Methodological Issues in Outcome Studies of At-Risk Infants

Glen P. Aylward, PhD  
School of Medicine, Southern Illinois University

Objective: To identify methodologic problems found in follow-up studies of infants at biologic and environmental risk and provide solutions and recommendations.

Methods: This article is a literature review.

Results: Problems fall into four groupings: (1) conceptualization/design issues, 2) subject population concerns, 3) procedural issues, and 4) measurement/outcome concerns.

Conclusions: Main-effect models are not useful; confounding and mediating variables must be identified. In addition, the following are needed: alternative analytic techniques, more precise subject selection and characterization of risk factors, geographically defined samples, broadened scope of outcome measures, and use of epidemiologic techniques.

Key words: developmental outcome; follow-up; high risk; infants; methodology.

Pediatric and clinical child psychologists are increasingly involved in follow-up of infants at biological and environmental risk (Aylward, 1997b). This is due in part to declining mortality rates and the resultant expanding numbers of surviving infants at increased risk for developmental morbidity. In addition to a focus on major disabilities (severe mental retardation, cerebral palsy, sensory impairments, and epilepsy), interest has also focused on the evaluation of more subtle, high prevalence/low severity dysfunctions: learning, attention, and behavioral problems; borderline intelligence; poor visual motor integration; deficits in spatial relations; reading and mathematics problems; language disorders; and deficits in “executive” behaviors.

Unfortunately, these follow-up studies often contain methodologic problems that compromise findings. Approximately 10 years ago, we undertook a meta-analysis of 80 follow-up studies published over the preceding decade (Aylward, Pfeiffer, Wright, & Verhulst, 1989) and pooled results of 4,006 infants < 2500 g or smaller and 1,568 controls. Eleven major problems were identified in these low birthweight follow-up studies. Similar problems have been cited in more recent reviews (McCormick, 1997a; Tyson & Broyles, 1996) with the inclusion of several recurrent issues: (1) use of gestational age versus birthweight, (2) changes in test instruments, (3) use of neuropsychological “batteries,” (4) changes in medical procedures (e.g., surfactant, steroids, ventilation), (5) application of sensitivity and specificity measures, and (6) the need to include quality of life measures. Comparability across studies is further reduced because of a lack of central focus or framework for actual data collection due to the diversity of purposes for follow-up (Johnson, 1997).

The following discussion highlights problems found in follow-up of at-risk infants and contains...
suggestions for improvement. These problems fall into four broad areas: (1) conceptualization/design issues, (2) subject populations, (3) procedural issues, and (4) measurement/outcome.

Conceptualization and Design Issues

Cause-Effect Inferences

In developmental follow-up studies, cause-effect inferences must be tempered by alternative explanations of observed effects that could be produced by confounding influences. Random assignment is not possible in studies of “naturally occurring” conditions such as extremely low birthweight (ELBW). In most follow-up studies, the predominant conceptualization of causal inference is (1) a condition (e.g., ELBW, drug exposure) leads to (2) some type of neurodevelopmental outcome. Unfortunately, this simplistic attribution of causality often is flawed on two ends.

First, multiple factors are associated with conditions such as ELBW (McCormick, 1997a). These include severity of neonatal course (days in hospital, other conditions), sociodemographic factors (socioeconomic status [SES] and social support, race), subsequent illness (asthma, hospitalizations), maternal physical and mental health, and environmental exposures to positive and negative experiences (lead, smoking in household, intervention). In fact, some suggest birthweight is best conceptualized as a marker of concomitant factors that influence outcome. Second, at the outcome end there are neurodevelopmental, cognitive, behavioral, health, and social issues. Various studies have shown that front-end factors that influence outcome vary, depending on the time of assessment, type of early risk factor, and type of outcome measured. In sum, the situation is far more complex than a main-effect, “A → B” model.

Spurious correlations may add much uncertainty to the model; this is particularly concerning given that the data in most outcome studies are correlational or descriptive. Depending on interpretation, Type I or Type II errors could result. Inclusion of confounding variables will reduce the error term and thereby decrease Type I errors. Moreover, measures selected to represent potential confounds must be reliable and valid in and of themselves, because measurement error in the control variables may detract from the validity of any causal inferences that can be drawn (Jacobson & Jacobson, 1996). When a perinatal variable and a potential confound such as environmental risk are included in multivariate outcome analysis, a portion of the variance will be attributed to the more reliable predictor solely because it was measured more accurately. Conversely, even if a confounding variable such as environment is very influential, if it is measured inaccurately, its impact will be underestimated. In general, unreliable measurement at the front end may produce Type II errors, whereas improper measurement of a confounding outcome variable may increase variability and produce spurious correlations (Type I error). Type II error is a particular problem when investigators fail to detect subclinical behavioral or developmental deficits because of insensitive test instruments.

This situation argues for use of measures with demonstrated reliability and validity. Moreover, care must be taken to separate potential confounding from mediating variables. Although both may reduce the attributable influence of a particular perinatal variable on outcome, interpretation of this influence depends on a priori categorization of which variables theoretically are expected to function as mediators and which as potential confounders; treatment of a mediator as a confounding variable may lead to the incorrect inference of a spurious correlation or a Type II error (e.g., Jacobson & Jacobson, 1996). Both may be tested by adding a control variable to the multivariate analysis. However, in the case of VLBW and environmental influences, interpretation of a reduction in the effect of birthweight on some outcome measure after environment is added depends on the hypothesized function of this control variable (environment).

Selection of control variables should be determined both on a conceptual basis and by univariate correlations that indicate at least a weak relationship between the variable and the outcome measure of interest. Multiple regression (stepwise and hierarchical) is often useful, but is problematic when multicollinearity exists due to highly correlated predictor variables, or when many correlated outcome variables are measured. Structural equation modeling (interrelations among composite “latent” variables are derived from multiple measures), partial least squares methods (sometimes considered a variant of structural equation modeling: permits detection of basic underlying patterns of association between constructs [Carmichael-Olson, Streissguth, Bookstein, Barr, & Sampson, 1994]), the use of
Subject Populations

Birthweight and Gestational Age

Prior to the 1990s, infants were primarily grouped by birthweight versus gestational age because of the uncertainty of the obstetric estimation of gestational age and the questionable utility of the postnatal assessment, particularly in very small infants (Hack & Fanaroff, 1988). However, fetal ultrasound has improved gestational age estimation, and gestational age is a stronger determinant of organ/system maturation and viability than is birthweight. Moreover, infants of very low birthweight may be (1) extremely premature babies (gestational age) with AGA birthweights, (2) less premature babies with SGA (<3rd percentile) birthweights, or (3) older preterm and term infants with extreme SGA birthweights (Touwen, 1986). This distinction is necessary because the ultimate survival and outcome of infants included in these groups can vary markedly. Therefore, both birthweight and gestational age should be considered in outcome studies, with particular care taken to ensure that only AGA infants are included in specific birthweight categories if this is the benchmark used for grouping subjects.

Medical/Biologic Risk

More precise characterization of the neonatal medical experience or biologic risk is necessary to compare outcomes across hospitals, for benchmarking, and to control for population differences. Various risk scores and neonatal admission severity scores for physiologic status and intensity of therapeutic intervention have been developed (e.g., Pollack et al., 2000). These scores may be used to stratify subgroups or can be controlled statistically in regression or ANCOVA analyses. Because the three major sources of morbidity in the neonatal period are intracranial events, pulmonary immaturity, and infections (McCormick, 1989), severe ultrasound abnormality, sepsis, necrotizing enterocolitis, chronic lung disease/bronchopulmonary dysplasia (BPD), hyperbilirubinemia, apnea of prematurity, retinopathy of prematurity, and indicators of asphyxia, such as seizures, should be included in the selected risk index. The number of days of hospitalization after birth is often used as a marker variable; however, it is disproportionately affected by smaller infants. If used, this variable should be adjusted for birthweight or gestational age.

Control Groups

When inferences are made regarding the outcome of infants from an identified group or those receiving a particular medical intervention, these infants should be compared to some other group to make such inferences meaningful (Kiely & Paneth, 1981). Traditionally, a full-term control group is used, drawn from similar geographic and social circumstances. However, choice of the type of comparison group depends on the purpose of the study and the hypotheses being tested. For example, when considering the incidence of disability in infants born at <800 g, use of a full-term comparison group would not be very informative. A comparison group of infants with birthweights between 800 and 1000 g, drawn from the same population, would be more appropriate. Determination of a control group when evaluating the efficacy of a new procedure is more straightforward.

In the case of ELBW infants, within-group comparisons could be employed, based on arrays of medical/biologic pre- and perinatal factors, or contrasts between those who have done well on a particular outcome measure versus those who have not. Sample stratification can be used when a high degree of confound exists between a particular perinatal variable and one or more background variables. This has been successfully accomplished in studies where medical risk and biological risk are dichotomized (high/low), thereby yielding four possible stratifications. “Oversampling” of infants manifesting a condition under consideration that occurs less frequently (e.g., Grade IV intraventricular hemorrhage [IVH]), also assists in comparisons and decreases the possibility of Type II errors. Conversely, oversampling may be misleading if one were to consider the impact of the risk factor on the overall population.

LISREL in path analyses (how well a hypothesized model fits the actual data), and growth curve modeling are recognized techniques used to delineate more specifically the role of confounding and mediating variables (Keith, 1993; Landry, Smith, Miller-Loncar, & Swank, 1997). Growth modeling is of particular interest: individual differences in development are examined in terms of rate of change (slope) and changes in the actual rate of change (curvature); this technique could allow determination of whether mediating variables differentially affect growth of cognitive or other abilities.
Medical/biologic variables fall into three broad areas: admission status (how “sick” the infant is, typically measured by variables such as birthweight, gestational age, Apgar score), medical response or intervention (ventilation, tertiary/secondary level of care), and sequelae at discharge (need for oxygen, neurosensory deficit, chronic illness). Each of these areas should be considered in follow up. Postdischarge medical status should also be noted, as subsequent hospitalizations are associated with lower verbal, visual-perceptual and visual motor scores, and less positive teacher ratings.

Sources of Bias

Samples. Small, single-hospital samples may yield data with limited applicability because of the variations in routine medical care. For example, the incidence of cerebral palsy can vary fourfold between different neonatal intensive care units (NICUs), and outcomes may differ in terms of whether the NICU is located in a hospital with a training program (marker for teaching hospital) and the volume of babies admitted (proxy for experience) (McCor- mick, 1997b). Although use of control groups drawn from the same hospital population can minimize this effect to some degree, pooling data from a geographically defined sample is more appropriate. Geographically defined studies are sounder because the numbers are larger, inferences are more secure, and hospital selection bias is minimized. Regional data, or those derived from nationwide collaborative networks, are most useful (Vermont-Oxford Trials Network, 1993). The importance of proper selection of the patient population cannot be underestimated, as the incidence of any outcome strongly depends on the “denominator” (i.e., study population) used (Escobar, Littenberg, & Petitti, 1991).

Age Cohort. The age cohort is important due to rapidly evolving changes in medical interventions (Hack & Fanaroff, 1999). For example, 30- to 40-year-old data on asphyxia obtained from the National Perinatal Collaborative Study have questionable relevance today. In terms of contemporary long-term follow-up, by the time school-age data on a particular cohort are collected and analyzed, practice changes in treatment may have occurred (e.g., assisted ventilation in the delivery room, surfactant, and prenatal and postnatal steroids). This argues for clear delineation of medical practices at the time of enrollment in follow-up studies and timely data analyses.

Subject Loss. Subject loss can bias the estimation of rate of handicap in follow-up studies (Tyson & Broyles, 1996). Dropout rates as high as 40% to 50% have been reported over the first year in indigent populations. Risk for dropout increases in larger, less sick babies; those from lower SES households; babies born to single, young mothers; and those not born at a tertiary care hospital. Caretakers of infants with identified problems or disabilities are more compliant with regard to follow-up attendance (Aylward, Hatcher, Stripp, Gustafson, & Leavitt, 1985; Campbell et al., 1993), thereby potentially inflating rates of disability in samples with a high dropout rate. Subject loss of 10% per year should be anticipated, this arguing for power analyses to secure ample subject samples. In addition, the convergent validity of other potentially useful data, such as those provided by home health visitors, primary care physicians, and parent report, should be explored as a means of reducing subject loss (John- son, 1997).

Procedural Issues

Environmental Factors

The Hollingshead Index (1975) was used in approximately a third of the meta-analytic studies, whereas maternal education was the most frequently used single marker variable. However, many children are exposed to both biologic and environmental risk, and this combination is sometimes referred to as “double jeopardy” or “double hazard” (Escalona, 1982; Parker, Greer, & Zuckerman, 1988). Here, nonoptimal biologic and environmental risks work synergistically to affect later functioning (Aylward, 1990, 1992). However, there is a ceiling effect in which a severe biologic risk will minimize environmental influences. Stated differently, the sickest infants are least responsive to environmental influences (Aylward, 1996; Bendersky & Lewis, 1994).

SES (maternal education and occupational status) is an insufficient marker for environmental quality. Social support, which includes tangible components (e.g., housing) and intangible components (attitudes, encouragement), should also be considered. The environment involves both process (proximal aspects experienced most directly; mother-infant interaction) and status features (distal and broader, involving aspects experienced more indirectly; social class; location of residence). Process or proximal environmental variables are more
predictive early on; status or distal factors are more predictive later (Aylward, 1992). Environmental effects become increasingly apparent between 18 and 36 months, with 24 months cited frequently. Environmental variables influence verbal and general cognitive outcome whereas medical/biologic factors are more strongly related to neurolologic and perceptual-performance function (Aylward, 1996; Bendersky & Lewis, 1994). Medical/biologic factors tend to determine whether a developmental problem occurs, but environmental factors temper or exacerbate the degree of problem (Hunt, Cooper, & Tooley, 1988).

Negative components of the environment have a synergistic or additive effect on infants who are biologically vulnerable vis-à-vis the transactional (Sameroff & Chandler, 1975) or “risk-route” models (Aylward & Kenny, 1979). Procedurally, infants can be stratified on some environmental measure (e.g., by quartiles), or environmental effects can be partialled out in statistical procedures. If possible, process and status aspects need to be measured. Because of the changing complexity and composition of contemporary environments, valid, more recently developed measures comparable across studies and administered quickly should be employed (see Aylward, 1997).

**Correction for Prematurity**

The consensus is that correction for prematurity should occur, arguably up to 2 years of age (Hunt & Rhodes, 1977). However, some investigators suggest that correction not be utilized or that that it be applied in an incremental fashion (e.g., half correction), depending on the infant’s gestational age, age at time of measurement, and area of function being assessed (Blasko, 1989; Miller, Debowitz, & Palmer, 1984). Arguments for incremental correction currently are not convincing. Imprecise gestational age estimation, concomitant medical issues, and a lack of consensus whether to correct to 37 or 40 weeks are additional confounds. Until a “correction algorithm” is devised, correction through 2 years is recommended.

**Measurement/Outcome**

**Selection of Outcome Measures**

Because there is no true “gold standard” in developmental assessment, terms such as “sensitivity” and “specificity” are misapplied. Instead, “co-positivity” and “co-negativity” are more appropriate in situations where scores on one test are compared to those obtained on a reference standard. The Bayley Scales of Infant Development (BSID; Bayley, 1969) and the more recent BSID-II (Bayley, 1993) are traditionally considered the best criterion measures. However, analogous to changes in medical procedures having an effect on outcome studies, changes in outcome measures have a similar effect by limiting comparisons. Mean IQ/DQ scores on a given test are estimated to increase 3 to 5 points per decade (Flynn, 1999). Therefore, the mean score of a test developed three decades ago conceivably could increase by as much as 15 points. Such adjustments must be considered when comparing scores in different age cohorts or longitudinally with different versions of the same test, because change could be due to real improvement or be an artifact of the different properties of the two instruments. Therefore, use of outdated tests such as the Stanford-Binet Form LM in current protocols, is not appropriate, nor is it useful to compare this with the more contemporary version of the same test.

Additional controversy surrounds the BSID-II itself (Gauthier, Bauer, Messinger, & Closius, 1999; Matula, Gyurke, & Aylward, 1997; Ross & Lawson, 1997; Washington, Scott, Johnson, Wendel, & Hay, 1998). If corrected age is used to determine the beginning item set, scores tend to be lower because the child is not automatically given credit for passing the earlier item set. The potential to generate several alternative developmental index scores may limit comparability across studies that use BSID-II scores in research protocols.

**Length/Duration of Follow-Up**

A minimum of at least 3 years’ follow-up appears necessary to identify problems of moderate severity and to measure IQ. However, subtler, high prevalence, low severity learning difficulties may not become apparent until later, making follow-up into early school age the most desirable practice. School entry is also an attractive end point because health and other problems can be better defined at this age (McCormick, 1989; Vohr & Msall, 1997). If the end point of follow-up is in itself a time of significant change and variability (such as 12 months of age), outcome measurement might be further compromised (e.g., the child who walks at 15 versus 12 months). In such situations, it is difficult to separate a delay, disorder, or deficit.
Selection of Outcomes

Traditionally, major handicaps were the primary focus in outcome studies. Interest then shifted toward more subtle learning, attention, and behavioral dysfunctions, and borderline IQ. Most recently, there has been a major emphasis on a broader, multidimensional conceptualization of outcome and health, including functional abilities, health status, and health-related quality of life (HRQL; McCormick, 1989, 1997a; Saigal et al., 1996). Children at early biologic risk subsequently have poorer health (e.g., bronchopulmonary dysplasia), related restrictions in ability to engage in usual childhood activities, slower physical growth, and poorer social-emotional development—all of which are not “traditional” morbidity measures yet translate into compromised school performance and other sequelae. As mentioned previously, documenting the child’s health postdischarge is also critical (Tyson & Broyles, 1996).

To evaluate outcome more precisely, in addition to “traditional” measures, profiles of the following areas need to be documented: health status; physical issues/limitations due to health; functional status or quality of life including adaptive behavior and day-to-day living; behavioral problems; social competency; gross, fine, and visual motor skills; and academics. Emphasis on functional measures is relatively new, due in part to difficulty in defining and measuring functional limitations and then relating these limitations to performance status at school, home, and the community (Msall, DiGaudio, & Duffy, 1993). However, during infancy and early childhood, parents must act as a proxy for the child, thereby increasing the possibility of bias (Hack, 1999).

As a result, issue-specific outcome measures should be folded into a basic outcome framework that could be compared across studies. This basic framework should include a follow-up protocol with standardized age at assessment, areas covered, and techniques used. However, more study-specific, narrow-band foci could also be employed. This approach would allow for investigation of specific deficits pertinent to the purpose of the follow-up study in conjunction with more “standard” cognitive, behavioral/social, functional, and health-related outcomes that would be of interest across studies (e.g., Taylor, Klein, Schatschneider, & Hack, 1998). The challenge is to accomplish this in a reasonable amount of time and at an acceptable cost.

Outcome Analyses

The correlation coefficient often is used in descriptive investigations relating perinatal variables and outcomes, or between two developmental scores obtained at different times. Unfortunately, this statistic is subject to the problem of restriction of range, where a fairly homogeneous distribution of scores can produce a low correlation. Moreover, correlations do not provide information regarding individual developmental patterns. Siegel (1985) emphasizes the need to predict ranges of scores, rather than exact scores. Correlations are misleading in that regard, as they assume a level of measurement precision not achieved in psychologic tests, environmental measures, or biomedical variables. Additionally, if risk factors and outcomes have differing distributions, an artificial cap may be placed on correlations and variance.

Because group means may mask individual patterns of cognitive development, and biologic risk groups are heterogeneous in terms of biomedical and sociodemographic variables, cluster analysis is attractive (Koller, Lawson, Rose, Wallace, & McCartney, 1997; Liau & Brooks-Gunn, 1993). This technique allows identification of homogeneous subsets of children with similar developmental patterns. These clusters of infants could be compared on variables “internal” to the cluster (e.g., risk or cognitive scores); variables “external” to the clusters (biomedical and sociodemographic) could be compared across clusters (Koller et al., 1997). This type of analysis has been used for cognitive development and holds promise with neuromotor, functional, and health outcomes as well.

Other useful outcome analyses relating to measures of effect are derived from developmental epidemiologic studies (Scott, Mason, & Chapman, 1999). Here, the interest is on differences in proportions of cases rather than differences in means or variance accounted for. This approach yields qualitatively different information about relationships among risk factors and developmental outcomes than is obtained through more traditional analyses. The risk ratio, typically used in prospective, longitudinal cohorts, reflects the relative increase in the probability of a negative outcome when the infant experienced a risk condition (e.g., ELBW with IVH versus ELBW without IVH, compared in terms of spastic diplegia). Effect of a risk factor (ELBW and IVH) is compared to some other referent group (ELBW).

The odds ratio is typically used in case-
controlled, retrospective studies in which infants are chosen based on whether they exhibit the outcome of interest (spastic diplegia), and data are gathered regarding previous exposure to a risk factor (IVH). The ratio is the increased odds of a negative outcome in infants who experienced a risk factor, relative to those who did not experience the risk factor. This is particularly useful when a condition is relatively rare (e.g., Grade IV IVH). Logistic regression could be employed in this analysis. Both of these techniques are sometimes considered “relative risk,” although this is not universally endorsed; if the incidence of an outcome is rare (<2%), the risk and odds ratio values become similar (Scott et al., 1999). These measures of effect are inherently different from regression/ANOVA models, as small differences in means between two groups can nonetheless lead to a larger difference in the proportion of extreme cases in these groups (i.e., a factor associated with a small mean decrease in IQ nonetheless may account for a larger number of children with mental retardation).

Receiver operating characteristic (ROC) curves can provide a qualitative measure of a test’s diagnostic performance (Centor & Schwartz, 1985) or the accuracy of a variable or grouping of variables in predicting outcome. Here the true-positive ratio rate is plotted against the false-positive rate for different threshold values. Points along the diagonal line indicate an equal chance of positive/negative outcome; the higher the ROC curve is from this line, the better the prediction. The area under the curve (AUC) is a quantitative measure of this discrimination, ranging from .5 (chance) to 1.0 (perfect discrimination). This technique has only recently been used in developmental outcome studies (Pollack et al., 2000).

Effect sizes need to be reported in outcome studies, as a small p value does not necessarily imply an important finding—it simply indicates the null hypothesis is not true (McCartney & Rosenthal, 2000). Traditional p values should be accompanied by estimates of both the size and direction of an effect. Two types of effect size estimates exist: r Family (assessed via correlation) and d Family (comparison of group means). Both are applicable in outcome studies and are more practically useful than binary decisions based on significance/nonsignificance. Effect sizes can be biased by measurement error, methodological choices (e.g., within- versus between-subjects designs) or by minimizing error terms (see McCartney & Rosenthal, 2000).

**Criteria**

There is a lack of consistency in terms of diagnostic criteria. Arguments are made both for and against viewing data as categorical or continuous. For example, in the meta-analysis of LBW studies, had a binary “normal/not normal” categorization been employed, no group differences would have been detected. However, viewing the data in a continuous fashion yielded a 6-point difference between LBW and control infants. It would appear that analyses of continuous data require decisions to include or delete severely involved infants; either option would alter results. Use of categorical methods allows inclusion of these babies but masks more subtle findings. Floor effects and missing data are particularly problematic. “Outliers” whose raw scores cannot be converted into scaled scores (as in the case of a BSID-II score <50) often are “censored” or excluded, and a high frequency of censoring may occur in populations with severely affected infants. As a solution, imputed values may be used (e.g., a score of 49 is recorded to indicate an unscalable score) and data analyzed using standard methods (see Lindsey, O’Donnell, & Brouwers, 2000). Means, corrected for censoring, can be compared to means based on inclusion of imputed values to verify that imputation is appropriate. With the BSID-II, extrapolated raw scores may be considered as an alternative (see Black & Matula, 2000). It is recommended that the mean IQ (and SD) and effect sizes and confidence intervals for each group, the proportion of mental retardation and borderline intelligence, and the proportion of major disorders (CP, blind, deaf) be reported. Comparisons excluding children with major handicaps provide insight as to how children who survive without major handicap fare.

**Conclusions**

A promising area of research includes relating routine brain imaging techniques, such as cranial ultrasound, and less frequently employed techniques, such as cerebral blood flow (PET or SPECT), oxygen or glucose metabolism (PET), and functional activity of the brain (echoplanar or FMRI), to outcomes. Use of biochemical markers such as pro-inflammatory cytokines and protective oligotrophins and neurotrophins (Dammann & Leviton, 1999) may enhance predictive accuracy in outcome stud-
ies. These investigations will require expanded collaboration among pediatric psychologists and other disciplines. It appears that a national consensus is needed with regard to elements and procedures necessary in basic follow-up protocols, to allow comparability across studies and pooling of data. In summary, those involved in developmental follow-up must realize that their data will influence actual decisions made for real children; it is imperative that these data be accurate and based on sound methodology.

References


Acknowledgments

Special thanks to Steven J. Verhulst for his helpful manuscript review.

Received December 1, 1999; revisions received July 1, 2000; accepted October 1, 2000


