Chapter 3
Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines

Rob R. Meijer, A. Susan M. Niessen, and Marvin Neumann

Abstract When it comes to decision-making based on psychological and educational assessments, there is compelling evidence that statistical judgment is superior to holistic judgment. Yet, implementing this finding in practice has proven to be difficult for both academic and professional psychologists. Knowledge transfer from research findings to practitioners and other stakeholders in psychological assessment is a necessary condition to close this gap. To obtain insight into how academic specialists in psychological testing disseminate knowledge about research findings in this area, we investigated how textbooks on testing and guidelines on test use report on, or do not to report on, decision-making in psychological and educational assessment. Second, we discuss some commonly encountered misunderstandings, and third we argue for a broader and more in-depth dissemination of research findings on this topic in textbooks and test standards; to this end we provide some suggestions.

3.1 Psychological and Educational Testing and Decision-Making: The Lack of Knowledge Dissemination in Textbooks and Test Guidelines

For decades, many Dutch psychology students' first acquaintance with psychometrics included studying the book by Drenth (1965, 1975) or the more recent editions by Drenth and Sijtsma (1990, 2006). Although, at some Dutch universities, this book

---

R. R. Meijer (✉) · A. S. M. Niessen · M. Neumann
Department Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands
e-mail: r.r.meijer@rug.nl; a.s.m.niessen@rug.nl; m.neumann@rug.nl

has since been replaced by more recent books, its influence on psychological testing in the Netherlands is significant. In our discussion with practitioners and academics, the book is still often mentioned as an authority textbook on test design and test use.

We still use the 2006 edition for our lectures to Dutch students, and one of the best features of the book is that it contains a chapter (Chap. 9) about "The contribution of a test in the decision-making process." As we discuss and illustrate in this chapter, there are not many introductory textbooks on test theory or psychological and educational testing that devote much attention, let alone a whole chapter, to test use and decision-making. Most textbooks pay close attention to topics like reliability, validity, and types of tests, but test *use*, that is, the basic principles on how professionals should use tests when they make decisions, is often not discussed. Also, on conferences where psychometric research is presented, such as those of the National Council on Measurement in Education (NCME), the International Test Commission, or the Psychometric Society, presentations on test use and decision-making are almost nonexistent.

This is perhaps not that surprising because, as discussed in van der Linden (1991), although the practice of testing is firmly rooted in the field of decision-making (educational selection, selection for the military and companies), test theory or psychometrics has been mainly developed as a measurement theory. There are a few exceptions: the well-known work by Taylor and Russel (1939) and the book by Cronbach and Gleser (1965); this latter work provided a theoretical basis for test-based decision-making. Thus, in courses on psychometrics, students learn about measurement theories like the principles of classical test theory, item response theory, and factor analysis and in more advanced courses about the development of different psychometric models, parameter estimation procedures, fit statistics, and the application of these models to empirical data. But, in psychological testing or related courses, test *use* is not really instructed. While most textbooks on psychological testing discuss the decision-making perspective (e.g., Taylor-Russel tables) and some focus on utility models, there is a lack of focus on usage, that is, how to combine test scores with other information, as we discuss below.

This underrepresentation of knowledge and skill in test use in academic education is problematic. As future professionals, most of our students will mainly use psychological tests as a decision-making tool. In most applied settings, psychological tests are part of an assessment used to make judgments and predictions about behavior of individuals (Kuncel, 2008). For example, consider the following two scenarios.

A parole board consisting of different professionals, including two clinical psychologists, has to decide about temporary or permanent release of a prisoner before the expiry of the sentence, on the promise of good behavior. This decision has important consequences for the prisoner and for society, and many factors determine the prisoner's future behavior. One of the standardized instruments that can be used to make this important decision is the Level of Service/Case Management Inventory (Andrews et al., 2004). This instrument assesses static and dynamic factors linked to recidivism risk based on 43 items, divided into 8 major categories. The total score provides information on the risk posed by the offender, and the

subcategories indicate individual characteristics that increase the risk of recidivism (i.e., criminogenic needs). The total score is used to determine the offender's initial risk level on a five-point ordinal scale ranging from very low risk to very high risk. Importantly, individual assessors can often override the initial risk level to create a final risk level when they see reasons to do so (see Guay & Parent, 2018 for more details). An important question is: Is it wise to override the initial risk level, for example, on the basis of professional expertise and experiences with a delinquent?

A hospital is searching for a consultant occupational physician. Requirements are "enthusiastic to continue the success of the team with innovative ideas, a careful decision maker, always putting the patients first, an excellent communicator, able to influence others positively and supportively, able to demonstrate leadership in a multi-professional environment" (these requirements were taken from an actual ad). A search team under the supervision of an I/O psychologist is advising the management which of 18 applicants is most suited for the job. They use an intelligence test, a situational judgment test, and an interview to decide which candidate is most suited for the job.

How should the information from the tests and the interview be combined to optimize the predictive validity of the decision? Should management review the scores on these three assessments and make a global judgment or should they compute a weighted average of the scores on these assessments and hire on the basis of this weighted average?

These two examples demonstrate test use by professional psychologists in (highly) consequential contexts. Other examples are deciding what diagnosis is the most suitable for a client, whether a client is eligible for a particular treatment, whether an athlete belongs to the 10% most capable athletes for a sports team, or whether a child needs extra training in particular subject matters in school.

Such decisions are rarely made using a single assessment tool. For example, in personnel selection, ability tests and interviews are used because these assessments are easy to administer and are expected to increase the criterion-related validity for later job performance, compared to only using one of these assessment tools (Schmidt & Hunter, 1998). Similarly, diagnoses and treatment recommendations in clinical psychology are often made based on a combination of tests, observations, biographical information, and clinical interviews. Therefore, it is not only important for professionals to know *what* information to use when making decisions (what are valid predictors and how can they best be measured) but also to know how to *combine* information from different sources to optimize prediction.

Many studies have been conducted to investigate how information can best be combined to optimize prediction. A major topic of investigation in this respect has been the distinction between holistic and statistical prediction. In holistic (or clinical, impressionistic, intuitive, informal) prediction, information is combined "in the head" of the decision-maker. Conversely, in statistical (or actuarial, mechanical) prediction, information is combined based on formal weighting procedures. In a classic review of 20 studies, Meehl (1954, inspired by Sarbin, 1943) showed that statistical prediction resulted in better predictions than holistic prediction. Many other studies confirmed these findings ever since (e.g., Grove et al., 2000).

Using statistical prediction is arguably one of the most effective ways to improve predictions and decisions in practice (Milkman et al., 2009). However, statistical prediction is not popular among professionals (e.g., Arkes, 2008; Highhouse, 2008; Meijer et al., 2020; Kuncel et al., 2013; Ryan & Sackett, 1987; Terpstra & Rozell, 1997; Vrieze & Grove, 2009).

There are several explanations for the underutilization of statistical prediction in practice, such as lack of perceived autonomy and fear of losing professional status (Highhouse, 2008; Nolan et al., 2020; Neumann et al., 2021b, 2021c). One important prerequisite, however, is knowledge. Without having knowledge about how to best combine information, psychologists will not use statistical decision-making (Neumann et al., 2021a). Therefore, in the present study, we first discuss a number of important characteristics of statistical prediction.

Second, we investigated how research findings on holistic and statistical prediction are disseminated. Textbooks are meant as summaries of academic research that synthesize findings and translate them into accessible information for students and professionals. Through studying how textbooks discuss holistic and statistical prediction, we learn about how research in this area is disseminated, which elements are unclear, and what misconceptions and controversies still exist. This knowledge is useful for two reasons (1): it may help improve the dissemination of research findings and (2) it provides input for research that is aimed at closing the science-practice gap (see Neumann et al., 2021b, for a research agenda).

Besides textbooks, test standards play an important role in disseminating information about evidence-based test use. Therefore, third, we describe if and how test standards disseminate knowledge on this topic. As we discuss below, test standards do not seem to be aimed at discussing or prescribing how test information can best be combined to optimize decision-making. We provide arguments for the importance of including research findings on information combination and decision-making to optimize test use in psychological practice. We want to emphasize that our aim is not to point fingers at authors of the textbooks and guidelines we reviewed, but to improve the dissemination of important research findings with respect to decision-making and prediction to strengthen psychology as an evidence-based, applied science.

3.1.1   Theory of Social Representation

To better understand how textbooks and test standards represent scientific theory of decision-making, we used the theory of social representation as discussed and used in Roulin and Bangerter (2012). They investigated the science-practice gap by studying how the use of structured interviews was diffused to practitioners in practitioner-oriented advice books. As they discussed "the theory of social representations (. . .) seeks to describe the social processes by which scientific knowledge is transformed into everyday knowledge used by laypersons" (p. 150). An interesting phenomenon is that laypersons often integrate new theories in

existing schemes or ideas. This is called anchoring. Second, this theory suggests focusing on the intermediary actors that translate scientific findings into social representations.

Authors of textbooks are the intermediary actors that delve into expert knowledge with the intention of diffusing it to students and professionals. They thus play a key role in the potential transformation of scientific findings, because (1) they may have different understandings of concepts than the experts they cite; and (2) they are designing their message to fit their audience's knowledge (Clark & Murphy, 1982). Compared to journalists and mass media, authors of textbooks are intermediary actors that stand much closer to the original research (Krathwohl, 1998, pp. 54–55) and are often specialists on the topic of their books.

3.2 Using Tests to Make Decisions

3.2.1 Basic Distinctions: Data Collection and Data Combination

For professionals that use assessment results for decision-making or prediction, which are almost all professionals in psychology and related disciplines, it is important to have knowledge about the way information can best be combined. Below we first provide descriptions of holistic and statistical prediction given by Meehl (1954, p. 3) and some later remarks given in Dawes et al. (1989) and Grove and Meehl (1996) because these articles are often cited in textbooks we discuss below. Meehl (1954, p. 3) discussed statistical prediction in the context of diagnosing persons for therapeutic sessions as follows:

"We may order the individual to a class or set of classes on the basis of objective facts concerning his life history, his scores on psychometric tests, behavior ratings or check lists, or subjective judgments gained from interviews". The mechanical combination of information for classification purposes, and the resultant probability figure which is an empirically determined relative frequency, are the characteristics that define the actuarial or statistical type of prediction.

Three important elements of statistical prediction are (1) both "objective" and "subjective" (but quantified) impressions can be considered; (2) there is a mechanical combination rule; and (3) the rule is based on empirically established relations between the combined scores and observations and the behavior we want to predict. So, statistical prediction is not restricted to psychological test use; an assumption sometimes made in textbooks as we discuss below.

Holistic prediction is described as follows by Meehl (1954, pp. 3–4):

On the basis of interview impressions, other data from the history, and possibly also psychometric information of the same type as in the first sort of prediction, we formulate, as in a psychiatric staff conference, some psychological hypothesis regarding the structure and the dynamics of this particular individual. On the basis of this hypothesis and certain reasonable expectations as of the course of outer events, we arrive at a prediction of what

is going to happen. This type of procedure has been loosely called the clinical or case-study method of prediction.

Importantly, in holistic (clinical) decision-making, a prediction is made by "thinking about" the available information, not by using a pre-defined rule or on the basis of explicit empirically established relations. Relatedly, Dawes et al. (1989) described holistic and statistical predictions as

in the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rests solely on empirically established relations between data and the condition or event of interest.

Furthermore, Dawes et al. (1989) noted that

Virtually any type of data is amenable to actuarial interpretation. For example, interview observations can be coded quantitatively (patient appears withdrawn: [1] yes, [2] no). It is thereby possible to incorporate qualitative observations and quantitative data into the predictive mix. Actuarial output statements, or conclusions, can address virtually any type of diagnosis, description, or prediction of human interest.

Thus, in short, statistical prediction is about the way information is *combined*, not about *what* information is used to make decisions.

3.2.2   Statistical Prediction Is Superior to Holistic Prediction

As mentioned above, many empirical studies and meta-analyses convincingly showed that following structured decision rules results in better prediction than combining information "in the head" (Meehl, 1954; Kuncel et al., 2013; Grove et al., 2000; Karelaia & Hogarth, 2008; Ægisdóttir et al., 2006; Morris et al., 2015). More specifically: Dawes et al. (1989) cited almost 100 comparative studies and found that the statistical method performed better than the holistic method. Grove et al. (2000) analyzed 136 studies from medicine, education, and clinical psychology, where professionals predicted outcomes such as academic performance, job success, medical or psychiatric treatment success, criminal recidivism, and suicide. They concluded that "Even though outliers can be found, no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction were identified." Grove and Meehl (1996, p. 26) discussed that, from a theoretical perspective, this conclusion should be expected:

From a theoretical viewpoint the issue may be rather uninteresting, because it is trivial. Given an encodable set of data – including such first-order inferences as skilled clinicians' ratings on single traits from a diagnostic interview – there exists an optimal formal procedure (actuarial table, regression equation, linear, nonlinear, configural, etc.) for inferring any prespecified predictand. This formula, fallible but best (for a specific clinical population), is known to Omniscient Jones but not to the statistician or clinician. However, the statistician is sure to approximate it better, if this is done properly. If the empirical comparisons had consistently favored informal judgment, we would have considerable explaining to do.

The argument that statisticians should do (and do) a better job at approximating the optimal way to combine information for prediction, and the sections in definitions of statistical prediction by Meehl (1954) and Dawes et al. (1989) that emphasize using statistical rules based on *empirically established* relations between information and the behavior we want to predict, reveal the most significant practical challenge for the application of statistical prediction in practice. They require the availability of data to design empirically based statistical prediction rules.

3.2.3 Robustness of Simple Rules

So, ideally, large datasets based on representative samples of the target population are collected to estimate optimal weights for each variable (e.g., in regression analysis), and the results are cross-validated. Clearly, this is often not possible in practice because such datasets are not available. Effective methods to tackle this steep hurdle are described by Dawes (1979). He discussed that, instead of using optimal weights derived from large, primary data, using the same weight for all variables (i.e., unit weighting) or even using randomly chosen but consistent weights in mechanical procedures still often results in better predictions than using holistic prediction.

However, under particular conditions, unit weighting can result in less valid predictions compared to using the single best predictor alone (Murphy, 2019; Sackett et al., 2017). A simple rule was discussed in Murphy (2019): avoid using predictors (i.e., give them a zero weight instead of a unit weight) that correlate more strongly with the other predictors than with the criterion. Moreover, this advice holds when decisions are made holistically as well, since adding such information could "dilute" the most predictive information (Dana et al., 2013).

3.2.4 People Are Bad at Identifying Exceptions to the Rule

When statistical rules are used in practice, they typically serve as decision aids that can be overruled when professionals believe that is appropriate (e.g., Guay & Parent, 2018). Importantly, research shows that overriding a statistical prediction because a certain specific case is believed to be an exception to the rule is a bad idea: people are not very good at correctly identifying these exceptions (Guay & Parent, 2018; Dietvorst et al., 2018; Dawes, 1979). This conclusion can be logically derived from the findings that statistical prediction outperforms holistic prediction; if people were good at identifying exceptions, holistic procedures would outperform mechanical procedures (see Dana et al., 2013 for a similar remark).

A question that arises from the above is whether psychologists can learn to match the predictive accuracy of statistical rules through experience. Kahneman and Klein (2009) discussed this question in depth and concluded that professionals

in psychology have a hard time to match the accuracy of their holistic predictions to the accuracy of decision rules, because (1) the environment in which psychologists act is difficult to predict and (2) feedback is absent or incomplete and delayed at best, which both seriously hinder learning. The biggest problem, however, is that these findings are in conflict with the *perceptions* of making accurate predictions that many professionals have when making decisions. As Kahneman discussed "If people can construct a simple and coherent story, they will feel confident regardless of how well grounded it is in reality" (Kahneman & Klein, 2010, p. 4).

3.2.5 Transparency

Another important characteristic of statistical prediction as defined above that we would like to mention is their transparency. By combining information in a pre-defined, transparent rule, we can replicate decisions, evaluate our policies, and adapt decision rules accordingly, *because we know exactly what we did*. In contrast, that is not the case when decisions are made holistically, because it cannot be directly observed how an assessor combines information "in the head." This makes it harder to evaluate and improve our decisions.

3.3 What Textbooks Communicate About Test Use and Data Combination

We investigated the following research questions:

1. Do textbooks on psychological testing discuss statistical/holistic decision-making?
2. Which references to sources do they use as the basis of their treatment of this topic?
3. Are their conclusions in line with the literature on this topic? In particular, we investigated five criteria: (3a) Is the overall conclusion in line with the empirical literature: statistical prediction should be preferred over holistic prediction? (3b) Do textbooks make a distinction between data collection methods (e.g., tests, interviews, observation) and data combination methods (according to a rule or in the head)? (3c) Is there a discussion about the robustness of using non-optimal weights? (3d) Do textbooks mention exceptions to the rule, and do they correctly discuss how to handle them? (3e) Is there a discussion about transparency of decision making? Although we consider transparency a very important aspect of decision-making, it is not often discussed in the statistical/holistic literature and therefore we did not take this aspect into account when evaluating criterion 3a.

3.4 Method

3.4.1 Sample

We conducted a broad search of textbooks on psychological testing. We started with an electronic search using the library search engine *SmartCat* with the search term "books on psychological testing" with restriction that books should be written in English and published after 1995. This date was a bit arbitrary; we were interested in how statistical versus holistic prediction using tests is discussed in the more recent literature. This resulted in 3031 hits. The first author of this study then selected books using the following inclusion criterion: the books should be broad introductory books on psychological testing. Books on specific topics, such as books exclusively on intelligence testing or test use in minority groups were excluded. This strongly reduced the number of hits. The third author independently selected books using the same search engine and based on the same criteria discussed above as the first author, and he found one book that was not identified by the first author, which was added to the list. This resulted in a selection of 13 textbooks (Table 3.1).

3.4.2 Coding

In each book, we analyzed the content of the text to evaluate if and how statistical and holistic prediction were presented. Because textbooks contain a large amount of information (often several hundred pages), we first looked at the index and the references to identify potentially useful sections. Index terms we used were *clinical*, *holistic*, *actuarial*, *mechanical*, *statistical prediction*, and *decision making*. Authors we looked for in the references were *Meehl* and *Dawes*. When these references did not provide any results, we also checked *Highhouse* and *Kuncel* and *Grove*. However, this did not provide additional information as all textbooks referring to Highhouse, Kuncel, or Grove also referred to Meehl or Dawes.

Two independent raters (first and third author) searched the books and coded the texts on the basis of the five research questions mentioned above under 3(a)-3(e). The two raters checked the text passages on the basis of the five criteria discussed above. Each criterion was rated on a four-point scale: (0) no description at all; (1) description is wrong; (2) there is some description, but lacks important points; and (3) fair, accurate description.[1] The two raters first coded the textbooks independently and then discussed any score differences between them until consensus was reached.

---

[1] We agree with an anonymous reviewer of this chapter that, although technically the ratings are nominal, the coding scheme we used may suggest that they are ordinal. An ordinal interpretation would imply that a wrong description is "better" than no description, which is not the case. As this reviewer correctly remarked "One may argue that the reverse is true, which is reflected in the opening lines of the great must-see movie The Big Short: 'It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.' – also see https://quoteinvestigator.com/2018/11/18/know-trouble/"

Table 3.1 Scores that reflect the way textbooks discuss different criteria

	Conclusion in line with literature	Data collection/combination	Robustness weights	Exceptions to the rule	Transparency
Anastasi and Urbina (1997)	1	1	0	1	0
Aiken (2003)	2	2	0	0	0
Murphy and Davidshofer (2005)	2	2	3	2	0
Kline (2005)	0	0	0	0	0
Domino and Domino (2006)	2	2	0	0	0
Reynolds and Livingston (2012)	2	0	0	2	0
Kaplan and Saccuzzo (2013)	2	0	0	0	0
Gregory (2013)	2	0	0	0	0
Hogan (2015)	1	0	1	0	0
Miller et al. (2015)	2	1	0	0	0
Cohen and Swerdijk (2015)	1	2	0	0	0
Furr (2018)	0	0	0	0	0
Cooper (2019)	0	0	0	0	0

Note: 0 = not discussed, 1 = incorrect description; 2 = description lacks important points; 3 = fair description

3.4.3 Results

In Table 3.1 we provide an overview of the textbook literature. Note that Kline (2005), Furr (2018), and Cooper (2019) did not discuss mechanical versus holistic prediction. Below we summarize the most important findings.

1. Most textbooks on psychological testing discuss statistical versus holistic prediction using a limited number of pages (between 1 and 9 pages, mostly 1–3 pages). There was no textbook that wholeheartedly endorsed the main conclusion from the empirical literature that statistical prediction should be preferred over holistic prediction. Some textbooks only mentioned the empirical results found, without drawing any conclusions or mentioning implications. Almost all textbooks suggested a middle-of-the-road compromise, where they indicate that a rule can be used in some cases, but that there are situations in which that is not possible or desirable. Most reasoning is of the form: *Meehl (1954) or some other meta-analysis found that statistical prediction is superior to clinical prediction. We generally agree with this conclusion, but there are conditions where clinical prediction is preferred (because there are exceptions, because you cannot use tests in all cases, because it is difficult to formulate a rule)*. For example, Murphy and Davidshofer (2005) provided an elaborate summary of the research on statistical versus holistic decision-making, but they also conclude:

   > However, in the long run, the automation of clinical prediction would limit the accuracy of clinical predictions, since it would preclude the use of behavioral observation data or the selection of appropriate tests to optimally assess the status of the individual patient. (p. 529)

There is, however, no reason why *quantified* behavioral observations could not be incorporated in statistical predictions. Furthermore, the "selection of appropriate tests to optimally assess the status of the individual patient" is still possible under mechanical decision-making.

In many passages, there was no explicit distinction between "the nature of information" and "how to combine information." Textbooks rarely explicitly described this distinction. Many passages provide examples of holistic versus statistical prediction which incorrectly suggest that statistical decision-making is tied to using tests and holistic decision-making is tied to using other information (sometimes in addition to tests). For example, Miller et al. (2015, p. 419) discussed that: "For more than 50 years, researchers have debated the accuracy of making diagnoses using the unstructured interview (called the clinical method) compared with using structured psychological tests (called the statistical method). In 1954, Meehl published the results of his examination of 20 studies that compared clinical and statistical predictions (Meehl, 1954). His conclusion was that statistical methods were as accurate as, and often more accurate than clinical methods."

2. Only optimal regression models are described as superior to holistic decision-making. The advantages of suboptimal rules such as unit weighting or expert weighting are not discussed. If authors mention specific examples, they often

come from the clinical context. An interesting example on the use of the MMPI is provided by Gregory (2013, pp. 487–493). Gregory (2013) discussed that "computerized narrative test reports should use existing actuarial formulas to determine the likelihood of various psychiatric diagnosis" (p. 491). However, Gregory (2013) also discussed that a drawback of statistical prediction is that when the rules are applied to a new client population, new rules should be determined because they will perform less well in a new population. Ideally, this would indeed be the case, at least when sufficiently large samples would be available. However, this remark ignores the empirical results that suboptimal weights generally do a better job than holistic combinations (Murphy et al., 2013; Yu & Kuncel, 2020).

Also, Hogan (2015, p. 177) noted that:

> *Can we replace clinicians with formulas? Sometimes yes, sometimes no. Development of formulas requires an adequate database. When we have an adequate database, we should rely on it. But we do not always have an adequate database. In that case, we must rely on clinical judgment to make the best of the situation.*

This is an often-encountered misunderstanding that despite articles like those by Grove and Meehl (1996) and Dawes and Corrigan (1974) seems to be ineradicable. As we discussed above, research showed that picking a number of valid predictors and choosing reasonable weights based on empirical research (e.g., meta-analysis) will often result in more accurate decisions than holistic judgment. If textbooks keep communicating that adequate databases are a necessary condition to be able to use statistical prediction, it is no wonder that practitioners almost exclusively use holistic judgment, because adequate data are rarely available.

3. Some textbooks state that, sometimes, holistic methods should be preferred. These are perhaps the most interesting passages because most of the time, no references are provided to support those statements; they seem to rely on "common sense" or "authority" arguments. Most importantly, there is no evidence that holistic methods should be preferred over mechanical procedures in *any* situation.

Some authors seem to imply that we do not know which decision-making method is superior. For example, Kaplan and Saccuzzo (p. 554) noted "Further, the question remains as to whether computer interpretations can ever be as good as, let alone better than, those of the clinician." Sometimes references are used, but then the content of these references is refuted by more recent articles, or the original articles are misinterpreted. For example, Aiken (p. 337) discussed that "under certain circumstances trained practitioners employing data from a variety of sources (case history, interview, test battery, and the like) are better than actuarial formulas (Goldberg, 1970; Holt, 1970; Wiggins & Kohen, 1971)." This is incorrect, because Goldberg (1970) showed the opposite, namely, that statistical rules created based on decisions made by the assessors were better than assessors themselves. Additionally, Holt (1970) is sometimes used as a reference in favor of holistic prediction, but Holt (1986, p. 378) himself conceded that statistical judgment is superior when he wrote:

> *Maybe there are still lots of clinicians who believe that they can predict anything better than a suitably programmed computer; if so, I agree that it is not only foolish but at times unethical of them to do so . . . If I ever accused him [Paul Meehl] or Ted Sarbin of "fomenting the controversy", I am glad to withdraw any implication that either deliberately stirred up trouble, which I surely did not intend.*

3.4.4 Conclusion on Decision-Making as Discussed in Textbooks

The way textbooks on testing discuss decision-making based on a combination of information is mostly not in agreement with the empirical literature. It seems as if authors of textbooks anchor mechanical decision-making to pre-existing schemes, as the theory of social representation would predict. These pre-existing schemes consist of ideas of how we make decisions in daily life: holistically. For example, Anastasi (p. 520):

> *A major contribution of the clinical method for example is that data are obtained in areas where satisfactory tests are unavailable through interviewing and observations of behavior. The clinical method is also better suited than the statistical method to the processing of rare and idiosyncratic events whose frequency is too low to permit development of statistical strategies.*

This remark seems to be based on "common sense," but not on results from the empirical literature which showed the opposite, namely, that people have a hard time in identifying valid idiosyncrasies. As a result, we speculate that many textbook authors (unintendedly) mix empirical findings in the literature with their own experiences. Furthermore, because the topic is more complex than many textbook authors perhaps realize, not enough space is devoted to carefully and accurately explaining the literature.

3.5 What Test Standards Communicate on Decision-Making with Tests

We investigated the following research questions:

1. Do test standards on psychological testing discuss statistical/holistic prediction?
2. Are their conclusions in line with the literature on this topic?[2]

There are different guidelines on test use. Internationally, the most important ones are the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014; in the remaining of this article referred to as the

---

[2] In contrast to the textbook research questions, we did not research which references were used because test standards include very few references.

Standards) and the International Test Commission Guidelines on Test Use (2013; in short, the ITC guidelines). The latter is available in many languages. Both guidelines fulfill an important role to transfer scientific assessment research to professional practice and contain important and very useful information.

3.5.1 Standards for Psychological and Educational Testing

To answer the first research question, it is important to first look at the mission of the Standards. On p. 1 it says

> *The purpose of the standards is to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test use. Although such evaluations should depend heavily on professional judgment, the standards provide a frame of reference to ensure that relevant issues are addressed.*

Furthermore, on p. 2 it is noted that

> *Although the principles and concepts underlying the standards can be fruitfully applied to day-today decisions – such as when a business owner interviews a job applicant, a manager evaluates the performance of subordinates, a teacher develops a classroom assessment to monitor student progress to an educational goal, or a coach evaluates a prospective athlete – it would be overreaching to expect that the standards of the educational and psychological testing field would be followed by those making such decisions. In contrast, a structured interviewing system developed by a psychologist and accompanied by claims that the system has been found to be predictive of job performance in a variety of settings falls within the purview of the standards. Adhering to the Standards becomes more critical as the stakes for the test taker and the need to protect the public increases.*

From these quotes it is clear that decisions made by persons not being a psychologist are considered beyond the scope of the Standards. It may also be inferred that the Standards are particularly concerned with the quality of individual assessment tools. However, decisions are seldom made based on one individual test or instrument. The Standards (p. 198) indeed discuss "In educational settings, a decision or characterization that will have major influences on a student should take into consideration not just scores from a single test, but other relevant information." How this may be done is discussed on p. 170.

> *In some instances, test information is used in a mechanical, automated fashion. This is the case when scores on a test battery are combined by formula and candidates are selected in strict top-down rank order, or when candidates above specific cut scores are eligible to continue subsequent stages of a selection system. In other instances, information from a test is judgmentally integrated with information from other tests and with nontest information to form an overall assessment of the candidate.*

Thus, the *Standards* discuss the difference between mechanical and judgmental (what we call holistic) decision-making, indicating that this is considered a topic of relevance for users of psychological tests. However, the Standards do not mention that mechanical judgment leads to more reliable and valid judgments

than holistic combinations of information. Second, the Standards incorrectly imply that mechanical decision-making can only be used when decisions are based exclusively on test scores and that taking information derived from other sources than standardized tests (such as interviews, biodata) into account requires holistic decision-making.

3.5.2 International Test Guidelines

The aim of the ITC test guidelines is described as follows (p. 7):

> *The Test Use guidelines relate to the competencies (knowledge, skills, abilities and other personal characteristics) needed by test users. These competencies are specified in terms of assessable performance criteria. These criteria provide the basis for developing specifications of the evidence of competence that would be expected from someone seeking qualification as a test user. Such competencies cover such issues as professional and ethical standards in testing, rights of the test taker and other parties involved in the testing process, choice and evaluation of alternative tests, test administration, scoring and interpretation, and report writing and feedback.*

Furthermore, we encountered several statements that encourage using multiple sources of information and thus indicate that information will need to be combined (listed below, with original reference numbers). However, no explicit statement on *how* to combine information was found.

> *2.1.4 Seek other relevant collateral sources of information.*
> *2.1.6 Ensure that full use is made of all available collateral sources of information.*
> *4. Make clear that the test data represent just one source of information and should always be considered in conjunction with other information.*

Thus, although the potential utility of testing in an assessment situation is discussed in the ITC guidelines, statistical versus holistic combination is not discussed. Furthermore, the statement in the ITC guidelines that "collateral information" is useful seems to imply that more information is better. However plausible this may sound, this is not true in general and can encourage problematic decision-making. For example, information from unstructured interviews when combined with valid grades can lower predictive validity compared to using grades alone, but at the same time increase the *feeling* of a valid decision (e.g., Dana et al., 2013).

3.5.3 Conclusion on Decision-Making as Discussed in Test Guidelines

Both guidelines pay little attention to obtaining reliable and valid judgments and decisions based on a combination of different sources of information (e.g., tests, interviews, questionnaires). In the vast majority of cases, psychological tests are

used with the main aim to aid decision-making about an individual, but the research literature on this issue is not discussed and its influence is minimal.

3.6   Concluding Remarks

The findings on how decisions can best be made based on a combination of information are exceptionally robust and should be highly consequential for psychological and educational practice, as well as other fields such as medicine and law (e.g., Arkes et al., 2008; Guay & Parent, 2018; Hanson & Morton-Bourgon, 2009; Schwab, 2008). Professionals and academic psychologists have a hard time accepting the superiority of statistical over holistic decision-making. Since Meehl (1954), a number of articles (e.g., Dawes, 1979; Grove & Meehl, 1996; Highhouse, 2008) addressed different types of objections with insightful explanations why these objections were unwarranted. As our results showed, 67 years after Meehl's publication, time has not resulted in a good understanding or appreciation of this topic in textbooks on psychological testing.

In some textbooks it is remarked that ethical guidelines of psychologists do not allow to completely rely on statistical decision-making. But as Murphy and Davidshofer (2005) discussed: "there are few excuses for not at least considering what a statistical model would say" (p. 530). Furthermore, using a statistical decision-making procedure does not imply that the psychologist is not responsible for the appropriateness of the procedure. As a reviewer remarked, The responsibility lies in selecting the relevant predictors, and setting up the rule to combine the information, but not so much second guess what the outcome is, every time the professional gets a 'hunch.' In fact, a psychologist should closely monitor the outcomes of statistical decision-making, use pilot studies, and intervene when things go wrong, for example, by excluding less valid predictors or adjusting the weights of a statistical rule. In fact, Dawes (2005) argued, and we agree, that it is unethical to not use a method that optimizes valid prediction.

If we take psychology as an applied science seriously, textbooks and test guidelines cannot stay behind in promoting an important finding in our field. Test guidelines form the link between scientific psychometrics and practice. It is thus the place where scientific findings can be disseminated to a wider audience. If we do not translate important empirical findings into guidelines for practice, our scientific findings will have very limited merit. When it comes to decision-making based on test scores, we think we can and should do a better job than we are doing at the moment.

When professionals do not adopt evidence-based procedures for test use, there are at least four possible reasons.

1. They do not have sufficient knowledge about the most appropriate procedures (Neumann et al., 2021a).
2. They do not believe in the evidence presented in scientific studies.

3. They know about and believe in the evidence presented in scientific studies, but do not act upon the evidence because of internal conflict (e.g., need for autonomy, Nolan & Highhouse, 2014).
4. They know about and believe in the evidence presented in scientific studies, but do not act upon them because of external pressures (e.g., stakeholder perceptions, being valued in their work, Nolan et al., 2020).

Including guidelines on test use and decision-making in test standards can help relieve all of these reasons. They can communicate the existing evidence to overcome reason 1, they can discuss common misconceptions and invalid counterarguments to overcome reason 2, and they serve to set a standard to resist both internal and external pressures that hinder using evidence-based prediction procedures. Therefore, we ask authors of textbooks and test guidelines to pay more attention to statistical decision-making.

As a final note, we started this chapter with observing that Drenth and Sijtsma (2006) devoted a whole chapter to the contribution of a test to the decision-making process. Statistical versus holistic prediction is part of this decision-making process, and the question the reader may have now is: How did they reflect on mechanical versus holistic prediction? Well, they did a good job. In response to the question, how should we combine the results of different tests? They noted that (p. 414; our translation):

First, this can be done via a statistical process of weighing test scores and possibly calculation of probabilities, and secondly via an intuitive, not statistical process of weighing and prediction. In this intuitive way it often concerns the different weighting across different cases; the process is less formalized, one does not follow a fixed strategy like in a statistical procedure.

Furthermore, they discussed that:

An evaluation of the many research findings in this context was in agreement with the original conclusions by Meehl that the statistical procedure is superior to the holistic method

And their explanation is (p. 414):

This result can be understood as follows. In a holistic combination of objective data, such as obtained in assessments with tests to predict an objective criterion, all kinds of biases, stereotypes, and unfounded assumptions play a role besides knowledge of the professional literature. One determines often on the basis of intuition the different weights, often in an inconsistent way. In this way some test scores are weighted too heavily, some are getting too few weights and per case and across different measurements there are fluctuations and inconsistencies.

Although they did not tick all the boxes in their chapter as suggested by us in Table 3.1, this phrasing of the main message of the statistical prediction literature was perhaps the most accurate description we found in the textbooks on psychological testing on statistical prediction. [3]

---

[3] *References marked with an asterisk indicate works included in the review.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Aiken, L. R. (2003). *Psychological testing and assessment*. Pearson Education Group.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

*Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *The level of service/case management inventory*. Multi-Health Systems.

Arkes, H. R. (2008). Being and advocate for linear models of judgment is not an easy life. In J. I. Krueger (Ed.), *Modern pioneers in psychological science: An APS-Psychology press series. Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 47–70). Psychology Press.

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2008). The influence of a physician's use of a diagnostic decision aid on the malpractice verdicts of mock jurors. *Medical Decision Making, 28*(2), 201–208. https://doi.org/10.1177/0272989X07313280

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J.-F. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). North-Holland Publishing Company.

*Cohen, R. J., & Swerdijk, M. E. (2015). *Psychological testing and assessment: An introduction to tests and measurement*. McGraw-Hill Education.

*Cooper, C. (2019). *Psychological testing, theory and practice*. Routledge.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. University of Illinois Press.

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 512–520.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology, 61*(10), 1245–1255. https://doi.org/10.1002/jclp.2-180

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95–106. https://doi.org/10.1037/h0037613

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

*Domino, G., & Domino, M. L. (2006). *Psychological testing an introduction*. Cambridge University Press.

Drenth, P. J. D. (1965). *De psychologische test* [The psychological test]. Van Loghum Slaterus.

Drenth, P. J. D. (1975). *De psychologische test* [The psychological test] (2nd ed.). Van Loghum Slaterus.

Drenth, P. J. D., & Sijtsma, K. (1990). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu van Loghum.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Bohn Stafleu van Loghum.

*Furr, R. M. (2018). *Psychometrics, an introduction*. SAGE.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin, 73*(6), 422–432. https://doi.org/10.1037/h0029230

*Gregory, R. J. (2013). *Psychological testing, history, principles, and applications* (7th ed.). Pearson Education.

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*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

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*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/CMI. *Criminal Justice and Behavior, 45*(1), 82–100. https://doi.org/10.1177/0093854817719482

Hanson, R. K. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1–21. https://doi.org/10.1037/a0014421

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(3), 333–342. https://doi.org/10.1111/j.1754-943

*Hogan, T. P. (2015). *Statistical testing, a practical introduction*. Wiley.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction. *American Psychologist, 25*(4), 337–339. https://doi.org/10.1037/h0029481

Holt. (1986). Clinical and statistical prediction. A retroperspective and would be integrative perspective. *Journal of Personality Assessment, 50*(3), 376–386. https://doi.org/10.1207/s15327752jpa5003_7

International Test Commission. (2013). *ITC guidelines on test use version 1.2*. Retrieved from: https://www.intestcom.org/files/guideline_test_use.pdf

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., & Klein, G. (2010). Strategic decisions: When can you trust your gut? *McKinsey Quarterly*, from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/strategic-decisions-when-can-you-trust-your-gut?cid=other-soc-lkn-mip-mck-oth-1912%2D%2D&sid=2972122698&linkId=79428658#

*Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological assessment and theory. Creating and using psychological tests. International Edition*. Central Learning.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

*Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. SAGE. https://doi.org/10.4135/9781483385693

Krathwohl, D. R. (1998). *Methods of educational and social science research: An integrated approach*. Longman.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology, 1*(3), 343–346. https://doi.org/10.1111/j.1754-9434.2008.00059.x

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*(6), 1060–1072. https://doi.org/10.1037/a0034156

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

Meijer, R. R., Neumann, M., Hemker, B. T., & Niessen, A. S. M. (2020). A tutorial on mechanical decision-making for personnel and educational selection. *Frontiers in Psychology, 10*, 3002. https://doi.org/10.3389/fpsyg.2019.03002

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science, 4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

*Miller, L. A., McIntire, S. A., & Lovler, R. L. (2015). *Foundation of psychological testing. A practical approach* (5th ed.). SAGE.

Morris, S. B., Daisley, R. L., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationship between individual assessments and job performance. *Journal of Applied Psychology, 100*(1), 5–20. https://doi.org/10.1037/a0036938

Murphy, K. R. (2019). Understanding how and why adding valid predictors can decrease the validity of selection composites: A generalization of Sackett, Dahlke, Shewach, and Kuncel (2017). *International Journal of Selection and Assessment, 27*(3), 249–255. https://doi.org/10.1111/ijsa.12253

*Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing, principles and applications* (6th ed.). Pearson Prentice Hall.

Murphy, K. R., Deckert, P. J., Kinney, T. B., & Kung, M. C. (2013). Subject matter expert judgments regarding the relative importance of competencies are not useful for choosing the test batteries that best predict performance. *International Journal of Selection and Assessment, 21*(4), 419–429. https://doi.org/10.1111/ijsa.12051

Neumann, M., Hengeveld, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021a). Education increases decision-rule use: An investigation of education and incentives to improve decision making. *Journal of Experimental Psychology: Applied*. Advance online publication. https://doi.org/10.1037/xap0000372

Neumann, M., Niessen, A. S. M., & Meijer, R. R. (2021b). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review, 11*(3), 205–239. https://doi.org/10.1177/2041386620983419

Neumann, M., Niessen, A. S. M., Tendeiro, J. N., & Meijer, R. R. (2021c). The autonomy-validity dilemma in mechanical prediction procedures: The quest for a compromise. *Journal of Behavioral Decision Making*. Advance online publication. https://doi.org/10.1002/bdm.2270

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance, 27*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Nolan, K. P., Dalal, D. K., & Carter, N. (2020). Threat of technological unemployment, use intentions, and the promotion of structured interviews in personnel selection. *Personnel Assessment and Decisions, 6*(2), 38–53. https://doi.org/10.25035/pad.2020.02.006

*Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing*. Pearson.

Roulin, N., & Bangerter, A. (2012). Understanding the academic-practitioner gap for structured interviews: "Behavioral" interviews diffuse, "structured" interviews do not. *International Journal of Selection and Assessment, 20*(2), 149–158. https://doi.org/10.1111/j.1468-2389.2012.00588.x

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40*(3), 455–488. https://doi.org/10.1111/j.1744-6570.1987.tb00610.x

Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology, 102*(10), 1421–1434. https://doi.org/10.1037/apl0000235

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*(5), 593–602. https://doi.org/10.1086/219248

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schwab, A. P. (2008). Putting cognitive psychology to work: Improving decision-making in the medical encounter. *Social Science & Medicine, 67*(11), 1861–1869. https://doi.org/10.1016/j.socscimed.2008.09.005

Taylor, H. C., & Russel, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*(5), 565–578. https://doi.org/10.1037/h0057079

Terpstra, D. E., & Rozell, E. J. (1997). Why some potentially effective staffing practices are seldom used. *Public Personnel Management, 26*(4), 483–495. https://doi.org/10.1177/009102609702600405

Van der Linden, W. J. (1991). Applications of decision theory to test-based decision making. In R. K. Hambleton et al. (Eds.), *Advances of educational and psychological testing: Theory and applications*. Springer Science + Business Media.

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40*(5), 525–531. https://doi.org/10.1037/a0014693

Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*(1), 100–106. https://doi.org/10.1037/h0031147

Yu, M. C., & Kuncel, N. R. (2020). Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions, 6*(2), 1–10. https://doi.org/https://scholarworks.bgsu.edu/pad/vol6/iss2/2