The Ethics of Using or Not Using Statistical Prediction Rules in Psychological Practice and Related Consulting Activities

Robyn M. Dawes†
Carnegie Mellon University

Professionals often believe that they must "exercise judgment" in making decisions critical to other people's lives. The relative superiority (established in roughly 150 studies) of statistical prediction rules (SPR's) to intuitive judgment for combining incomparable sources of information to predict important human outcomes leads us to question this personal input belief. Some professionals hence use SPR's to "educate" intuitive judgment, rather than replace it. In psychology in particular, such amalgamation is not justified. If a well-validated SPR that is superior to professional judgment exists in a relevant decision making context, professionals should use it, totally absenting themselves from the prediction.

Scientific theory and understanding cannot be reduced to a set of prediction equations, but ability to predict—even what will be found about past events (as in an autopsy or archaeological dig)—is the criterion for meeting the "show me" demand of all scientific assertions. Moreover, applied work also demands predictive capacity. A practitioner can influence change in a desirable direction only when able to predict what is likely to occur as the result of various possible interventions, or none at all.

While some physical processes follow deterministic principles (at certain levels of analysis anyway), and while some analysts accept as a matter of faith that processes in psychology and other social sciences would "ultimately" be deterministic "if only we know enough," most such processes

†Send requests for reprints to the author, Department of Social and Decision Sciences, 208 Porter Hall, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: rd1b+@andrew.cmu.edu.

Philosophy of Science, 69 (September 2002) pp. S178–S184. 0031-8248/2002/69supp-0016$10.00
Copyright 2002 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/341844 Published online by Cambridge University Press

are for all practical purposes probabilistic. Moreover, medicine is as well—at least until encountering a "pathognomic" sign or symptom (one that occurs *only* in the presence of a particular pathology). Thus, it is not surprising that a number of investigators in these areas have developed Statistical Prediction Rules (SPR's); some of these rules are situation specific, some developed to be valid across a variety of situations.

Once an SPR has been developed (and validated to apply to a particular situation) practitioners in that situation face three possible ways of making a prediction.

Method #1: Relying on past experience and training to make an intuitive prediction.
Method #2: Relying wholly on the SPR developed to be used in that situation.
Method #3: Taking account of the output of the SPR but possibly modifying it on the basis of professional experience and intuition.

In his classic 1954 book, Paul Meehl summarized approximately 20 studies comparing the predictive validity on methods #1 and #2. In no case was Method #1 superior. By 1996, Grove and Meehl were able to summarize about 135 studies comparing Methods #1 and #2. In less than five percent was there evidence for the superiority of Method #1. Recently, Swets, Dawes, and Monahan (2000) have compared all three methods—in the areas of psychology and medicine. In the former area (e.g., predicting future violence, success on parole, success in a professional or academic program), Method #2 was clearly the best. In medicine, sometimes Method #2 was superior and sometimes Method #3 was. In no case was Method #1 superior to *either* of the others. Swets, Dawes, and Monahan used the signal detection statistic of the area under Received Operating Curves (ROC's) as their criterion of predictive validity. Other investigators have used correlation coefficients and accuracy of categorization—without any differences in the direction of the results and their implications.

The research supporting Methods #2 and #3 provides both practitioners and theorists with incremental validity; that is, the research and the resulting methods allow people to improve their performance—either when they are attempting to make a practical prediction or to test a theory (the latter by incorporating the theory into an SPR). I suggest that a provision of such incremental validity is a criterion of value for what Meehl (1967, 1978) terms the "soft" areas of psychology and other social science (which in my opinion is almost universally "soft"—despite the pretensions of neo-classical economic theory).

But there is often a problem with acceptance of incremental validity. A prime example is the research indicating the invalidity and potential bias of using interviews in predicting behavioral criteria such as success on

https://doi.org/10.1086/341844 Published online by Cambridge University Press

parole, future violence, or success in an academic or training program. Here, validity can be incremented by abandoning such interviews, especially in favor of more *accurate* assessments of past behavior or current aptitude (based on tests that have been carefully devised and evaluated). People, particularly those involved in such procedures and often paid handsomely for their involvement, are, however, convinced that they work and hence don't abandon them (see Dawes 1994, chap. 3). Thus, ethical research that truly could change "intuitively compelling" practice for the better can be met with hostility. That reaction is a structural necessity if indeed we are to provide incremental validity in areas of human endeavor where people already have their preconceptions. There is no such problem, for example, in doing research in neural transmitters (at least until it comes to the conclusion that there is no such thing as an "addictive personality").

The statistical versus clinical controversy is just one example of the problem. Consider, also, the "dodo bird" finding about "relationship" psychotherapies (i.e., those that are not followed according to some protocol, such as many behavioral therapies). The gold standard of comparison of randomly constituted experimental versus control groups yields the conclusion that these types of psychotherapies are successful. That same standard leads to the conclusion that the credentials and experience of the therapist are irrelevant to success; see Dawes 1994, chap. 2. (So much for faith in the "gurus" who not only become our therapists, but appear on television almost daily—especially after some catastrophe—to tell us how to think and feel.)

Another example is sterile needle exchange programs for IV drug users. They do not, in fact, increase drug usage, and of course no one can catch HIV or Hepatitis C from a sterile needle (see Moses 1994). Whatever arguments against these programs may be made on deontological bases, the ones that state that they are worthless either because HIV-infected drug users would not use them, or because they will encourage IV drug use, or—somewhat surprisingly—both, are simply not true. But here many people have such strong preconceptions that they regard the data as irrelevant.

Another example concerns memory. While the subjective experience of recall may be that of "replaying" what actually happened—much like replaying a rewound videotape—belief about what happened is a highly reconstructive process. (See, for example, Bartlett 1932 for examples involving material presented to subjects and Pearson, Ross, and Dawes 1991 for materials in the subject's own life history.) Thus, it is possible to have people create images and subsequently to convince them that these images refer to what actually happened in their lives, rather than to retrospective illusions based on "giving free reign to the imagination" (Poole et al. 1995), creating vivid images of what *might* have happened, or complying with

https://doi.org/10.1086/341844 Published online by Cambridge University Press

bias introduced by the wording of questions. (See Loftus 1994 for a review of her work and the work of others.)

A final instance concerns the recent "flap" over findings about the long-term effects of child sexual abuse (Rind et al. 1998). As the authors of a careful meta-analysis argued, the definition of such abuse is generally so broad that it includes many types that in fact do not often lead to lasting negative effects. Nevertheless, their conclusion and the article itself were denounced—with the help of the American Psychological Association, which published it (see Fowler letter of June 9, 1999 to Representative Tom Delay)—by Congress (106th Congress, 1999). Here, there is a corollary that is extraordinarily important, and again which tends to be contrary to general public belief. Most people get over most things on their own. That's true even of addictions (see Cohen et al. 1989).

If indeed "soft" areas of psychology and other social sciences have value insofar as they provide incremental validity, and if, as argued, such validity often is alienating to people with strong preconceptions (such as members of Congress), then the research that has the *most* value will be the *least* accepted. At the extreme, it is sometimes labeled as "nonsensical" or "unethical," but more often it is simply ignored. For example, I was a member of the National Academy CBASS-AIDs committee that attempted to obtain support for an experimental evaluation of sterile needle exchange programs (*prior* to implementing them; see Turner, Miller, and Moses 1989). I persuaded professorial colleagues in physical sciences to write letters of support to Congress, but I could persuade *none* from the social or behavioral sciences, even though it was "our own" research that led to the conclusion that such programs would save (a considerable number of) lives without the negative side effect of encouraging drug use.

Now I would like to return to the problem of using statistical prediction rules. As Paul Meehl, who started it all, wrote some fifteen years ago (Meehl 1986, 372–373):

> There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one. When you are pushing 90 investigations [today about 150], predicting everything from the outcomes of football games to the diagnosis of liver disease and when you can hardly come up with a half dozen studies showing even a weak tendency in favor of the clinician, it is time to draw a practical conclusion.

Thus, a major implication of all the research is that to practice ethically, the practitioner in these areas **must** employ SPR's (either Method #2 or #3) when they are available. Moreover, the practitioner claiming to use his or her own intuition to "improve" an SPR has an ethical obligation to keep track of outcomes to see if modification really does result in im-

https://doi.org/10.1086/341844 Published online by Cambridge University Press

provement (and must be wary of confounded judgments, self-fulfilling prophesies and other challenges to the validity of evaluating such feedback). These ethical mandates, following from the research, are quite different from those commonly believed to be in operation, where "responsibility" is interpreted as "accepting it" by "making the predictive judgment oneself." Rather, responsibility implies relinquishing some, or total, control over that judgment. Following this ethical mandate is extraordinarily difficult.

People object that it is not just *practically* but *ethically* irresponsible to substitute the "decision" of an SPR for one's own judgment (a substitution often considered to constitute "epistemic irresponsibility," according to Bishop (2000), who supports it). These SPR's are, after all, based on examining a large sample of people and then assuming that what is true of the sample is true of the individual. We accept that type of inference quite often in areas such as medicine; for example, most of us have our children inoculated against polio, even though the efficacy of the vaccine was determined by comparing an entire experimental group with an entire control group of children years ago. In addition, knowing that one medical procedure is twice as likely to be successful as another results in another "no-brainer" choice between the two—even though when converted to a correlation coefficient between prediction and outcome the result is a phi value of .33 (with equal base rates! Without, the phi value, which is the same as a Pearson correlation coefficient between zeros and ones, is even lower). We would not, however, accept the idea that since most people in our income brackets are Republicans, we must ourselves have conservative political leanings. (What an insult!)

Clearly, the degree to which we are willing to substitute an SPR for our own judgment about our ability to understand particular people and situations "in all their uniqueness" varies from context to context. Here, however, we have reviewed a wide variety of contexts and found that the use of the SPR with Methods #2 and #3 is indeed superior to the use of unaided intuition (Method #1). Here I have attempted to make the case that the class of context in which we should use the SPR is broad. Perhaps one way of characterizing it is that it is the class in which we are responsible for making judgments about *others* (which would even include their political judgments).

In addition to the inference from the group to the person, statistical judgments bypass what we believe makes human beings "special," our highly developed cerebral cortex. Due to past scientific discoveries, we have had to relinquish the idea that our habitat is the center of the universe, and then that we are unique among all animals, but have instead evolved throughout history just as others have. In fact, recent studies indicate that we are not even uniquely "the tool making animal" or "the

https://doi.org/10.1086/341844 Published online by Cambridge University Press

social animal." Well, at least we are supposed to be able to reason and judge logically—even if we do not do so automatically (see Dawes 2001). It is that unique ability, which is associated with our cerebral cortex, that appears to make us human. Unfortunately, however, the main difference between ourselves and chimpanzees lies not in the development of the cerebral cortex, but in the number of sensory cells devoted to the visual system (see Sinsheimer 1971).[1]

We gladly relinquish our *perceptual* abilities to some outside aid when there is something deficient about these. We wear reading glasses, we wear hearing aids, we use crutches if we have broken bones, and ultimately many of us even accept intravenous feeding in our (understandable from an evolutionary perspective) obsession to live forever. It is a little more difficult to accept outside aids to our perceptual and motor abilities when there is "nothing wrong." We do so, however, when we know that we should land Boeing 727's by instruments when flying from a dark space into a lighted area—rather than relying on our vision, which together with the shape of the cockpit creates an illusion that we are higher above the landing strip than we in fact are. Relinquishing our intellectual abilities to the external aid of an SPR meets with more resistance. But nobody ever said that behaving ethically was easy.

REFERENCES

106th Congress. (July 12, 1999). House Concurrent Resolution 107, sponsored by Matt Salmon.

Bartlett, F. C. (1932), *Remembering.* Cambridge: Cambridge University Press.

Bishop, M. A. (2000), "In Praise of Epistemic Irresponsibility: How Lazy and Ignorant Can You Be?", *Synthese* 122: 179–208.

Cohen, S. et al. (1989), "Debunking Myths about Self-Quitting", *American Psychologist* 44: 1355–1365.

Dawes, R. M. (1994), *House of Cards: Psychology and Psychotherapy Built on Myth.* New York: The Free Press.

——— (2001), *Everyday Irrationality: How Pseudo-Scientists, Lunatics, and the Rest of Us Systematically Fail to Be Rational.* San Francisco, Calif.: Westview Press.

Fowler, R. D. (June 9, 1999), *Letter to Tom DeLay.*

Loftus, E. F. (1994), "The Repressed Memory Controversy", *American Psychologist* 49: 443–445.

Meehl, P. E. (1967), "Theory Testing in Psychology and Physics: A Methodological Paradox", *Philosophy of Science* 34: 103–115.

——— (1978), "Theoretical Risks and Tabular Risks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology", *Journal of Consulting and Clinical Psychology* 4: 806–834.

1. Dawes (2001) argues that it is the *combination* of social interaction with the cerebral cortex that leads to our intelligence. We are able to scrutinize and criticize each other's ideas—especially after the invention of the printing press and even more modern communication devices (e.g., e-mail courtesy of the computer). The particular virtue of communication is that it is easier for us to *recognize* good arguments and conclusions than to *generate* them (and conversely, it is easier to recognize flaws once pointed out to us by others than to avoid generating them ourselves).

https://doi.org/10.1086/341844 Published online by Cambridge University Press

——— (1986), "Causes and Effects of My Disturbing Little Book", *Journal of Personality Assessment* 50: 370–375.

Moses, Lincoln E. 1994, *Proceedings: Workshop on Needle Exchange and Bleach Distribution Programs.* Washington, D.C. : National Academy Press.

Pearson, R. W., M. Ross, and R. M. Dawes (1991), "Personal Recall and the Limits of Retrospective Questions in Surveys", in J. M. Tanur (ed.), *Questions about Questions: Inquiries into the Cognitive Bases of Surveys.* New York: Russell Sage Foundation, 65–94.

Poole, D. A., D. S. Lindsay, A. Memon, and R. Bull (1995), "Psychotherapy and the Recovery of Memories of Childhood Sexual Abuse: U.S. and British Practitioners' Opinions, Practices, and Experiences", *Journal of Consulting and Clinical Psychology* 63: 426–437.

Rind, B., P. Tromovitch, and R. Bauserman (1998), "A Meta-Analytic Examination of Assumed Properties of Child Sexual Abuse Using College Samples", *Psychological Bulletin* 124: 22–53.

Sinsheimer, R. L. (1971), "The Brain of Pooh: An Essay on the Limits of Mind", *American Scientist* 59: 20–28.

Swets, J. A., R. M. Dawes, and J. Monahan (2000), "Psychological Science Can Improve Diagnostic Decisions", *Psychological Science in the Public Interest* 1: 1–26.

Turner, C. F., H. G. Miller, and L. E. Moses (1989), *AIDS: Sexual Behavior and Intravenous Drug Use.* Washington, D.C.: National Academy Press.

https://doi.org/10.1086/341844 Published online by Cambridge University Press