Psychological Assessment
2000, Vol. 12, No. 1, 31-39

DOI: 10.1037//1040-3590.12.1.31

Computers Will Become Increasingly Important for Psychological Assessment: Not That There's Anything Wrong With That!

Howard N. Garb
Pittsburgh Veterans Affairs Medical Center and University of Pittsburgh

Though one can expect that computer programs will become increasingly important for psychological assessment, current automated assessment programs and statistical-prediction rules are of limited value. Validity has not been clearly established for many automated assessment programs. Statistical-prediction rules are of limited value because they have typically been based on limited information that has not been demonstrated to be optimal and they have almost never been shown to be powerful. Recommendations are made for building and evaluating new computer programs. Finally, comments are made about the ethics of using computers to make judgments.

Despite the allusion in the title of this article to the television show *Seinfeld*, one need not be concerned that this article is about nothing. In fact, this article is about something important: the anticipated growth of a new technology and its effect on psychological assessment.

Computers can be used for several purposes in psychological assessment. For example, they can be used to collect data, score protocols, and make judgments and decisions. I limit my discussion to the use of computers for making judgments and decisions. Comments are made about both computer-based test interpretation programs (sometimes called automated assessment programs) and statistical-prediction rules.

Computers are likely to become increasingly important for psychological assessment because mental health professionals are not good at some judgment tasks (Garb, 1998). For example, though interrater reliability has been good for at least some clinicians when the task has been to describe psychiatric symptoms and make diagnoses (e.g., Sartorius et al., 1993), it has frequently been poor when behavior therapists have set treatment goals (e.g., Felton & Nelson, 1984), clinicians have described clients using psychoanalytic theory (e.g., Collins & Messer, 1991), and psychiatrists have made decisions about whether a client should receive antidepressant medicine, electroconvulsive treatment, or psychotherapy (e.g., Keller et al., 1986). As another example, biases (e.g., race bias, social class bias, and gender bias) sometimes occur when clinicians make judgments for some tasks (e.g., Garb, 1997).

By using computers to make judgments, some of the problems associated with clinicians' judgments could be avoided. For example, in contrast with judgments made by clinicians, judgments made by computers have perfect test–retest reliability. Also, if clinicians could agree on what statistical-prediction rule should be used for a particular task, then interrater reliability would be perfect. Put another way, if all of the clinicians at a hospital could agree that judgments should be made by using a particular statistical-prediction rule, then interrater reliability would be perfect for the clinicians at that hospital. In addition, the use of computers can help to minimize the occurrence of errors and biases that sometimes occur when clinicians make judgments, though bias can still occur when computer-based test interpretation programs and statistical-prediction rules are used. With regard to computer-based test interpretation programs, judgments and decisions will be biased if the expert judge who had written a program is biased. With regard to statistical-prediction rules, judgments are unlikely to vary as a function of race, gender, or any other client characteristic unless the client characteristic has been related to the behavior or trait that is being predicted or described. On the other hand, one can infer that a statistical-prediction rule is biased if it predicts better for one group than another—even if the same rule is used for both groups (Cleary, Humphreys, Kendrick, & Wesman, 1975). Thus, to ensure that computer programs are unbiased, the effects of client characteristics (e.g., race, gender, etc.) need to be investigated.

Another reason computers are likely to become increasingly important for psychological assessment is due to how clinicians learn from experience. Mental health professionals frequently have difficulty learning from their clinical experiences (Garb, 1989; also see Meehl, 1997). This difficulty occurs for several reasons. For example, clinicians have difficulty learning from experience when they do not receive accurate feedback. Unfortunately, mental health professionals typically do not receive accurate feedback on whether their judgments are valid or invalid. Also, clinicians have trouble learning from experience when their cognitive processes are inadequate (i.e., when they remember information incorrectly). Because clinicians frequently have difficulty learning from experience, they may use information that has little or no validity (Garb, 1984; Garb, Florio, & Grove, 1998).

Compared with how clinicians learn from experience, statistical-prediction rules, but not computer-based test interpretation programs, have a decided advantage. Statistical-predictions rules are usually based on accurate feedback. That is, when deriving statistical-prediction rules, investigators frequently obtain criterion

Howard N. Garb, Department of Behavioral Health, Pittsburgh Veterans Affairs Medical Center, and Department of Psychiatry, University of Pittsburgh.

Correspondence concerning this article should be addressed to Howard N. Garb, Department of Behavioral Health (116A-H), Veterans Affairs Medical Center, 7180 Highland Drive, Pittsburgh, Pennsylvania 15206–1297. Electronic mail may be sent to garb.howard@pittsburgh.va.gov.

scores, at least for tasks like making behavioral predictions and diagnoses. Obtaining good criterion scores is normally too expensive for clinicians in the course of their clinical practice. Similarly, expert clinicians who write computer-based test interpretation programs do not usually collect criterion information. Instead, when writing computer-based test interpretation programs, expert clinicians draw on their clinical experiences and their knowledge of clinical lore and research results—they do not actually collect criterion scores to empirically derive rules. In general, statistical-prediction rules would do well because they make use of the inductive method: A statistical-prediction rule would do well to the extent that one can generalize from a derivation sample to a new sample.

I have two additional comments on the promise of using computers to make judgments. First, it is important to note that computers and computer programs are becoming increasingly powerful. The development of neural network models, described in this Special Section (Price et al., 2000), serves as one example of progress in computer science. Second, the results of the meta-analysis presented earlier in this Special Section (Grove, Zald, Lebow, Snitz, & Nelson, 2000) demonstrated that statistical prediction is typically more accurate than, or as accurate as, clinical judgment. In fact, on average, statistical-prediction rules were about 10% more accurate than clinical judgments.

The thesis of this article is that although clinicians sometimes make poor judgments and we can expect computers to become more powerful, other problems (methodological, reluctance of clinicians) need to be overcome if computer programs are to transform psychological assessment. Several topics are addressed, as follows. First, the quality of present-day computer programs are described. Next, methodological recommendations are made for building more useful computer programs. Finally, issues related to the acceptance of computer programs are discussed.

Critique of Present-Day Computer Programs

Computer-Based Test Reports

Many automated assessment programs for interpreting psychological test results are not validated (Adams & Heaton, 1985; Garb, 1998; Garb & Schramke, 1996; Matarazzo, 1986; Moreland, 1985; Snyder, Widiger, & Hoover, 1990; but also see the contribution by Butcher, Perry, & Atlis, 2000, to this Special Section). Automated assessment programs may also be biased. For example, interpretations may be correct more often for White clients than for African American clients. Finally, though test–retest reliability is perfect for computer-based test reports (given a particular test protocol, the same test report will always be printed), interrater reliability may be poor for judgments made by clinicians who use the computer-based test reports. Clinicians generally use computer-based test reports with other information (e.g., history information). One should not assume that the interrater reliability of their judgments is good.

Statistical-Prediction Rules

Clinical versus statistical prediction has been the subject of debate for decades (e.g., Dawes, Faust, & Meehl, 1989; Goldberg, 1968, 1974; Grove & Meehl, 1996; Holt, 1958, 1970; Kleinmuntz, 1990; Meehl, 1954, 1967, 1986; Sawyer, 1966; Sines, 1970; Wiggins, 1973, 1981). On the one hand, statistical predictions are usually as accurate as, or more accurate than, clinical judges. Statistical-prediction rules have even been more accurate than clinicians when mental health professionals have been given (a) assessment data that was used as input information for a statistical rule and (b) the predictions made by the statistical rule (Goldberg, 1968; Leli & Filskov, 1981, 1984; Moxley, 1973; Shagoury & Satz, 1969; Young, 1972; also see Stricker, 1967). In these studies, clinicians were not able to determine when a rule was likely to be wrong: Given the predictions of a statistical rule plus the assessment data used by the statistical rule, they could not make predictions that were more accurate than the predictions made by the statistical rule (also see Meehl, 1957). On the other hand, when clinicians have been given more information than was given to the statistical-prediction rule, they have sometimes been more accurate (see Grove et al., 2000, this Special Section).

Though statistical-prediction rules have the potential to transform psychological assessment, most current rules are of limited value. Although there have been many review articles on clinical versus statistical prediction (see references in previous paragraph), specific statistical-prediction rules have rarely been described and recommended for use in clinical practice. For example, in a recent review (Grove & Meehl, 1996), objections to statistical prediction were rebutted. Although their arguments in favor of statistical prediction are persuasive, they never described specific statistical-prediction rules and never said what statistical rules they believe should be used in clinical practice.

Present-day statistical-prediction rules are of limited value for two reasons (Garb, 1994). First, for some tasks (e.g., making diagnoses, describing symptoms or personality traits), investigators have not usually collected psychological test data and history, interview, and observation information and then tried to identify the best predictors. In fact, for these tasks, results from only a single psychological test have usually been used as input data for the statistical-prediction rules. Second, and even more to the point, present-day statistical-prediction rules have rarely been shown to be powerful. For example, statistical-prediction rules have not been powerful when the task has been to make diagnoses or describe personality and psychopathology. For these tasks, criterion scores have been made by clinicians using information that is usually obtained in clinical practice (e.g., history, observation, and interview data). Being able to predict ratings that clinicians can already make, using information that is readily available, is not very useful. Also, for these tasks, statistical-prediction rules were compared with clinicians who were given limited information (usually results from only a single psychological test). The finding that statistical rules have been able to outperform clinicians who are given results from only a single psychological test is not impressive.

To demonstrate that current statistical-prediction rules are of limited value, several of the best-known statistical-prediction rules are described and discussed. They are the Goldberg (1965, 1972) linear rules for diagnosing mental disorders and the Halstead Impairment Index (Halstead, 1947; modified by Reitan & Wolfson, 1985) for diagnosing neurological impairment.

Goldberg linear rules. The Goldberg (1965) linear rule makes use of Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1942) scale scores to discriminate be-

tween neurotic and psychotic clients. To compute a score using this rule, one adds and subtracts MMPI *T* scores using the following formula: Lie (*L*) + Paranoia (*Pa*) + Schizophrenia (*Sc*) – Hysteria (*Hy*) – Psychasthenia (*Pt*). Goldberg used a cutoff score of 45 to discriminate between neurotic and psychotic profiles, but he recommended that cutoff scores be derived in each setting that the index is used. Using data collected by Meehl (1959), the hit rate for the Goldberg index was 74% and the hit rate for the average clinician was 68% (results reported in Goldberg, 1965). However, the Goldberg index is famous not because it was able to do better than a group of clinicians, but because it was more accurate than more complex statistical rules, including regression equations, profile typologies, Bayesian techniques, density estimation procedures, the Perceptron algorithm, and sequential analyses (Dawes & Corrigan, 1974; Goldberg, 1969; Meehl & Dahlstrom, 1960). Goldberg's simple linear rule was more accurate than the above-mentioned complex statistical rules even though the relation between MMPI results and the differential diagnosis of neurosis versus psychosis is generally thought to be complex.

Although the Goldberg (1965) index is one of the best-known statistical-prediction rules, one can question whether it should be used by itself in clinical practice to make final diagnoses of neurosis versus psychosis. For example, according to Graham (1993),

> It is important to note that the index is useful only when the clinician is relatively sure that the person being considered is either psychotic or neurotic. When the index is applied to the scores of normal persons or those with personality disorder diagnoses, most of them are considered to be psychotic. (p. 225)

One can conclude that when making a differential diagnosis of neurosis versus psychosis, the only score or index from the MMPI that should be used is the Goldberg index. However, before using this index, one needs to rely on either clinical judgment or another statistical-prediction rule to rule out diagnoses of "normal" and personality disorder.

Goldberg (1972) later constructed a series of linear statistical rules that in addition to discriminating between neurosis and psychosis can also be used to classify individuals as "normal" or sociopathic. However, these rules should also not be used by themselves to make judgments. Goldberg had used these rules to classify mean MMPI profiles that were obtained for groups of participants (each mean profile was obtained for a group of participants who all shared the same criterion diagnosis). When the rules were used to classify MMPI profiles for individual participants, results were "mixed" (Zalewski & Gottesman, 1991, p. 566) with hit rates as low as 26% for psychotic, neurotic, character disorder, and indeterminate participants (Pancoast, Archer, & Gordon, 1988). Undoubtedly, results were disappointing because reliability is better for mean MMPI profiles that are obtained for a group of participants than for separate MMPI profiles that are obtained for individual participants.

Halstead Impairment Index. The best-known statistical-prediction rule in the area of clinical neuropsychology is the Halstead Impairment Index (Halstead, 1947). Using this index, one makes diagnoses of brain damage versus no brain damage. Originally, a client's score on the index was computed by counting the proportion of Halstead's 10 index tests that were in the brain-damaged range, but by general agreement 3 of the tests have been eliminated from the index because they were found to be less accurate (Reitan & Wolfson, 1985). No other statistical-prediction rule developed for the detection of brain damage is more accurate than the Halstead Impairment Index, even though the Halstead Impairment Index is over 50 years old (Russell, 1995).

Though the Halstead Impairment Index is the best-known statistical-prediction rule in the area of clinical neuropsychology, and though there is no other statistical rule that is more accurate, the Halstead Index should not be used by itself to make diagnoses of brain damage versus no brain damage. The reason it should not be used by itself to make final ratings is because ratings made by using the Halstead Index have not been as accurate as judgments made by neuropsychologists (Russell, 1995; also see Goldstein, Deysach, & Kleinknecht, 1973). In his review of the literature, Russell (1995) reported that the average hit rate for the Halstead Index has been 76%. In contrast, when neuropsychologists are given neuropsychological test protocols (but no other information), their average hit rate for this task is about 85% (Garb & Schramke, 1996; Russell, 1995).

There are several reasons that may explain why neuropsychologists outperform the Halstead Index. First, the neuropsychologists had more information available: The Trail Making Test and the Aphasia Screening Test do not contribute to the Halstead Index even though they make up part of the Halstead–Reitan battery. Second, neuropsychologists may be able to recognize patterns within tests and across tests that may relate to the presence of neurological impairment. Interactions are not recognized by the Halstead Index. Third, neuropsychologists may be able to make use of specific or pathognomonic signs. Tests are weighed equally by the Halstead Index.

Another reason the Halstead Impairment Index should not be used to make final diagnoses of brain damage is because one should not attend only to neuropsychological test data when making diagnoses. For example, when making a diagnosis of brain damage, a neuropsychologist needs to know if a client is anxious or depressed and if the client's performance is affected by a medical disorder, medication, or impaired energy level due to poor health or other causes. Similarly, if brain damage has already been documented (e.g., by a neuroimaging procedure), then one will need to make a diagnosis of brain damage even if the Halstead Impairment Index indicates that brain damage is not present (this could occur if a client had a high level of functioning before suffering brain damage). In conclusion, the Halstead Impairment Index should not be used to make final judgments, but instead it should be used as a source of input information for clinical judgments.

Recommendations for Building and Evaluating New Computer Programs

If computer programs are to transform psychological assessment, then advances need to be made in the way that programs are written and evaluated. Recommendations are limited to improving statistical prediction. Different recommendations are made for different judgment tasks.

Describing Traits and Symptoms

One recommendation for improving statistical prediction is that the information with the greatest validity should be used to make

judgments and decisions. As already noted, for the task of describing personality traits and psychiatric symptoms, the amount and type of input information given to statistical-prediction rules has almost always been limited. Typically, results from only a single psychological test have been used as input data for these statistical-prediction rules. Investigators should collect a large amount of input information, and optimal sets of predictors should be chosen by deriving statistical-prediction rules.

A reason why optimal information has not been used as input for statistical rules for the task of describing traits and symptoms can be described. Criterion ratings for these tasks have usually been made by clinicians who used information that is normally available in clinical practice. Criterion contamination can occur if information used by criterion judges is also used as input information for statistical-prediction rules.

To use optimal information and avoid criterion contamination, one needs to use new methods for obtaining criterion scores. Methods that can be used include behavior sampling, psychological testing, structured interviewing, and the act-frequency approach (see Garb, 1998, pp. 22–23). For example, behavior-sampling methods can be used to collect observations of a person's behaviors, psychometrically sound questionnaires can be used to assess a client's attitudes, structured interviews can be used to evaluate the completeness of descriptions of psychopathology, and the act-frequency approach can be used to evaluate descriptions of traits, motivations, and needs. Using the act-frequency approach (Buss & Craik, 1986), one follows clients over time to determine whether the clients exhibit behaviors that exemplify particular traits, motivations, and needs. By using these methods for obtaining criterion scores, psychological test protocols and interview, history, and observation information can all be used as input data for statistical rules without the risk of criterion contamination.

Psychodiagnosis

As with the description of traits and symptoms, innovative methods need to be used to obtain construct measurements of diagnoses (that can be used as criterion scores) so that psychological test protocols and interview, history, and observation information can be used as input data without the risk of criterion contamination. Several methods can be used including the longitudinal, expert, and all data (LEAD) procedure (Spitzer, 1983), an approach using the E. Robins and Guze (1970; also see L. N. Robins & Barrett, 1989) criteria, and a more general method based on construct validation (Cronbach & Meehl, 1955; Widiger, 1993a, 1993b). For an extended description of these methods, see Garb (1998, pp. 44–53).

With the LEAD procedure, to make criterion diagnoses (or more properly, imperfect construct measurements), clinicians collect longitudinal data (clients are followed for a period of time), expert clinicians make diagnoses, and all available data are given to the expert clinicians. One can then compare the diagnoses made by a statistical-prediction rule with the diagnoses made by using the LEAD procedure.

To validate diagnoses using the E. Robins and Guze (1970) criteria, one conducts follow-up studies, family studies, and laboratory studies. For example, a follow-up study can be conducted to evaluate how well treatment response and course of psychopathology are related to diagnoses. By conducting family studies, one can evaluate how well a set of diagnoses is related to results obtained from controls and family members of clients. Finally, a laboratory study can be conducted to determine if performance on laboratory tasks and results from biological measures are related to a set of diagnoses.

Using the construct validation approach, diagnoses are placed in a theoretical context. With this approach, one can still evaluate the validity of a set of diagnoses by conducting follow-up, family, and laboratory studies, but one could also conduct additional types of studies as long as diagnoses are placed in a theoretical context. For example, using the construct validation approach, one might investigate whether diagnoses made for White clients are more valid than diagnoses made for Black clients. Thus, one could try to determine if diagnoses made by a statistical-prediction rule are less likely to be biased (e.g., with regard to race bias) than diagnoses made by a group of mental health professionals. One should be aware that even though a statistical rule may make diagnoses for White clients and Black clients the same way (using the exact same rule for both groups), the rule could still be biased (Funtowicz & Widiger, 1995; Widiger & Spitzer, 1991). The rule would be biased if diagnoses for one group (e.g., Whites) were more valid than the diagnoses for another group (e.g., Blacks). Conversely, using different decision rules across race (or gender or another client characteristic) does not necessarily denote bias (Cleary et al., 1975). If a single statistical-prediction rule is the best rule for making diagnoses for one group of clients (e.g., Whites) but not another group of clients (e.g., Blacks), then more than one rule should be used.

Behavioral Prediction

Even when good criterion scores are obtained, it can be difficult to construct powerful statistical-prediction rules. For example, for the task of predicting violence, good criterion scores can be obtained. However, for this task, investigators have concluded that "we are not prepared to recommend these methods [statistical rules] for clinical use" (Gardner, Lidz, Mulvey, & Shaw, 1996, p. 609).

For the prediction of violence, the problem is not so much obtaining good criterion scores as finding valid predictors. For example, in an ongoing study at the Western Psychiatric Institute and Clinic at the University of Pittsburgh (Gardner et al., 1996; Lidz, Mulvey, & Gardner, 1993), mental health professionals at the psychiatric emergency department predicted whether patients would become violent in the next 6 months. Criterion scores were obtained by interviewing the patients and informants (people whom the patients identified as knowing the most about them) over the following 6 months. Individuals were most likely to become violent if they were young, were heavy drug abusers, reported recent urges to harm others, had extensive histories of violence, had a history of serious violent behavior, and did not have a thought disorder. However, the addition of other information to prior violence did not significantly improve the accuracy of predictions. Thus, instead of deriving a statistical-prediction rule, one could simply use past behavior to predict future behavior. A similar conclusion was reached by Mossman (1994) in a meta-analysis of results on the prediction of violence. To improve our predictions of violence, we need to identify more valid predictors.

To do this, we may need to develop a better understanding of the causes of violence.

Statistical rules for the prediction of suicide have not been clinically useful. Predicting suicide is even more difficult than predicting violence because of problems with low base rates (Elwood, 1993; Finn & Kamphuis, 1995; Meehl & Rosen, 1955; Wiggins, 1973, pp. 248–253). For example, in one study (Pokorny, 1983), 4800 Veterans Affairs psychiatric inpatients were followed for 5 years. During this 5-year period, 67 of the participants committed suicide. Using information that is strongly related to the occurrence of suicide (e.g., severity of depression, history of suicide attempts), a discriminant function analysis was able to make correct predictions for 74% of the participants. However, using this statistical rule, correct predictions were made for only 35 of the 67 participants who committed suicide, and false-positive predictions were made for 1,206 of the participants who did not commit suicide. When base-rate information was used to set the cutoff score for the discriminant analysis rule, only one prediction of suicide was made, and this prediction was incorrect.

It is difficult to say how we can make more accurate predictions of suicide. One can argue that short-term predictions of suicide are more likely to be accurate than long-term predictions of suicide. However, because the base rate for suicide is so low, one would probably need to collect data on tens of thousands of patients to look at the short-term prediction of suicide. One could also argue that more valid predictors need to be identified. However, given that prediction is difficult when base rates are low, it is not clear how we could identify predictors that would allow us to obtain a clinically meaningful level of accuracy.

Causal Judgments

The most difficult type of judgments to make are causal judgments. In my book on judgment research and psychological assessment (Garb, 1998), I concluded that "A review of the validity of causal judgments [made by mental health professionals] did not reveal any task for which validity was good or excellent" (p. 101). For example, reliability has frequently been poor when clinicians have described clients using psychoanalytic theory (e.g., Collins & Messer, 1991). Validity has also been disappointing for tasks that do not require clinicians to use psychoanalytic theory. For example, in a study on behavioral assessment (O'Brien, 1995), when psychology graduate students were given a client's self-monitoring data, they were able to identify controlling variables that were most strongly correlated with symptoms only 51% of the time.

I now recognize that validity has been good when statistical-prediction rules have been used to make causal judgments, but this result has been limited to behavioral assessment (Schlundt & Bell, 1987; also see Shiffman, 1993). As observed by Schlundt and Bell (1987),

> when clients keep a self-monitoring diary, a large amount of data is often generated. Typically, clinicians review the records and use clinical judgment to identify patterns and draw inferences about functional relationships among antecedents, behaviors, and consequences. Although the clinical review of self-monitoring records provides data that might not be otherwise obtained, clinical judgment is known to be subject to inaccuracies . . . and statistical prediction is typically more accurate and reliable than clinical judgment (p. 216)

Sequential and conditional probability analyses can be used to analyze self-monitoring data. This has been done to help in the treatment of a variety of problems including smoking addiction and the modification of eating behavior in bulimia, hypertension, and obesity.

For many causal judgment tasks, the use of statistical-prediction rules is not promising. This is due to the difficulty of obtaining criterion scores. For example, it would be especially difficult to obtain criterion scores to evaluate the validity of psychodynamic interpretations. However, though the use of statistical-prediction rules is not promising for many tasks, computer programs can be used to help clinicians formulate causal models (Haynes, Leisen, & Blaine, 1997). Thus, we may be able to use computers to improve clinical judgment, even when we cannot use them to make statistical predictions. Finally, though the use of *ABAB* interrupted time-series research designs in applied behavior analyses do not involve using computers to make judgments, it should be noted that their use can clarify the nature of causal relations (Kazdin, 1992).

Treatment Decisions

Statistical-prediction rules are rarely, if ever, used to make treatment decisions. Statistical predictions are not even used as input information for clinicians.

In contrast to statistical-prediction rules, treatment guidelines, established by committees of expert mental health professionals, have become increasingly important. For example, beginning in 1992, the National Institute of Mental Health and the Agency for Health Care Policy and Research funded the Schizophrenia Patient Outcomes Research Team to develop and disseminate guidelines for the treatment of schizophrenia based on existing scientific evidence (Lehman, Steinwachs, & the coinvestigators of the PORT Project, 1998). Similarly, the National Institute of Mental Health launched the Depression Awareness, Recognition, and Treatment Program to, in part, educate mental health professionals about the appropriate use of psychotropic and psychological interventions for the treatment of depression (Regier et al., 1988). As a third example of the development and dissemination of treatment guidelines, a task force organized by the Division of Clinical Psychology of the American Psychological Association has listed, beginning in the Winter 1995 issue of *The Clinical Psychologist* and continuing in subsequent issues, treatment interventions that have been empirically validated. Thus, efforts have been made to improve decisions, but they have involved setting up treatment guidelines, not deriving statistical-prediction rules.

Though statistical-prediction rules are rarely, if ever, used to make treatment decisions, their use may prove to be useful in the future. For example, to make treatment decisions, psychologists conduct functional analyses and interpret psychological tests (see the Special Section titled "Assessment in Psychological Treatment—A Necessary Step for Effective Intervention," which appeared in the December 1997 issue of *Psychological Assessment*). In the future, psychologists may use statistical-prediction rules to integrate information when conducting functional analyses. Statistical-prediction rules may also be used to make decisions on the basis of psychological test protocols and other types of data. To evaluate the validity of statistical-prediction rules, one will need to

evaluate whether the use of statistical-prediction rules is related to improved treatment outcome.

Neuropsychological Assessment

One of the most important things neuropsychologists do, and presumably what they do best, is describe clients' cognitive strengths and deficits. It seems unlikely that statistical rules will ever be used to combine test scores with other information for this task, just as it is unlikely that statistical-prediction rules would ever be used to estimate general intelligence by combining the results of an intelligence test with other information. The tests already yield scores that are estimates of cognitive strengths and deficits, and because the tests are psychometrically sound, there is no need to enter the test results and other information into a statistical-prediction rule to describe cognitive strengths and deficits. To improve the validity of descriptions of cognitive strengths and weaknesses, one would try to build better tests, not enter test results and other information into statistical-prediction rules.

Statistical-prediction rules are likely to be more valuable for making diagnoses and for describing the etiology and course of disorders (e.g., McKinzey, Podd, Krehbiel, Mensch, & Trombka, in press; McKinzey & Russell, 1997; Mittenberg, Rotholc, Russell, & Heilbronner, 1996; Tenhula & Sweet, 1996). History, interview, and observation information can be very important for these tasks (e.g., it is important to know a client's medical history), and it is unlikely that statistical-prediction rules using only psychological test results will be as accurate for these tasks as clinicians using both psychological test results and other types of data. To avoid criterion contamination, when deriving statistical-prediction rules that use psychological test information and history, interview, and observation data, one cannot base criterion scores on the same history, interview, and observation data that was given to the statistical-prediction rule. Thus, innovative methods need to be used to obtain construct measurements. For diagnosis, it is sometimes possible to evaluate the validity of clinical judgments by comparing the judgments to results obtained from autopsies (e.g., for the diagnosis of dementia) or to results obtained by following a patient over time (e.g., if cognitive functioning improves in response to being placed on an anti-depressant, then this implies that a patient was depressed and not demented). Prognostic judgments can, of course, be evaluated by following patients over time. Judgments about etiology can sometimes be evaluated by autopsy (e.g., dementia attributed to Alzheimer's disease can be differentiated from dementia attributed to vascular disease). When statistical-prediction rules are given psychological test data and history, interview, and observation information, it may be especially difficult to evaluate judgments made about malingering, but it may be possible to also do this by following patients over time.

Neuropsychologists also make behavioral predictions and treatment decisions. For example, they sometimes predict whether clients can manage their finances, and they sometimes make recommendations regarding placement. To obtain criterion scores for these tasks, one can collect longitudinal data to evaluate the validity of behavioral predictions and the utility of treatment decisions.

Acceptance of Computers by Clinicians

Clinicians may have ethical objections to using computers to make judgments. For example, if a clinician strongly believes that a patient should be placed on haldol but the computer recommends that the patient be placed on lithium, the clinician may believe that it would be unethical to not place the patient on haldol. However, computers should be used to make final judgments if they have been shown to be more valid than clinicians' judgments, and if clinicians are unable to say when computers are likely to be right or wrong. To do otherwise would be unethical.

Many mental health professionals may believe that they are more accurate than the average clinician. Though they may agree that statistical prediction rules are more accurate than the average clinician, they may decide to not use computers to make judgments. However, it is important to remember that Grove et al. (this Special Section) found that level of clinical experience was not related to how well clinicians do in comparison to prediction rules. There is no evidence that clinicians who think that they are more accurate than other clinicians really are more accurate, but there is evidence that presumed experts are frequently no more accurate than other clinicians (Garb, 1989).

Clinicians may also resist using computers for baser reasons (Grove & Meehl, 1996; Meehl, 1986). For example, they may be against the use of computers because they are afraid of losing their jobs or because the use of computers may hurt their self-esteem. Many of their fears may be unjustified. However, whether their fears are justified or not, it is important to remember that ultimately we must be concerned with relieving the pain and suffering of our clients. If we can make more accurate judgments and treatment decisions with statistical-prediction rules, then this is what we must do.

Summary and Discussion

Using computers to make judgments and decisions in personality assessment can lead to dramatically improved reliability, a decrease in the occurrence of biases, and an overall increase in validity and utility. Unfortunately, many automated assessment programs for interpreting psychological test results are not validated, and most current statistical-prediction rules are of limited value.

Strong advocates of statistical prediction argue that clinicians should already be using statistical-prediction rules to make judgments and decisions. For example, Gardner et al. (1996) were able to derive statistical-prediction rules for predicting violence that were more accurate than the predictions of a group of clinicians. Though they concluded that "we are not prepared to recommend these methods [statistical prediction rules] for clinical use" (p. 609), one could argue that their statistical-prediction rules should be used in clinical practice because they were more accurate than clinicians and because we know that statistical-prediction rules are, on average, 10% more accurate than clinicians (Grove et al., this Special Section). There is merit to this argument, especially because the statistical-prediction rule was compared with clinicians who were given all of the information that they usually use in clinical practice. However, the argument is less strong for other tasks. For example, for describing personality traits and psychiatric symptoms and for making diagnoses, clinicians were not given all

of the information that they usually have available in clinical practice. Also, statistical-prediction rules have almost never been used to make causal judgments and treatment decisions, therefore with few exceptions (Schlundt & Bell, 1987), one cannot argue that clinicians should be using present-day rules for these tasks.

Recommendations vary for the different judgment tasks. For describing traits and symptoms and making diagnoses, new methods can be used to collect criterion information. This would allow investigators to draw on a larger pool of assessment data without leading to criterion contamination. It would also enable them to better establish the validity of their rules. For behavioral prediction tasks, we can already collect good criterion information. Perhaps for this reason, some of the most powerful statistical-prediction rules available today involve the prediction of behavior. Rules for the prediction of violence are likely to be among the first statistical-prediction rules to enjoy widespread use in clinical practice. With regard to making causal judgments, statistical rules for understanding the relations among antecedents, behaviors, and consequences should also enjoy widespread use. However, for other casual judgment tasks (e.g., for evaluating the validity of a psychodynamic interpretation), obtaining criterion scores continues to be difficult. Finally, statistical-prediction rules have not been used to make treatment decisions, but this should eventually change because we can obtain criterion scores by studying treatment outcome.

In conclusion, the use of statistical-prediction rules will transform the practice of psychological assessment. As statistical-prediction rules become more powerful, their use will indeed become important.

References

Adams, K. M., & Heaton, R. K. (1985). Automated interpretation of neuropsychological test data. *Journal of Consulting and Clinical Psychology, 53*, 790–802.

Buss, D. M., & Craik, K. H. (1986). Acts, dispositions, and clinical assessment: The psychopathology of everyday conduct. *Clinical Psychology Review, 6*, 387–406.

Butcher, J. N., Perry, J. N., & Atlis, M. M. (2000). Validity and utility of computer-based test interpretation. *Psychological Assessment, 12*, 6–18.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30*, 15–41.

Collins, W. D., & Messer, S. B. (1991). Extending the plan formulation method to an object relations perspective: Reliability, stability, and adaptability. *Psychological Assessment, 3*, 75–81.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989, March 31). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review, 13*, 409–419.

Felton, J. L., & Nelson, R. O. (1984). Inter-assessor agreement on hypothesized controlling variables and treatment proposals. *Behavioral Assessment, 6*, 199–208.

Finn, S. E., & Kamphuis, J. H. (1995). What a clinician needs to know about base rates. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 224–235). New York: Oxford University Press.

Funtowicz, M. N., & Widiger, T. A. (1995). Sex bias in the diagnosis of personality disorders: A different approach. *Journal of Psychopathology and Behavioral Assessment, 17*, 145–165.

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review, 4*, 641–655.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*, 387–396.

Garb, H. N. (1994). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior, 10*, 377–394.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice, 4*, 99–120.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Garb, H. N., Florio, C. M., & Grove, W. M. (1998). The validity of the Rorschach and the MMPI: Results from meta-analyses. *Psychological Science, 9*, 402–404.

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin, 120*, 140–153.

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence by patients with mental illnesses. *Journal of Consulting and Clinical Psychology, 64*, 602–609.

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs, 79* (9, Whole No. 602).

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*, 483–496.

Goldberg, L. R. (1969). The search for configural relationships in personality assessment: The diagnosis of psychosis versus neurosis from the MMPI. *Multivariate Behavioral Research, 4*, 523–536.

Goldberg, L. R. (1972). Man versus mean: The exploitation of group profiles for the construction of diagnostic classification systems. *Journal of Abnormal Psychology, 79*, 121–131.

Goldberg, L. R. (1974). Objective diagnostic tests and measures. *Annual Review of Psychology, 25*, 343–366.

Goldstein, S. G., Deysach, R. E., & Kleinknecht, R. A. (1973). Effect of experience and amount of information on identification of cerebral impairment. *Journal of Consulting and Clinical Psychology, 41*, 30–34.

Graham, J. R. (1993). *MMPI-2: Assessing personality and psychopathology* (2nd ed.). New York: Oxford University Press.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Halstead, W. C. (1947). *Brain and intelligence*. Chicago: University of Chicago Press.

Hathaway, S. R., & McKinley, J. C. (1942). *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press.

Haynes, S. N., Leisen, M. B., & Blaine, D. D. (1997). Design of individualized behavioral treatment programs using functional analytic clinical case models. *Psychological Assessment, 9*, 334–348.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1–12.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25*, 337–349.

Kazdin, A. (1992). *Research design in clinical psychology*. New York: Macmillan.

Keller, M. B., Lavori, P. W., Klerman, G. L., Andreasen, N. C., Endicott,

J., Coryell, W., Fawcett, J., Rice, J. P., & Hirschfeld, R. M. A. (1986). Low levels and lack of predictors of somatotherapy and psychotherapy received by depressed patients. *Archives of General Psychiatry, 43*, 458–466.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*, 296–310.

Lehman, A. F., Steinwachs, D. M., & the coinvestigators of the PORT Project. (1998). Translating research into practice: The Schizophrenia Patient Outcomes Research Team (PORT) Treatment Recommendations. *Schizophrenia Bulletin, 24*, 1–10.

Leli, D. A., & Filskov, S. B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B Form I. *Journal of Clinical Psychology, 37*, 623–629.

Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology, 40*, 1435–1441.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of the American Medical Association, 269*, 1007–1011.

Matarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41*, 14–24.

McKinzey, R. K., Podd, M. H., Krehbiel, M. A., Mensch, A. J., & Trombka, C. C. (in press). Detection of malingering on the Luria–Nebraska Neuropsychological Battery: An initial and cross-validation. *Archives of Clinical Neuropsychology*.

McKinzey, R. K., & Russell, E. W. (1997). Detection of malingering on the Halstead–Reitan Battery: A cross-validation. *Archives of Clinical Neuropsychology, 12*, 585–589.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*, 268–273.

Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology, 6*, 102–109.

Meehl, P. E. (1967). What can the clinician do well? In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 594–599). New York: McGraw-Hill.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375.

Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4*, 91–98.

Meehl, P. E., & Dahlstrom, W. G. (1960). Objective configural rules for discriminating psychotic from neurotic MMPI profiles. *Journal of Consulting Psychology, 24*, 375–387.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

Mittenberg, W., Rotholc, A., Russell, E., & Heilbronner, R. (1996). Identification of malingered head injury on the Halstead–Reitan Battery. *Archives of Clinical Neuropsychology, 11*, 271–281.

Moreland, K. L. (1985). Validation of computer-based test interpretations: Problems and prospects. *Journal of Consulting and Clinical Psychology, 53*, 816–825.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*, 783–792.

Moxley, A. W. (1973). Clinical judgment: The effects of statistical information. *Journal of Personality Assessment, 37*, 86–91.

O'Brien, W. H. (1995). Inaccuracies in the estimation of functional relationships using self-monitoring data. *Journal of Behavior Therapy and Experimental Psychiatry, 26*, 351–357.

Pancoast, D. L., Archer, R. P., & Gordon, R. A. (1988). The MMPI and clinical diagnosis: A comparison of classification system outcomes with discharge diagnoses. *Journal of Personality Assessment, 52*, 81–90.

Pokorny, A. D. (1983). Prediction of suicide in psychiatric patients: Report of a prospective study. *Archives of General Psychiatry, 40*, 249–257.

Price, R. K., Spitznagel, E. L., Downey, T. J., Meyer, D. J., Risk, N. K., & El-Ghazzawy, O. G. (2000). Applying artificial neural network models to clinical decision making. *Psychological Assessment, 12*, 40–51.

Regier, D. A., Hirschfeld, R. M. A., Goodwin, F. K., Burke, J. D., Lazar, J. B., & Judd, L. L. (1988). The NIMH Depression Awareness, Recognition, and Treatment Program: Structure, aims, and scientific basis. *American Journal of Psychiatry, 145*, 1351–1357.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead–Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry, 126*, 107–111.

Robins, L. N., & Barrett, J. E. (Eds.). (1989). *The validity of psychiatric diagnosis*. New York: Raven Press.

Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization in neuropsychology. *Neuropsychology Review, 5*, 1–68.

Sartorius, N., Kaelber, C. T., Cooper, J. E., Roper, M. T., Rae, D. S., Gulbinat, W., Ustun, T. B., & Regier, D. A. (1993). Progress toward achieving a common language in psychiatry: Results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in *ICD-10*. *Archives of General Psychiatry, 50*, 115–124.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Schlundt, D. G., & Bell, C. (1987). Behavioral assessment of eating patterns and blood glucose in diabetes using the self-monitoring analysis system. *Behavior Research Methods, Instruments, & Computers, 19*, 215–223.

Shagoury, P., & Satz, P. (1969). The effect of statistical information on clinical prediction. *Proceedings of the 77th Annual Convention of the American Psychological Association, 4*, 517–518.

Shiffman, S. (1993). Assessing smoking patterns and motives. *Journal of Consulting and Clinical Psychology, 61*, 732–742.

Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry, 116*, 129–144.

Snyder, D. K., Widiger, T. A., & Hoover, D. W. (1990). Methodological considerations in validating computer-based test interpretations: Controlling for response bias. *Psychological Assessment, 2*, 470–477.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*, 399–411.

Stricker, G. (1967). Actuarial, naive clinical, and sophisticated clinical prediction of pathology from figure drawings. *Journal of Consulting Psychology, 31*, 492–494.

Tenhula, W. N., & Sweet, J. J. (1996). Double cross-validation of the Booklet Category Test in detecting malingered traumatic brain injury. *The Clinical Neuropsychologist, 10*, 104–116.

Widiger, T. A. (1993a). Issues in the validation of the personality disorders. In L. J. Chapman, J. P. Chapman, & D. C. Fowles (Eds.), *Progress in experimental and psychopathology research. Volume 16* (pp. 117–136). New York: Springer.

Widiger, T. A. (1993b). Validation strategies for the personality disorders. *Journal of Personality Disorders, 7*, 34–43.

Widiger, T. A., & Spitzer, R. L. (1991). Sex bias in the diagnosis of personality disorders: Conceptual and methodological issues. *Clinical Psychology Review, 11*, 1–22.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Wiggins, J. S. (1981). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior, 10*, 377–394.

Young, R. C. (1972). Clinical judgment as a means of improving actuarial prediction from the MMPI. *Journal of Consulting and Clinical Psychology, 38*, 457–459.

Zalewski, C. E., & Gottesman, I. I. (1991). (Hu)Man versus mean revisited: MMPI group data and psychiatric diagnosis. *Journal of Abnormal Psychology, 100*, 562–568.

Received July 16, 1998
Revision received December 2, 1998
Accepted January 28, 1999 ■