola, U. Broms, G. W. Montgomery, J. C. Wang, A. Agrawal, et al. 2007. Genetic linkage to

chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples. *American Journal of Human Genetics* 80 (5): 856–66.

133. Morley, K. I., S. E. Medland, M. A. Ferreira, M. T. Lynskey, G. W. Montgomery, A. C. Heath, P. A. Madden, and N. G. Martin. 2006. A possible smoking susceptibility locus on chromosome 11p12: Evidence from sex-limitation linkage analyses in a sample of Australian twin families. *Behavior Genetics* 36 (1): 87–99.
134. Sham, P. 1997. *Statistics in human genetics*. London: Hodder Arnold.
135. Li, M. D., J. Z. Ma, R. Cheng, R. T. Dupont, N. J. Williams, K. M. Crews, T. J. Payne, R. C. Elston, and Framingham Heart Study. 2003. A genome-wide scan to identify loci for smoking rate in the Framingham Heart Study population. *BMC Genetics* 4 Suppl. 1: S103.
136. Goode, E. L., M. D. Badzioch, H. Kim, F. Gagnon, L. S. Rozek, K. L. Edwards, and G. P. Jarvik. 2003. Multiple genome-wide analyses of smoking behavior in the Framingham Heart Study. *BMC Genetics* 4 Suppl. 1: S102.
137. Vink, J. M., D. Posthuma, M. C. Neale, S. P. Eline, and D. I. Boomsma. 2005. Genome-wide linkage scan to identify loci for age at first cigarette in Dutch sibling pairs. *Behavior Genetics* 36 (1): 100–111.
138. Duggirala, R., L. Almasy, and J. Blangero. 1999. Smoking behavior is under the influence of a major quantitative trait locus on human chromosome 5q. *Genetic Epidemiology* 17 Suppl. 1: S139–S144.
139. Lewis, S. J., S. Zammit, D. Gunnell, and G. D. Smith. 2005. A meta-analysis of the MTHFR C677T polymorphism and schizophrenia risk. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 135 (1): 2–4.
140. Lerman, C., and W. Berrettini. 2003. Elucidating the role of genetic factors in smoking behavior and nicotine dependence. *American Journal of Medical Genetics Part B, Neuropsychiatric Genetics* 118 (1): 48–54.
141. Li, M. D. 2006. The genetics of nicotine dependence. *Current Psychiatry Reports* 8 (2): 158–64.
142. Batra, V., A. A. Patkar, W. H. Berrettini, S. P. Weinstein, and F. T. Leone. 2003. The genetic determinants of smoking. *Chest* 123 (5): 1730–9.
143. Li, M. D., J. Z. Ma, and J. Beuten. 2004. Progress in searching for susceptibility loci and genes for smoking-related behaviour. *Clinical Genetics* 66 (5): 382–92.
144. Munafó, M. R., T. G. Clark, E. C. Johnstone, M. F. G. Murphy, and R. T. Walton. 2004. The genetic basis for smoking behavior: A systematic review and meta-analysis. *Nicotine & Tobacco Research* 6 (4): 583–98.
145. Tyndale, R. F. 2003. Genetics of alcohol and tobacco use in humans. *Annals of Medicine* 35 (2): 94–121.
146. Al Koudsi, N., and R. F. Tyndale. 2005. Genetic influences on smoking: A brief review. *Therapeutic Drug Monitoring* 27 (6): 704–9.
147. Greenbaum, L., K. Kanyas, O. Karni, Y. Merbl, T. Olender, A. Horowitz, A. Yakir, D. Lancet, E. Ben-Asher, and B. Lerer. 2006. Why do young women smoke? I: Direct and interactive effects of environment, psychological characteristics and nicotinic cholinergic receptor genes. *Molecular Psychiatry* 11 (3): 312–22, 223.
148. Bierut, L. J., P. A. Madden, N. Breslau, E. O. Johnson, D. Hatsukami, O. F. Pomerleau, G. E. Swan, et al. 2007. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human Molecular Genetics* 16 (1): 24–35.
149. Uhl, G. R., Q. R. Liu, T. Drgon, C. Johnson, D. Walther, and J. E. Rose. 2007. Molecular genetics of nicotine dependence and abstinence: Whole genome association using 520,000 SNPs. *BMC Genetics* 8:10.
150. Spence, J. E., L. A. Corey, W. E. Nance, M. L. Marazita, K. S. Kendler, and R. M. Schieken. 1988. Molecular analysis of twin zygosity using VNTR DNA probes. Abstract. *American Journal of Human Genetics* 43 (3): A159.
151. Eaves, L. 1976. A model for sibling effects in man. *Heredity* 36 (2): 205–14.
152. Neale, M. About Mx. http://www.vcu.edu/mx/about-mx.html (accessed December 24, 2008).
153. Kendler, K. S., M. C. Neale, C. J. MacLean, A. C. Heath, L. J. Eaves, and R. C. Kessler. 1993. Smoking and major depression. A causal analysis. *Archives of General Psychiatry* 50 (1): 36–43.