CONTENTS

1

Introduction to Prediction Statistics in Psychology

This introduction contains three main sections. The first part situates this work in the recent history of theory and practice of psychological assessment. It addresses why you may not already know this material and why it is important to learn it now. The second section describes the distinct features of statistical prediction tools and how they are different from other psychological measures. The final section of the Introduction provides of an overview of the types of information provided by statistical prediction tools, which are the core topics of this work.

A BRIEF HISTORY OF STATISTICAL PREDICTION IN PSYCHOLOGICAL ASSESSMENT

In 1954, Paul Meehl published his disturbing little book titled *Clinical Versus Statistical Prediction* (Meehl, 1954, 1986). His conclusion, that statistical prediction was as good as or better than the opinions of respected experts in psychology, was unsettling to those who relied on experts for guidance. It was even more disturbing to the experts. Can we all be replaced by a formula? Should we be?

Meehl's book sparked a generation of research and analysis on the conditions under which professional judgment adds value beyond the opinions of otherwise intelligent laypersons and when the experts can be outdone by a formula. This work showed that, in general, statistical risk tools outperform

https://doi.org/10.1037/0000275-001
Prediction Statistics for Psychological Assessment, by R. K. Hanson

professional judgment when (a) the number of relevant predictor variables is large, (b) the effect of any particular variable is small, and (c) we do not receive rapid feedback concerning the accuracy of our decisions (see Kahneman & Klein, 2009). Most often, when we make predictions about people, all three of these factors apply. Will an individual with a history of criminal behavior reoffend? Will my client relapse if we end treatment today? Will this marriage last? Does this athlete have what it takes to excel in the big leagues? Human existence is complex. Rarely is a single factor determinant of any course of action, and outcomes may not be known for years after the decisions have been made. Under such conditions, we cannot expect to develop intuitive expertise (so-called professional judgment) concerning the likelihood of future behavior. We can, however, increase our predictive accuracy by using statistics.

The use of statistics to predict outcomes for individuals is a relatively recent practice, having gained traction only in the latter part of the 20th century. Since the 1800s, actuaries have routinely used statistical models to estimate life expectances and insurance risks—for example, how your car insurance premium is calculated. The application of statistical models to human behavior and disease outcomes came much later. Even for ancient, basic problems related to games of chance, there was no mathematic analysis of probability before 1650, despite the obvious incentives. To quote the historian of ideas Ian Hacking (1975), "Someone with only a modest knowledge of probability mathematics could have won himself the whole of Gaul in a week" (p. 3). We have been adding, subtracting, and multiplying numbers for millennia, but we only began to think in terms of statistical probabilities within the past few centuries. At first the practice was restricted to a few members of the intellectual elites; now statistical probabilities are taught in elementary school.

Using statistical probabilities to make decisions does not come naturally; it requires commitment, training, and practice. In his later years, Meehl frequently bemoaned the lack of influence his book had on the routine activities of psychologists. When I was trained in clinical psychology in the 1980s, we read Meehl and his critics. We were not, however, trained on any actuarial risk tools. Few were available. Instead, we were taught that assessing the likelihood of future outcomes for individuals, such as the likelihood of violent recidivism, was beyond the scientific expertise of the profession.

Much has happened since then. I like to think that Meehl's ghost is resting more easily given developments since his passing in 2003. In contemporary correctional psychology and forensic mental health (my areas of practice), statistical prediction tools are ubiquitous. After sporadic use in the 1990s, empirically derived risk tools are now widely used and routinely inform decisions with significant consequences. Dozens of empirically derived risk assessment tools have been introduced during the past 20 years, including tools for general criminal recidivism, violent recidivism, sexual recidivism, intimate partner violence, and arson. Almost all U.S. correctional systems have risk tools for institutional placement, release decisions, and community supervision

FIGURE 1.1. Paul Meehl (1920–2003)

Note. A strong voice for statistical prediction in psychology. Used with permission from Leslie Yonce-Meehl.

(Desmarais & Singh, 2013). Recent trainees in forensic psychology are expected to know how to select, score, and interpret a number of crime and violence risk tools.

From Risk Prediction to Risk Management

The increased use of prediction tools for crime and violence was, in part, motivated by new and better risk tools. In particular, the emerging risk tools responded to evaluators' responsibility to facilitate risk management. Writing in the 1990s, Jim Bonta (1996) described three generations of risk assessment tools in corrections. The first generation was unstructured professional judgment. Before the widespread use of prediction tools, decisions concerning the likelihood of recidivism were based on case analysis and evaluators' implicit theories concerning the causes and correlates of antisocial behavior. For example, I remember case conferences where each of the team members were asked, in turn, "Do you think he is dangerous?" The ensuing, open discussion guided the ultimate decision by the team leader. The problem with this process was not that the decisions were unrelated to the outcome. They were, if weakly. Nor was there a problem with unstructured risk assessment not being accepted by decision makers. The courts in both the United States and Canada have had a wide tolerance for what qualifies as expert opinion concerning risk assessment. The major problem was that the opinions formed by designated experts were no different from those made by otherwise intelligent laypeople (Quinsey & Ambtman, 1979). Individuals who repeatedly reoffended were considered high risk; individuals who committed an isolated offense at an advanced age were considered low risk. A further difficulty with

unstructured professional judgment was its lack of transparency. If Johnny Knuckles was designated high risk, he did not know why he was considered dangerous. Disagreements between opposing experts were often decided based on which expert had the most impressive credentials.

Bonta's (1996) second generation of risk tools were actuarial measures based on static, historical risk markers, such as criminal history and age. The risk tools specified, in advance, the risk factors to consider, and how these factors were to be combined into an overall evaluation of risk (often by counting the number of risk factors present). Well-designed second-generation risk tools were more accurate than unstructured professional judgment. Furthermore, Mr. Knuckles could now see why he was deemed high risk. Unfortunately, he still had no idea what he could do about it. Neither did his therapists or case managers. The items on second-generation risk tools documented risk-relevant history without illuminating the psychological and social problems responsible for risk.

Bonta's third generation of risk tools were actuarial risk tools that included dynamic risk factors. Dynamic risk factors can be changed through deliberate intervention, and when the number and severity of such risk factors decrease, so does the likelihood of the outcome. In the context of corrections, dynamic risk factors are often called *criminogenic needs*. Examples of criminogenic needs are negative attitudes toward authority, association with criminal peers, aimless use of leisure time, substance abuse, and lifestyle impulsivity. The risk levels assigned by third-generation risk tools came with explanations. Evaluators using risk tools with dynamic risk factors can easily identify what needs to change to reduce risk and tailor their interventions accordingly.

Although second-generation risk tools are still widely used in corrections and forensic mental health, since the 1990s, there has been increased attention to risk tools that support case formulation and intervention. These tools include an expanding set of validated third-generation risk tools. These tools also include structured professional judgment (SPJ) measures, which are widely used. With SPJ measures, the risk and protective factors are specified in advance, but there is no explicit method of combining the factors into an overall evaluation of risk. Instead, the factors identified inform a case formulation, which is then used to infer overall risk levels. The case formulations and risk levels in the commonly used SPJ measures are primarily intended to inform risk management decisions and only secondarily speak to the likelihood of recidivism. The ideal risk tool, of course, would do both well (see Chapter 17).

The Steady Rise of Prediction Tools in Psychology

The rising influence of statistical prediction tools has been striking in the assessment of risk for crime and violence, but it is far from absent in other areas of applied psychology. Educational psychologists have long been concerned with predicting school success using IQ tests. The University of Waterloo screens potential students using an algorithm that includes empirically derived weights

for discounting student grades given by specific high schools (Cain, 2018). Given the widespread availability of high-quality data and cheap data processing, we should expect a steady increase in statistical prediction tools relevant to the concerns of psychologists.

COMMON CRITICISMS OF PREDICTION TOOLS

Even though actuarial risk tools are widely used and many have gained general acceptance in the professional, scientific, and legal communities, they are not without their critics. The following are some common concerns expressed about the use of prediction tools in psychology.

Prediction Tools Are Unnecessary

When we can trust self-report, prediction tools add little and are rarely used. In many contexts, conscious intentions are a good indicator of future behavior (Fishbein & Ajzen, 2010, Chapter 2). Prediction tools in psychology have value only when there are strong situational demands to say the right things (job interviews, parole board hearings) or when individuals lack the insight, knowledge, and judgment to make reliable self-assessments. In such contexts, prediction tools can effectively summarize a wide range of small contributing factors, the influence of which would be otherwise difficult to weight appropriately.

Individuals Are Not Statistics

Some critics challenge the fundamental assumption that statistics can and should be used to predict the behavior of individuals. There are several forms of this argument, some of which are addressed in more detail in subsequent chapters. A common form was expressed by Sherlock Holmes:

> While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but the percentages remain constant. So says the statistician. (Doyle, 1890/2000, p. 37)

And to this day, it is easy to find statisticians and other contemporary experts who concur. I don't. Although there are academic traditions that support Holmes's view, I doubt that many adherents arrived at this conclusion after a study of the philosophy of probability, such as I review in Chapter 2. Instead, many people are attracted to such views because of deeply held beliefs in the independence, autonomy, and dignity of the individual. We sense that statistics have deep roots in governments' attempts at regulation and social control of the untrustworthy (the colonies) and the undesirable (*Les misérables*; see Hacking, 1990).

Statistical prediction has inherent limitations. Human behavior is not fully determined by any set of conditions. One of my teachers said that a prediction scheme that accounted for more than 50% of the variance was probably misspecified (e.g., the outcome was somehow included in the predictors). Most risk and protective factors in psychology are small things that nudge people in certain directions. Some of these influences are external, such as peers, disease, job requirements, the laws of the land, and social and cultural identities. Some of these influences are the habits, memories, and expectations that we all carry with us. Nobody, however, seriously questions whether there are genuine differences in how individuals behave. And no psychologist would question that these differences can be quantified.

Prediction Tools Can Be Biased

This criticism accepts the premise of statistical prediction for individuals but finds fault with its application. In particular, assessment tools in education and criminal justice have been criticized as biased toward already marginalized ethnic and racialized groups. One form of this argument is that the same indicators mean different things based on different sociocultural histories. Having unstable employment, transient housing, and negative attitudes toward authority may be valid indicators of psychopathy for White males in Canada. For the Indigenous peoples of Canada, however, these are predictable consequences of colonization, systemic discrimination, and intergenerational trauma. Similarly, risk indicators can mean different things for members of the dominant socioeconomic classes compared with individuals who have been marginalized based on race, religion, or gender identification. Assessment tools that aim to promote transparent and equitable decision making can be unintentionally reinforcing the prejudices that these measures were intended to protect against. Such concerns were sufficient for the Supreme Court of Canada to stop the Correctional Service of Canada from using any actuarial risk tools for Indigenous inmates until the tools have been specifically validated for this racial group (*Ewert v. Canada*, 2015, 2018).

NUMERACY IS NEEDED

Regardless of one's position on the preceding issues, serious participation in these debates requires understanding the statistics used to create prediction schemes. Psychologists are expected to understand the measures they use. This requires numeracy in prediction statistics. If a credible research study finds that the area under the curve (discussed in Chapter 8) for Indigenous males was .68 for Risk Tool X, should we be reassured or alarmed? Most actuarial risk tools can be scored without specialized training in prediction statistics; however, scoring an actuarial risk tool is not a complete risk assessment. Scores need to be interpreted. This requires understanding the quantitative information

implied by risk scores, such as percentile ranks, recidivism rates, and risk ratios. It also requires understanding what research says about the strengths and weaknesses of any specific prediction tool. Expertise in prediction statistics will never be evenly distributed in the population. It is my conviction, however, that raising the level of numeracy of the users and consumers of statistical prediction tools is both possible and desirable.

Stanislav Andreski (1972) told of a debate about the existence of God between the great Swiss mathematician Euler (on the side of God) and members of the court of Catherine II of Russia (on the side of Voltaire, the skeptic). After going on for some time, Euler got a blackboard and wrote the following: $(x + y)^2 = x^2 + 2xy + y^2$, therefore God exists. As Euler expected, the members of the court did not understand the formula, and because they did not want to look stupid, they conceded the argument to Euler. This is a bad way of making decisions.

The importance of statistical numeracy is not limited, of course, to psychology. In his commentary on Gigerenzer et al.'s (2007) review of the importance of statistical literacy for evidence-based medicine, John Monahan (2007) summarized the evidence as follows:

> marshalling one study after another, [Gigerenzer and colleagues (2007)] demonstrate that, across widely varying samples of health professionals, patients, and policymakers, in all countries studied, statistical illiteracy reigns supreme—often with catastrophic consequences for individual and public health. (p. i)

It is not uncommon for individuals to consider a drug with serious side effects in 15 out of 100 cases to be less risky than another drug that had serious side effects in 128 out of 1,000 cases (this is called *denominator neglect*; Yamagishi, 1997). Harsher criminal justice decisions are made when outcomes are presented as two out of 10 individuals like Jack will reoffend compared with presenting the same information as a 20% recidivism rate (Slovic et al., 2000). It doesn't have to be this way. Although there is considerable variation across fields, there is increasing recognition of the value of using statistics to make predictions about individuals' outcomes in medicine, business, and sports. Thanks to Michael Lewis (2003), *moneyballing* is now a verb.

THE DISTINCTIVE FEATURES OF PREDICTION TOOLS

There is no shortage of books and articles concerned with developing prediction equations in psychology. A very common approach is to designate one variable as the variable of interest (i.e., the dependent variable) and then use the remaining variables (the independent variables) to predict the dependent variable using multiple regression. The primary intent of such studies, however, is either theory development or the description of groups (e.g., are males better at mental rotations than females, after controlling for education?). Some of these studies include longitudinal data, but, again, the primary interest is rarely on predicting the outcome for individuals. The literature on

how to use a preexisting prediction scheme for applied decision making is even smaller.

The Conceptual Foundations of Prediction Tools

Prediction tools are created to address questions concerning the likelihood of a future event, behavior, or outcome. The basic assumptions supporting prediction tools are relatively straightforward:

1. Prediction tasks are ubiquitous in psychology. The constructs we use to describe people frequently carry implications for future actions. People who score high on conscientiousness, for example, would be expected to keep their promises.
2. We routinely make judgments about the expected conduct of specific individuals. Not only do we assert that conscientious people are apt to keep their promises, we also assert that *Mary* is likely to keep *her* promise because she is conscientious. This is a prediction for a specific case. In certain contexts, the outcome is of sufficient consequence that there is value in quantifying its likelihood as a probability or rate. There are no conceptual barriers to assigning numeric values to likelihoods for individual cases, any more than there are barriers to using numbers to describe any other feature or characteristic of anybody (see Chapter 2).
3. Group data are extremely useful when forming professional judgments concerning likelihoods for individual cases. Statements concerning likelihoods for individual cases have increased accuracy when informed by the base rate of the outcome in the reference group that the case most closely resembles.
4. Along with the base rate, risk assessments need to consider case-specific features that increase or decrease the likelihood of the outcome. Strong evidence that a characteristic is a risk or protective factor is provided by empirical associations observed in follow-up studies.
5. Given that they are a large number of potentially risk-relevant factors for most outcomes of interest, comprehensive evaluations should consider a range of risk factors, hence the need for risk prediction tools. Risk assessment tools identify, in advance, the risk factors to consider and how they should be organized into an overall evaluation of risk.

Creating useful prediction schemes is both science and art. It requires evidence, but it also requires matching the prediction scheme to the applied decisions and to the decision makers. Research is required to establish a tool's scientific credentials. Statistical prediction tools will not be used, however, unless the decision makers perceive them as improvements over the alternatives. Prediction schemes increase in value when they are easy to learn, easy to use, and provide valuable information that is more difficult or costly to obtain in other ways.

Identifying Risk and Protective Factors

Credible prediction schemes are rarely developed by a single study. All samples, no matter how large and representative, have unique features that do not generalize to other samples and settings. Consequently, prediction equations developed on any specific sample will generally work best on that sample. Single-sample equations are vulnerable to overfitting (i.e., mistaking random features of a data set for basic reality). In other words, don't put too much trust in the results of any single study. Effects that are significant at $p < .05$ or even $p < .001$ may not be found in any other samples. Conversely, the lack of a significant effect in one data set does not mean that it will not be found in others. The reason for the between-study variability is often difficult to determine. Some of the variation will be related to random sampling; however, arbitrary features of the study design can have hidden consequences, which may never be fully understood.

The most generalizable risk tools are based on programs of research that identify factors that are related to the outcome from multiple independent samples and cull factors that seem like they should be related to the outcome but are not. A strong conceptual model that explains why these factors are related to the outcome is desirable but not always necessary (see Chapter 17). A thorough and informed search for valid risk factors minimizes the likelihood of missing major factors.

Combining Predictive Factors

Once a sufficient pool of predictive factors has been identified, the next step is combining them into an overall score. The best way of combining risk factors remains a topic of debate in scientific and professional communities. There are currently three broad approaches guiding research and practice: (a) actuarial risk tools based on relatively simple additive models (e.g., Violence Risk Appraisal Guide—Revised; Harris, Rice, et al., 2015); (b) SPJ tools in which the items are specified in advance but the overall evaluation is based on a unique, case-specific formulation (e.g., HCR-20^{v3}; Douglas et al., 2013); and (c) complex statistical prediction tools (e.g., Classification of Violence Risk; Monahan, 2021). Although certain measures do better in certain studies, none of these three broad approaches have demonstrated overall superior predictive accuracy, at least not in the area of crime and violence where risk tools are routinely used (Brennan, 2017; Campbell et al., 2009; Tully et al., 2013; Yang et al., 2010). Consequently, evaluators' decisions concerning the type of prediction tool to use must be based on other considerations.

My own research and professional practice have privileged measures based on relatively simple additive predictive models. My work has focussed on sexual recidivism risk assessment; however, simple additive models are likely to work well in many areas of psychological assessment because the causal constructs of interest are correlated with each other and are measured with error. Under such conditions, statistical interactions, which add incremental

predictive power to complex statistical models, are rarely observed (Busemeyer & Jones, 1983; McClelland & Judd, 1993). I have also not been convinced that integrating risk factors using professional judgment is more accurate than simpler, mechanical methods (Hanson & Morton-Bourgon, 2009).

Overall, the way of combining predictive factors appears less important than ensuring that relevant factors are considered. In his 1979 article "The Robust Beauty of Improperly Specified Regression Equations," Robin Dawes reviewed research demonstrating that prediction schemes with simple weights do as well as schemes with complex weights. For example, a simple count of the risk factors present (yes = 1, no = 0) usually works as well as fractional weights developed through regression (e.g., score = –1.04 [age in years] + 2.34 [prior failure] + 1.54 [lack of social support] + 1.15 [negative peer influences] + 1.87 [negative attitudes]). Canadian researchers Daryl Kroner, Jeremy Mills, and John Reddon (2005) even found that randomly selecting items (with their original weights) from a diverse set of validated criminal recidivism risk tools did as well as any of the original prediction tools. Although it is almost always possible to improve prediction by refining the item weights, the incremental gains are small. In contrast, meaningful gains can be achieved by considering new, risk-relevant information.

DIAGNOSIS AND PROGNOSIS

Another concern with much of the discourse on prediction in psychology is the blurring of the tasks of diagnosis and prognosis. In *diagnosis*, the task is estimating the likelihood that the individual currently has a specific condition. In *prognosis*, or true prediction, the evaluator wants to know whether an individual who does not currently have the condition will develop it at some future date. For example, a simple test can determine whether a woman is currently pregnant (diagnosis), whereas a completely different set of factors need to be considered to assess the likelihood that she will become pregnant in the next 5 years. Although certain statistics are useful for both, they are interpreted differently. For example, logistic regression can be used to estimate the likelihood of brain cancer given neuroimaging results. Decision thresholds can be set and tested by comparing the decisions implied by the diagnostic measure to a highly credible measure of the true state of affairs for this patient (e.g., biopsy). The diagnostic decisions implied by regression equations can either be correct or incorrect for a specific individual. Using aggregated cases, a diagnostic procedure based on logistic regression can be evaluated in terms such as hits, false positives, positive predictive value, and negative predictive value.[1]

Logistic regression can also be used to create a prediction scheme; however, there is no gold standard against which to evaluate a prognosis for a specific individual. The outcome of interest is not present and may never occur. At the time of assessment, it is impossible to determine whether the assessment is correct or incorrect. Nobody has the outcome of interest at that time. More

[1]For a definition of these terms, see Chapter 7 (Dichotomous Predictors) or the Glossary.

than that, the actual outcome for any individual will never indicate that the prediction was wrong, except in the very unusual case where the prediction was zero or 100%. The validity of a prediction scheme can only be evaluated in aggregate data.

Prediction Schemes Assess the Likelihood of the Outcome

Instead of saying whether the outcome will be present or not, prediction schemes assess its likelihood. These likelihoods can range from very small to very large and are often expressed as probabilities between zero and one (various ways of expressing likelihoods are covered in Chapter 4). When weather forecasters announce that the likelihood of rain today is 50%, they are not expressing ignorance. Instead, they are communicating that they expect it to rain on 5 out of 10 days like today. Similarly, criminal recidivism risk tools distinguish between groups of individuals whose expected recidivism rates can range from less than 5% to well over 80%, with many gradations in between.

The language of hits and false alarms is well suited to diagnostic tests but poorly aligned with the information provided by prediction tools. At the time of assessment, all statements that the individual will have the outcome are necessarily false (the outcome is not present in anybody). Using diagnostic language, everybody in the high-risk groups are "false alarms" at the start of follow-up. Over time, some proportion of the sample will turn into hits, that is, individuals with the outcome who are correctly identified as having the outcome by some decision rule. However, the proportion of hits depends more on the length of follow-up and the base rate of the outcome than on the accuracy of the prediction scheme.

Consider a prediction tool that perfectly ranks employees on the likelihood of advancing within the organization, and assume that the top 200 (20%) of 1,000 employees are designated as high-flyers, that is, those likely to advance. If 150 (15%) of employees advance in any 1 year, then 150 of the 200 high-flyers would advance in the first year of follow-up, which corresponds to a hit rate of 75%. There would be, however, 50 high-flyers who did not advance, which corresponds to a false alarm rate of 25%. When follow-up is extended another year, an additional 150 individuals advance, bringing the total to 300 out of 1,000 (30% base rate). Now the hit rate would be 100% (all the high-flyers have advanced), and there are no false alarms. However, there is another problem: The decision rule failed to identify 100 individuals who did advance (see *specificity* in the Glossary). The point is that diagnostic statistics say little about the accuracy of prediction tools because they are heavily influenced by the decision threshold and the absolute frequency of the outcome.

The Causal Order Is Different for Diagnostic Versus Prognostic Indicators

Diagnostic indicators are typically the results of the disorder (e.g., the fever is caused by the infection). When the disease abates, so do the symptoms. The relationship between the symptoms and the disorder is usually probabilistic

(not every symptom is present in every positive case, and sometimes the symptom is present without the disorder), so there may be uncertainty in diagnoses. The strength of the relationship between the disorder and the indicators (or symptom) is often expressed as a statistic called a likelihood ratio (see the Glossary; Akobeng, 2007b). In the context of diagnosis, the probabilistic relationship between the disorder and the indicator is expected to remain constant across samples and settings (e.g., the chances of observing fever when a patient has malaria is roughly the same for patients in Toronto and Mozambique, even though the rates of malaria would be drastically different).

In the context of prediction, the causal order is reversed: the indicators are intended to influence the likelihood of the outcome, not the other way around. The outcome does not influence the indicators because it does not exist at the time of assessment. For example, poor instruction precedes and increases the likelihood of school failure. Humans are goal-oriented creatures, however, so the distinction between predictor and outcome is not 100% pure (e.g., teachers can teach poorly because they expect certain students to fail). Nevertheless, the logic of prediction schemes is that they are using information at Time 1 to predict outcomes at Time 2. What happens at Time 2 cannot change what has already happened at Time 1. Whether a student succeeds cannot change the fact that the student previously received a good or a poor education. Unlike the context of diagnosis, individuals will acquire the outcome without any change in their initial status on the indicator and predictor variables. The consequence is that evaluating prediction tools requires a different point of view and different statistics than those used to evaluate diagnostic schemes. Even when the same statistics are used appropriately in the context of diagnosis and of prognosis, they mean different things and need to be interpreted accordingly.

For applied prediction, however, the distinction between diagnosis and prediction is rarely absolute. In practice, evaluators use historical or current conditions to predict future events. Identifying individuals' past and current functioning is a diagnostic task (i.e., what risk-relevant characteristics apply in this case?). For example, a prediction scheme that uses intelligence to predict school success (a prediction task) must first assess intelligence (a diagnostic task). Notice that a variable such as prior grades is both a diagnostic indicator of intelligence and a predictor of school success. Even when risk factors are identified purely on the basis of their empirical relationship with the outcome, they must also be indicators of risk-relevant latent constructs, even if the nature of these latent construct is unknown. Consequently, in psychological assessment, risk factors typically function as both diagnostic and prognostic indicators. What is being assessed in risk assessment is discussed further in Chapter 2 and Chapter 17.

THE NUMERIC INFORMATION PROVIDED BY PREDICTION TOOLS

Just as different statistics are appropriate for evaluating an assessment measure as a prediction tool rather than as a norm-referenced measure, so too are the inferences that can be made from scores. The quantitative inferences from

norm-referenced measures typically concern relative placement in some normative groups (e.g., top 5% of graduate students). In contrast, the quantitative inferences from prediction tools concern the likelihood of the outcome of interest (e.g., 85% chance of successfully completing a doctoral degree). Notice that certain measures can function both as norm-referenced measures and as prediction tools (e.g., academic aptitude tests).

Discrimination and Calibration

Prediction schemes provide two main types of quantitative information: whether certain classes of individuals are more likely to have the outcome than others (relative risk) and the expected rate of the outcome (absolute risk) for individuals "like this." Consequently, statistical prediction schemes are properly judged against these two general criteria. The relative risk information is also called *discrimination*, or the extent to which the individuals who eventually have the outcome look different from individuals who do not have the outcome. The second criterion is often referred to as *calibration*, or the extent to which the observed rates in replication studies match the rates expected by the prediction tool. In other words, if 15 out of 100 individuals like Fred are expected to fail graduate school within 2 years, how stable are these rates against samples and settings?

The most accurate prediction tools are high on both discrimination and calibration. It is possible, however, for measures to still be useful if their calibration is poor. For example, imagine a measure that accurately ranks students in their potential to graduate on time (good discrimination); however, the actual graduation rates vary considerably across universities and programs (poor calibration). In this case, administrators may still want to use the tool to select those most likely to succeed in a timely manner (keeping in mind the local graduation rates). The measure would not provide accurate information about the proportion of students who will graduate (that is determined by the programs); it would, nevertheless, correctly identify who is more likely to graduate.

Quantifying Likelihoods

As previously stated, numeracy is needed to evaluate research findings and integrate new knowledge into practice. For users of statistical prediction tools, numeracy is essential because much of the information provided by these tools is numeric. Although it is possible to talk of likelihoods in words (e.g., low, moderate, high), actuarial prediction tools have the benefit of *quantifying* likelihoods, for example, expressing a low likelihood of rain as a 10% chance. The following section reviews the central numeric information provided by statistical prediction tools.

Dichotomous Outcomes

This book focuses on outcomes that are either present or absent, that is, dichotomous outcomes. Examples of dichotomous outcomes include passing

a test, quitting a job, being arrested for a crime, or recovering from depression. Much of the material can be generalized to other types of outcomes, such as annual income, degree of happiness, and the number of criminal convictions. The focus, however, is on socially valued, clearly defined, dichotomous outcomes. Quantifying dichotomous outcomes may seem straightforward but several options are available and widely used (see Chapter 4). It is best not to confuse them. As well, because nobody has the outcome at the time of assessment, all estimates of likelihood must consider time: likelihoods only make sense when paired with a follow-up period (graduate within 5 years, recover from depression within 3 months, quit before the end of basic training).

Relative and Absolute Risk

In addition to quantitative estimates of the likelihood of the outcome, prediction tools also provide information about how risky individuals are compared with others (an aspect of discrimination). It is quite possible to quantify relative risk while having no firm opinion about the absolute likelihoods (see Part III of this volume). For example, the Centers for Disease Control and Prevention (2020) stated that the risk of lung cancer is 25 times greater for men who smoke compared with nonsmokers. This tells you that smoking is bad for your health, but other information is also relevant to judging the health risks.

I have a personal example of the importance of considering both relative risk and absolute risk. When my eldest son was 2, he had a febrile seizure. For those of you who have never seen one in a loved one, this is terrifying. We took him to a neurologist, who, after some basic developmental tests (all normal), asked us whether we wanted to send him for further diagnostic testing. The further tests involved electrodes and sleep deprivation and were not considered particularly informative at this age (poor discrimination). I then asked if febrile seizures increase the likelihood of epilepsy in adulthood. The doctor replied that they double the chances. Notice that he replied in terms of relative risk (a risk ratio, an indicator of discrimination). I then asked about the base rate of epilepsy in adults. He replied that it was about 1% to 2%. In our situation, the base rate information (calibration) was much more important than the discrimination accuracy of the diagnostic indicator. We decided on no further tests.

Percentile Ranks

Another quantitative indicator provided by statistical predictions tools is the percentile rank: compared with some normative group, how unusual is this score? For example, the results of a university entrance exam may be in the 97th percentile, indicating that it is unusually high. This indicates that the candidate is among the top 3% most likely to succeed in university (if the exam is valid), and, for universities using this exam, it is also a good indicator that the candidate could be admitted. In contrast, a candidate whose scores correspond to a percentile rank of 48 (near the middle of the pack) would be less likely to be admitted to the university and less likely to succeed if admitted.

Percentile ranks and related scores (e.g., *Z*-scores, *T*-scores) are probably the most common method of reporting the results of psychological tests. Even though percentile ranks are not unique to prediction tools, they are included in this book because they are a quantitative indicator that is easily calculated, easily understood, and has the potential of enhancing risk communication (see Chapter 16).

PREDICTION TOOLS MAY OR MAY NOT MEASURE LATENT CONSTRUCTS

In most psychological testing, percentile ranks are used to indicate where an individual should be placed, compared with other individuals, on the psychological feature of interest (e.g., impulsivity, neuroticism). These features are considered latent constructs because they are not directly observed. Instead, they are inferred from indicators. For example, the latent construct of neuroticism may be inferred from self-reports of worry, distressed facial expressions, and high autonomic reactivity. The items in most psychological measures are all intended to be indicators of the same construct. Individuals for whom more of the items apply are considered to rank higher on the underlying, latent dimension.

In contrast, the most accurate risk tools consider multiple risk-relevant constructs measured by diverse sources of information. Consequently, percentile ranks on prediction tools rarely lend themselves to easy psychological interpretation, nor do prediction tools aspire to high internal consistency (e.g., omega > .80). Instead, the most accurate prediction tools aim to comprehensively sample all the risk-relevant domains, with no major contributors neglected.

It is sometimes possible, however, to use prediction tools to infer psychological constructs: Specifically, when the construct is dimensional and includes the propensity for the outcome as a core feature, there will be strong overlap between the items in a prediction tool and the items in a measure of the corresponding psychological construct. Chronic rule violation and callous disregard for others are both indicators of antisocial personality disorder as well as efficient indicators of future criminal behavior. Measures designed to predict success in school (past grades) will also assess intelligence (a latent construct). Consequently, relative placement on the risk tool can inform relative placement on some latent psychological constructs.

Even when a risk tool is used purely for prediction purposes, evaluators need some understanding of what is being assessed (Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, 2014, p. 13). Credible causal theories minimize the likelihood that the observed associations between the indicators and the outcome are confounded by variables that are not appropriate for the assessment task. For example, few evaluators would consider it appropriate to ask

about parents' income on a college admissions assessment, even if it is correlated with educational achievement. Further discussion of what should be assessed by prediction tools can be found in Chapter 17.

META-ANALYTIC PERSPECTIVE

This book approaches statistics from the perspective of effect sizes and meta-analysis (Cumming, 2013). Null hypothesis significance tests are occasionally presented but deemphasized. The assumption is that evaluators want to know more than whether a prediction tool works better than chance (a very low threshold). Instead, they want to know the strength of the relationship between the predictor and the outcome, and the plausible range of the size of that relationship (i.e., confidence intervals for effect sizes). The meta-analytic framework is privileged because it is rare for any single prediction study to have sufficient credibility to justify changes in applied practice. Confidence in findings increases when substantively similar effects are found in diverse samples and setting. Meta-analysis provides the framework for estimating average effects across studies and for examining the extent to which the variation across studies is more than would be expected by chance alone.

OVERVIEW OF THE CHAPTERS

This book provides a broad conceptual model supporting the use of prediction tools in psychological assessment. At its heart, however, it is a statistics textbook, accompanied by the requisite explanations, equations, and exercises. The core chapters cover the essential statistics used for evaluating and interpreting prediction tools: proportions and likelihoods (Part II), discrimination (Part III), and calibration (Part IV). Part V (Chapter 16) addresses percentile ranks, which are a useful quantitative indicator for diverse assessment measures. The statistics chapters are followed by a discussion of other, largely nonnumeric, indicators of the quality of prediction tools (Chapter 17) and some suggestions on advancing risk communication (Chapter 18). Because many readers will be unfamiliar with the technical terms used in the book, I have included a Glossary and an Appendix with a brief refresher on formulas and graphs.

As a teaching text, the statistics chapters start easy and get harder. The beginning of the chapters should be accessible to almost everybody interested in the topic; the latter sections assume increased numeracy. Do not be discouraged if you stop reading partway through certain chapters. The book is not intended to be read cover to cover (except, of course, by my students for whom it is required reading). For some of you, it will be light reading, almost conversational. For the statistically minded reader, this would be an introductory text. For all readers, I hope that it increases your confidence with prediction statistics and your motivation to go further.

FIGURE 1.2

"Are you just pissing and moaning, or can you verify what you're saying with data?"

Note. By Edward Koren. From The Cartoon Bank of The New Yorker Collection. Copyright 1999 by Condé Nast. Reprinted with permission.

I do not expect readers to permanently master the material the first time they learn it. I have learned and forgotten this material many times. I was taught logarithms and exponents in high school, then had to relearn them again during my BA, during my PhD, and several more times during my professional career as a researcher. Our brains are finite networks, and we regularly decommission pathways to make room for current concerns. Logits fade away if not frequently revisited.

The working knowledge of communities of practitioners is similarly impermanent. The founding motivation for the first modern university in 1088 in Bologna was to revive the lost learning of the Romans. The great achievements of our civilizations cannot be sustained without continual education of the next generation. Even mundane achievements, such as buttons or hopscotch, need support to survive. One of the recurrent themes in science fiction are worlds filled with advanced technology that nobody understands. These stories resonate with anybody who has marvelled at airplanes, electric lights, and plastic toothbrushes. One contemporary variation on the lost knowledge theme is the Dr. Stone manga series, which chronicles the struggles of a young genius to recreate basic science and technology after waking up in a

primitive, dystopian future. How many of us would know how to make paper from scratch, let alone a computer monitor?

Empirical probabilities and statistical prediction are valuable innovations, and their application to the problems of human affairs is a relatively recent development. When used appropriately, they have considerable potential to improve decision making. These innovations can only achieve their potential, however, if they are part of the working knowledge of practitioners. May this book increase the number of people who could create risk tools from scratch.