# The National Merit Twin Study[1]

*Robert C. Nichols*

This chapter will be concerned with some of the methodological problems and some early results of the National Merit Twin Study which is still in progress and when finished will be reported in detail elsewhere. Analyses completed so far have been primarily concerned with the relative effects of heredity and environment on ability measures, which will therefore be the major focus of this chapter.

Of course, both heredity and environment are necessary for any behavior to occur. If heredity had not produced a behaving organism there would be no behavior, and if environment had not provided the appropriate situation for the development of the organism there would also be no behavior. Both are necessary and completely interdependent influences, and it is not possible to say that one is more important than the other. If, in an experimental situation, either heredity or environment were completely controlled, variation in any behavior could be produced by variation of the other. However, this is not the question we are asking. We are interested in accounting for variance in human behavior as it is observed in the natural social setting, and it is a perfectly legitimate question to ask how much of this variance is due to variation in heredity and how much is due to variation in the environment. The question "If everyone had the same heredity, how much would the variance in behavior be reduced?" is a convenient simplification of the question we are asking. It is a simplification because there are undoubtedly many instances in human behavior of a phenomenon which has been demonstrated in animals that the effect of environ-

[1] This study is part of the research program of the National Merit Scholarship Corporation and is financed by grants from the National Science Foundation, the Carnegie Corporation of New York and the Ford Foundation.

ment on behavior varies with the genetic characteristics of the animal.

## Estimation of Heritability

Twins raised together in the same family form a natural experiment in which the two kinds of twins vary in hereditary similarity, but are presumably about equal in similarity of environment. The influence of heredity can be inferred from the greater similarity of identical twins than of fraternal twins. The estimation of the proportion of the variance in the trait which can be attributed to heredity poses some additional problems, but with certain assumptions it is possible to estimate the proportion of variance of a trait which is due to variation in heredity.

Figure 1 shows a schematic representation of the sources of variance in twin data. The left-hand vertical line represents the total variance of a trait in identical twins and the right-hand line the total variance in fraternal twins. The possible sources of varia-

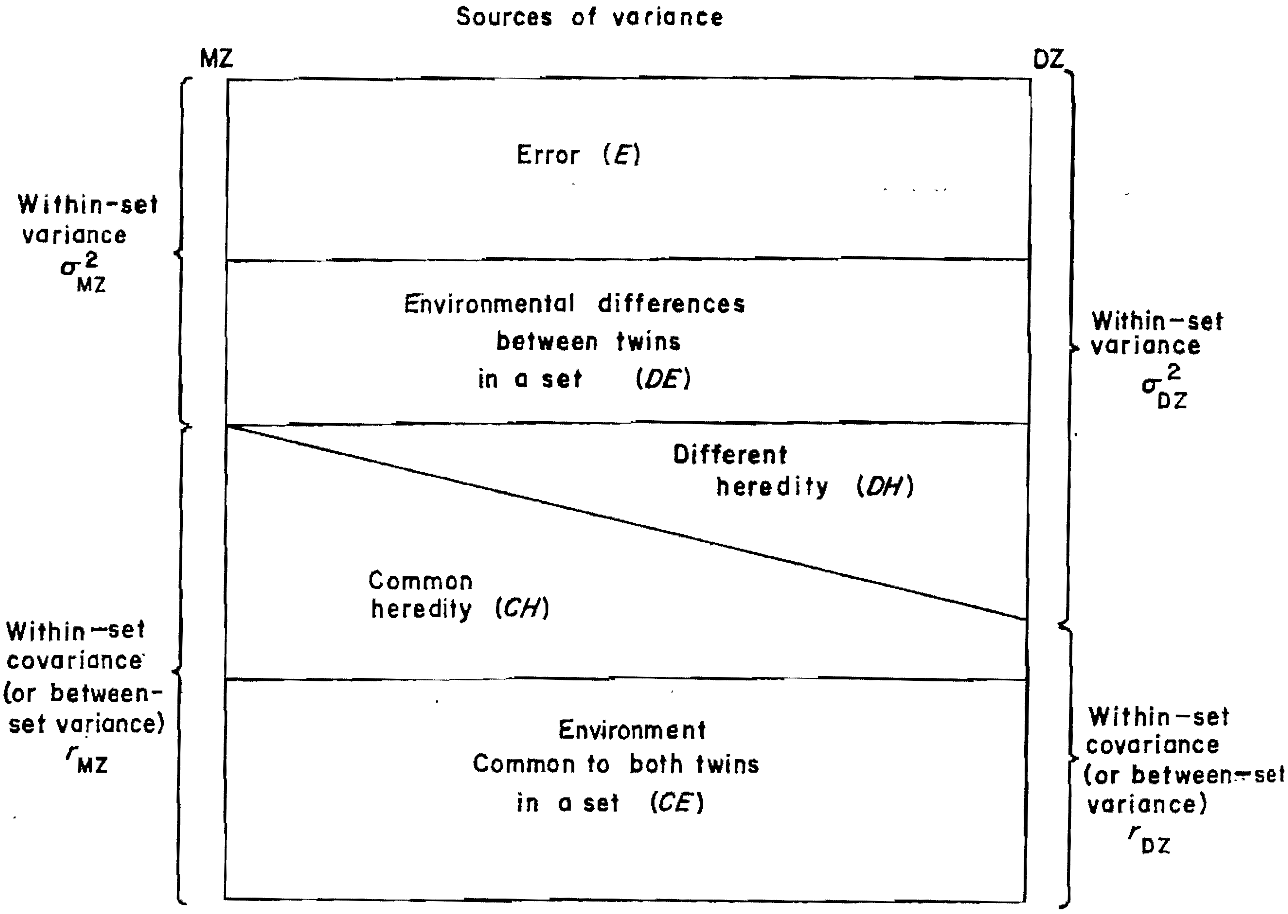
Sources of variance
MZ
DZ
Within-set variance $\sigma^2_{MZ}$
Error (E)
Environmental differences between twins in a set (DE)
Within-set variance $\sigma^2_{DZ}$
Different heredity (DH)
Common heredity (CH)
Within-set covariance (or between-set variance) $r_{MZ}$
Environment Common to both twins in a set (CE)
Within-set covariance (or between-set variance) $r_{DZ}$

Fig. 1. Schematic representation of sources of variance in twin data.

tion are listed between the two lines. Both hereditary and environmental variance is divided into that common to twins of a set and that which is different for the two twins of a set.

The particular power of the twin method lies in the fact that the difference between the intraclass correlations for identical and fraternal twins is equal to the proportion of the total variance due to hereditary differences between fraternal twins. Since fraternal twins have on the average half their genes in common, this is half the hereditary variance in the trait. This fact can be used to construct heritability coefficients from twin correlations which give estimates of the proportion of the variance in a trait attributable to heredity. The coefficient $h^2$ proposed by Holzinger (1929) is the ratio of half the hereditary variance to the variance within sets of fraternal twins.

$$h^2 = \frac{\sigma^2_{\mathrm{DZ}} - \sigma^2_{\mathrm{MZ}}}{\sigma^2_{\mathrm{DZ}}} = \frac{r_{\mathrm{MZ}} - r_{\mathrm{DZ}}}{1 - r_{\mathrm{DZ}}} = \frac{DH}{DH + DE + E}$$

Since error of measurement enters into the within-set variance the correlations are usually corrected for attenuation. Another coefficient which we have developed for use in our twin study is called

$$HR = \frac{2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}})}{r_{\mathrm{MZ}}} = \frac{2DH}{CH_{\mathrm{MZ}} + CE}$$

*HR*. This is the ratio of the hereditary variance to variance due to heredity and environment common to both twins of a set. If one is willing to assume that the major environmental influences on a trait, at least those which might be measured or manipulated, are common to both twins of a set, this ratio is the proportion due to heredity of the variance attributable to heredity and major environmental variables. This ratio also offers the advantage of not including error variance and thus not requiring correction for unreliability of the measuring instruments.

The schematic representation in Fig. 1 and the logic behind the heritability coefficients make certain assumptions which may or may not hold in a given study or with respect to a given trait being investigated. The four major assumptions are the following: (*a*) that the similarity of environmental influence is the same for fraternal and identical twins; (*b*) that for the trait in question there is no correlation between parents due to assortive mating (although if the correlation between parents is known it can be corrected for);

(*c*) that hereditary variance in a continuous trait being studied shows no dominance or interaction effects; (*d*) that hereditary and environmental influences are not correlated (although small correlations make little difference).

Relatively minor violations of these assumptions may not be serious, since the heritability coefficients are subject to large sampling fluctuations and even with large samples can give only rough approximations of the heritability of a trait. Since the heritability coefficients depend on ratios between correlations, they are much more stable with high correlations than when the correlations are relatively low. Thus, in any given sample of twins we may be able to get stable estimates of the heritability of intelligence measures where the correlations tend to be high, but much less reliable estimates of the heritability of personality traits where the correlations tend to be low.

## Sampling Procedures

As a part of National Merit's annual talent search, a 3-hour scholastic aptitude test is administered to about 600,000 high school juniors each spring. In 1962, we included an item on the test asking if the student was a twin. From those checking they were twins, the 1507 pairs who met all of the following criteria were selected: (*a*) same sex, (*b*) attended same high school, (*c*) same last name, and (*d*) same home address. This procedure missed many sets of twins, probably those with the greatest differences in ability, those with differences in educational or health histories, and those who were separated. Assuming approximately one twin birth in every 100 of which two-thirds are fraternal, about 4000 sets of like-sex twins would be expected from a total sample of 600,000, and less than half that many were obtained. Yet this procedure yielded a large number of twins of the same age for whom test scores were available. The exclusion of twins with markedly different experiences is desirable for most research objectives.

## Twin Diagnosis

The first problem, and a crucial one for our study, was to develop an adequate method for the diagnosis of zygosity. The older studies of twins have almost always diagnosed twins as identical or fraternal on the basis of similarity of appearance. Recently, however, blood typing has made these diagnoses much more precise, since blood can be typed with great accuracy for a number of independent, categorical, genetically determined characteristics which have high penetrance. Identical twins are of course alike on all blood groups and it is possible to calculate the probability of getting any combination of blood groups alike by chance in fraternal twins (Smith & Penrose, 1955). If all available blood groups are used an average accuracy of around 98% can be obtained. Since we had a large sample of twins which was geographically dispersed, we were very interested in finding an easy method of diagnosis which had sufficient accuracy to allow us to use our large sample. If we had to rely on blood typing the sample size would be greatly reduced. Most investigators have tended to assume that diagnoses based on physical similarity are of questionable accuracy, but nevertheless we decided to attempt the construction of a diagnostic index based on physical similarity.

We sent all of our twins a four-page questionnaire asking questions about similarity of physical appearance, and replies were received from both twins of 79% of the sets and from at least one twin of 82%. By contacting twins who were in college through the college health services, we were able to obtain usable blood samples on both twins of 124 sets. The blood was typed for 22 anti-sera, and all analyses were run twice using anti-sera from different suppliers. On the basis of the blood analyses, 42 sets were diagnosed fraternal and 82 were diagnosed identical. All 42 of the fraternal sets and an equal number of the identical sets were set aside for cross-validation. We used the remaining 40 identical sets and some fraternal sets diagnosed by marked differences in hair and eye color to develop a preliminary diagnostic index on the basis of the questionnaire physical similarity data. The index which was developed is applied on two levels. The first-level diagnoses are based on a single marked dissimilarity in height, weight, eye or hair color or on marked similarity as indicated by mistaken identity. The second-level diagnoses are based on combinations of several less striking indications of similarity or difference. Table I shows how well this preliminary index worked on the blood-diagnosed cross-validation cases. About half of the cases were diagnosed at the first level and about half at the second level. The over-all accuracy of diagnosis was 93%. The index was revised on the basis of the experience with the cross-validation sample and all cases were diagnosed using the revised index. As a further indication of the validity of the index diagnoses, intraclass correlations for the NMSQT (National Merit Scholarship Qualifying Test) composite score were calculated for identical and fraternal twins diagnosed by various methods. Less accurate diagnoses would be expected to lead to a higher intraclass correlation for fraternal twins and a lower correlation for identical twins. As can be seen from Table II, the correlations for identical

TABLE I

Cross-Validation of Preliminary Diagnostic Index

| | Blood diagnosis | |
|---|---|---|
| Index diagnosis | MZ($N$ = 42) | DZ($N$ = 42) |
| Level one | | |
| MZ | 15 | 2 |
| DZ | 1 | 21 |
| Undiagnosed | 26 | 19 |
| Level one and level two | | |
| MZ | 38 | 4 |
| DZ | 1 | 35 |
| Undiagnosed | 3 | 3 |
| Index with intuitive diagnosis of uncertain cases | | |
| MZ | 40 | 4 |
| DZ | 2 | 38 |

TABLE II

Intraclass Correlations of the NMSQT Composite Score for MZ and DZ Twins Diagnosed by Various Methods

| | MZ | | DZ | |
|---|---|---|---|---|
| Method of Diagnosis | Correlation | $N$ | Correlation | $N$ |
| Blood, hair, and eye (index development and cross-validation cases) | .88 | 82 | .68 | 199 |
| Index level one | .89 | 92 | .65 | 137 |
| Index level two | .87 | 513 | .59 | 146 |
| Undiagnosed by index, Diagnosed intuitively | .87 | 50 | .77 | 20 |

twins were about the same for all methods of diagnosis, and the correlations for fraternal twins actually decreased (but not significantly so) from the blood diagnoses to the less certain index diagnoses. Those twins who could not be diagnosed by the index—cases which gave no clear indication either of similarity or dissimilarity—were apparently mainly identical, but they were not included in any further analyses.

On the basis of experience with the development of the index, we have concluded that it is not difficult to diagnose twins, and that diagnoses almost as accurate as blood diagnoses can be made very simply. The feasibility of diagnosing twins by mail opens up many new possibilities for large-scale twin studies. All extensive testing programs should include an item inquiring about twin status, and diagnosis by the physical similarity index would provide large samples of twins on which test scores were already available.

## Inheritance of General and Specific Ability

The first analysis we have done with the twin data is a study of twin similarity on the National Merit Scholarship Qualifying Test, the three-hour aptitude test by which the twin sample was first identified. In this study, we were concerned with the heritability of general intelligence, which has been well researched, as well as the heritability of specific abilities, which has been little studied.

Most previous studies of the inheritance of ability have been concerned with general intelligence, although it is now generally recognized that intelligence is not a unitary trait. There are a number of different abilities, which tend to be positively correlated. There are two major explanations for this structure of intelligence. The first explanation, growing out of Spearman's theory of general and specific abilities, is that the general level of ability is determined mainly by heredity and that the ability to perform particular tasks is either facilitated by or is not facilitated by environmental experience producing specific abilities. The second explanation, growing out of Thompson's sampling theory, is that many very specific abilities are hereditarily determined and that the phenomenon of general intelligence arises from the use of the same specific abilities in many different tasks and from the generalized effects of environmental influences. Intelligence, like all behavior, is determined by some complex interaction of heredity and environment, but the question remains whether the herediatry component is mainly general—or mainly specific.

The NMSQT consists of five subtests measuring achievement in different academic areas. A factor analysis of these five subtests indicated that there is one general factor and that once the influence of this general factor is taken out, the residual variances of the five subtests are relatively independent. We wanted to study the heritability of general ability and of the specific abilities measured separately by each subtest. General ability is well represented by the composite score (the sum of the standard scores for the five substests). In order to get a measure of the specific factor measured by each subtest, we computed multiple regression equations for predicting each subtest score from the other four subtests, and then computed residual scores for each subtest which were independent of the other four subtests. The reliabilities of the subtests vary around .90 and the multiple correlations of the subtests with the other four vary around .75, so that there is a reliable portion of each subtest which does not overlap with the others. However, the residual scores contain a large proportion of error so that we would not expect large correlations.

The heritability of the composite score will show the inheritance of general intelligence, and the heritability of the residual scores will show the inheritance of certain fairly specific abilities which are independent of general intelligence. These heritability coefficients are shown in Table III.

As we would expect, the identical twins were much more alike than the fraternal twins with respect to the composite score. Both heritability coefficients indicate that about 70% of the variance in the composite score can be attributed to differences in heredity, and this is about the same value that has been obtained by other investigators, some of whom have used different methods of estimating heritability.

The twin correlations for general ability show a most encouraging consistency with the results of other twin studies. Table IV gives a summary of results of studies using a variety of tests conducted in different countries, in different languages, over a period of 35 years. The consistency of the findings is striking.

Now turning to the heritability of the residual scores, we can see

## TABLE III

### Intraclass Correlations and Heritability Ratios for NMSQT Subtest, Composite and Residual Scores

| | Intraclass correlation | | | | | Intraclass correlation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MZ | DZ | | | | MZ | DZ | | | |
| Test | $N = 315$ | $N = 209$ | t[a] | *HR* | $h^2$[b] | $N = 372$ | $N = 273$ | t[a] | *HR* | $h^2$[b] |
| Composite | .87 | .62 | 6.62[c] | .73 | .72 | .88 | .65 | 7.76[c] | .65 | .74 |
| Subtests | | | | | | | | | | |
| English Usage | .71 | .64 | 1.46 | .22 | .27 | .77 | .49 | 6.16[c] | .99 | .67 |
| Mathematics Usage | .74 | .42 | 5.69[c] | 1.16 | .80 | .70 | .47 | 4.48[c] | .87 | .65 |
| Social Studies Reading | .76 | .50 | 5.08[c] | .92 | .67 | .79 | .52 | 6.01[c] | .92 | .60 |
| Natural Science Reading | .69 | .52 | 2.96[c] | .60 | .51 | .66 | .48 | 4.48[c] | .68 | .50 |
| Word Usage (Vocabulary) | .85 | .64 | 5.32[c] | .60 | .69 | .86 | .64 | 6.78[c] | .62 | .75 |
| Residula Subtests | | | | | | | | | | |
| English Usage | .40 | .42 | −.25 | −.10 | | .48 | .25 | 3.37[c] | .96 | |
| Mathematics Usage | .48 | .21 | 3.47[c] | 1.14 | | .43 | .23 | 2.74[c] | .92 | |
| Social Studies Reading | .33 | .16 | 2.09[d] | 1.06 | | .29 | .17 | 1.60 | .83 | |
| Natural Science Reading | .27 | .32 | −.61 | −.36 | | .31 | .10 | 2.77[c] | 1.37 | |
| Word Usage (Vocabulary) | .55 | .37 | 2.50[d] | .65 | | .55 | .26 | 4.48[c] | 1.07 | |

[a] t-test of significance of difference between MZ and DZ correlations.

[b] $h^2$ was computed with correlations corrected for attenuation.

[c] $p < .01$

[d] $p < .05$

the instability of the heritability coefficient when dealing with low correlations. We have not reported Holzinger's $h^2$ for the residuals, since correcting the residual scores for attenuation would take us pretty far from reality. All residual subtests showed significant differences between fraternal and identical correlations for at least one test and all residual subtests showed significant heritability in the total sample of both sexes combined. The question now is whether we should attach any significance to the sex differences in the heritability coefficients. I am inclined to think that we should not, and that at least until the study is replicated (we are collecting another sample of twins with NMSQT scores next year) we should attribute the differences in the various heritability coefficients of the residual subtests to the unreliability of ratios between low correlations. If we are willing to assume that the sex differences and the differences between subtests in heritability are due to sampling fluctuations, the average heritability coefficient of the ten

TABLE IV

SUMMARY OF RESULTS OF STUDIES OF THE INTELLECTUAL RESEMBLANCE OF TWINS

| | MZ twin sets | | DZ twin sets | | |
|---|---|---|---|---|---|
| Study | Intraclass correlation | N | Intraclass correlation | N | Test |
| Holzinger (1929) | .88 | 50 | .63 | 52 | Binet IQ |
| Newman, Freeman, Holzinger (1937) | .91 | 50 | .64 | 50 | Binet IQ |
| Newman, Freeman, Holzinger (1937) | .92 | 50 | .62 | 50 | Otis IQ |
| Blewett (1954) | .75 | 26 | .39 | 26 | PMA Composite |
| Husen (1959) | .90 | 215 | .70 | 416 | Swedish Military Induction Test |
| Husen (1960) | .89 | 134 | .62 | 180 | Reading Achievement |
| Husen (1960) | .87 | 134 | .52 | 181 | Arithmetic Achievement |
| Erlenmeyer-Kimling and Jarvik (1963) | .87 | 14[a] | .53 | 11[b] | Various Intelligence Measures |
| Nichols (present study) | .87 | 687 | .63 | 482 | NMSQT Composite |

[a] Median of 14 studies.
[b] Median of 11 studies.

shown may be a good indication of the heritability of specific abilities such as those measured by the residual subtests. The average heritability of the residual subtsets is .75, about the same as that obtained for the composite score. This result supports the point of view that very specific abilities are independently inherited.

One of the assumptions of twin studies which has been most frequently questioned is that the similarity of environmental influences is the same for fraternal and identical twins. If identical twins have had more similar experiences, this might account for their greater behavioral similarity without any hereditary influence. In order to see if this could account in part for the differences shown in Table III, we asked the twins to report all periods of separation and any differences in experiences and illnesses. We repeated the analyses shown in Table III after excluding from the sample all sets of twins who reported separation for 1 month or more, those who reported that one twin had a major illness or disability not shared by the other, and those who reported a major environmental experience (such as marriage or special training) by only one twin. These criteria led to the exclusion of 18% of the identical twins and 25% of the fraternal twins. The correlations tended to be slightly higher for all scores when twins with different experiences were excluded, indicating that the differences in experience tend to produce differences in ability (or perhaps to reflect differences in ability rather than causing them), but correlations tended to be affected about equally for both fraternal and identical twins, so that the conclusions concerning the effects of heredity were not affected.

## Future Plans

In the National Merit Twin Study we have also collected a great deal of personality data on our sample of twins, but results of these data are not yet available. The following briefly outlines the sort of data that have been collected and the kinds of analysis which are planned.

All twins who were diagnosed by the physical similarity index were sent a packet of questionnaire materials which required about 3 hours to answer, and complete returns were obtained from 845 sets (about 72%). These materials included the California Psychological Inventory, the Holland Vocational Preference Inventory, an experimental behavior inventory, and a number of self-rating and life history items. In addition the twins' mothers completed a questionnaire concerning the early life and development of the twins with particular attention to differences between the twins in early environmental experiences. We also obtained teacher and peer ratings of personality differences between the twins for most of the sample. We have just completed a followup of the twins 1 year after the initial personality data were obtained and at the end of the freshman year for those twins attending college. In this followup some of the original measures were repeated and particular attention was given to differences in experiences of the twins during the intervening year.

In analysis of these data we shall be concerned with the heritabilities of the various personality measures and shall try to find a rough ordering of personality traits from those with a large hereditary component to those dependent mainly on environment. We shall attempt to relate personality differences between twins of a set to differences in environmental experiences, and we shall attempt to get some idea of the dimensions of hereditary influences by factor analyses of correlation matrices divided into hereditary and environmental components and by separate factor analyses of items with high and low heritabilities.

## References

Blewett, D. B. An experimental study of the inheritance of intelligence. *J. Mental Sci.*, 1954, **100**, 922–923.

Erlenmeyer-Kimling, L., & Jarvik, L. F. Genetics and intelligence: a review. *Science*, 1963, **142**, 1477–1479.

Holzinger, K. J. The relative effect of nature and nurture influences on twin differences. *J. Educ. Psychol.*, 1929, **20**, 241–248.

Husen, T. *Psychological twin research.* Vol. I. *A methodological study.* Stockholm: Almqvist & Wiksell, 1959.

Husen, T. Abilities of twins. *Scand. J. Psychol.*, 1960, **1**, 125–135.

Newman, H. H., Freeman, F. N., & Holzinger, K. J. *Twins: a study of heredity and environment.* Chicago: Univ. Chicago Press, 1937.

Smith, S. M., & Penrose, L. S. Monozygotic and dizygotic twin diagnosis. *Ann. Human Genet.*, 1955, **19**, 273–289.

## Discussion

**Cattell:** Well while we are on that topic, how do you feel about the reliability of mail questionnaires? Personally I have had a rather disappointing experience, where the same questionnaire administered verbally in a family group situation, was mailed to an equivalent group, the reliabilities dropped appreciably; and I presume that the individual is apt to be under suggestion from someone administering the questionnaire in the homes. This may be quite a serious thing, I feel, and I wonder what your experience is?

**Nichols:** The usual individual may be influenced, but we have had very good luck with it, perhaps because we have been working with very bright students. We also sent this questionnaire to a random sample of people who took the National Merit exam and got a lower response rate from them. In checking them for errors we also found that they made many more errors, e.g. marking their sex incorrectly.

**Cattell:** My circulation was to housewives and it appears that they usually asked their husbands, "Well, my dear, am I talkative?"

**Nichols:** We have tried to check the accuracy of reporting some objective things, and we found that if you ask a straightforward unambiguous question "Did you or did you not do this? What was your grade point average last semester?," you get very good reports. We've checked grade point averages against transcripts from college and found remarkable accuracy in reporting of these things. Now when it comes to attitude items and self-ratings you do run into kibitzers, etc., however, I do feel that the responses are valid.

**Gottesman:** Was it your intent to mail out things like the CPI (California Psychological Inventory) and have them filled out and mailed back to you? I wonder (*a*) if this is ethical and (*b*) if the Psychological Corporation and other companies that publish these tests for restricted use would be happy with that?

**Nichols:** I would answer yes to both of your questions. Gough has been really intrigued with our use of the CPI, and has supported the idea.

**Gottesman:** So there are a lot of CPI's floating around the United States right now.

**Nichols:** Yes, we sent some to Harvard as a matter of fact.