Adv in Health Sci Educ (2014) 19:409–427
DOI 10.1007/s10459-013-9453-9

REVIEW

Exploring the role of first impressions in rater-based assessments

Timothy J. Wood

Received: 16 January 2013 / Accepted: 12 March 2013 / Published online: 26 March 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Medical education relies heavily on assessment formats that require raters to assess the competence and skills of learners. Unfortunately, there are often inconsistencies and variability in the scores raters assign. To ensure the scores from these assessment tools have validity, it is important to understand the underlying cognitive processes that raters use when judging the abilities of their learners. The goal of this paper, therefore, is to contribute to a better understanding of the cognitive processes used by raters. Representative findings from the social judgment and decision making, cognitive psychology, and educational measurement literature will be used to enlighten the underpinnings of these rater-based assessments. Of particular interest is the impact judgments referred to as first impressions (or thin slices) have on rater-based assessments. These are judgments about people made very quickly and based on very little information. A narrative review will provide a synthesis of research in these three literatures (social judgment and decision making, educational psychology, and cognitive psychology) and will focus on the underlying cognitive processes, the accuracy and the impact of first impressions on rater-based assessments. The application of these findings to the types of rater-based assessments used in medical education will then be reviewed. Gaps in understanding will be identified and suggested directions for future research studies will be discussed.

Keywords First impressions · Rater-based assessment · Rater-cognition

Imagine for a moment you are a clinical supervisor and are meeting the incoming trainees for the first time at an informal “meet the resident” night. You meet one resident and after talking for a few minutes about where they received their degree and their clinical interests, you excuse yourself. As you walk away, you surmise that this trainee is going to be high maintenance and you are not looking forward to working with them. You then meet a second trainee, and after exchanging some pleasantries, you walk away thinking

T. J. Wood (✉)
Academy for Innovation in Medical Education (AIME), RGN2206, Faculty of Medicine,
University of Ottawa, Ottawa, ON K1H-8M5, Canada
e-mail: twood@uottawa.ca

"superstar". In both cases you made a judgment about the trainees' behavior based on small talk; a judgment that will shape your relationship with the person. You may also be highly confident in that judgment, but in neither case fully aware of what informed it. This quick judgment we make about other people, often based on very little information is something we all do in any social situation when we first meet people. What is unclear is what role these first impressions might play if we are assessing the competencies or skills of a learner.

Medicine has a long history of assessing the competence of learners by relying on the judgments of teacher and/or experts. This use of these people as raters is likely a reflection of two factors. First, the skills that make a good physician (e.g., problem solving, communication, clinical judgment, professionalism) do not necessarily lend themselves easily to non-rater based assessments methods like written examinations. The second factor relates to how physicians are trained. Learners are observed in clinical settings as part of their training, and this observation lends itself naturally to assessment tools based on those observations. More recently, there has been an increased push to adopt a competency-based framework to assess the skills of learners (Holmboe et al. 2010). This assessment framework emphasizes the use of feedback as well as workplace assessments, both of which require observation thus further highlighting the critical role of the rater.

Unfortunately, all humans have preconceived notions, biases and abilities that influence the quality of the judgments they make when assessing the competence of learners (Gigerenzer and Gaissmaier 2011; Hoyt 2000; Landy and Farr 1980; Saal et al. 1980, 1974; Williams et al. 2003). To ensure the assessments that are used in medical education have value (i.e., are valid and reliable), it is crucial that we understand the cognitive processes behind how people assign scores when assessing others. In fact, the collection of this type of information is considered part of the chain of validity evidence one should collect with regards to any assessment (AERA et al. 1999; Clauser et al. 2008; Cook and Beckman 2006; Downing and Haladyna 2009). The goal of this paper is to gain a better understanding of the cognitive processes used by raters when assessing the clinical skills of learners. Representative findings from the social judgment and decision making literature as well as the educational measurement literature will be used to enlighten the underpinnings of these rater-based assessments. Of particular interest is the impact of judgments often referred to as "first impression", "thin slice" or "zero acquaintance" judgments. Although these terms refer to slightly different research paradigms, they will be treated synonymously in this paper and will refer to any judgment about a person that is made quickly, is based on little information, and involves judgments of people a rater has never met (i.e. strangers).

First impressions

What do you think of the first time you meet someone? Is this person outgoing or shy, friendly or grumpy, organized or sloppy, competent or incompetent? These initial judgments about the person occur rapidly and without much conscious reflection as to how they were derived. They have important implications, however, because they guide how we initially interact with the person, what information we remember about the person and our predictions about future behavior. These judgments about others are called impressions, which are categories that we use to help us perceive, organize and integrate information about an individual's personality and behavior (Feldman 1981; Fiske and Neuberg 1990; Gingerich et al. 2011). First impressions are a type of impression that is made quickly, usually within 5 min of meeting someone for the first time. First impressions have been found to be surprisingly accurate given how quickly they form and the limited information on which they are based (Ambady and Rosenthal 1992; Ambady 2010; Harris and Garris 2008). They are potentially critical as they influence how we process information about other people, which may be highly relevant for assessment. Given the presumed importance of first impressions for social judgments and assessment, addressing the following questions should promote a better understanding of the subject (1) what are the underlying cognitive processes behind a first impression? (2) How accurate are they? (3) What factors influence this accuracy? (4) What are the implications of first impressions for assessment?

What are the cognitive processes behind a first impression?

Many cognitive activities including decision making, reasoning, categorization, and memory are thought to consist of two underlying processes; what have come to be known as System 1 and System 2 processes (Evans 2008; Uleman et al. 2008; Kahneman 2011). It is generally accepted that System 1 processes are rapid, effortless, non-analytic, automatic, and/or unconscious, whereas System 2 processes are slow, effortful, analytic, controlled, and/or conscious. To illustrate the difference between these two processes, read the following two words out loud: cat, dog. Presumably the correct pronunciation popped into your head without you having to devote much attention or conscious thought to pronouncing the words. This fast effortless method of reading illustrates what one would expect with a System 1 process. Now read the following two words out loud: parasitological, incudostapedial. Presumably your pronunciation was effortful and required you to explicitly sound out the syllables of the words, in other words what one would expect with a System 2 process. This example was designed to illustrate how the two processes manifest themselves in that they feel somewhat different. Both System 1 and System 2 processes can be used to perform many of the cognitive activities listed above; therefore the focus of research in cognitive psychology and social judgment and decision making is to try to understand how we coordinate these two processes (see Brooks 2005; DeNisi et al. 1984; Fiske and Neuberg 1990; Jacoby 1991; Kahneman 2011; Norman 2009; Schneider and Chein 2003 for examples in these areas).

First impressions are thought to reflect primarily System 1 processes because they are made quickly, effortlessly, and appear to require very few cognitive resources like memory or attention (Ambady 2010; Ambady et al. 2000). If this assumption about first impressions is true, then specific patterns of results should occur in first impression studies when manipulations designed to help or interfere with System 1 processes are used (Bargh 1992; Jacoby 1991; Schneider and Chein 2003). For example, for tasks that rely on unconscious processes, (processes that operate without an individual needing to devote large amounts of cognitive or attentional resources), it should be difficult for people to accurately verbalize how they performed a task or even be aware of what information they actually used. One just has to think of how hard it is to explicitly verbalize to a child the steps needed to ride a bicycle to realize this occurs.

There is an extensive history in cognitive pyschology of studying the level of awareness associated with System 1 processes (Hasher and Zacks 1979; Jacoby and Kelley 1990; Klein 2009; Logan 1992), but there have only been a few studies that have looked at the level of awareness associated with first impressions. Of the work that has been done, the focus has been on the relationship between accuracy and confidence. For example, (Smith et al. 1991) asked participants to view short (60–90 s) video clips and then answer questions related to interpersonal relations between people in the videos and provide confidence ratings of the accuracy of their answers. They found a low but positive relation between accuracy and confidence levels. They also found evidence that participants knew when they were guessing and when they were answering based on some kind of process, they just couldn't describe how they were making the judgment. Similarly, (Ames et al. 2010), asked participants to view either photographs or short 60 s videos of people and then make personality judgments, rate their confidence in the judgment and answer questions about their judgments. These judgments were compared to peer ratings of the people being rated. Ames et al. found low to moderate levels of accuracy for some personality dimensions and a low relationship between the accuracy of the judgment and the confidence level of the raters. Interestingly, the correlation between accuracy and confidence was highest for those raters with no confidence in their rating because they knew when they had no idea about a personality judgment. The conclusion from both of these studies is that, although people may be aware of the outcome of forming a judgment of others, they appear to have difficulty articulating how they did it and/or have little insight into how they actually made that judgment.

Although not explicitly a study of first impressions, Wigton (1980) reported relevant results related to awareness. Five first year medical students were trained to each present five different case presentations that varied in quality and in length (3–6 min). Videos of each presentation were made and physician raters ranked the quality of the case presentations. For each rater, however, the case presentation was held constant and only the actor portraying the part was changed. Wigton found that rankings of the videos were as influenced by the particular actor as by the quality of the case presentation suggesting that raters were influenced by more than just the case presentation skills of the student. Interestingly, Wigton also reported that most of the judges were highly confident in their ratings of the students despite low correlations between raters on the same video and content being controlled across actors, but could not articulate what the student variables were that led to their ratings; a result similar to that found by Smith et al. (1991) and Ames et al. (2010).

Recently, Biesanz et al. (2011) questioned the finding that a low relationship between confidence and accuracy in first impression judgments reflects a lack of awareness. They argued that people are aware at the time when individual first impression judgments are accurate even if they do not know how accurate their judgments are globally. A similar argument was made by Eva and Regehr (2011) when they made the distinction between self-assessment, which is based on a global judgment of one's ability, and self-monitoring, which is based on a moment-by-moment awareness of one's ability. Clearly more work needs to occur in this area in order to understand the degree the level of awareness around first impressions.

Another pattern of results that one should expect if a first impression reflects the use of System 1 processes is that the judgments should be made quickly. A good example of the speed of an impression was demonstrated by Willis and Todorov (2006). They found that people can produce as accurate an impression of the personality traits associated with a person in a photograph after 100 ms exposure as they do when viewing the same photograph with no time constraints. Others studies have investigated the speed of a first impression by comparing the accuracy of ratings made after short time periods to longer time periods. (Ambady and Rosenthal 1992, 1993; Babad et al. 2004; Tom et al. 2009). For example, Carney et al. (2007) asked raters to view video clips of couples talking but the clips were of various lengths (5, 20, 45, 60, or 300 s). Raters were asked to rate the couples on a range of personality traits. When the ratings across all the personality dimensions were pooled, the conclusion was that although viewing a 5 min clip produced the most accurate results, the increase compared to 60 s was so minimal that 60 s actually produced the best ratio of accuracy and time. An additional study using an Objective Structured Clinical Examination (OSCE) is also relevant. Dodson et al. (2009) asked examiners on an admissions OSCE to provide a rating of the examinee's abilities at the 5 min mark and then again at the 8 min mark. Ratings at 5 min were lower than ratings at eight minutes, but the correlation between ratings at the two time points was high (r = 0.82–0.91) with no drop in reliability for scores at the 5 min mark. Although Dodson did not use ratings based on <5 min of observation, the result is consistent with other first impressions studies in that ratings after a short time period are as accurate as those after a longer time period.

Another relevant line of research used to study the speed of making a first impression, has focused on using qualitative and mixed methods analyses to study when a first impression is made. Govaerts et al (2011; see also Govaerts et al. 2013) asked experienced and inexperienced examiners to watch two videos of a trainee with a patient. The cases varied in complexity. Raters were asked to stop the tape at the point in which they thought they could judge the performance of the trainees' abilities, make a rating of the trainee and then describe their rationale for the rating. For the simple case, experienced and inexperienced raters stopped the tape at 112 and 109 s respectively. For the complex case, experts stopped the tape at 260 s whereas inexperienced raters stopped the tape at 139 s. In all conditions, therefore, the examiners thought they could judge the performance in under 5 min. In another example, Yaphe and Street (2003) asked examiners who had just completed an oral examination with trainees to review a video of the examination. During the examination, two examiners asked a trainee a series of five questions related to a patient case in an attempt to assess the trainee's communication, professionalism and professional development skills. While viewing the video, the oral examiner was asked to stop the video whenever they thought something important happened in the examination. At this point, they were asked to discuss any thoughts that arose at that time. Analysis of the recall statements revealed that as soon as the candidate began to answer an oral question, examiners formulated a first impression of the examinee's abilities and then used this impression to guide follow-up exploratory questions while assigning a provisional grade. This provisional grade was then confirmed through another series of questions meant to confirm it. Although Yaphe and Street do not report any times associated with when the tape was stopped, it is clear from their results that a first impression would be formed immediately and based on very little information.

A third characteristic that one would expect if first impressions reflect System 1 processes is that the judgment should require few cognitive resources to operate. In cognitive psychology, one common method used to study the automaticity of a task has been to use a divided attention task; that is, introduce a second task that a rater must complete at the same time they make a required judgment (Jacoby 1991) and then measure the performance or accuracy of the judgment. The reasoning is that if a task is being performed using automatic processes, it will require few cognitive resources, and therefore an individual has extra resources to devote to the second task. By this logic, introducing a simultaneous task will have little impact on the accuracy or performance associated with the initial judgment. If the addition of the second task impairs the accuracy or performance associated with the initial judgment, then it suggests that the rater is relying primarily on a System 2 process to make the initial judgment. Because System 2 processes require extensive cognitive resources to operate, the addition of the second task can interfere with cognitive processes needed to perform the initial task. In a variation of this divided attention manipulation, some researchers have used instructional manipulations designed to impair automatic processes by focusing conscious attention on those processes. In other words, encouraging someone to use a System 2 process on a task that is otherwise done using a System 1 process can impair the functioning of the System 1 process.

An example of a study that combined both a divided attention task and an instructional manipulation to explore the use of first impressions was described by Patterson and Stockbridge (1998). Participants viewed a set of brief scenes (28–124 s in length) from videos of people and were asked to make judgments about the people in the scenes. Some participants were asked to base their judgments on their intuition (first impression group) whereas some were asked to pay special attention to non-verbal cues like facial expressions or tone of voice (deliberative group). In addition to manipulating the instructions, participants were placed in either a high cognitive load group in which they had to recall a nine item list of groceries while judging the videos, or were placed in a low cognitive load group in which they received no other instruction. Patterson and Stockbridge found that participants in the high cognitive load condition were more accurate in the first impression group compared to the deliberative group, a finding one would expect if a first impression judgment relied primarily on a System 1 process. Interestingly, in the low cognitive demand group, scores were actually higher for participants in the deliberative condition compared to the first impression group. Patterson and Stockbridge suggested that this latter finding occurred because either the first impression group had more time to think about their answers, which interfered with the formation of a first impression, or because those in the deliberative group and had more cognitive resources to devote to processing the scenes. A similar result using a divided attention task was found by Ambady (2010). Participants were asked to rate the overall effectiveness of a set of teachers based on 10 s video clips. Participants were assigned to one of four conditions: (a) a cognitive load condition in which participants counted backwards while watching the clip (b) a reasons analyses task in which participants noted their reasons for making their judgments before making the ratings (c) a control condition in which participants made their judgment immediately after watching each clip and (d) a delayed rating condition in which participants waited 1 min before making a rating. The initial ratings were compared to the end of course ratings made by students of the teacher. Ambady found a low correlation between the initial ratings and the final course ratings in the reasons condition ($r = 0.27$) but no differences between the control and delay conditions compared to the cognitive load condition ($r = 0.65$–0.71). This pattern is what would be expected if participants were relying primarily on a System 1 process to make their initial judgments.

There is also evidence that using a divided attention task to understand the coordination of System 1 and System 2 processes is not limited to instructional manipulations. In a study to examine the effect of mood on the accuracy of first impressions, Ambady and Gray (2002) asked a group of participants to watch 15 s silent video clips of couples talking to each other and rate the type of relationship the couples had. Reasoning that people in a sad mood might be more deliberative when making a judgment and therefore less likely to rely on a first impression, they manipulated mood by adding a divided attention task. One group of participants was put in a sad mood before watching the videos and a second group of participants were just asked to watch the videos and make the judgments without any mood inducing manipulations. There was also a third group of participants who were also put in a sad mood but were additionally asked to count backwards from 1000 while making a judgment in an attempt to disrupt any System 2 processes. Ambady and Gray found that the ability to judge the type of relationship the couples had was lowest for the sad group, with no difference between the sad group in the divided attention condition and the control group.

In summary, research looking at the underlying cognitive processes behind first impressions would suggest that first impressions likely reflect the reliance on a System 1 process because raters are typically unaware of how they created an impression, the impressions are made quickly, and they are not sensitive to manipulations that add competing attentional demands.

How accurate are first impressions?

Given that first impressions are made quickly, based on very little information and reflect the use of System 1 processes, how accurate are they? The answer to this question is debatable. Much of the literature on dual process models (Evans 2008; Croskerry 2009; Kahneman 2011; Tversky and Kahneman 1974) has focused on the increase in errors that occur when people rely on System 1 processes. The explanation for the increase is that when relying on System 1 processes, raters are more likely to be influenced by heuristics, memory retrieval or other cognitive biases which lead to errors compared to processes that are more deliberative and analytic. From this perspective, first impressions should be prone to errors, and judgments based on them should be avoided.

Despite evidence that System 1 processes can lead to more errors than System 2 processes, some researchers have challenged the generality of these results (Eva and Norman 2005; Gigerenzer and Gaissmaier 2011; Klein 2009). For example, (Norman 2009; also Sherbino et al. 2012), in a review of clinical reasoning studies, described situations in which errors were associated with slow responses to clinical problems, whereas fast responses were more accurate. No claim was made that System 1 processes were more accurate than System 2 processes, but Norman did question the assumption that slower processes were necessarily more accurate than faster processes. Other researchers have compared decisions based on deliberative conscious process versus faster unconscious processes and have described situations in which the faster unconscious processes lead to better decisions (Dijksterhuis et al. 2006; Johnston et al. 1997; Wilson and Schooler 1991). The Ambady (2010) and Patterson and Stockbridge (1998) studies described previously also demonstrated the advantage of making judgments based on a first impression compared to more deliberative processes.

Other researchers have suggested that social judgment research has focused too much on errors, which are defined as deviations from statistical models (Funder 1987; see also Gigerenzer and Gaissmaier 2011), or on laboratory research with artificial stimuli (Kenny and Albright 1987). These researchers have suggested that rating-based research needs to be focused on what people can do in more naturalistic settings or with more realistic stimuli because a wrong judgment in the laboratory may be a correct judgment in the real world. This distinction is best demonstrated by considering how the Müller-Lyer Illusion is interpreted in terms of System 1 and System 2 processes. As shown in Fig. 1, the illusion is that people see the bottom line as being longer even if they know this is not the case. From the traditional social judgment perspective that focuses on errors (e.g., Kahneman 2011, pp. 26–28), the illusion reflects a System 1 error and people must learn to avoid this. From a perspective that focuses on more naturalistic judgments (e.g., Funder 1987, pp. 79–80), the first line is interpreted to be closer to you in space than the second line and therefore the second line is correctly judged as being longer because it reflects the proper use of perspective. In other words, the illusion does not reflect an error of judgment.

Another comment about accuracy is needed. Accuracy is a relative concept in many social judgment studies because a gold standard that clearly defines right or wrong does not

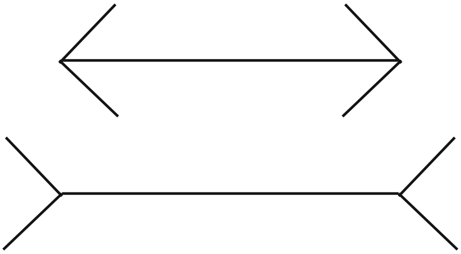

Fig. 1 Müller-Lyer effect

exist. Rather, accuracy is based on a judgment, which may rely on agreement or prediction (Funder 1987; Funder and West 1993; Kenny 1993). In the case of agreement, studies usually compare ratings made by a rater to those made by the target (self-other agreement) or to a rating made by other raters using the same criteria (consensus rating). For example, comparing a first impression rating of personality by a rater to the personality ratings that a target completes would provide a measure of accuracy. In the case of prediction, the ratings are used to see if they predict a future result based on either the same or different criteria. For example, these types of studies could compare a first impression rating of personality of a teacher by a student to an end of semester teacher evaluation. In either case, because there is often no definitive gold standard to measure accuracy and because the correlation can be influenced by properties of the rating scale and/or of the ratees (e.g. lack of variability), it is difficult to know how big a correlation should be to reflect a high degree of accuracy. As argued by Funder (1987), research in this area should study circumstances in which judgments become more or less accurate rather than focus on the magnitude of the correlation.

In light of the argument that accuracy is relative, studies of first impressions have shown considerable range in terms of the degree of accuracy. To cite a few examples, Barrick et al. (2010) found a moderate correlation between first impression based on a short rapport session and an interview score ($r = 0.42$). Borkenau and Liebler (1992) examined accuracy of people in judging personality factors based on short (90 s) videos of a target reading a weather report. They found highest accuracy for extraversion ratings ($r = 0.51$) with the other factors being less well judged (agreeableness = 0.35, conscientiousness = 0.25, neuroticism = 0.10, openness = 0.20). Colvin and Funder (1991) compared personality and behavior ratings made by participants who had watched a 5 min video of a target interacting with another person versus people who were actually acquaintances of the target. They found differences between strangers and acquaintances in terms of the accuracy of the personality judgments but a moderately high correlation ($r = 0.62$) between ratings of behavior. Findings similar to these led Ambady and Rosenthal (1992) to conduct a meta-analysis of 38 articles that looked at the accuracy of thin slices of behavior based on <300 s of exposure. They found a moderate effect (effect size 0.39, 95 % confidence interval of 0.34–0.48) of accuracy across the studies reviewed, suggesting that people are relatively accurate in their judgments despite making these judgments quickly.

In summary, a common perspective from dual code theorists is that reliance on System 1 processes, including first impressions, can lead to errors in judgment and that we need to be wary of relying on these processes when making judgments. Researchers have had two responses to this perspective. First, the pattern is not necessarily true in all cases and it has been shown that judgments based on slow deliberative processes can be more error-prone than those made on first impressions. Second, some researchers have questioned whether studies of factors that produce errors are of value, and point out that often errors made in the laboratory using System 1 processes are actually correct judgments when studied outside the laboratory. In addition, accuracy of most judgments, including first impressions, is relative because there is often no gold standard that determines right from wrong. Given this relativity, it may be more fruitful to study conditions that cause accuracy to increase or decrease rather than focus solely on whether one process leads to errors.

What factors modify the accuracy of a first impression?

So far the discussion of first impressions has focused on the underlying cognitive processes and to what degree judgments based on first impressions are accurate. An examination of other factors that could modify the accuracy of a first impression would be of value. Gingerich et al. (2011) has reviewed some of these factors within the larger impression formation literature and they include: mood of the rater, similarity to other people the rater knows, and seeing information in advance. Ambady et al. (2000) conducted a review of factors specific to first impressions and identified several factors related to the person being judged and to the rater. For example, with regard to the people being rated, observable personality traits like extroversion are typically judged more accurately than less observable traits like neuroticism or openness (Ambady et al. 1999; Borkenau and Liebler 1992; Lippa and Dietz 2000). For teachers, traits like expressiveness are judged most accurately but there may also be an impact of course difficulty on these and other teacher ratings (Babad et al. 2004).

With regard to rater-based factors, intelligence has been identified as a factor that could influence the accuracy of first impressions judgments. Lippa and Dietz (2000) asked people to view 30 s clips of men advertising eye glasses and then rate the men on three personality traits. They found that intelligence scores of the raters correlated significantly with the overall accuracy of their personality ratings. Gender has also been identified as a potential factor that can influence the accuracy of judgments based on first impressions (Ambady et al. 1995; Chan et al. 2011; c.f. Lippa and Dietz 2000). For example, Chan et al. (2011) asked a group of people to rate 5 min videos of females answering “get to know you” questions and then were asked to provide personality ratings. Although there was no gender difference in terms of distinctive accuracy (difference between the person being rated and the average person with that trait), there was a gender difference on normative accuracy (similarity between the person being rated and average person with that trait) with first impression ratings from female raters being more accurate than those from male raters. Another rater-based factor that can influence the accuracy of a first impression is the mood of the rater, with sad raters having less accurate first impressions than happy raters (Ambady and Gray 2002). Rater experience with a particular content area or type of participant may also be a factor that could lead to improved ratings, at least for judging sexual orientation (Ambady et al. 1999).

Another issue related to factors that could influence the accuracy of a judgment based on a first impression is related to impression management and the stability of the first impression. Impression management is most commonly studied in the job interview literature and refers to situations in which interviewees attempt to influence an interviewer by controlling the interaction between themselves and interviewer. Barrick et al. (2009) conducted a meta-analysis to determine the degree that job applicants can alter their image during an interview and found evidence that appearance (i.e. physical and professional), impression management (i.e., self-promotion, ingratiation to the interviewer, emphasizing positives, focusing attention on the interviewer), and verbal (voice) and non-verbal (smiling, eye contact) characteristics can all have an influence on the impression a rater may form. There is also evidence that an initial impression can be altered. Macan and Dipboye (1990) found that interviewers with negative pre-interview impressions are more likely to change their impressions during the context of an interview than interviewers with positive pre-interview impressions. Harris and Garris (2008) reviewed several studies that investigated how to overcome a negative first impression and concluded that important factors would be to ensure the rater is not distracted, ensure the rater is motivated to be accurate and highlight one's unique characteristics.

In summary, a number of factors were identified that can influence the accuracy of a first impression. Some of these factors are related to the person being judged: either unintentional factors like similarity to the rater or intention factors like those deliberately used by ratees to manage impressions raters may create. Other factors that influence accuracy like gender, intelligence, or mood are related to the rater. Finally, some traits like extraversion appear to be easier to judge than other traits.

What is the impact of first impressions for assessment?

The majority of studies of first impressions have focused on the ability of raters to make a personality judgment of some kind, rate the abilities of a teacher, or predict the success of a job interview. Little research in this area has focused directly on the types of assessments that are common in medical education. That said, there is some evidence that first impressions could be influential in the assessment of learners. The first area deals with a phenomenon called self-fulfilling prophecies or an expectancy effect. This phenomenon occurs when an initial impression influences subsequent interactions between the rater and the person being rated (Dipboye 1982; Harris and Garris 2008; Rosenthal 1994). For example, Snyder et al. (1977) led female and male participants to believe they were getting to know an opposite sex participant during a telephone conversation. Prior to the conversation, the men were given a description of a female participant along with a random photograph of an attractive or unattractive female. Females were told nothing about the males. After talking to each other for 10 min, both groups were asked to complete an impression rating of the other person. Males rated females who they thought were attractive higher on the scales than women they thought were unattractive. Of note, women rated men who thought they were attractive more positively than men who thought they were unattractive even though they had no knowledge of the information the men received. Snyder et al. also had blinded judges listen to tapes of the conversations and rate the participants. These judges rated the women who the males had thought were attractive higher than women the males had thought were unattractive despite being blinded to the conditions of the study. Snyder et al. concluded that if male participants had negative expectations, they treated the female participants in a less friendly manner, getting a negative reaction from the females. If the males had positive expectations of the female, they treated her in a friendlier manner, which led to more positive ratings from the women. In other words, expectations of the male participants resulted in the female participants behaving in ways consistent with those expectations. Similarly, in a study using job interviews, Dougherty et al. (1994) asked blinded judges to rate tapes of interviews for positions at a company. All of the interviewers had access to information about the applicant and were asked to create a first impression rating based on this information before actually meeting and interviewing the applicant. Dougherty et al. found that positive first impressions were related to more positive communication styles by the interviewer, increased likelihood to extend an offer, and more positive vocal style. They also found that interviewers with more positive first impressions were more likely to stress the positives of the company and asked fewer close-ended questions. Interviewees who received a positive first impression by the interviewer rated the rapport with the interviewer higher than those interviewees with negative first impressions. This pattern therefore suggesting that the interviewer altered the way they conducted the interview based on their first impression.

The second area to which first impressions could impact on assessment is a type of rater bias called a halo effect. A halo effect is thought to occur when a rater fails to discriminate among independent aspects of behavior when making a judgment about a person. Halo is typically manifested as either high average correlations across all dimensions being assessed, low average standard deviations across all dimensions being assessed, when a single factor accounts for all of the variability in scores across multiple dimensions, or when a significant rater × ratee interaction is found (Balzer and Sulsky 1992; Cooper 1981).

Several sources of halo have also been identified by researchers. The first source of halo occurs when a rater makes a judgment about a person based on a general impression (e.g. a first impression) that they form. This impression then influences all subsequent ratings or judgments about the person. For example, if a rater forms a first impression of a learner that is either positive or negative in nature, then this impression will guide the ratings on all dimensions being rated. The second source of halo occurs when a salient dimension or trait drives the ratings on other dimensions being judged. For example, a high or low rating on communication skills could influence ratings on other dimensions, even those that may be unrelated, like technical skills or knowledge. A third source of halo is an inadequate discrimination between dimensions being rated. This source of halo usually occurs when the dimensions being rated are ambiguous and raters end up grouping what are intended to be unrelated dimensions and providing similar ratings. Studies that have directly compared these three sources of halo (Fisicaro and Lance 1990; Woehr et al. 1998) have found that the general impression model tends to provide a better account of rating data even with instructions deliberately designed to encourage salient dimension and inadequate discrimination models (Lance et al. 1994). Cooper (1981) added three additional sources of halo. These include under sampling, which occurs when raters have insufficient experience sampling the person's behavior to make independent ratings and are therefore forced to rely on a general impression; insufficient effort, which occurs when raters are careless when providing ratings; and cognitive distortion, which occur when raters rely on memory and can be distorted depending on the beliefs of the raters.

Does the presence of a halo effect lead to accurate or inaccurate ratings? The presence of halo is considered to be due to a System 1 process (i.e., use of a general impression or memory of behaviors rather than independent ratings) and therefore, like the discussion around the accuracy of first impressions, the accuracy of judgments influenced by a halo is debatable. On the one hand, halo implies a decrease in the number of independent opportunities within a particular rating for a target to demonstrate proficiency, and thus feedback on an individual's strengths and weaknesses would be difficult to do. On the other hand, the high reliability that is often associated when halo is present has led some researchers to argue that halo can lead to accurate judgments (Cooper 1981; Goffin et al. 2003; Murphy et al. 1993) especially if trying to discriminate performance levels between ratees (but not within ratees).

The increased accuracy in judgments associated with a halo effect was demonstrated by Nathan and Tippins (1990). They asked supervisors to rate the performance of clerical workers on a rating scale and these ratings were then compared to results on a variety of performance tests related to clerical ability, verbal ability, and numerical ability. They found that when there was more halo associated with the ratings, there was also a higher correlation with the outcome of the related performance tests, suggesting that halo actually led to higher levels of validity. In a study contrasting rater training techniques, Bernardin and Pence (1980) studied the ability of certain training techniques to control halo. They trained three groups of student raters: one group trained by focusing on errors in their ratings, one group trained by focusing on accuracy, and a control group that did not have any training. The raters in each group then viewed vignettes for two teachers and provided ratings. The group that was trained by focusing on reducing their errors did display less evidence of halo- based and leniency errors, but they also produced less accurate ratings than the other two groups, indicating that the groups with more halo were also more accurate. There was another interesting finding in their study in that there was little difference in accuracy between the group of raters who were trained to focus on accuracy of judgments and those who were in the control group. In a later review of rater training studies, Woehr and Huffcutt (1994) reported that the decrease in accuracy for the rater error group in the Bernardin and Pence study was likely due to the type of error training that was used, and that other types of training can increase accuracy and decrease halo. However, Cook et al. (2008) found similar results as the Bernadin and Pence study in that there was little difference in halo and accuracy between raters who were trained and those in a control group. It would appear, therefore, that the relationship between halo and accuracy is an area that warrants further research to understand the conditions that influence this relationship.

There is another issue with halo that is unrelated to underlying cognitive processes but could impact on how halo is measured and interpreted. Much of the research on halo errors has made the distinction between true halo and illusory halo. True halo is the natural correlation that exists between items or dimensions being measured. Illusory halo is the correlation that exists over and above the true halo and is usually what people think of when they talk about halo errors. The issue, as described by Murphy et al. (1993), is that there is no easy way to separate the influence of the two types of halo; therefore, it is extremely difficult to accurately determine the degree that correlations between scores on items reflect one or the other types of halo.

In summary, first impressions may influence the types of assessments used in medicine in two ways. It could contribute to a self-fulfilling prophecy in which negative or positive first impressions influence the way a rater thinks about or interacts with a target. It could also contribute to the presence of a type of rater bias called the halo effect because one of the causal mechanisms behind halo is the use of a general impression by a rater when making a judgment about a target.

Conclusion, implications for assessment in medical education

Medicine relies heavily on examination formats that require raters to assess the skills and knowledge of examinees, for example OSCEs, oral examinations, and other simulation-based examinations. With the advent of competency-based education, rater-based assessments that occur in the workplace have increasing prominence as assessment tools (e.g. mini-CEX, portfolios, case-based discussions, multi-source feedback, etc.). Despite the increasing importance of rater-based assessments, there is a realization that raters may demonstrate inconsistencies or variability in the scores assigned. In medical education, there is a recent and growing literature that is exploring the underlying cognitive process addressing how raters assess learners within the context of examinations and workplace-based assessments (Govaerts et al. 2011; Kogan et al. 2011; Yeates et al. 2013). Factors related to impression formation (Gingerich et al. 2011), cognitive load (Tavares and Eva 2013; van Merriënboer and Sweller 2010; Wood 2013), familiarity with the examinee (Stroud et al. 2011), rater expertise (Berendonk et al. 2013) as well as rater-biases (Iramaneerat and Yudkowsky 2007; Williams et al. 2003) and overly structured assessments within competency-based frameworks (Ginsburg et al. 2010) have all been identified as influencing the way assessors assign scores. By reviewing the literature on judgments based on first impressions, this paper adds to the ongoing investigation of the underlying cognitive aspects of rater-based assessments in medical education.

Based on the findings that were reviewed, first impressions do contribute to how raters assign scores when making judgments of other people. Although they may have a role in some aspects of assessment related to self-fulfilling prophecies and a halo effect, the exact impact on assessment in medical education has not been clearly defined. For this reason, a research agenda based on the following questions is needed to more fully understand the role of first impressions in assessment.

1) To what degree will first impressions influence subsequent ratings within a particular assessment context or tool?

The basic finding, that first impressions are related to subsequent scores, is compelling but requires demonstration in a variety of contexts. A deeper understanding of the magnitude and potential reasons for this effect is required. For example, work by Yaphe and Street (2003) and Govaerts et al. (2011, 2013) would suggest that raters use first impressions to help guide their choice of questions and how they assign scores but a relationship between first impressions and subsequent scoring still needs to be demonstrated.

2) Do first impressions change within the context of a single assessment session and if so under what conditions?

Anecdotally, many physician examiners can describe an examinee that started off an OSCE station or oral examination badly and then recovered brilliantly. These stories suggest that impressions can change, but such anecdotal evidence must be supported by rigorous research. The stability of a first impression is particularly important for examinations like OSCEs, oral examinations or mini-CEXs that require a decision about testing time associated with individual cases. If a judgment about the examinee's ability is made within the first couple of minutes, and that judgment remains stable throughout the assessment despite a change in an examinee's performance, then longer assessments may not be adding anything to the quality of the rating that one cannot get within a few minutes. Evidence for this stability would have implications for test developers and may challenge assumptions about the fidelity of the cases.

3) How does the coordination of System 1 and System 2 processes influence the use of and the accuracy of a first impression?

Under some circumstances, System 1 processes, like first impressions, can lead to more accurate judgments than System 2 processes, but it is not clear under what conditions this may occur. A better understanding of how these processes are coordinated and how various factors influence these processes is needed. One such factor is the scoring method used. There is a considerable amount of literature on the advantages and disadvantages of using checklists versus rating scales for assessments (Hawkins and Boulet 2008; Van der Vleuten and Swanson 1990). A checklist is a highly deliberative scoring process so would likely reflect the use of System 2 processes. Rating scales, on the other hand, have more room for rater interpretation and are less rigid, so could allow a larger role for System 1 processes like first impressions to influence scoring. The purpose of the assessment (i.e., formative or summative assessment), is also important in terms of whether System 1 or System 2 processes should be favored. It is possible that an examination designed for formative feedback might favor a deliberative, analytical scoring process in order to provide feedback, whereas an examination designed solely for summative assessment may favor a more global, less analytical scoring process. Cognitive load is another factor that would likely influence the use of and accuracy of first impressions. Because some rating tasks require a higher degree of cognitive resources (i.e., attention) than other tasks, the resulting scores could start to mimic the results found under divided attention manipulations described earlier. For example, imagine a situation in which a rater must evaluate an examinee's history taking, communication skills and professionalism while they interact with a live patient in a busy Emergency Department. This would be a distracting environment with a high cognitive load that might make System 2 processes difficult to use. One might get a different result in a situation in which a rater only has to evaluate history taking skills of an examinee in a clinic.

4) To what degree does a self-fulfilling prophecy influence the ratings?

One of the ways that first impressions can influence assessments is by leading to a self-fulfilling prophecy. The presence of this phenomenon has not typically been studied within the context of assessment other than job interviews, so would require further investigation in different contexts. For example, in an OSCE, it is possible that even when the rater plays a passive role, like a physician examiner, positive or negative first impressions could be reflected in the body language of the examiner that could then be observed by the examinee and influence their performance.

5) What is the relationship between first impressions and the halo effect?

First impressions are thought to contribute to the presence of a halo effect. Under some circumstances, especially when one wants to identify specific strengths and weakness within a person, the presence of halo would make the assessment difficult. In other circumstances, especially when trying to discriminate abilities between individuals, the presence of halo may actually be a benefit due to the high reliability. What is unclear is what those circumstances are, and how manipulations that influence first impressions impact on the presence or absence of halo.

The following example illustrates how one might apply the questions above to study the role of first impressions on an examination, and in doing so, gain a better understanding of the cognitive processes associated with the raters scoring. Let's say a researcher wants to create a new OSCE at their medical school. One of the decision points pertains to station length. If the test administrator feels that an examination consisting of 30 min stations will be more realistic than the traditional 5–10 min stations, consideration of first impressions may impact the final scoring decision of the OSCE examiners. As a first step, the researcher should demonstrate the degree that examiners form and rely on first impressions when scoring the examinee. He or she could conduct a study in which examiners in a longer station provide a rating of the examinees' abilities within the first 30 s of the station. This rating could then be compared to the final score on the station in terms of both a correlation and to see if there are any examples of large differences between the first impression and the final score. If the correlation between the two measures is statistically significant, then it would support the existence of an important relationship between the two measures. If further analyses show evidence of halo effect, then it would also suggest that the longer station is not adding anything to the quality of the measure. On the other hand, even with a high correlation, there might be examples of discrepancies between the two measures, so another follow up study might involve identifying those cases when there is a large difference and then interviewing those examiners to find out what led to such a change. The researcher might want to consider additional factors that potentially influence the presence of and the accuracy of the first impression. One such factor is related to the scoring. First impression ratings could be compared to a condition in which examiners score examinees using a checklist versus a condition in which they use a rating scale. If rating scales support the use of System 1 processes and checklists facilitate System 2 process, one might find a larger correlation with the former scoring system. To determine if there is any evidence of a self-fulfilling prophecy, the stations could be videotaped and a blinded reviewer could watch the examiner's behavior to see if he/she displays any non-verbal cues that would be related to their first impression of the examinee and to the examinee's score.

The literature that was reviewed in this paper has focused on the role of first impressions with regards to the assessment of learners. However, there exists another area of medical education research in which a study of the role of first impressions could have an impact and that is in regards to clinical reasoning and decision making. The current debate about models of clinical reasoning is analogous to the debate about how people make judgments about the behavior and personalities of others. It is generally accepted that clinicians have available two underlying processes to help them reach a clinical decision, a fast automatic system, or System 1 processes, and a slower more reflective system, or System 2 processes (Croskerry 2009; Eva and Norman 2005, 2009; Norman and Eva 2010; Pelaccia et al. 2011). As part of the System 1 process, some researchers have introduced the concept of judgments based on pattern recognition or intuition, which are thought to reflect the fast automatic retrieval of other similar cases from memory and are based on limited information (Norman et al. 2007), much like a first impression. Given the similarities between first impressions and pattern recognition, it is likely that these are similar cognitive processes and therefore many of the questions and findings around first impressions that were reviewed would apply to pattern recognition. In fact, considerable research has already been done to study the role System 1 and System 2 processes that underlay clinical reasoning. For example, Sherbino and colleagues are studying the role of fast accurate clinical decisions versus slower less accurate decisions by encouraging participants to respond as fast as possible when making a diagnosis (Sherbino et al. 2012) versus slow and accurate (Norman et al. in press) and by introducing a divided attention task (Monteiro et al. unpublished manuscript) similar in nature to Ambady's work reviewed above (Ambady 2010). Other researchers have started to argue that focusing on errors due to System 1 biases may not be a fruitful line of research (Norman and Eva 2010), echoing comments made by Funder (1987) and Kenny and Albright (1987) with regards to errors in social judgments.

In conclusion, the studies reviewed in this paper suggest that first impressions influence our perceptions, predictions and theories about other people we meet. Given that rater-based assessments also involve the same judgments and influences as other social interactions, it makes sense that first impressions could have an influence on the assessment. As medical education researchers try to gain a better understanding of the underlying cognitive processes raters use when assessing the skills and competence of learners, it would be wise to consider how we use these impressions and how they influence our judgments. This paper is intended to be a step toward that goal.

Acknowledgments The author would like to thank Dr. Andrea Gingerich and Dr Sydney Smee for their helpful discussions that led to this project and Dr. Stan Hamstra, Dr. Susan Humprey-Murto, Kulamakan Mahan Kulasegaram, Sarah Lynch, Dr. Debra Pugh, Dr. Nikki Woods and the journal editor for their thoughtful review and suggestions regarding this manuscript.

References

AERA, APA, & NCME. (1999). Standards for educational and psychological testing (pp. 9–24). Washington, DC: American Educational Research Association.

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry, 21*(4), 271–278.

Ambady, N., Bernieri, F., & Richeson, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32*, 201–271.

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology, 83*(4), 947–961.

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology, 77*(3), 538–547.

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69*(3), 518–529.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431–441.

Ames, D. R., Kammrath, L. K., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. *Personality and Social Psychology Bulletin, 36*(2), 264–277.

Babad, E., Avni-Babad, D., & Rosenthal, R. (2004). Prediction of students' evaluations from brief instances of professors' nonverbal behavior in defined instructional situations. *Social Psychology of Education, 7*(1), 3–33.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975–985.

Bargh, J. A. (1992). The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects. *The American Journal of Psychology, 105*(2), 181–199.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *The Journal of Applied Psychology, 94*(6), 1394–1411.

Barrick, M. R., Swider, B. W., & Stewart, G. L. (2010). Initial evaluations in the interview: Relationships with subsequent interviewer evaluations and employment offers. *The Journal of Applied Psychology, 95*(6), 1163–1172.

Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: Assessors perspectives. *Advances in Health Sciences Education: Theory and Practice*. doi: 10.1007/s10459-012-9392-x

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*(1), 60–66.

Biesanz, J. C., Human, L. J., Paquin, A. C., Chan, M., Parisotto, K. L., Sarracino, J., et al. (2011). Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. *Social Psychological and Personality Science, 2*(5), 452–459.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*(4), 645–657.

Brooks, L. R. (2005). The blossoms and the weeds. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 59*(1), 62–74.

Carney, D., Colvin, C., & Hall, J. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*(5), 1054–1072.

Chan, M., Rogers, K. H., Parisotto, K. L., & Biesanz, J. C. (2011). Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality, 45*(1), 117–120.

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2008). Issues of validity and reliability for assessments in Medical Education. In E. S. Holmboe & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (pp. 10–23). Philadelphia: Mosby Elsevier.

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology, 60*(6), 884–894.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 166.e7–166.e16.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. A. (2008). Effect of rater training on reliability and accuracy of mini-cex scores: A randomized, controlled trial. *Journal of General Internal Medicine, 24*(1), 74–79.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218–244.

Croskerry, P. (2009). Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. *Advances in Health Sciences Education, 14*, 27–35.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behaviour and Human Performance, 33*, 360–396.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*(5763), 1005–1007.

Dipboye, R. L. (1982). Self-fulfilling prophecies in the selection-recruitment interview. *The Academy of Management Review, 7*(4), 579.

Dodson, M., Crotty, B., Prideaux, D., Carne, R., Ward, A., & De Leeuw, E. (2009). The multiple mini-interview: How long is long enough? *Medical Education, 43*(2), 168–174.

Dougherty, T. W., Turban, D. B., & Callender, J. C. (1994). Confirming first impressions in the employment interview: A field study of interview behaviour. *Journal of Applied Psychology, 5*(5), 659–665.

Downing, S. M., & Haladyna, T. M. (2009). Validity and its threats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 21–56). New York: Routledge.

Eva, K. W., & Norman, G. R. (2005). Heuristics and biases—A biased perspective on clinical reasoning. *Medical Education, 39*(9), 870–872.

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education: Theory and Practice, 16*(3), 311–329.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Applied Psychology, 66*(2), 127–148.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419–429.

Fiske, S., & Neuberg, S. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (23rd ed., pp. 1–75). San Diego: Academic Press Inc.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*(1), 75–90.

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: A introduction. *Journal of Personality, 61*(4), 457–476.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine, 86*(10), S1–S7.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*(5), 780–786.

Goffin, R. D., Jelley, R. B., & Wagner, S. H. (2003). Is halo helpful? Effects of inducing halo on performance rating accuracy. *Social Behaviour and Personality, 31*(6), 625–636.

Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education: Theory and Practice, 16*(2), 151–165.

Govaerts, M. J. B., Van de Wiel, M. W. J., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education: Theory and Practice*. doi:10.1007/s10459-012-9376-x.

Harris, M., & Garris, C. (2008). You never get a second chance to make a first impression. In N. Ambady & J. Skowronski (Eds.), *First impressions* (pp. 147–168). New York, NY: Guilford Press.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*(3), 356–388.

Hawkins, R. E., & Boulet, J. R. (2008). Direct observation: Standardized patients. In E. S. Holmboe & R. E. Hawkins (Eds.), *Evaluation of clinical competence* (pp. 102–118). Philadelphia, PA: Mosby Elsevier.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*(8), 676–682.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64–65.

Iramaneerat, C., & Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation and the Health Professions, 30*(3), 266–283.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

Jacoby, L., & Kelley, C. (1990). An episodic view of motivation: Unconscious influences of memory. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 2, pp. 451–480). New York, NY: Guilford Press.

Johnston, J. H., Driskell, J. E., & Salas, E. (1997). Vigilant and hypervigilant decision making. *The Journal of Applied Psychology, 82*(4), 614–622.

Kahneman, D. (2011). *Thinking, fast and slow*. Canada: Doubleday.

Kenny, D. A. (1993). A coming-of-age for research on interpersonal perception. *Journal of Personality, 61*(4), 789–807.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102*(3), 390–402.

Klein, G. (2009). *Streetlights and shadows: Searching for the keys to adaptive decision making*. Cambridge, MA: MIT Press.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*(3), 332–340.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72–107.

Lippa, R. A., & Dietz, J. K. (2000). The relation of gender, personality and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior, 24*(1), 25–43.

Logan, G. D. (1992). Attention and preattention in theories of automaticity. *The American Journal of Psychology, 105*(2), 317–339.

Macan, T. H., & Dipboye, R. L. (1990). The relationship of interviewrs' preinterview impressions to selection and recruitment outcomes. *Personnel Psychology, 43*(4), 745–768.

Monteiro, S. D., Sherbino J. D., Ilgen, J. S., Dore, K. L. Gaissmaier, W., Wood, T. J., et al. (unpublished manuscript). *Diagnosing Fast and Slow: The Effect of Interruptions on Speeded and Reflective Clinical Reasoning.*

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218–225.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology, 75*(3), 290–296.

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*, 37–49.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education, 44*, 94–100.

Norman, G. R., Sherbino, J., Dore, K. L., Wood, T. J. Ph. Young, M. E., Gaissmaier, W., et al. (in press). The etiology of diagnostic errors: A controlled trial of System 1 vs. System 2 reasoning. *Academic Medicine*.

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41*, 1140–1145.

Patterson, M. L., & Stockbridge, E. (1998). Effects of cognitive demand and judgment strategy on person perception accuracy. *Journal of Nonverbal Behavior, 22*(4), 253–263.

Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Medical Education Online, 16*, 5890.

Rosenthal, R. (1994). Interpersonal expectancy effects : A 30-year perspective. *Current Directions in Psychological Science, 3*(6), 176–179.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science, 27*(3), 525–559.

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Krueger, S., et al. (2012). On the relation between processing speed and diagnostic error. *Academic Medicine, 87*(6), 785–791.

Smith, H. J., Archer, D., & Costanzo, M. (1991). "Just a hunch": Accuracy and awareness in person perception. *Journal of Nonverbal Behavior, 15*(1), 3–18.

Snyder, M., Tanke, E., & Berscheid, E. (1977). Social perception and interpersonal behavior : On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology, 35*(9), 656–666.

Stroud, L., Herold, J., Tomlinson, G., & Cavalcanti, R. B. (2011). Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. *Academic Medicine, 86*(10), S8–S11.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education: Theory and Practice*. doi:10.1007/s10459-012-9370-3.

Tom, G., Tong, S. T., & Hesse, C. (2009). Thick slice and thin slice teaching evaluations. *Social Psychology of Education, 13*(1), 129–136.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology, 59*, 329–360.

Van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58–76.

Van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education, 44*(1), 85–93.

Wigton, R. (1980). The effects of student personal characteristics on the evaluation of clinical performance. *Journal of Medical Education, 55*, 423–427.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270–292.

Willis, J., & Todorov, A. (2006). Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181–192.

Woehr, D. J., Day, D. V., Winfred, A., & Bedeian, A. G. (1998). The systematic distortion hypothesis: A confirmatory test of the implicit covariance and general impression models. *Basic and Applied Social Psychology, 16*(4), 417–434.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.

Wood, T. J. (2013). Mental workload as a tool for understanding dual processes in rater-based assessments. *Advances in Health Sciences Education: Theory and Practice.* doi:10.1007/s10459-012-9396-6

Yaphe, J., & Street, S. (2003). How do examiners decide?: A qualitative study of the process of decision making in the oral examination component of the MRCGP examination. *Medical Education, 37*(9), 764–771.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently. *Advances in Health Sciences Education: Theory and Practice*. doi:10.1007/s10459-012-9372-1.