In Risk in Intellectual & Psychosocial Development,
ed. by Dale Farran + James McKinney
Academic Press, 1986

11

A Four-Dimensional Model of Risk Assessment and Intervention

KENNETH KAYE

Much of what has been written about the concept of risk is based on concepts of assessment and intervention that I believe are extremely problematic. One reason these two concepts are problematic is that they are usually discussed independently. The other reason is that each is viewed unidimensionally.

It is the thesis of this chapter that (1) the assessment of any sort of developmental risk, and (2) the availability of interventions to reduce it, are two dimensions of the same problem, and furthermore that each of those dimensions has to be viewed two-dimensionally. Therefore, all meaningful discussion of "risk," whether from a research or a policy point of view, requires a four-dimensional model.

Assessment and Intervention Are Inseparable Concepts

For some reason, clinicians and researchers alike often try to separate these two aspects of the same problem. They have tended to deal with assessment without regard to what interventions might be available—without regard to the question, "Assessment for what?" Or they have tried to evaluate intervention strategies with inadequate attention to the question, "Intervention with whom?" I illustrate these tendencies with three examples below. At this point, we need to see the problem of

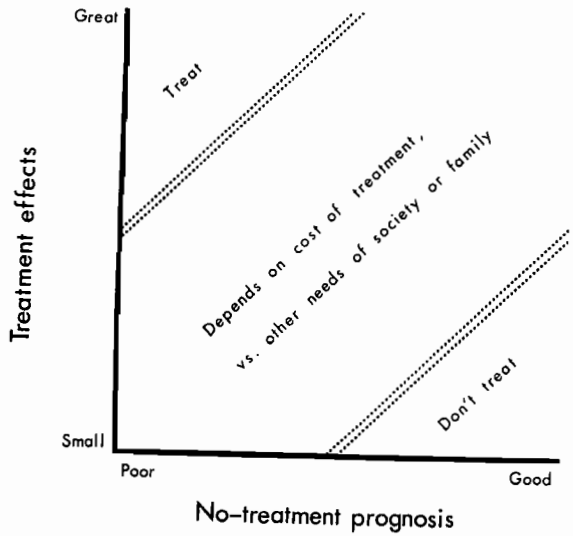

Figure 1. The decision to intervene, viewed two-dimensionally as the intersection of (1) the prognosis if no treatment is attempted and (2) the expected effects of the treatment.

developmental risk as portrayed two-dimensionally in Figure 1. I suggest that the phrase "at risk" really means "should be treated." The patient and society take a risk by failing to take advantage of some available intervention. Clinically, human beings are at risk when two things can be said about them:

1. They can be clearly identified as having a poor developmental prognosis, compared to the rest of the population, if left untreated.
2. A treatment exists that has been shown to improve that prognosis significantly when offered to this segment of the population. (In this context, *significantly* means not just a statistically significant difference, but to a significant extent.)

We all have a chance of getting cancer, but we do not say we are at risk for cancer. Smokers are at risk for lung cancer, not just because they have a higher likelihood of inducing it, but because there is something they can do to reduce that likelihood: quit smoking.

So not everyone with a poor, untreated prognosis is at risk: just those whom we have reason to believe an intervention can help. Similarly, not everyone who might gain from an intervention is at risk: just those whose prognosis is poor without it. The important implication, reiterating a crucial monograph by Cronbach and Gleser (1965), is that an assessor should always be looking for the individuals who need and will

profit from an existing intervention program or one that can be designed, and an intervener should always be designing a program for individuals who can be identified as needing it. This is why the two concepts are inseparable.

Assessment: A Signal-Detection Problem

A fundamental misconception in the field of behavior assessment is that classification is based on measurement theory. Instead, it has to be based upon decision theory (Cronbach & Gleser, 1965, Chapter 11). The measurement-theory approach assumes that people can be located along any given dimension, subject to the precision of available instruments. My assertion that assessment is usually unidimensional may seem strange in view of the fact that most research is clearly aimed at analyzing the multiple factors that put people in a risk category. The instruments themselves, including the neonatal assessment scales and the IQ tests discussed here subsequently, are indeed multifactorial. But the idea of what it means to classify someone along the abscissa of Figure 1 has been seen unidimensionally instead of as a two-dimensional problem in signal detection.

The decision process inherent in assessment (no matter how many factors are involved in the prognosis itself—one, two, or many) is conceptually a two-dimensional problem (Figure 2). The more reliable an instrument is for detecting nearly every instance of a given category (e.g., the category of all newborns at risk for failure-to-thrive and treatable by intervening with their families), the more the instrument can be expected to err in the direction of false positives, misclassifying as risk cases some individuals who do not really belong in that category. Conversely, the more reliable an instrument is in the sense that nearly every case selected will really belong in the category, the more it can be expected to err in the direction of false negatives, by failing to identify some cases that also belong in that category. These two kinds of reliability have been defined as r_β and r_α, respectively (Kaye, 1980).

Figure 1 indicated that the question of whether to intervene would often depend on the cost to society, or to the family itself, of doing so versus not doing so. Now we can see this as having an extra dimension, because the assessment has to consider the likelihood and cost of an alpha error (false positive) versus the likelihood and cost of a beta error (false negative). A single validity coefficient, such as a correlation between the assessment instrument and some outcome criterion, is almost meaningless in this context. In the first place, there are the two kinds of

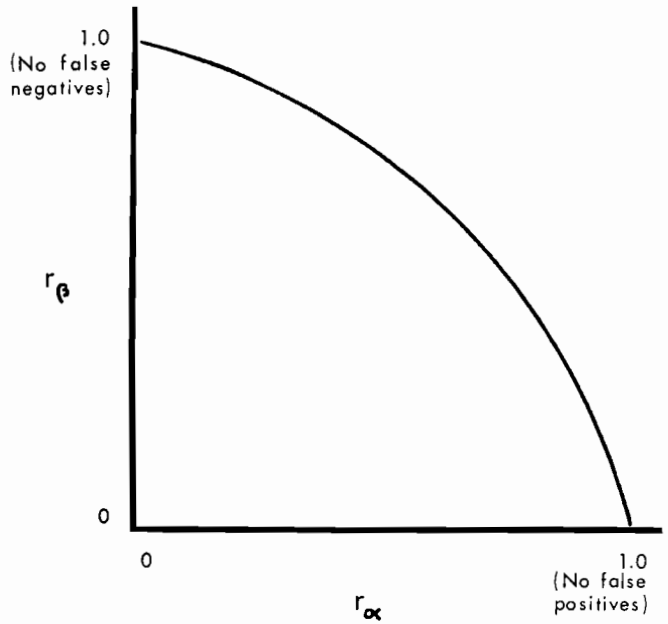

Figure 2. Risk assessment, viewed two-dimensionally. When false positives are eliminated (high r_β), there are more false negatives (low r_α), and vice versa.

reliability, with a trade-off between them that has to be evaluated anew in any particular application. Secondly, the shape of Figure 2 will vary at different points along the abscissa of Figure 1: the reliability of any test is greater, the smaller the selection ratio. For example, the same test can select the top or bottom 10% of a normally distributed sample much more reliably than it can select the top or bottom 25%, and it can do the latter more reliably than it can classify cases as belonging in the top or bottom half. It is not an exaggeration to say that any assessment scale that merely reports a linear validity coefficient has oversimplified the problem to such an extent as to be practically useless.

Intervention: A Cost–Benefit Problem

Even if the untreated prognosis can be stated with certainty, and an intervention exists, there is still a cost–benefit decision to be made. A coma patient with no brain wave may be regarded as at risk if one believes that the costs of continuous intravenous life support are outweighed by the moral or scientific benefits, or may be regarded as legally

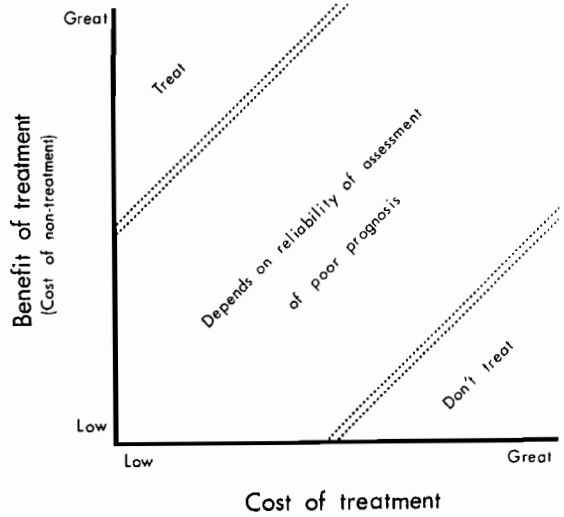

Figure 3. A four-dimensional model of assessment and intervention: benefit, cost, and the two dimensions (shown in Figure 2) of reliability of assessment of poor prognosis.

dead if one does not. Figure 3 needs to be substituted for the ordinate in Figure 1.

Now, we have a four-dimensional model—r_α, r_β, cost, and benefit—which I suggest is really at the root of all policy discussions about intervention as well as all methodological discussions about assessment. Sometimes the issues seem less complex only because we make certain assumptions that simplify the model. For example, we know that all rational people should refrain from smoking. The assessment is easy—you smoke or you don't—and so is the intervention: quit. But the problem becomes this simple only if we ignore the benefits many people apparently believe they get from smoking (relaxation, image enhancement) and only if we accept a high false positive rate in our assessment (the many smokers whose health does not suffer). Smokers themselves have an intuitive understanding of these oversimplifications, so they ignore the surgeon general's warnings. More sophisticated assessments, involving how much of what kind of cigarettes can be smoked by which people at precisely what costs to their health, would give people a more convincing basis for deciding whether the benefits of quitting are worth the physical, mental, and emotional cost of doing so.

The Johnson administration's Project Head Start presented a good example of the cost–benefit question. The real question was not "How much can we boost IQ and scholastic achievement?" (Jensen, 1969), but "What is the cost of boosting it by any given amount?" and "In what

terms should the benefits be evaluated?" It should be obvious that these were not questions that psychologists, educators, and other human development experts could answer. They were political questions as much as technological ones. All questions about intervention are inherently political questions, as are all questions about assessment.

We can turn now to a few examples of how researchers and practitioners concerned about developmental risk have failed to include all four dimensions in their thinking. (Note that I am merely advocating this four-dimensional model conceptually; from a mathematical point of view, more than four factors will often be required.)

IQ Testing: Assessment with Confusion about "Why?"

We have developed in the United States, and continue to maintain despite all the controversy of the 1970s, a system in which nearly every school child is administered an average of four nationally standardized aptitude and achievement tests per year, or about 50 over the course of Grades 1 through 12. At most, 2 or 3 of those 50 tests will be used to track a child through programs loosely designed (but inadequately evaluated) for children of different cognitive capacities; for a small number of children, another couple of testings may be used as part of a diagnostic battery aimed at identifying learning disabilities for which specific interventions may or may not exist. For most children, all such tests—and for all children, most such tests—are administered without any purpose except the perpetuation of a multi-million-dollar industry and the statistical comparison of different demographic groups (Kaye, 1973).

When the justification for a test is supposed to be selection—that is, when it is a placement test—then there ought to be research matching the program into which a group of children are to be placed with the criteria on which they are selected. Let me limit the argument to IQ tests, though much of this applies to standardized achievement tests too. IQ tests usually constitute *mensor gratia mensoris*, measurement for its own sake. They certainly are not tests for risk, unless it can be shown that children of a certain age with IQs in a certain range have a poor prognosis in normal classes and a *better* prognosis—considering all social, emotional, and cognitive costs and benefits—in special school programs. That was the subject of many lawsuits against local and state school authorities in the 1970s and 1980s: the burden of proof fell upon those who purport to assess ability, and that assessment process has repeatedly been ruled inseparable from the question of what interventions are available (Jensen, 1980, Chapter 2).

In the cost–benefit analysis, a crucial issue is the cost of the assessment itself. And the benefit of the assessment must be measured in the value of whatever information it provides beyond what would already have been known without it. Sechrest (1963) called this the "incremental validity" of a test. For example, the so-called validity coefficient of IQ, .80 to .85, gives an indication of how much better the tests predict school achievement than a random throw of dice would predict it. We tend to forget that the previous year's school performance will predict this year's performance very well; IQ only raises the multiple correlation by about 5%. How much is that incremental validity worth? It depends how specifically the information can be used in tracking children into more effective special programs. The reality is that IQ scores, as actually used in schools, give teachers no worthwhile information that could not be more usefully derived from each year's actual school achievement. Even Jensen (1980) takes this position at the same time that he defends the reliability, validity, and lack of bias in IQ tests.

Behavioral Assessment of Newborns

Now, let us extend these same considerations to the high-risk infant. When are newborns at risk? When their developmental prognosis is relatively poor, and when a cost-effective intervention exists that can significantly improve that prognosis. One of the most important factors that can make an intervention cost-effective is selecting all and only the babies and parents who need the program, so that its resources are not wasted on those who do not need it or cannot profit from it. Hence, the assessment of newborns, or of newborn–parent interaction, should be conceptualized as the development of placement tests for actual or feasible intervention programs.

That was one of the explicit aims of Brazelton's (1973) scales: to be a clinical instrument for decisions about the treatment of young infants and their parents. (The other goal was a research instrument.) However, exactly as was the case with general-ability or IQ scales, the clinical uses of newborn assessment—that is, the specific intervention programs for which it could be regarded as a placement test—were not explicitly involved in the creation of the assessment instrument. Perhaps Brazelton and his colleagues felt that not enough was yet known about psychological treatment of high-risk infants and parents, or perhaps they naively believed that an all-purpose assessment could be possible. Consequently, I argue, the Neonatal Behavioral Assessment Scales (NBAS) have extremely limited utility for clinicians concerned with neonatal risk.

One possible use in this area would be a source of research evidence that a population thought to be at risk—because of medical history, socioeconomic status (SES), obstetric medication, and so forth—does or does not show deficits at birth. Among 31 published studies of different samples of newborns using the NBAS (Sostek, 1978), 60% were studies of this kind. (That proportion is even higher, I suspect, among the much larger number of unpublished NBAS studies.) Sometimes the known risk does lead to NBAS deficits as predicted; sometimes it does not. For example, Aleksandrowicz and Aleksandrowicz (1974) found effects of obstetrical medication upon NBAS performance; Tronick, Wise, Als, Adamson, Scanlon, and Brazelton (1976) did not. More important, however, is the fact that even consistent results in this type of study tell us nothing about the risk for such populations. Infants may show severe deficits in the first month of life, yet have a perfectly good developmental prognosis, or they may show no deficits on the NBAS, yet have a terrible prognosis.

A better use of such assessments is to identify which individual infants within a presumed high-risk population are the ones actually at risk. For example, premature infants are considered at risk because samples of premature infants tend to differ significantly from full-term infants on several long-term measures (Sameroff & Chandler, 1975), but this is only because some premature infants have serious neurological deficits and/or cumulative transactional deficits with caretakers. Only that subpopulation is really in need of intervention, so a function of newborn assessment would be to select those babies out of the total premature population. (If all premature infants needed intervention, the assessment scales would be superfluous.) Of the studies reviewed by Sostek, only four, or 13%, tested the NBAS's predictiveness (incremental validity) over and above the known risk factors. The NBAS helped to predict 12-month Bayley developmental scores, whether infants were premature, full-term, or postmature (Field, Hallock, Ting, Dempsey, Dabiri, & Shuman, 1978). Tronick and Brazelton (1975) showed that the NBAS produced fewer false predictions of neurological deficit over a 7-year follow-up than did standard neonatal neurological tests. In terms of Figure 2, the NBAS's r_β was as good as, and its r_α was better than, the neurological test. This is the only published study of which I am aware that treats newborn assessment as a two-dimensional decision problem rather than a unidimensional measurement problem.

Unfortunately, the studies just mentioned had to do only with assessment, not with all four dimensions that are necessarily involved in the concept of risk. Only three, or 10% of the published studies up to the time of Sostek's 1978 bibliography, looked at the intervention question.

Two of those had very short-term follow-ups, simply showing that th NBAS itself was sensitive to the results of extra stimulation with low birthweight infants. The only study with a 12-month follow-up showe significant Cattell IQ score effects for a stimulated group of low-birth weight infants, regardless of NBAS performance. In other words, th intervention was successful with the presumed high-risk group, but th newborn behavioral assessment was completely superfluous.

That, I think, is a fair summary of the status of newborn assessment a a clinical tool: the cost of administering the scales would often be greate than the cost of giving treatment (for example, greater vestibular stim ulation in the nursery) to all infants in a presumed high-risk group. Onl where there might be great cost, danger, or a social stigma attached t the treatment would it be worth developing newborn assessment pr cedures, and then the question in need of research is the four-dimen sional one that has barely been broached.

Until then, perhaps our presumption should be that *every* newborn i at risk. Devoting limited research funds to improving the way society a a whole, and families in general, prepare new human organisms fo personhood makes more sense than devoting those funds to assessmen tools that lack any specific utility for treatment.

The fields of IQ testing and newborn behavioral assessment have i common the fact that far too much energy was wasted on the develo ment of instruments with insufficient attention to the question of wh they were going to be used for. We have in IQ a fairly good predictor school achievement, but equally good political, moral, economic, an pedagogical reasons for ignoring that predictor and giving children th opportunity to disconfirm it. The treatment choices for which Binet assessment was originally designed—whether to leave children in th Paris school system or to remove them—have been replaced with larger and slightly more sophisticated range of alternatives, each which requires its own set of placement tests based on the four-dime sional considerations. As intervention programs for the parents or car takers of young infants become more sophisticated, they too will ea require their own placement tests, for which the NBAS will not serv

Child and Family Therapy: Intervention with Vagueness about "Whom?"

We can also point to areas of research in which more or less th opposite error has prevailed: concentration on intervention technique for children known to be at risk, with inadequate attention to the pro

lem of precisely when a given treatment is indicated. This complaint has been leveled against much psychotherapy research (Epstein & Vlok, 1981; Kauffman, 1977), for example, where different ways of treating schizophrenia are evaluated as though the disease were a precise and unitary phenomenon as unvarying and as well understood as a physical illness, like scurvy. Scurvy is a hemorrhaging condition that results from a vitamin C deficiency. When it is suspected, biochemical diagnostic methods are available, of which the ultimate and best test is whether the symptoms respond to vitamin C. There are few mental or psychosomatic illnesses for which that kind of specific biochemical and/or trial-and-error drug test is possible. Instead, we have diagnostic categories based on a combination of symptomatology and etiology. Like schizophrenia, many of our categories undoubtedly confuse several different underlying causes, which we hope eventually to distinguish from one another. The way to make scientific progress in that direction is to be constantly differentiating types and subtypes. A general principle, as valid in relation to disorders of psychological development as it is in medicine, is that progress in the refinement of diagnostic categories is inseparable from progress in methods of treatment. They must always be two facets of one and the same research program.

This has not always been clear to psychotherapy researchers, who sometimes accept diagnostic categories like hyperactive, autistic, psychotic, depressed, or acting-out as though they were labels for integral entities rather than convenient, conventional, often arbitrary and always blurry dividing points along multidimensional continua. Once the category has been reified, the investigator proceeds to compare different treatment models in terms of their efficacy with this type of patient. Suppose that a sample of children with diagnosis D are randomly assigned to two treatment conditions. If one model is effective with 60% of the patients treated, and its competitor helps only 40%, this should not make the first method the treatment of choice for all future patients with that particular diagnosis. Among the other possibilities, it may mean that 60% of the children classified as D really have a different problem than the other 40%.

To take a specific example, a group of family therapists has reported success in treating the families of children with chronic severe asthma (Liebman, Minuchin, & Baker, 1974). The alternative to family therapy was individual child therapy, attempting to address the emotional problems that seemed to contribute to the severity and chronicity of the asthma, and continued dependence upon steroids, allergic desensitization, and bronchodilation exercises. The authors presented a theoretical rationale for structural family therapy and reported seven case studies.

Each child had visited the emergency room frequently and had bee hospitalized at least three times in the previous year. After 5 to 1 months of family therapy, the symptoms went into remission, and at th time of publication, having been followed for another 10 to 22 month subsequent to this therapy, none of the seven cases had suffered acut attacks.

These were impressive results. No competing therapeutic model ha been as successful, to date. Similar results have been reported for struc tural family therapy with anorexia nervosa, intractable diabetes, an severe chronic gastric disorders in children (Minuchin, Baker, Rosman Liebman, Milman, & Todd, 1975). However, the applicability of thei work is severely limited by the lack of any systematic research on whe family therapy is actually indicated for chronic psychosomatic illnes The Minuchin group implies that it is always indicated, but proponent of other forms of therapy (e.g., psychodynamic approaches to anorexi will not be convinced by the published studies, in which the way the te samples were selected is not discussed. It appears likely that the famil therapists had some excellent intuitions about which cases would r spond to their preferred mode of treatment. Before others can achiev the same success, we need some assessment tools to help predict whic children's illnesses have been caused, maintained, or exacerbated by th kinds of dysfunctional family interactions that respond to family therap

Although it is too sweeping a generalization, I believe that as crude our treatment methods are in clinical child psychology, they have bec developed to a technical level that far outstrips our methods of asses ment. Yet every treatment plan entails a diagnosis. It can ultimately b no more effective than the diagnosis is accurate. Most importantly, w cannot make much improvement in the efficacy of therapy without s multaneous advances in the classification of illnesses. Those advance in turn, cannot be on some arbitrary descriptive or even etiological bas alone: they have to be on the basis of what is treatable. (It is mo important to subdivide hyperactive children, for example, according what interventions they will respond to than according to the cons quences of their hyperactive behavior in the classroom, or according the prenatal traumas that may have induced it.) Etiology, in fact, is clinical importance only because it may provide clues as to some de cit—whether vitamin C or a nurturant parent—that can still be made u for. A theory about the cause can sometimes, but not necessarily, lead a treatment plan.

Furthermore, the research that we require, basing diagnostic categ ries upon treatment models, and evaluating treatment in terms of mo refined classifications of behavioral disorders, will have to involve all

the preceding four dimensions. (See Sroufe, 1975, for a similar critique of drug therapy with hyperactive children.)

It may occur to the reader that I have made no distinction between the diagnosis of a childhood behavior disorder and the concept of risk. Indeed, I can think of no difference. When we diagnose a child as requiring treatment, we are saying, "There is a high risk that these problems will not go away with the passage of time, that they will have a negative impact on the child's education and social development, that the parents' relationship with this child will suffer, or that the family as a whole will develop dysfunctional patterns of interaction." We are also saying that a treatment exists, with a reasonable likelihood of improving that prognosis. On the other hand, even if there is something wrong with a child, if it does not present a developmental risk for the child or family, either because it is a passing phase (like the "terrible twos") or because it is untreatable (like tone deafness), then it is unnecessary and unethical to pretend to intervene.

Discussion

Thinking about the problem of developmental risk four-dimensionally leads to some criticism of much work, both in the field of risk assessment and in the field of intervention. I have tried to argue that there is no justification for creating and routinely administering assessment scales, at any age, unless the scores on those tests are known to predict a "differential payoff" (Cronbach & Gleser, 1965) in different economically and politically feasible treatment conditions. On the other hand, no intervention can be defended without addressing precisely whom it can help, by how much, and what their fate would be without it. A treatment of a high-risk sample cannot be evaluated simply in terms of significant before–after differences, or even treatment–control group differences. Proper evaluation must be in terms of a cost–benefit payoff matrix. That, in turn, becomes more than just an evaluation of the specific program. Program evaluation is, in fact, a phase in the evolution of theories about the problem itself (Cronbach & associates, 1980).

A final point concerns the concept of assessment and intervention conceived of as society's tampering with its own internal variance. When a group is selected from the lower end of the distribution of predicted developmental outcomes and given special treatment, the goal is not so much to increase the population mean or median as it is to reduce the variance. On the other hand, there are circumstances in which we select individuals for special treatment from the upper end of

the distribution, as in college placement examinations and in elementary school programs for gifted children. Their aim is actually to increase the population variance. The concept of risk can be inverted somewhat to apply here: The untreated prognosis for the brightest students is better than for any other students, but still not as superior as it can be made to be with extra investment of society's resources. So the parents of those children regard them as at risk in regular school classes: at risk of not remaining superior.

Suppose, instead, that we were to consider the goal of public programs to be not necessarily decreasing or increasing the variance among children's developmental attainments, but raising the mean or median for the whole population. Alternatively, suppose we decide that it is more important to help certain demographic groups than it is to help others. Such considerations would add even more dimensions to the model. We would need to weigh the relative value of an intervention aimed at one group compared with the value of all other interventions that the same money could buy for some other groups. That may seem, and indeed is, far more complex than the way developmental psychologists have thought about these issues in the past. Yet it is inevitably the way politicians, bureaucrats, and even private philanthropists have to make decisions.

Does this mean that a scientific approach to the issues addressed in this book is impossible? No, but I think it means that a model strictly based in our own disciplines, those concerned with human development, is impossible. The concept of developmental risk, as something that can be identified and ameliorated, involves economics, history, sociology, and political science: It is about as interdisciplinary as a concept can be.

References

Aleksandrowicz, M., & Aleksandrowicz, D. (1974). Obstetrical pain-relieving drugs as predictors of infant behavior variability. *Child Development, 45,* 935–945.

Brazelton, T. B. (1973). *Neonatal behavior assessment scale.* Philadelphia: Lippincott.

Cronbach, L., & Associates. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Cronbach, L., & Gleser, G. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Epstein, N., & Vlok, L. (1981). Research on the results of psychotherapy: A summary of evidence. *American Journal of Psychiatry, 138,* 1027–1035.

Field, T., Hallock, N., Ting, G., Dempsey, J., Dabiri, C., & Shuman, H. (1978). A first-year follow-up of high-risk infants: Formulating a cumulative risk index. *Child Development 49,* 119–131.