# 1

# Introduction

## *The Importance of Multiple Rater Data*

### CHAPTER OVERVIEW

"How are you?" is one of the most often asked questions in the world. An answer to this question requires a self-report. In psychological research, a response to this question is often obtained in a more structured way by presenting adjectives like *happy* or *sad* and providing a numeric response scale whose end points are often anchored with words like *not at all* and *very much*. If, for example, a psychotherapist asks the question, she is usually convinced that an answer to this question provides important information that cannot be replaced by any other information. What a person thinks about her or his inner feelings is only accessible to the self and can only be communicated by the self. However, the psychotherapist might have another view of the client's state given her clinical experience. Her rating from an observer's perspective might differ. From a clinical perspective, both ratings are important and present different views on the same phenomenon. Which answer is the correct one? This question cannot be appropriately answered without understanding why the two raters differ (if ever). Even if different psychotherapists are asked to rate the state of the same client, it is likely that their ratings will differ. Are clinical diagnoses then reliable and valid? These are important questions, not only for psychological research but also for applied psychology. A psychotherapist might want to get a clearer answer to this question and might measure the blood pressure, assesses the hormonal status, analyze the mobility behavior using GPS data of the smartphone, and so on. However, all these measures would add new interesting information and might help to better understand the client, but they are not able to replace the self-report of the client and the informant report of the psychotherapists.

This does not mean that self- and other reports are the gold standard method for all research questions. Psychology as an empirical science has developed a variety of assessment methods that are continuously refined and supplemented (e.g., Eid & Diener, 2006). Modern technologies such as mobile sensing, for example, offer new ways of assessing behavior, attitudes, and feelings that would not have been conceivable 50 years ago (Mehl, Eid, Wrzus, Harari, & Ebner-Priemer, 2024). However, these methods have not resulted in self- or other reports becoming obsolete. Instead, it is often recommended to additionally assess self- and/or other reports to better understand and to contextualize these more objective measurements. For instance, one suggestion is to combine mobile sensing with ambulatory self-report assessments (Ebner-Priemer & Santangelo, 2024).

It is a well-known fact that different raters usually do not perfectly agree when they assess the same characteristic (e.g., Letzring & Spain, 2021). The causes and the consequences of this phenomenon have occupied psychometricians and applied researchers over the history of scientific psychology. Starting in the early years of the last century, researchers were interested in understanding the differences between the raters and how these differences can be reduced by selecting a sufficient number of appropriate raters and by optimizing the rating process (e.g., Rugg, 1921; Webb, 1915). Furthermore, researchers were also interested in why ratings are correlated across distinct traits (Thorndike, 1920). This has led to the development and refinement of different error concepts and causes of rater bias over the last century (e.g., Cronbach, 1955, 1970; Guilford, 1954; Letzring & Spain, 2021; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Saal, Downey, & Lahey, 1980; Thorndike, 1920; Webb, 1915).

Moreover, researchers have been interested in using statistical methods for analyzing the reliability and validity of ratings. Webb (1915), for example, recommended obtaining at least two ratings. He emphasized that “the extent to which the two estimates of each quality agree with each other represents the degree of ‘reliability’ that can be placed upon the particular pair of judgments” (p. 25). He estimated the reliability by correlating the ratings of two different raters. Furthermore, he suggested specifying a lower bound of the reliability that would have to be obtained before pooling the ratings (if their reliability passes the minimum) to get a better measure of a trait that then can be related to other measures. Although Webb’s recommendation is still applied today, psychometric models for analyzing the reliability and validity of multiple ratings have been developed over the last century that allow a more fine-grained analysis of these basic research questions. The present book stands in this tradition. It aims at presenting up-to-date psychometric measurement models of structural equation modeling to analyze multiple rater data. These models can be applied to analyze the reliability and validity of multiple ratings. Moreover, they allow measurement of rater-specific effects. The models are based on the fact that not all raters are interchangeable but that there can be structural differences between different rater groups. A self-report differs from a peer-report; a teacher report is not exchangeable with a parent report. However, there might be raters that *are* interchangeable, for example, multiple students rating the teaching quality of the same professor. These distinctions are introduced in the next chapter and provide organization for Chapters 3 to 10 in which specific measurement models are presented. Because these models focus on multiple raters, we will give a short overview of the importance and limitations of self- and other reports in psychology in this chapter.

## 1.1 ADVANTAGES AND LIMITATIONS OF SELF-REPORTS

The self-report method is one of the most widely applied methods of psychological assessment (e.g., Lucas & Baird, 2006). Its advantages and limitations are well-documented and discussed in detail in many handbooks and textbooks on psychological assessment. For the assessment of many constructs in psychology and other social sciences the self-report seems to be the most appropriate assessment method. Paulhus and Vazire (2007), for example, mention self-efficacy, self-esteem, well-being, personal projects, and life goals as constructs that can be best assessed by self-report. This list can be easily extended by other inner phenomena that are often exclusively accessible to the person such as feelings, emotions, intentions, explicit attitudes, etc. Self-reports are not only obtained for assessing inner phenomena but have a much broader area of application. Lucas and Baird (2006), for example, distinguish between three types of self-reports: (a) self-reports of "objectively verifiable phenomena like behaviors and events" (p. 30), (b) self-reports of "psychological constructs (e.g., beliefs, intentions, and attitudes)" (p. 30), and (c) self-reports "as a form of behavior, in and of itself" (p.30) that can be used for predictive reasons. This wide use of self-reports is due to the many advantages self-reports have. According to Lucas and Baird (2006, p. 29), self-reports are "simple, quick, inexpensive, flexible, and often provide information that would be difficult or impossible to obtain in other ways." In addition, Paulhus and Vazire (2007) highlight that self-reports can often be easily interpreted as they are provided in everyday language. They emphasize that nobody else has more information (in quantity and breadth) about a person and her or his intrapsychic state than the self. They also stress that people are often highly motivated to provide detailed information about their personality, and that these self-perceptions are—regardless of whether they are correct or not—a causal force as they have a strong influence on how people behave in the world.

Yet, self-reports also have significant limitations and disadvantages that have been well researched and documented. The most prominent limitations are response sets and styles such as self-presentation (e.g., impression management, self-deception) or the selection of specific response categories (e.g., acquiescence, extreme responding, choice of the middle category) (e.g., Bandalos, 2018; Lucas & Baird, 2006; Paulhus & Vazire, 2007). In particular, in social cognition and survey research many influences have been detected that can be a threat for validity throughout the whole response process (item interpretation, response generation, response reporting, response editing; Sudman, Bradburn & Schwarz, 1996); these are discussed in detail by Bandalos (2018), Lucas and Baird (2006), Schwarz and Sudman (1996), and Torangeau, Rips, and Rasinksi (2000). Finally, the self-knowledge of an individual and her or his introspection might be limited, for example, by biases, blind spots, and self-delusion (Paulhus & Vazire, 2007; Spain, 2021; Vazire & Carlson, 2011). Motivational factors such as selfenhancement, self-protection, and self-verification might also play an important role (e.g., Bollich-Ziegler, 2021). Given that self-reports are influenced by so many factors, the accuracy of self-reports has been questioned. *Accuracy* is a central concept of rater studies. According to Funder (1999, p. 3) accuracy "refers to the relation between what is perceived and what is." Referring to Funder (1999), Osterholz, Breil, Nestler, and Back (2021, p. 46) define *accuracy* as the "level of correspondence between observers' personality judgments and targets "true" personality." Although the term accuracy is more often used in studies on informant assessments, it also can be applied to self-reports (e.g., Bollich-Ziegler, 2021). In Section 1.3 we will discuss the accuracy concept and its relations to other concepts in more detail. However, many ways to increase the accuracy of self-reports have been developed and resulted in optimizing self-report data (e.g., Bandalos, 2018; Bollich-Ziegler, 2021; Lucas & Baird, 2006). Moreover, given all the results of the many studies on the validity of self-reports, it appears that self-reports have "considerable validity" (Spain, 2021, p. 169), and that they are an indispensable data source in psychology. However, given their limitations it might be worthwhile to supplement self-reports by other assessment methods to analyze their validity and to overcome limitations unique to self-reports.

## 1.2 ADVANTAGES AND LIMITATIONS OF OTHER REPORTS

Beside self-reports, other reports have also become a standard assessment method in psychology (e.g., Neyer, 2006). In personality psychology, they are the major assessment method for personality judgments (Letzring & Spain, 2021). In personnel psychology, leadership behavior is typically assessed by supervisor and subordinate reports (e.g., Blackman, 2021; Harris & Schaubroeck, 1988; Lee & Carpenter, 2018; Mahlke et al., 2016). In educational psychology, peer and teacher reports are widely used (e.g., Li et al., 2016). In developmental psychology, parent reports and observer reports are important sources of information to assess the behavior and temperament of children (e.g., De Los Reyes et al., 2015; Duhig, Renk, Epstein, & Phares, 2000). In clinical psychology, multiple informants are used for the assessment of psychopathology (e.g., Achenbach, Krukowski, Dumenci, & Ivanova, 2005). The list could be easily continued.

Informant reports are obtained for several reasons. First, informant ratings are often obtained when the targets are not able to provide the information by themselves such as the quality-of-life assessment of elderly people suffering from dementia (Robertson et al., 2017). Second, overt behavior is observed and rated for scientific or applied assessment reasons such as the assessment of preschool childrens' attachment (e.g., Deneault, Bureau, Yurkowski, & Moss, 2020). Third, the perception of a target and how characteristics of others are perceived by strangers or informed raters are of scientific interest (e.g., impression formation, interpersonal perception; e.g., Ambady & Skowronski, 2008; Kenny, 2020; Zebrowitz, 1990). Fourth, the informant's attitude toward, and relations with, a target are of interest such as in romantic relationships (e.g., Luo &

Watson, 2021). Fifth, the informants are the experts whose evaluations of the targets are the basis for a decision such as assessment center ratings (e.g., Woehr & Arthur, 2003) or performance ratings (e.g., Johnson, Penny, & Gordon, 2008; Lane & Iwatani, 2016). Sixth, informants might provide more valid information about the target than the targets themselves due to blind spots and lack of self-knowledge (e.g., Vazire, 2010; Vazire & Carlson, 2011; Vazire & Mehl, 2008). Moreover, informant reports have the advantage that they can be easily collected (e.g., via the internet), they are cheap to collect, and they can be aggregated to reduce the impact of rater-specific effects (Vazire, 2006). Therefore, other reports are a unique data source in the social and behavioral sciences that cannot be replaced by other methods.

Like every assessment method, informant reports have their own limitations. These limitations are partly linked to those of self-reports. Response styles such as socially desirable and dishonest responding also play a role for other reports (e.g., Vazire, 2006). Many response errors have been documented such as the error of leniency, the error of central tendency, the halo effect, the logical error, the contrast error, and the proximity error (Guilford, 1954; Saal et al., 1980). Guilford, for example, defines the *error of leniency* as "a general, constant tendency for a rater to rate too high or too low for whatever reasons" (p. 278), and the *error of central tendency* as a tendency "to displace individuals in the direction of the mean of the total group" (p. 279). The term *halo effect* was introduced into science by Thorndike (1920). According to Saal et al. (1980), the halo effect is "consistently conceptualized as a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behavior" (p. 415). The *logical error*, going back to Newcomb (1931), is "due to the fact that judges are likely to give similar ratings for traits that seem logically related in the minds of the raters" (Guilford, 1954, p. 279). The *contrast error* (introduced into psychology by Murray, 1938) indicates "a tendency for a rater to rate others in the opposite direction from himself in a trait (Guilford, 1954, pp. 279–280). According to the *proximity error*, detected by Stockford and Bissell (1949), "adjacents traits on a rating form tend to intercorrelate higher than remote ones, their degrees of actual similarity being presumably equal" (Guilford, 1954, p. 280). Furthermore, other ratings can be influenced by positivity and negativity biases and current mood (Podsakoff et al., 2003).

Because other raters also have to provide a response on a rating scale the threats for validity that are present at different steps of the response process and have been mentioned for self-reports in the last section also play a role for other reports. Moreover, informants might not have sufficient information about those constructs for which self-reports are particularly important and valid (e.g., inner feelings). For example, the *self–other knowledge asymmetry (SOKA) model* (Vazire, 2010) postulates that the self-report method is superior to other reports when traits show a low observability and a low evaluativeness (desirability), whereas other reports (in particular given by close others) shall be more appropriate if traits are high on evaluativeness. However, not all predictions of the model can be supported by empirical results (for an overview, see Bollich-Ziegler, 2021).

### 1.2.1 Models of accuracy of other ratings

These limitations of other reports question the accuracy of other reports. The accuracy of other reports is an important topic of personality judgment and person perception research (for an overview, see Funder, 1999; Kenny, 2020; Letzring & Spain, 2021). In many studies, the conditions of accuracy of interpersonal perceptions have been analyzed and several theoretical models have been developed to integrate the major findings into a general theoretical framework. According to Letzring and Spain (2021), the major theoretical models are the *lens model* (Brunswik, 1956), the *realistic accuracy model* (Funder, 1995), the *social relation model* (Kenny & Albright, 1987), and the *social accuracy model* (Biesanz, 2010).

According to the *lens model* (Brunswik, 1956; Osterholz et al., 2021), accuracy depends on cues that are emitted by the target and observed by the other rater. In order to ensure accuracy, relevant cues are needed for the trait being judged (*cue validities*), and the relevant cues have to be used by the other raters for the formation of their judgment (*cue utilization*). A sensitive rater uses the valid cues and ignores the invalid cues in judging a trait. Inaccurate judgments can be the result of missing relevant emitted cues and/or inappropriate use of emitted cues in judgment formation.

The *realistic accuracy model* (Funder, 1995; Letzring & Funder, 2021) divides the cue validity part of the lens model in two stages: relevance and availability. *Relevance* means that the cues must be relevant for the trait being judged. This stage is mainly under the control of the target (Letzring & Funder, 2021). *Availability* is given when the cues are available (e.g., visible) in the relevant context. This stage primarily depends on the interaction between the observer and the target (Letzring & Funder, 2021). Moreover, the cue utilization part of the lens model is split into two stages: detection and utilization. *Detection* means that the relevant and available cues must be noticed by the rater and is, therefore, primarily affected by the rater (Letzring & Funder, 2021). *Utilization* is also primarily due to rater abilities and comprises the adequate distinction between relevant and irrelevant cues and the appropriate use of the relevant ones (Letzring & Funder, 2021). Inaccurate judgments can be due to failures at each stage.

The *social relation model* (Kenny, 2020; Kenny & Albright, 1987; Malloy, 2021) is a componential model that is strongly influenced by Cronbach's (1955) distinction of different types of accuracy. It is a social relation model because it assumes that a rating is based on a social interaction between two individuals of a dyad (a target and a rater), and that each part of a dyad can be target and rater. In Kenny and Albright's approach, a rater's judgment of a target is decomposed into four components: (a) a *constant* that represents the average level of a construct as rated by a population of raters across all targets of a population, (b) the *actor effect* of a rater which characterizes the tendency of a rater to rate targets with respect to the construct in a certain way, (c) the *partner effect* which indicates the way in which a specific target (partner to be rated) is generally rated by other raters, and (d) the *relationship effect* which shows the way in which a specific target-rater pair deviates from the expected rating given the other effects. Malloy (2021) adds *random error* (residual) as fifth component, which provides a formal decomposition of an observed rating score that resembles the decomposition of a score in a two-way analysis of variance model. Accuracy can be considered with respect to each of the effects (a) to (d) if criterion variables for these effects representing the "true" values of a construct are available (e.g., a measure assessing a specific behavior in dyads). According to Kenny and Albright (1987) the following accuracy types can be defined: *Elevation accuracy* is given when the average rated level of construct (across all possible targets and raters) equals the average level measured by the criterion variable. *Response set accuracy* requires that the actor effect equals the construct value of the target measured by the criterion variable. *Individual accuracy* necessitates that the average rated construct value of a target (across different raters) equals the construct value of the target measured by the criterion variable (e.g., the average level of the observed behavior across dyads). *Dyadic accuracy* will be present if the rater correctly rates the construct value of the target with respect to the specific dyad (e.g., that the target behaves to the specific rater differently than to the average of other raters). Inaccuracy can occur with respect to all four components.

The *social accuracy model* (Biesanz, 2010, 2021) is also a componential model that allows estimating rater, target, and dyadic effects, and distinguishes different types of accuracy. The model requires different raters rating different traits of a target and validity measures for the construct of the targets to be rated. Depending on the design considered (see Chapter 2), an observed rating variable (representing a rating of a target on a specific item by a specific rater) is decomposed into different components by regressing the observed rating variable on two independent variables in a random coefficient regression analysis: (a) an average validity measure for each observed variable (item) that is calculated as the average of the validity measure across all targets, and (b) a centered target-specific validity measure of an item that is calculated as the difference between the item-specific validity measure of a target and the average validity measure across all targets. The distinction between these two validity measures allows the assessment of two different types of accuracy, *normative* and *distinctive* accuracy (Biesanz, 2021). *Normative accuracy* is given if the average person's profile of trait (or item) scores obtained by the validity measure equals those obtained by the perceived score (averaged across all targets). *Distinctive accuracy* is present if the deviations of a target from the average target (with respect to the different traits or items considered) obtained by the validity measures equal those obtained by the perceived scores. The random intercept and the random regression slopes are decomposed into target-, rater- and dyad-specific random effects. The variances of these random effects reflect the degree to which targets, raters, and dyads differ, for example, in the way the observed ratings are determined by the average validity measure (knowledge about the "average person" with respect to the construct considered) and the target validity measure (deviation of the target from the "average person"). The random effects can be explained in extended models by including possible moderator variables. Biesanz (2021) discusses this model in more detail and provides an empirical illustration.

### 1.2.2 Sources of accuracy of other ratings

In general, *raters* can differ strongly in characteristics that are important for obtaining accurate responses ("good judge," Funder, 1995; for an overview, see Colman, 2021), and "bad judges" might provide invalid ratings. For example, according to Neyer (2006, p. 53), a good judge "needs a strong sensitivity to what is happening in his or her social environment," "can make a connection between the observed behaviors and the personality traits underlying them," and "needs to be objective, rational, and unconcerned with the opinions of others when making judgments." Furthermore, characteristics of the *target* can have an influence on accuracy. In an overview of studies on the "good target," Mignault and Human (2021) concluded that targets who show high scores on characteristics such as psychological well-being, personality coherence, self-knowledge, power, physical attractiveness, extraversion, emotional expressiveness, social skills, and who are liked, seem to be more accurately judgeable. Moreover, accuracy can depend on characteristics of the *relationship between the raters and the targets*. Raters who know the targets better appear to show higher distinctive and normative accuracy, whereas raters who like the targets better show lower distinctive and normative accuracy (Wessels, Zimmermann, Biesanz, & Leising, 2020). With respect to *trait* characteristics as a moderator of accuracy, it is, for example, well known that traits that have a higher visibility (e.g., extraversion) are better judgeable than traits showing lower visibility (e.g., neuroticism) (Funder & Dobroth, 1987). There are other moderators of accuracy and interactions between moderators of accuracy that are discussed elsewhere in the literature (e.g., Letzring & Funder, 2021; Letzring & Spain, 2021; and Podsakoff et al., 2003). Research on the limitations of other reports has led to many ways to improve the quality of other reports (Letzring & Spain, 2021). Like self-reports, other reports are indispensable in psychology and other empirical sciences.

## 1.3 USEFULNESS OF MULTIPLE RATER STUDIES

Because both self- and other ratings have specific limitations it is worthwhile to integrate multiple raters in the research and/or assessment process. There are two major advantages of multiple rater studies. First, they allow us to analyze the validity of ratings. Second, they can increase the validity of conclusions drawn from ratings. Because multiple raters can be considered as multiple methods (Kenny, 1995), validity issues of multiple rater studies can be discussed in the framework of the multitrait–multimethod (MTMM) analysis (Campbell & Fiske, 1959).

### 1.3.1 Analyzing the validity of ratings

According to Messick (1989, p. 13) "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment.” If, for example, conclusions are drawn from the self-report of an individual about her or his extraversion, the question arises whether these conclusions are adequate and appropriate. If the self-report converges with an informant report of the extraversion of this individual, there is convergent evidence that the inferences might be adequate and appropriate. If the self-reports of extraversion do not correlate with self-reports of intelligence (because extraversion and intelligence are rather unrelated; Wolf & Ackerman, 2005), then there is discriminant evidence that the inferences might be appropriate (because there are, for example, no halo effects). Analysis of convergent and discriminant correlations as a way to examine the appropriateness of conclusions has a long tradition in psychology and is the core of Campbell’s and Fiske’s (1959) validation approach for multimethod studies that builds the theoretical validity framework of the present book.

### *1.3.1.1 Convergent and discriminant validity*

According to Campbell and Fiske (1959, p. 81) the validation process is characterized by the following four aspects:

> 1. Validation is typically convergent, a confirmation by independent measurement procedures. (…)
> 2. For the justification of novel trait measures (…) discriminant validation as well as convergent validation is required. (…)
> 3. Each test or task employed for measurement purposes is a trait-method unit. (…)
> 4. In order to examine discriminant validity, and in order to estimate the relative contributions of trait and method variance, more than one trait as well as more than one method must be employed in the validation process.

With respect to multiple rater studies, this approach requires that at least two raters and at least two different traits be considered. This is in contrast to many multiple rater studies of personality judgments in which a single trait has been the focus of accuracy research (Letzring & Funder, 2021). Integrating multiple traits in the validation process, however, has several advantages. It allows us to consider discriminant validity and, therefore, enriches the database for judging the validity of inferences. Furthermore, it permits analyzing to which degree rater-specific influences (e.g., rater biases) are specific to a trait or generalize across traits (Eid, 2006). Because informants usually judge different traits at the same time in real life (Letzring & Funder, 2021), analyzing different traits can contribute to a better understanding of the judgment process. Moreover, it allows analyzing *profile accuracy,* which refers to the ordering of a target’s values on different traits (Osterholz et al., 2021).

*Convergent validity* is given when the ratings of the same trait provided by multiple raters are strongly correlated. *Discriminant validity* requires that the ratings of different traits that are conceptually independent are uncorrelated. The concept of convergent validity is related to two other concepts of multiple rater studies—consensus and accuracy.

#### *1.3.1.2 Consensus and accuracy*

According to Shrout (1995, p. 81) the term *consensus* is "usually interpreted to mean agreement between raters" but he mentions that is also used synonymously with terms such as "*congruence*, *concordance*, *consistency*, *correlation*, and/or *reliability*" (p. 81). He emphasizes that the terms *agreement* and *congruence* are reserved in psychometrics to indicate "absolute interchangeability" (p. 81), which signifies *perfect* agreement (Choudhary & Nagaraja, 2017). That means that two raters agree when their ratings are identical. Perfect agreement differs from concepts such as consistency, concordance, and correlation which indicate *relative* agreement, for example, rank order agreement (Fiske, 1978; Shrout, 1995). Moreover, Shrout (1995) emphasizes that it is unreasonable to use the term *reliability* interchangeably with consensus, as reliability refers to the reproducibility of single rater measurements. Whereas unreliability is due to unsystematic measurement error influences, disagreement is due to systematic rater-specific effects (Shrout, 1995). *Convergent validity* assessed via the correlation coefficient is, therefore, related to consistency and relative consensus and is less strict than consensus and agreement in its narrow sense of perfect agreement.

*Accuracy* is a much more loaded word (Funder, 1999). We have already pointed out that accuracy means "the extent to which a perception matches the truth" (Kenny, 2020, p. 15). The problem that is linked to the concept of accuracy in psychology is that the "truth" is typically not known and not measurable by a gold standard measurement. Moreover, there are also philosophical positions that question the "real" existence of personality traits. We will not go into these philosophical details that are more broadly discussed elsewhere (e.g., Funder, 1999; Kruglanski, 1989; Letzring & Funder, 2021; Malloy, 2021; Slaney, 2017) and that are not relevant for understanding and interpreting the measurement models presented in this book. Given the problems that are related to the identification of the "true" personality, the analysis of accuracy follows a more pragmatic approach. Accuracy is typically analyzed by comparing a rating with a criterion measure (Malloy, 2021). Criterion measures can, for example, be self-reports (self–other agreement), expert ratings, and behavioral observations (Neyer, 2006). However, the interpretation of the association between a rating and a criterion measure as accuracy depends on the validity of the criterion measure itself which often cannot be proved.

*Consensus* and *accuracy* refer to different facets of convergence. Different raters can show perfect consensus/agreement but there might be no accuracy if the ratings are not valid (Shrout, 1995). Moreover, raters might differ in their accuracy, resulting in partial accuracy but no consensus (Shrout, 1995). However, if all raters are accurate, there has to be consensus (Shrout, 1995). Because there are typically no gold standard measures of a trait to which a rating can be compared, the distinction between consensus and accuracy is often unclear in validity studies using multiple raters. Therefore, the interpretation of the associations of a rating with other ratings or criterion measures as a proof of validity often depends on plausibility assumptions. Nevertheless, analyzing the consensus (and accuracy) is an important methodological tool. If there are no consensus and no criterion-related associations at all, this strongly questions the validity of inferences based on rating data.

### 1.3.2 Improving the validity of inferences by multiple ratings

Multiple ratings can be used to improve the validity of inferences, at least in two different ways. First, the different ratings can be aggregated. This is usually done to reduce the influence of rater-specific effects and is a rather old strategy (Rugg, 1921; Webb, 1915). If rater-specific effects are mainly due to response styles, aggregation is a way of eliminating (averaging out) the influence of response styles. The aggregated values are then less affected by response styles and often show stronger associations with other criterion variables than the single ratings (e.g., Back & Nestler, 2016). Second, if raters mainly differ because they have different access to life domains, these different views might complement each other in predicting criteria and enhance hereby criterion-related validity. In this case, the different ratings can, for example, be entered as independent variables in a regression analysis to predict a criterion variable. In such an analysis, the different ratings would be differentially weighted to predict a criterion variable in an optimal way. This would be different from the aggregation approach, because in the aggregation approach each rating would get the same weight.

## 1.4 THE ROLE OF MEASUREMENT MODELS

Because of the high importance of rating data for different areas of psychological research, the analysis of the quality of rating data is essential for ensuring high-quality research and appropriate assessment. As in other areas of psychological research and assessment, quality criteria such as reliability and validity have to be fulfilled. These quality criteria can be analyzed in psychometric models that allow (a) separating unsystematic measurement error from systematic rater-specific effects, (b) analyzing the generalizability of rater-specific effects across different traits, (c) measuring trait differences that are free of measurement error and—if possible—rater-specific effects, and (d) integrating covariates and criterion variables for explaining rater-specific effects and examining criterion validity. In the present book, we present such models for different types of raters and different types of measurement designs.

## 1.5 CHAPTER SUMMARY

Self- and other reports belong to the most widely used research and assessment methods in the behavioral and social sciences that cannot be replaced by any other method. They allow unique insights into behavior, feelings, thoughts, attitudes, personality, and many other constructs that often cannot be obtained by other methods. Moreover, they are important for understanding self-construal, social perception, and social functioning in general. Self- and other reports have many advantages but also serious limitations. One way to overcome some limitations of single rater data is the consideration of multiple raters that represent different rater groups and have access to different sources of information about a target. The validity (accuracy) of multiple rater data can be investigated in empirical studies that draw on models of accuracy that have been primarily developed in social perception research. Multiple rater studies can enhance the validity of conclusions, for example, via the aggregation approach or by integrating the different views in a regression analysis. Analyzing the quality of multiple rater data requires special psychometric models that take the specific nature of multiple rater data into account.

## 1.6 SUGGESTED FURTHER READINGS

Readers interested in historical aspects of rater studies in psychology are referred to the texts of Webb (1915), Thorndike, (1920), Guilford (1954), and Cronbach (1955, 1970). Lucas and Baird (2006) as well as Paulhus and Vazire (2007) discuss the most relevant advantages and limitations of self-reports. Letzring and Spain (2021) give a comprehensive overview of advantages and limitations of other reports as well as the most important theoretical approaches and the results of empirical studies related to personality judgment based on self- and other reports. The fundamentals of social perception and person perception theories as well as social relation models are extensively described by Funder (1999), Kenny (2020), and Malloy (2018). Interpersonal accuracy with respect to many different constructs of psychological research are presented in the edited book of Hall, Schmid Mast, and West (2016). For a deeper discussion of the validity concept, see Kane (2013).

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*(3), 177–193.

Cronbach, L. J. (1970*). Essentials of psychological testing*. New York: Harper & Row.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego: Academic Press.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hall, J. A., Schmid Mast, M., & West, T. V. (Eds.). (2016). *The social psychology of perceiving others accurately.* Cambridge: Cambridge University Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kenny, D. A. (2020). *Interpersonal perception: The foundation of social relationships* (2nd ed.). New York: Guilford Press.

Letzring, T. D., & Spain, J. S. (2021). *The Oxford handbook of accurate personality judgment*. Oxford: Oxford University Press.

Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington: American Psychological Association.

Malloy, T. E. (2018). *Social relations modeling of behavior in dyads and groups*. Amsterdam: Elsevier Press.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York: Guilford Press.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Webb, E. (1915). Character and intelligence: An attempt at an exact study of character. *British Journal of Psychology, Monograph Supplements, 1*(3), 1–99.

Guilford Publications
info@guilford.com
www.guilford.com