Biometrika (2006), **93,** 3, *pp.* 573–586

Printed in Great Britain

Differential effects and generic biases in observational studies

BY PAUL R. ROSENBAUM

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, U.S.A.

rosenbaum@stat.wharton.upenn.edu

SUMMARY

There are two treatments, each of which may be applied or withheld, yielding a 2×2 factorial arrangement with three degrees of freedom between groups. The differential effect of the two treatments is the effect of applying one treatment in lieu of the other. In randomised experiments, the differential effect is of no more or less interest than other treatment contrasts. Differential effects play a special role in certain observational studies in which treatments are not assigned to subjects at random, where differing outcomes may reflect biased assignments rather than effects caused by the treatments. Differential effects are immune to certain types of unobserved bias, called generic biases, which are associated with both treatments in a similar way. This is explored using several examples and models. Differential effects are not immune to differential biases, whose possible consequences are examined by sensitivity analysis.

Some key words: Differential effects; Differential unobserved bias; Double pairs design; Generic unobserved bias; Observational study; Sensitivity analysis.

1. INTRODUCTION: MOTIVATION, OUTLINE, LIMITATIONS

1·1. *Can some unobserved biases be removed*?

An observational study is an attempt to estimate treatment effects in a context in which it is unethical or infeasible to perform a randomised experiment (Cochran, 1965). In the absence of random assignment, subjects receiving one treatment condition may differ before treatment from subjects receiving another, so that differing outcomes after treatment may reflect biased assignment rather than an effect caused by the treatments. Biases that are accurately recorded in pretreatment covariates may often be removed by analytical adjustments such as matching or covariance adjustment (Rosenbaum, 2002a), but biases that are not measured, or unobserved biases, present substantial problems in observational studies. Various tactics in research design and analysis attempt to address unobserved biases; see Rosenbaum (2002b, § 4, §§ 6–10; 2004).

Here, an additional tactic is examined, namely the removal of generic unobserved biases when examining differential effects. This tactic is not new: it motivates two clever study designs, namely Evans' (1986) 'double pairs design' in research on injury control, and Speilman & Ewens' (1998) sibship transmission/disequilibrium design in genetic epidemiology. A problem in studying accidents is that car crashes vary markedly in severity, and adequate data about speeds, road traction and forces are rarely if ever available. Seat belts may reduce injury, but individuals who wear seat belts may drive

more cautiously than those who do not, in such a way as to reduce the forces of the crash. Evans' double pairs design eliminates certain biases without measuring them; see § 2·1. Similarly, an association between a disease and a genetic marker may be indicative of a gene that causes the disease, or it may merely result from some ethnic, racial, geographic or other population structure associated with the disease. The sibship transmission/disequilibrium design removes such population structures without measuring them by comparing the genes of siblings with and without the disease.

Differential effects are immune to certain unobserved biases, called generic unobserved biases because they affect different treatments in a similar way. The study of differential effects is one tactic for partially blunting certain common concerns about the conclusions of an observational study, and it is important to consider briefly what such tactics can and cannot accomplish.

1·2. *Tactics to address unobserved bias: Can they work?*

One or two careful randomised experiments often end scientific debate, but, even with the best tactics, observational studies are rarely as decisive, and appropriately so. A sensitivity analysis, such as that in Cornfield et al. (1959) for smoking and lung cancer, may indicate that only extremely large unobserved biases could explain an observed association, and it is certainly useful to know this, but nonetheless some observational studies have been affected by extremely large unobserved biases; see Rosenbaum (1988). Campbell (1969) suggested comparing a treated group to two different control groups known to differ on a specific unobserved covariate; however, this useful tactic addresses bias from just one covariate, and there are always others to be considered. Tactics that exploit a highly specific pattern of anticipated treatment effects, as in Trochim (1985), may usefully reduce sensitivity to unobserved bias (Rosenbaum, 2004), but the right pattern of unobserved biases can also produce the anticipated pattern of associations. These tactics reduce, eliminate or quantify an aspect of the uncertainty from unobserved biases, but other aspects remain.

Unobserved biases may replicate along with effects. If one study in Rome is followed by others in Montana and Tokyo, it is likely that, in each successive study, drivers who wear seat belts will drive more cautiously than those who do not. A new observational study reduces uncertainty about unobserved biases if and only if it adequately addresses some aspect not addressed in previous studies (Rosenbaum, 2001); it need not address every aspect. In thinking about tactics to address unobserved biases, the appropriate question is what they contribute to a sequence of studies, rather than whether or not they are definitive in a single study. Evans' double pairs design makes a notable contribution by eliminating the bias due to cautious driving that other designs do not adequately address, even though it is certainly open to other concerns or objections.

2. EXAMPLES OF DIFFERENTIAL EFFECTS AND GENERIC BIAS

2·1. *Effectiveness of seatbelts*

Table 1 is a $2 \times 2 \times 9$ table from Evans (1986) together with its 2×2 marginal table formed by summing over the nine age strata. Consider, first, the 2×2 marginal table. Each unit consists of two people in the front seat of a car during a crash with at least one fatality, so the crash was recorded in the Fatal Accident Reporting System. The table records only crashes in which exactly one of these two people died and exactly one was wearing a seat belt. In Table 1, in the $300 = 189 + 111$ crashes in which the driver was

Table 1. *Mortality in the front seat of a car with one person belted, with age-imbalance strata. Data from Evans* (1986)

		Driver	
		Not belted	Belted
	Age stratum s	Passenger	
	(Driver, Passenger)	Belted	Not belted
Driver died, passenger survived	$s=1$	75	36
Driver survived, passenger died	(16–24, 16–24)	22	92
Driver died, passenger survived	$s=2$	6	6
Driver survived, passenger died	(16–24, 25–34)	4	20
Driver died, passenger survived	$s=3$	2	4
Driver survived, passenger died	(16–24, $\geqslant$35)	2	17
Driver died, passenger survived	$s=4$	12	8
Driver survived, passenger died	(25–34, 16–24)	6	15
Driver died, passenger survived	$s=5$	22	24
Driver survived, passenger died	(25–34, 25–34)	17	30
Driver died, passenger survived	$s=6$	3	6
Driver survived, passenger died	(25–34, $\geqslant$35)	6	21
Driver died, passenger survived	$s=7$	4	8
Driver survived, passenger died	($\geqslant$35, 16–24)	0	8
Driver died, passenger survived	$s=8$	5	9
Driver survived, passenger died	($\geqslant$35, 25–34)	2	16
Driver died, passenger survived	$s=9$	60	52
Driver survived, passenger died	($\geqslant$35, $\geqslant$35)	52	144
Driver died, passenger survived	All strata	189	153
Driver survived, passenger died	combined	111	363

unbelted and the passenger was belted, odds were $\frac{189}{111}=1{\cdot}7$ to 1 that the unbelted driver died, whereas, in the $516=153+363$ crashes in which the driver was belted and the passenger was unbelted, odds were $\frac{363}{153}=2{\cdot}4$ to 1 that the unbelted passenger died.

In Table 1, driver and passenger were in the same car in the same crash. Whether or not the car exceeded the speed limit, whether or not the driver kept a safe distance from the car ahead, whether the driver braked quickly or was distracted, in all cases, these unmeasured characteristics are the same for the driver and the passenger. Obviously, the risk in the driver's seat may always differ from the risk in the passenger seat, but both circumstances are represented in Table 1, so this may be studied. Quite probably, different crashes place driver and passenger at different risks, and this is not recorded. Unbelted drivers with belted passengers may drive differently from belted drivers with unbelted passengers, so there may be differential biases in addition to the generic biases that are removed by the double pairs design.

Table 1 does not provide information about two unbelted or two belted individuals. In a severe crash, an unbelted person may be thrown about, becoming a hazard to others, and Table 1 does not speak to this issue. Quite probably, a car with driver and passenger both belted is driven more safely than a car with driver and passenger both unbelted, so those comparisons are likely to be most biased by unmeasured characteristics of the crash.

Norvell & Cummings (2002) apply a double pairs design to study the effects of helmet use on mortality in fatal motorcycle crashes in which two people were riding on the same motorcycle. Levitt & Porter (2001) use a different but somewhat related strategy in

comparing the cost-effectiveness of airbags and seat belts: they restrict analysis to crashes in which someone died in another vehicle involved in the same crash. As the Fatal Accident Reporting System only records accidents and fatalities, there are certain problems of ascertainment; these do not affect tests of no effect, but they do affect confidence intervals, which require methods similar to those in Rosenbaum (2005). Ascertainment problems arise only in this first example, so they will not be discussed further.

2·2. *Alzheimer's disease and nonsteroidal anti-inflammatory drugs*

The familiar pain-relief medication, ibuprofen, is a nonsteroidal anti-inflammatory drug. Biological reasoning and certain empirical studies suggest that regular use of such drugs may reduce the risk of Alzheimer's disease (in 't Veld et al., 2002). Problems are evident, however. For instance, in 't Veld et al. (2002) comment that, if people who are becoming cognitively impaired are less aware of pain, then a reduced use of these drugs may be a consequence of the undetected early stages of Alzheimer's disease rather than a protective factor. If so, a comparison of users of these drugs with nonusers might mislead.

Acetaminophen is a pain reliever but is not a nonsteroidal anti-inflammatory drug. To the extent that people who are becoming cognitively impaired may be less prone to take medication for pain relief, there is a generic bias, that is a bias which would be expected to produce a negative association between pain relief medication and Alzheimer's disease. Anthony et al. (2000, Table 2) examined various pain medications in relation to risk of Alzheimer's disease: they found that the group that used just nonsteroidal anti-inflammatory drugs had about half the frequency of Alzheimer's disease compared to the group that used just analgesic compounds that were not nonsteroidal anti-inflammatory drugs. The generic bias does not explain the differential effects of these two types of pain relief medication. To explain away the differential effect, there would need to be a differential bias, one that depressed the use of ibuprofen to a much greater extent than it depressed the use of acetaminophen. This finding about differential effects of nonsteroidal anti-inflammatory drugs and other analgesics does not eliminate all concerns, but it does blunt one specific concern.

2·3. *Vitamins and chronic diseases*

Lawlor et al. (2004) demonstrate generic bias in studies of vitamins. They note that observational studies often suggest that vitamins reduce risk of chronic diseases, but the findings are often not reproduced in randomised experiments. They offer several possible explanations; see also Vandenbroucke (2004). Lawlor et al. (2004) show that individuals with high levels of vitamin C or E in their blood tend to have higher socioeconomic status, to eat less fat and more fibre, smoke less, exercise more and are less often obese. Many advantages and good behaviours are packaged together. They show that many of the positive associations they found between vitamins and individual measures of socio-economic status do not disappear when adjustments are made for other measures of socioeconomic status; each measure seems to say that the bias has not been fully controlled using the other measures. This is, by the way, precisely the pattern anticipated from a single, unmeasured generic bias with which these observable quantities are positively associated, a pattern called conditional association; see Holland & Rosenbaum (1986).

The differential effect of two nutrients, vitamins C and E say, compares two treatments: high E with low C and high C with low E. The differential effect is obviously not the main effect: both vitamins might be beneficial to the same degree, so substantial main

effects cancel to leave no differential effect. Nonetheless, as shown later, the differential effect is resistant to certain generic biases. Vitamins with similar sociological associations may have different biological effects, and, when this is true, examination of differential effects helps to identify the biological effects.

2·4. *Drug abuse among pregnant women*

The examples in §§ 2·1 and 2·2 made plausible but unverifiable conjectures about unobserved covariates. The next example looks at certain observed covariates which, in some other study, might not have been observed. This sheds light on the idea of generic bias, but it is not helpful for the particular observational study, because these covariates were in fact observed, so adjustments could be made for them. Therefore, only the issue of covariate balance is examined here. One might conjecture that women who use crack cocaine regularly while pregnant might also be more likely to act in other ways that place the foetus at risk than do women who do not use crack regularly while pregnant. However, one might expect the same pattern of behaviour among women who used marijuana regularly while pregnant. To what extent are these conjectures correct?

The National Institute on Drug Abuse (NIDA, 1999) collected data about 1048 infants born live in 1992 at one of eight hospitals in Washington. In particular, they collected data about illegal drug use and birth outcomes, such as prematurity and low birthweight. With 62 infants excluded whose mothers who may have used heroin, the focus will be on mothers who regularly used either crack or marijuana during pregnancy, or regularly used neither drug, or both. By regular use is meant a response of 'at least once a week'.

In terms of the six covariates in Table 2, the marijuana-only and crack-only groups are much more similar to each other than either is to the women who used neither marijuana nor crack. Almost all of the marijuana and crack users smoked cigarettes during pregnancy, whereas fewer than a quarter of the women in the 'neither' group smoked cigarettes. Similarly, the two user groups were more likely to drink alcohol, to receive only later prenatal care or no prenatal care, to have less than a high school education, to be unmarried and to not want the pregnancy. Many of the differences are of substantial magnitude.

Again, Table 2 is included to illustrate one basic point. Imagine that the covariates in Table 2 had not been measured, and an analysis was performed of the 2×2 factorial of treatments, (marijuana use) $\times$ (crack use). The main effects of crack use and marijuana use are both strongly confounded with all of the covariates in Table 2, and several of these,

Table 2. *Four groups of babies in the U.S. National Institute on Drug Abuse's study of live births in the District of Columbia*

	Regular Crack/marijuana use			
	Neither	Marijuana	Crack	Both
	Number of babies			
	931	11	39	5
Cigarette use during pregnancy (%)	22	100	92	100
Alcohol use during pregnancy (%)	23	64	74	100
Late or no prenatal care (%)	10	27	48	40
Less than high school education (%)	23	54	49	60
Married (%)	34	0	3	0
Did not want to get pregnant (%)	32	46	59	80

such as cigarette use and alcohol use, are known to affect birth outcomes. However, these same covariates, even if they were not measured, do much less to bias the comparison that estimates the differential effect of crack use alone versus marijuana use alone. Obviously, the differential effect is not a main effect: crack and marijuana may both be harmful, perhaps to the same degree. The differential effect of two treatments may be more or less interesting than the main effect of one of the treatments; that depends on the context. The point in Table 2 is simply that, in some contexts, differential effects may be less affected by generic biases.

3. STRUCTURE: A 2×2 FACTORIAL WITH BIASED TREATMENT ASSIGNMENT

3·1. *Notation: Strata, treatments, treatment effects and observed responses*

A finite population of N units is divided into S strata, $s = 1, \ldots, S$, defined by observed pretreatment covariates, x, with n_s units, $i = 1, \ldots, n_s$, in stratum s, so that $N = \sum_{s=1}^{S} n_s$. If there is no observed covariate, take $S = 1$, whereas, for matched pairs, $n_s = 2$ for $s = 1, \ldots, S$. Write $\mathscr{X}$ for the array of observed covariates, x_{si}, for $i = 1, \ldots, n_s$ and $s = 1, \ldots, S$. Strata are homogeneous in the observed covariates, so $x_{si} = x_{sj}$ for all s, i and j. In addition to the observed covariates, in an observational study, there is typically concern about certain specific covariates that were not observed; however, it is not possible to stratify on these. There are two factors, each at treatment or control levels. Write $Z_{sik} = 1$ if the ith unit in stratum s receives treatment k, for $k = 1, 2$, and write $Z_{sik} = 0$ otherwise. A treatment combination specifies the level of both factors, $Z_{si1} = a$, $Z_{si2} = b$ say. In § 2·1, where the unit consists of the two occupants of the front seat of a car, $Z_{si1} = 1$, $Z_{si2} = 0$, indicates that the driver was belted and the passenger was not. In § 2·2, $Z_{si1} = 1$, $Z_{si2} = 0$ indicates use of a nonsteroidal anti-inflammatory drug, ibuprofen say, and not acetaminophen. Write Z for the $N \times 2$ array of (Z_{si1}, Z_{si2}) with the rows in the lexicographical order.

Write $\pi_{absi} = \mathrm{pr}(Z_{si1} = a, Z_{si2} = b)$. For instance, in a completely randomised 2×2 factorial experiment, one would have $\pi_{11si} = \pi_{10si} = \pi_{01si} = \pi_{00si} = \frac{1}{4}$ for all s, i. The π_{absi} describe a unit's chance of exposure to the treatment in the population. Treatment assignments in the population for distinct units (s, i) will be assumed to be independent, with dependence introduced later through conditioning.

Each unit i in each stratum s has a potential response under each treatment combination, four potential responses in total, $(r_{11si}, r_{10si}, r_{01si}, r_{00si})$, of which only one is observed, r_{absi} being observed if $Z_{si1} = a$ and $Z_{si2} = b$; see Neyman (1923) and Rubin (1974). Here, each of the four potential responses, r_{absi}, may itself be a vector. In § 2·1, r_{absi} is a pair of binary variables indicating whether the driver and front passenger would live or die under seat-belt pattern a, b. In § 2·2, r_{absi} is a binary variable indicating Alzheimer's disease under treatment pattern a, b. Write $\mathscr{R}$ for the $N \times 4$ array containing the potential responses $(r_{11si}, r_{10si}, r_{01si}, r_{00si})$ for the N units under the four treatment patterns each unit might receive.

The effect on the ith unit in stratum s of treatment combination a, b as opposed to combination a', b' is a comparison of the potential response r_{absi} and the potential response $r_{a'b'si}$. For instance, if, in § 2·1, in crash (s, i), $r_{11si} = (1, 1)$ and $r_{10si} = (1, 0)$ then both driver and passenger would have lived if both had been belted, whereas only the driver would have lived if only the driver had been belted. In § 2·2, if $r_{01si} = 1$ and $r_{10si} = 0$, then patient i would develop Alzheimer's disease with regular use of just acetaminophen but not with regular use of just ibuprofen.

In particular, the differential effect is the comparison of r_{10si} and r_{01si}. It is the comparison of one treatment in lieu of the other. Of course, the differential effect is only a part of the whole story, often an interesting part, but always just a part. In particular, there might be no differential effect, $r_{10si} = r_{01si}$, and yet a substantial effect, $r_{10si} > r_{00si}$ and $r_{01si} > r_{00si}$. Ibuprofen and acetaminophen might both be beneficial, but equally so. Crack and marijuana might be equally harmful. In § 2·1, in a severe crash, a single unbelted person might be lethal to both people, so that $r_{00si} = r_{10si} = r_{01si} = (0, 0)$ and yet $r_{11si} = (1, 1)$.

Each unit receives some one treatment combination, $Z_{si1} = a$ and $Z_{si2} = b$ say, and the response r_{absi} under that pattern is observed. Write R_{si} for the one observed response of the ith unit in stratum s, so that, formally,

$$R_{si} = Z_{si1}Z_{si2}r_{11si} + Z_{si1}(1 - Z_{si2})r_{10si} + (1 - Z_{si1})Z_{si2}r_{01si} + (1 - Z_{si1})(1 - Z_{si2})r_{00si}.$$

In randomisation inference, as developed by Fisher (1935, Ch. 2), the array of potential responses, $\mathscr{R}$, is a fixed feature of the finite population of N subjects, whereas quantities that depend on the treatment assignment, (Z_{si1}, Z_{si2}), such as R_{si}, are random variables. To say that a quantity is fixed is to say that all probabilities implicitly condition on its value. To be specific, to say that randomisation ensures that $\pi_{11st} = \pi_{10si} = \pi_{01si} = \pi_{00si} = \frac{1}{4}$ for all si implicitly means that the chance that unit si receives a treatment does not vary with the potential responses $(r_{11si}, r_{10si}, r_{01si}, r_{00si})$ that unit si would exhibit under alternative treatments; this is, of course, true in a randomised experiment, but perhaps not in an observational study. To emphasise, the π_{absi} are implicitly conditional probabilities given the observed and unobserved covariates and potential responses $\mathscr{R}$.

3·2. *Models for treatment assignment when there is no unobserved bias*

Treatment assignment is said to be free of unobserved bias, or ignorable, if π_{absi} does not depend on i, so that

$$\pi_{absi} = \zeta_{abs}, \tag{1}$$

say, for all a, b, s, i. This means that two units, i and j, in the same stratum s have the same chance ζ_{abs} of receiving treatment ab, although that chance may vary from stratum to stratum. This is true by design in a completely randomised experiment, as $\pi_{absi} = \frac{1}{4}$ for all a, b, s and i. In an observational study free of unobserved bias where (1) is true, the ζ_{abs} may vary with s, that is, with the observed covariates. In Table 1, for instance, the strata specify the ages of driver and passenger, where $s = 1$ if both are between 16 and 24, $s = 7$ if the driver is at least 35 and the passenger is between 16 and 24, and $s = 9$ if both are at least 35. If the study is free of unobserved bias, $\pi_{absi} = \zeta_{abs}$, then the chance of exposure to a combination of treatments varies with s, that is with age in Table 1, but π_{absi} does not vary with variables that were not measured.

When treatment assignment is free of unobserved bias (1), inferences can be drawn about the effects of the four treatments in a straightforward way, simply by adjusting for the strata. Write m_{abs} for the number of units receiving treatment combination ab in stratum s, so that $m_{10s} = \sum_{i=1}^{n_s} Z_{si1}(1 - Z_{si2})$ for example, and write $m_s = (m_{11s}, m_{10s}, m_{01s}, m_{00s})$ and $m = (m_1, \ldots, m_S)$. If treatment assignment is free of unobserved bias, then m is sufficient for the unknown $\pi_{absi} = \zeta_{abs}$, and the conditional distribution $\mathrm{pr}(Z = z|m)$ of treatment assignments Z given m does not depend on the unknown π_{absi}. Indeed, if the study is free

of unobserved bias, $\text{pr}(Z=z|m)$ is a randomisation distribution:

$$\text{pr}(Z=z|m)=\left(\prod_{s=1}^{S}\frac{n_s!}{m_{11s}!m_{10s}!m_{01s}!m_{00s}!}\right)^{-1} \tag{2}$$

for all Z such that m_{abs} units in stratum s receive treatment ab. For instance, (2) is the randomisation distribution in a stratified randomised experiment yielding, for example, the randomisation distribution of the aligned rank test of Hodges & Lehmann (1962) as described by Tardif (1981). In words, if the study is free of unobserved bias, it may be analysed as if it arose from a stratified randomised experiment in which m_{abs} subjects were picked at random to receive treatment ab, $a=0,1$, $b=0,1$ in each stratum s; see Rosenbaum (2002b, § 8.3.2) for detailed discussion.

In § 2·1, one worries that (1) might be false, that age is not the only concern, because a disposition to wear seat belts may go hand-in-hand with other safe-driving practices, such as sobriety, moderate speed, greater distance from other cars, and so on. There were similar concerns about unobserved covariates in §§ 2·2–2·4.

3·3. *Models for treatment assignment when there is generic unobserved bias*

Suppose that there is a relevant, unobserved covariate, u_{si}, which should have been controlled by stratification, but was not controlled because it was not measured and recorded. Write $\mathscr{U}$ for the u_{si} $(i=1,\ldots,n_s, s=1,\ldots,S)$. If the treatment assignment probability, π_{absi}, varied not only with the stratum, s, but also with u_{si}, then (1) is not true; units in the same stratum differ in terms of π_{absi} and u_{si}. For instance, in § 2·1, u_{si} might measure a disposition to drive safely, while in § 2·2 u_{si} might measure awareness of pain. The purpose of this section is to distinguish generic unobserved biases from differential unobserved biases, both of which involve u_{si}.

Consider as illustrations the following models, in which (1) is false because π_{absi} varies with u_{si}. The first is one of the simplest and best known latent variable models for binary variables due to Georg Rasch (1980); see also van der Linden & Hambleton (1997). Here

$$\pi_{absi}=\frac{\exp\{a(\kappa_{s1}+\phi_s u_{si})\}}{1+\exp(\kappa_{s1}+\phi_s u_{si})}\frac{\exp\{b(\kappa_{s2}+\phi_s u_{si})\}}{1+\exp(\kappa_{s2}+\phi_s u_{si})}. \tag{3}$$

If $\phi_s=0$ for all s, then (3) implies (1), but, if $\phi_s\neq 0$ for some s, then (1) is false if u_{si} varies in stratum s. The second model is a bivariate logit model of a type discussed by Cox (1970, § 7.6):

$$\pi_{absi}=\frac{\exp\{a\kappa_{s1}+b\kappa_{s2}+ab\kappa_{s12}+\phi_s(a+b)u_{si}+\psi_s abu_{si}\}}{1+\exp(\kappa_{s1}+\phi_s u_{si})+\exp(\kappa_{s2}+\phi_s u_{si})+\exp(\kappa_{s1}+\kappa_{s2}+\kappa_{s12}+2\phi_s u_{si}+\psi_s u_{si})}, \tag{4}$$

where again (1) would be true if $\phi_s=\psi_s=0$ for all s, but is not true if $\phi_s\neq 0$ or $\psi_s\neq 0$ for some stratum s if u_{si} varies in that stratum. For instance, in § 2·2, if u_{si} measures awareness of pain, one would expect $\phi_s>0$ and $\psi_s>0$, so that greater awareness of pain promotes greater use of pain relievers, such as ibuprofen and acetaminophen; see Chamberlain (1980) for a related model.

A third model is related to ideas of Tversky & Sattath (1979) about preference trees. They imagine that choices among several alternatives are made in a hierarchical or tree-structured manner. For instance, in § 2·2, the decision to seek pain relief might come first,

followed by a selection of a specific analgesic; that is, one might first decide whether or not to use analgesics regularly, and, if a decision is made to use analgesics, then decide to use a nonsteroidal anti-inflammatory drug or another type of analgesic or both. Suppose that the initial decision to use analgesics has probability t_{si} and the second decision to use the specific analgesics (a, b) with $a+b>0$ has conditional probability ς_{absi}, so that $\pi_{00si}=1-t_{si}$, $\pi_{01si}=t_{si}\varsigma_{01si}$, $\pi_{10si}=t_{si}\varsigma_{10si}$ and $\pi_{11si}=t_{si}\varsigma_{11si}$. The concern in § 2·2 was that the early, undiagnosed stages of cognitive impairment might lead person (s, i) to be less aware of pain, and hence less likely to use analgesics of all kinds, so that the first stage of the two-stage decision might be biased by an unobserved variable u_{si} directly related to the outcome under study; however, there was no obvious reason why early cognitive impairment would lead to a specific choice of analgesic, so perhaps the second stage might not be biased. Define the 'preference tree model' to say that t_{si} may vary with both the stratum s and the individual i, but ς_{absi} varies only with a, b and s, and not with i: $\varsigma_{10si}=\exp(\kappa_{s1})$ and $\varsigma_{01si}=\exp(\kappa_{s2})$. Note that the preference tree model can also describe settings in which one cannot choose both treatments by setting $\varsigma_{11si}=0$.

Write $\rho_{si}=\pi_{10si}/\pi_{01si}$ for the odds of treatment pattern $(Z_{si1}=1, Z_{si2}=0)$ relative to pattern $(Z_{si1}=0, Z_{si2}=1)$. In § 2·1, ρ_{si} is the odds of a belted driver and unbelted passenger relative to an unbelted driver and belted passenger. In § 2·2, ρ_{si} is the odds of use of ibuprofen alone versus acetaminophen alone, and so on. Under the Rasch model (3), the bivariate logit model (4) and the preference tree model, the odds are $\rho_{si}=\exp(\kappa_{s1}-\kappa_{s2})$. This motivates the following definition.

DEFINITION 1. *There are said to be only generic unobserved biases if ρ_{si} varies with s but not with i, that is, if*

$$\rho_{si}=\frac{\pi_{10si}}{\pi_{01si}}=\lambda_s, \tag{5}$$

for all s and i.

The three models serve solely to motivate Definition 1. Under models (3) and (4) and the preference tree model, (1) is typically false, but (5) is always true; that is, these models typically are not free of unobserved bias, yet they only lead to generic unobserved biases.

Since (5) plays a central role, it is useful to express it in another way. Recall from § 3·1 that $\pi_{absi}=\mathrm{pr}(Z_{si1}=a, Z_{si2}=b|\mathscr{R}, \mathscr{X}, \mathscr{U})$. If (1) were true, then $(Z_{si1}, Z_{si2}) \perp\!\!\!\perp (\mathscr{R}, \mathscr{U})|\mathscr{X}$, in Dawid's (1979) notation for conditional independence, so that (1) implies strongly ignorable treatment assignment in the sense defined in Rosenbaum & Rubin (1983). Now (1) may be false when (5) is true; that is, there may be bias from the unobserved u_{si}, but only generic unobserved bias. In particular, if (5) is true, then

$$(Z_{si1}, Z_{si2}) \perp\!\!\!\perp (\mathscr{R}, \mathscr{U})|(\mathscr{X}, Z_{si1}+Z_{si2}=1), \tag{6}$$

which might be described as strong ignorability for the differential effect. It is not difficult to check that, not only does (5) imply (6), but they are equivalent when $\pi_{10si}>0$ and $\pi_{01si}>0$ (Rosenbaum, 1987, § 3·1).

3·4. *Models for treatment assignment when there is differential unobserved bias*

There is differential unobserved bias if (5) is false. For instance, in § 2·2, if people suffering cognitive decline were less aware of pain, and therefore less inclined to use pain relief medication of all kinds, then (5) might be true; however, if their cognitive decline made

them specifically less inclined to use ibuprofen and more inclined to use acetaminophen, then (5) would be false. For instance, in place of (4), the alternative model

$$\pi_{absi}=\frac{\exp(a\kappa_{s1}+b\kappa_{s2}+ab\kappa_{s12}+\phi_s au_{si}+\tilde{\phi}_s bu_{si}+\psi_s abu_{si})}{1+\exp(\kappa_{s1}+\phi_s u_{si})+\exp(\kappa_{s2}+\tilde{\phi}_s u_{si})+\exp\{\kappa_{s1}+\kappa_{s2}+\kappa_{s12}+(\phi_s+\tilde{\phi}_s+\psi_s)u_{si}\}} \tag{7}$$

yields a differential bias if $\phi_s \neq \tilde{\phi}_s$ because $\rho_{si}=\exp\{\kappa_{s1}-\kappa_{s2}+(\phi_s-\tilde{\phi}_s)u_{si}\}$ then depends on u_{si} and varies with i.

The model for sensitivity to differential bias says that two units, i and i', in the same stratum s, may have different values of $\rho_{si}=\pi_{10si}/\pi_{01si}$ and $\rho_{si'}=\pi_{10si'}/\pi_{01si'}$, but the difference is by at most a factor of $\Gamma \geqslant 1$, that is

$$\frac{1}{\Gamma} \leqslant \frac{\rho_{si}}{\rho_{si'}} \leqslant \Gamma. \tag{8}$$

The sensitivity of inferences to violations of (5) will be appraised by computing bounds on inferences assuming (8) for several values of Γ.

4. INFERENCE

4·1. *Preliminaries: Notation and distributions*

Write $L_{si}=Z_{si1}+Z_{si2}$, $L=(L_{11}, L_{12}, \ldots, L_{S,n_S})^{\mathrm{T}}$, so that

$$\mathrm{pr}(Z_{si1}=1|Z_{si1}+Z_{si2}=L_{si})=\begin{cases}0, & \text{if } L_{si}=0,\\ \dfrac{\pi_{10si}}{\pi_{10si}+\pi_{01si}}=\dfrac{\rho_{si}}{1+\rho_{si}}, & \text{if } L_{si}=1,\\ 1, & \text{if } L_{si}=2.\end{cases}$$

Consider the conditional distribution $\mathrm{pr}(Z=z|L)$ of treatment assignments Z given L. For 'concordant' units with $L_{si}=0$ or $L_{si}=2$, treatment assignment (Z_{si1}, Z_{si2}) is determined by L_{si}, whereas for 'discordant' units with $L_{si}=1$ the treatment combination may be $(Z_{si1}, Z_{si2})=(0, 1)$ or $(Z_{si1}, Z_{si2})=(1, 0)$. For this reason, Table 1 excluded the concordant units. Since the concordant units are degenerate given L, it is convenient to have a separate notation for the discordant units. Write $\bar{n}_s$ for the number of discordant units with $L_{si}=1$ in stratum s, renumber these as $j=1, \ldots, \bar{n}_s$, writing $(\bar{Z}_{sj1}, \bar{Z}_{sj2})$ for the treatment combination for the jth such unit, necessarily $(0, 1)$ or $(1, 0)$, and write $\bar{Z}$ for the matrix with $\bar{N}=\sum \bar{n}_s$ rows and two columns containing the $(\bar{Z}_{sj1}, \bar{Z}_{sj2})$ in lexicographical order. Then, because given L the concordant pairs are degenerate, it follows that $\mathrm{pr}(Z=z|L)=\mathrm{pr}(\bar{Z}=\bar{z}|L)$ is

$$\mathrm{pr}(\bar{Z}=\bar{z}|L)=\prod_{s=1}^{S}\prod_{j=1}^{\bar{n}_s}\left(\frac{\rho_{sj}}{1+\rho_{sj}}\right)^{\bar{z}_{sj1}}\left(\frac{1}{1+\rho_{sj}}\right)^{1-\bar{z}_{sj1}}, \tag{9}$$

for $\bar{z}_{sj1}=0$ or $\bar{z}_{sj1}=1$ and $\bar{z}_{sj2}=1-\bar{z}_{sj1}$, for $s=1, \ldots, S$ and $j=1, \ldots, \bar{n}_s$.

4·2. *Inference assuming generic biases only*

If there is unobserved bias, so that (1) is false, but it is only generic bias, so that (5) is true, then, although the generic biases are not observed, they may be removed in the following way. Write $h_s = \sum_{j=1}^{\bar{n}_s} \bar{Z}_{sj1}$ for the number of discordant units in stratum s that received only treatment $k = 1$, so that the remaining $\bar{n}_s - h_s$ discordant units in the stratum received only treatment $k = 2$. Write $h = (h_1, \ldots, h_S)^{\mathrm{T}}$. If there is only generic unobserved bias (5), then $\rho_{si} = \lambda_s$ and (9) equals

$$\prod_{s=1}^{S} \left(\frac{\lambda_s}{1+\lambda_s} \right)^{h_s} \left(\frac{1}{1+\lambda_s} \right)^{\bar{n}_s - h_s}. \tag{10}$$

Note that h is sufficient for $\lambda = (\lambda_1, \ldots, \lambda_S)^{\mathrm{T}}$ in (10).

Consider the conditional distribution $\mathrm{pr}(Z = z|L, h) = \mathrm{pr}(\bar{Z} = \bar{z}|L, h)$; it is a known distribution, free of the unknown parameter λ. Write Ω for the set of possible values, $\bar{z}$, of $\bar{Z}$ given L, h and write $|\Omega|$ for the number of elements $\bar{z} \in \Omega$. Here, $\mathrm{pr}(\bar{Z} = \bar{z}|L, h)$ in (10) attaches the same probability $|\Omega|^{-1}$ to each of the

$$|\Omega| = \prod_{s=1}^{S} \binom{\bar{n}_s}{h_s}$$

different $\bar{z} \in \Omega$ corresponding to the ways of assigning h_s of the $\bar{n}_s$ discordant units in stratum s to $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ and the remaining $\bar{n}_s - h_s$ discordant units to $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$, for $s = 1, \ldots, S$; that is,

$$\mathrm{pr}(\bar{Z} = \bar{z}|L, h) = 1 \Big/ \prod_{s=1}^{S} \binom{\bar{n}_s}{h_s} \tag{11}$$

for all $\bar{z} \in \Omega$. This is exactly the distribution of treatment assignments that would arise if h_s of the $\bar{n}_s$ discordant units in stratum s were randomly assigned to $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$, and the rest to $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$, with independent assignments in distinct strata. This randomisation distribution arises in the randomised crossover studies described by Gart (1969) and Hodges & Lehmann (1973). Alternatively, it is the two-treatment, stratified randomisation distribution in which the two competing treatments are $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ and $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$.

In short, if the study is free of unobserved bias, in the sense that (1) is true, then one obtains a stratified randomisation distribution $\mathrm{pr}(Z = z|m)$ in (2) for all four treatment combinations, and the four treatment combinations may be compared as if the study were a stratified randomised experiment. If there are unobserved biases, because (1) is false, then the randomisation distribution (2) for all four treatment combinations is no longer available, and comparisons of all four treatments may be biased by failure to control u_{si}. However, if there are only generic unobserved biases, so that (1) may be false but (5) is true, then one obtains a randomisation distribution (11) for comparing two treatment combinations, specifically $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ and $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$, that is for inference about the differential effect. For instance, this randomisation distribution could be used to test the null hypothesis of no differential effect, $H_0 : r_{01si} = r_{10si}$ for all s and i, and the test could be inverted to set confidence limits; see Rosenbaum (2002b, § 3).

4·3. *Inference with differential biases*

In § 4·2, generic unobserved biases (5) could be removed when studying differential effects among units receiving exactly one of the two treatments. In fact, conditioning on

a sufficient statistic yielded a randomisation distribution, namely the randomisation distribution for the stratified two-treatment comparison. In parallel, if there are differential unobserved biases, that is when (5) is false as measured by Γ in (8), then the generic component of bias is removed when studying differential effects, and the appropriate sensitivity analysis for differential biases is identical to the sensitivity analysis for a stratified two-group comparison of units receiving exactly one of the two treatments.

In particular, it is not difficult to show (Rosenbaum, 1995, § 1.2; 2002b, § 4.2) that, if we use (8) and condition on L, h in (9), we obtain

$$\text{pr}(\bar{Z} = \bar{z}|L, h) = \frac{\exp(\gamma \sum_{s=1}^{S} \sum_{j=1}^{\bar{n}_s} \bar{z}_{sj1} u_{sj})}{\sum_{v \in \Omega} \exp(\gamma \sum_{s=1}^{S} \sum_{j=1}^{\bar{n}_s} v_{sj1} u_{sj})} \tag{12}$$

with $0 \leqslant u_{sj} \leqslant 1$ for $j = 1, \ldots, \bar{n}_s$ and $s = 1, \ldots, S$, where $\gamma = \log(\Gamma)$. The bounds on u_{sj} in (12) derive from (8). Here, u_{sj} may be thought of as an unobserved covariate that was controlled neither by the stratification nor by the focus on discordant units, and therefore contributes to differential bias. If $\Gamma = 1$ so that $\gamma = \log(\Gamma) = 0$, then (12) becomes the randomisation distribution on Ω. When $\Gamma > 1$ so that $\gamma > 0$, the distribution in (12) is unknown because the u_{sj} are unknown; however, (8) or (12) defines a measured departure from the randomisation distribution. For fixed $\Gamma > 1$, the sensitivity analysis determines upper and lower bounds on inference quantities, such as significance levels; then Γ is varied to display the sensitivity of the inference. Although derived in a new way from (8), the resulting distribution (12) is familiar, and the necessary calculations are given in Rosenbaum (2002b, § 4).

4·4. *Example, continued: Effectiveness of seatbelts*

In § 2·1, it seemed implausible that wearing seat belts was unrelated to other, unmeasured safe-driving practices; that is, (1) seemed implausible. The model of generic unobserved bias (5) seems less implausible, for it says that, for both the passenger and the driver, wearing seat belts is positively related to other safe driving practices, and related in a similar way for passenger and driver. However, (5) is far from certain: there could be differential biases as in (8), because the driver controls the car and the passenger does not. Perhaps a car with an unbelted driver and a belted passenger is driven less safely than a car with a belted driver and an unbelted passenger. A differential bias in the opposite direction is also plausible: perhaps the unbelted parent with a belted child is a more cautious driver than the belted parent with an unbelted child.

Table 1 compares crashes with a belted driver and unbelted passenger $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ to crashes with an unbelted driver and a belted passenger, $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$. The null hypothesis of no differential effect, $H_0: r_{01si} = r_{10si}$ for all s and i, says that, if exactly one person is belted, it does not matter who that is: the same bivariate pattern, R_{si}, of fatalities will occur with $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ and with $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$, so that, for example, only the driver will die regardless, $R_{si} = r_{01si} = r_{10si} = (1, 0)$. As discussed in § 4·2, if there are only generic unobserved biases (5), then this comparison may be based on a known permutation distribution (11). If both passenger and driver die, or if neither passenger nor driver dies, then the permutation distribution (11) receives a constant contribution from such a crash, so that these concordant crashes do not appear in Table 1; see Gart (1969) for closely parallel considerations. Under the null hypothesis of no differential effect, the null distribution of Table 1 from (11) is the distribution of $S = 9$ independent hypergeometric distributions. One suitable test is then the Mantel & Haenszel

(1959) test (Birch, 1964), which uses the total number, in this case 189, of unbelted driver fatalities in Table 1 as the test statistic. Since the margins of the hypergeometric distribution are fixed, this test statistic yields the same significance level as the number of unbelted passenger fatalities, and similarly for the total number of unbelted fatalities. If there were only generic unobserved biases (5) and the null hypothesis were true, 128·7 unbelted driver fatalities would have been expected, rather than the 189 observed, the discrepancy being highly significant, with a Normal deviate of 9·0. Wearing seat belts may be associated with unmeasured safe-driving practices, but if these biases are generic (5), affecting passenger and driver in the same way, then there is no doubt that seat belts save lives.

How large would differential biases have to be to alter this conclusion? Suppose a belted driver with an unbelted passenger can be up to $\Gamma \geqslant 1$ times more or less likely than an unbelted driver with a belted passenger depending upon a covariate u_{si} that was not measured, as in (8); that is, the belt pattern $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (1, 0)$ may be up to Γ times more or less likely than $(\bar{Z}_{sj1}, \bar{Z}_{sj2}) = (0, 1)$ depending upon a covariate u_{si} that was not measured, as in (7) for instance. For a $2 \times 2 \times S$ table, upper bounds on the significance level from (12) are obtained from the convolution of S extended hypergeometric distributions; see Rosenbaum (1995; 2002b, § 4). The analysis just performed based on (5) is the same as the analysis based on (8) with $\Gamma = 1$. For $\Gamma = 2$, 3 and 4, the upper bounds on the significance levels are, respectively, 0·0000071, 0·043 and 0·55. A differential bias of magnitude $\Gamma = 3$ could not explain the pattern seen in Table 1, since for all biases (8) with $\Gamma = 3$ the significance level is 0·043 or less; however, a bias of $\Gamma = 4$ could explain Table 1. In other words, Table 1 cannot be explained away in terms of generic biases of any magnitude, and, to explain Table 1 as something other than a protective effect of seatbelts, one would need to postulate quite substantial unobserved differential biases.

Acknowledgement

This work was supported by a grant from the U.S. National Science Foundation.

References

Anthony, J. C., Breitner, J. C., Zandi, P. P., Meyer, M. R., Jurasova, I., Norton, M. C. & Stone, S. V. (2000). Reduced prevalence of AD in users of NSAIDs and H_2 receptor antagonists. *Neurology* **54**, 2066–71.

Birch, M. W. (1964). The detection of partial association, I: the 2×2 case. *J. R. Statist. Soc.* B **26**, 313–24.

Campbell, D. T. (1969). Prospective: Artifact and control. In *Artifact in Behavioral Research*, Ed. R. Rosenthal and R. Rosnow, pp. 351–82. New York: Academic Press.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Rev. Econ. Studies* **47**, 134–55.

Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *J. R. Statist. Soc.* A **128**, 234–65.

Cornfield, J., Haenszel, W., Hammond, E. et al. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

Cox, D. R. (1970). *Analysis of Binary Data*. London: Chapman and Hall.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *J. R. Statist. Soc.* B **41**, 1–31.

Evans, L. (1986). The effectiveness of safety belts in preventing fatalities. *Accident Anal. Prev.* **18**, 229–41.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Gart, J. J. (1969). An exact test for comparing matched proportions in crossover studies. *Biometrika* **56**, 75–80.

Hodges, J. L. & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Ann. Math. Statist.* **33**, 482–97.

Hodges, J. L. & Lehmann, E. L. (1973). Wilcoxon and t-tests for matched pairs of typed subjects. *J. Am. Statist. Assoc.* **68**, 151–8.

Holland, P. W. & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Statist.* **14**, 1523–43.

IN 'T VELD, B. A., LAUNER, L. J., BRETELER, M. M. B., HOFMAN, A. & STRICKER, B. H. C. (2002). Pharmacologic agents associated with a preventive effect on Alzheimer's disease. *Epidemiol. Rev.* **2**, 248–68.

LAWLOR, D. A., SMITH, G. D., BRUCKDORFER, K. R., KUNDU, D. & EBRAHIM, S. (2004). Those confounded vitamins: What can we learn from differences between observational versus randomized trial evidence? *Lancet* **363**, 1724–7.

LEVITT, S. D. & PORTER, J. (2001). Sample selection in the estimation of air bag and seat belt effectiveness. *Rev. Econ. Statist.* **83**, 603–15.

MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719–48.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments (in Polish), *Roczniki Nauk Roiniczych*, Tom X, pp. 1–51. Reprinted in English in *Statist. Sci.* 1990, **5**, 463–80.

NIDA (1999). Washington DC Metropolitan Area Drug Study (DC*MADS), 1992. U.S. National Institute on Drug Abuse: ICPSR Study No. 2347. http://www.icpsr.umich.edu.8080/ICPSR-STUDY/02347.xml

NORVELL, D. D. & CUMMINGS, P. (2002). Association of helmet use with death in motorcycle crashes: A matched-pair cohort study. *Am. J. Epidemiol.* **156**, 483–7.

RASCH, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

ROSENBAUM, P. R. (1987). Comparing item characteristic curves. *Psychometrika* **52**, 217–33.

ROSENBAUM, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika* **75**, 577–81.

ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Am. Statist. Assoc.* **90**, 1424–31.

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Am. Statistician* **55**, 223–7.

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies (with Discussion). *Statist. Sci.* **17**, 286–327.

ROSENBAUM, P. R. (2002b). *Observational Studies.* New York: Springer.

ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153–64.

ROSENBAUM, P. R. (2005). Attributable effects in case^2-studies. *Biometrics* **61**, 246–54.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

SPEILMAN, R. S. & EWENS, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–8.

TARDIF, S. (1981). On the almost sure convergence of the permutation distribution for aligned rank test statistics in randomized block designs. *Ann. Statist.* **9**, 190–3.

TROCHIM, W. M. K. (1985). Pattern matching, validity and conceptualization in program evaluation. *Eval. Rev.* **9**, 575–604.

TVERSKY, A. & SATTATH, S. (1979). Preference trees. *Psychol. Rev.* **86**, 542–73.

VANDENBROUCKE, J. P. (2004). When are observational studies as credible as randomized experiments? *Lancet* **363**, 1728–31.

VAN DER LINDEN, W. J. & HAMBLETON, R. K. (Eds). (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

[*Received July* 2004. *Revised November* 2005]