Organizational Behavior and Human Decision Processes 147 (2018) 26–47

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

A ratings pattern heuristic in judgments of expertise: When being right Looks wrong

Gerri Spassova[a,*], Mauricio Palmeira[b], Eduardo B. Andrade[c]

[a] Department of Marketing, Monash Business School, Melbourne, Australia
[b] SKK Graduate School of Business, Sungkyunkwan University, Seoul, Republic of Korea
[c] Brazilian School of Public and Business Administration at FGV, Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Expertise judgments
Ratings pattern
Rating variance
Uniformity

ABSTRACT

We propose a "ratings pattern heuristic" in judgments of expertise—that is, people's tendency to undervalue critics who assign the same rating to multiple options, overlooking diagnostic information which would clearly justify the uniform ratings. The heuristic is driven by a strong association between discrimination and expertise and a focus on summary ratings. People "punish" uniform (vs. varied) raters even when (a) uniform ratings are acknowledgedly more likely (studies 1a and 1b), (b) the uniform rater's past performance is superior (studies 2 and 3), and (c) the uniform rater also reports varied sub-ratings (study 4a), unless participants are prompted to assess the sub-ratings prior to choosing a critic (studies 4b and 5). Study 6 reveals that critics are less aware than judges of the impact of the pattern of their ratings on others' perceptions.

1. Introduction

Imagine that Bill and Rodney – managers at a consulting company – are evaluating three projects that their company could take on. According to Bill, the three projects look equally promising and he assigns all of them the same rating. Rodney, on the other hand, thinks the projects vary in revenue potential and assigns a different rating to each one. Given this information, who would you say is better at judging the projects' potential? Most of us would probably choose Rodney, as he seems to be able to better differentiate among the options. Now assume that the company takes on all three projects. A year later, they have performed equally well, as Bill predicted. How would this additional information affect your preferences? Would you still prefer Rodney, given that Bill clearly has the performance advantage?

In the current research, we propose that people infer expertise from the pattern of ratings given by a critic – a phenomenon we refer to as "ratings pattern heuristic." Individuals tend to judge critics who give the same rating to multiple options (hereafter referred to as "uniform critics") as less expert than critics who give different ratings (hereafter referred to as "varied critics"). Importantly, we show that people rely on this heuristic even when it conflicts with objective information indicating that the rated options are more likely to be of equal, rather than different, quality, or with other expertise-diagnostic cues favoring the uniform critic. Further, an asymmetry is also observed. Whereas those judging expertise are highly sensitive to a uniform ratings pattern, those providing the ratings do not seem to anticipate this effect.

1.1. What influences perceptions of expertise

Prior research has identified several factors that influence perceptions of expertise. One such factor is past performance – critics with a superior track record or greater task-relevant experience are considered to have more expertise and their opinions are weighted more heavily (Birnbaum & Stegner, 1979; Feick & Higie, 1992; Harvey & Fischer, 1997; Gershoff, Broniarczyk, & West, 2001). When there is limited access to objective information about the critic's past performance or experience, people may resort to heuristics such as the perceived similarity of the critic to the self (Feick & Higie, 1992; MacKie, Gastardo-Conaco, & Skelly, 1992; Yaniv & Milyavsky, 2007) or the critic's expressed confidence (Price & Stone, 2004; Sniezek & Van Swol, 2001). Critics who are perceived to be more similar to the self, or who provide more confident judgments, are preferred over less similar or less confident ones.

Research shows that the reviews and ratings given by a critic can also influence other people's perceptions of the critic (Amabile, 1983; De Langhe, Fernbach, & Lichtenstein, 2016; Floyd, Freling, Alhoqail, Cho, & Freling, 2014; Rosario, Sotgiu, De Valck, & Bijmolt, 2016). Giving negative and critical reviews has been associated with perceptions of greater intelligence and competence (Amabile, 1983), but recent findings suggest that online reviewers who give low ratings are

* Corresponding author.
E-mail addresses: gerri.spassova@monash.edu (G. Spassova), mauricio.palmeira@skku.edu (M. Palmeira), eduardo.b.andrade@fgv.br (E.B. Andrade).

https://doi.org/10.1016/j.obhdp.2018.05.003
Received 28 April 2017; Received in revised form 14 February 2018; Accepted 6 May 2018

seen as less credible than those who give high ratings (Lim & Van Der Heide, 2015; Wang, Cunningham, & Eastin, 2015).

In our research, we examine the impact of another aspect of ratings on perceptions of expertise: the variance of the ratings assigned by a single critic. We argue that people associate uniform ratings with inferior expertise. For example, a financial analyst who has given a 4-star rating to three investment funds is judged as less knowledgeable than one who has given 4-, 3-, and 5-star ratings. In the absence of other diagnostic information, inferring lower expertise from uniform ratings may not be unreasonable, as it will be detailed below. Our investigation, however, focuses on cases where diagnostic information suggesting otherwise is available.

1.2. When heuristics fail

Heuristics have been defined as strategies that "ignore part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods" (Gigerenzer & Gaissmaier, 2011; pg. 454). According to dual-process models of judgment and decision making, heuristics belong to System 1 processes which tend to be more rapid, low-effort, and automatic, relative to System 2 processes which tend to be slower, high-effort, and controlled (Kahneman & Frederick, 2002; Stanovich, 1999; see Evans, 2008 for a review).

When applying dual-process models to social judgement, researchers have made similar distinction between heuristic processing, which is cognitively frugal, rule- or category-based, and systematic processing, which is more effortful, analytic, and comprehensive (Brewer, 1988; Chaiken, 1980; Chen & Chaiken, 1999; Fiske & Neuberg, 1990; Petty & Cacioppo, 1981). Social judgments formed on the basis of heuristic processing tend to reflect salient and easily processed cues (e.g., the person's gender, race, age, or profession), rather than more complex, individualistic, or particularistic judgment-relevant information (e.g., the person's unique attributes or behavior).

Heuristics function reasonably well in many situations, particularly in environments characterized by high uncertainty when only part of the relevant information is known (Gigerenzer & Gaissmaier, 2011). However, they perform worse when preferred over equally, or even more diagnostic, judgment-relevant cues. For example, people continue to rely on confidence as an expertise cue even when it is not adaptive to do so, in the presence of more diagnostic cues such as accuracy or outcome information (Keren & Teigen, 2001; Price & Stone, 2004; Van Swol & Sniezek, 2005).

According to the heuristic-systematic model of information processing (Chen & Chaiken, 1999), people continue to rely on heuristics even when the latter are no longer appropriate, in an effort to strike a balance between the goals of minimizing cognitive effort and achieving accuracy. Heuristics offer such a compromise as they involve highly accessible and easy to process cues, which are reasonably relevant to the task. As long as these cues are deemed to produce sufficiently accurate judgments, they predominate over less salient or more difficult to process information (Chen & Chaiken, 1999). The duration heuristic (Yeung & Soman, 2007) is one such example. Although the duration of a service does not determine its value, people who believe in a positive correlation between the two use duration as a cue since it can be easily measured on an objective scale, whereas value is often difficult to assess. However, people continue to apply the "longer is better" rule even in situations where it is no longer applicable (e.g., evaluating more highly a locksmith that opens a door in 20 rather than 2 min; Yeung & Soman, 2007).

The above discussion suggests that heuristics based on highly accessible and easy to process cues can be quite resistant to the presence of objective information that challenges their applicability (Chen & Chaiken, 1999; Keren & Teigen, 2001; Price & Stone, 2004; Yeung & Soman, 2007). In the current research, we propose that the reliance on the ratings pattern, and specifically the extent to which the ratings display variance, is one such heuristic that remains robust in the face of more diagnostic expertise information. We propose that this is the case because both rating variance and summary ratings are highly accessible and easy to process expertise-relevant cues. The next two sections detail the reasoning behind this proposition.

1.3. The "discrimination ability – expertise" association

Discrimination is a critical feature of expertise (Hammond, 1996; Shanteau, Weiss, Thomas, & Pounds, 2002) and a number of studies across different domains have shown that experts, relative to novices, are able to make finer distinctions between members of a category. For example, musicians detect pitch changes faster and more accurately than non-musicians (Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2005), experts in natural categories such as birds or cars are better than novices at distinguishing among members of these categories (Alba & Hutchinson, 1987; Gauthier, Skudlarski, Gore, & Anderson, 2000), and people are better at recognizing faces of their own race (Malpass & Kravitz, 1969). Research on categorization also shows that as one's expertise in a domain increases, the basic level at which they categorize items in that domain becomes more specific, i.e., they spontaneously make finer discriminations (Dougherty, 1978; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). This suggests that the link between discrimination and expertise is likely to be highly accessible in people's mind. However, the fact that experts can identify finer and finer differences among options does not mean that the options should also differ in quality. For example, wines may differ along subtle dimensions that only a connoisseur can identify, but still be of the same overall quality and deserve the same rating. A product line may include a wide range of options, many of which can still perform equally well. In fact, with the ever increasing speed of global competition and technological innovation, more and more products and services are comparable in overall quality and differ only in subjective or peripheral features. In some situations, information such as statistics or performance results may explicitly indicate that the options are of equal quality. For example, if it is known that investment funds have performed equally well over a period of time, these funds should receive equal ratings. Yet, we propose that the association between discrimination and expertise is so strong that people would still expect giving varied ratings to reflect higher expertise.

Furthermore, in the context of ratings, discrimination can (a) be deduced simply from the variance in the critic's ratings, even if this variance is not diagnostic of expertise (as when the critic discriminated randomly or on the wrong attributes) and (b) overshadow equally, or even more diagnostic but harder to process information, as it offers a compromise between cognitive efficiency and accuracy (Chen & Chaiken, 1999).

The above properties should lead people to rely on discrimination as a heuristic for expertise even in the presence of contradicting diagnostic information. Lack of discrimination—that is, uniform assessments across options—should thus negatively impact perception of expertise. In the context of critic ratings, we expect giving uniform ratings to be seen as a sign of inferior expertise even when the evaluated options are more likely to be of equal, rather than different, quality. Put formally:

Hypothesis 1.. *Uniform ratings will negatively impact perceptions of expertise even in the presence of information indicating that the rated options are more likely to be equal rather than different.*

We test this hypothesis with our first batch of studies (study 1a to 3). In studies 1a and 1b we contrast judgments about the likelihood of an outcome to judgments about expertise. We show that even when uniform ratings are considered much more likely than varied ratings, the critic giving varied ratings is considered more expert than the one giving the uniform ratings. Study 2 provides a second test for this hypothesis using a sequential design. Preference for a financial analyst with superior track record is substantially reduced when participants

learn that he has rated three funds equally, even though information about the actual performance of the funds suggests that uniform ratings were more appropriate. In study 3, similarly, preference for a more accurate project manager is significantly reduced when participants are told that he had given equal ratings to three projects. Study 3 also rules out confidence as an alternative explanation.

1.4. The overwhelming power of summary ratings

The second reason for the robustness of the ratings pattern heuristic lies in the fact that, irrespective of pattern, summary ratings themselves are highly influential and often weighted more heavily than other diagnostic information in decision making (Chintagunta, Gopinath, & Venkataraman, 2010; De Langhe et al., 2016; Kostyra, Reiner, Natter, & Klapper, 2016). Mean user rating has been shown to be one of the strongest predictors of sales, decreasing the influence of other traditionally important cues such as brand name or price (Chintagunta et al., 2010; De Langhe et al., 2016; Kostyra et al., 2016).

Summary ratings are also processed easier than other diagnostic information such as text reviews (Chen, Hong, & Liu, in press) which could be equally or even more informative but take longer to read. It is plausible that options could differ at the level of individual dimensions or attributes, yet if relative weaknesses on some attributes are compensated by relative strengths on other, the critic should still give equal ratings. For example, job candidate A could have superior interpersonal skills, but candidate B could be more knowledgeable. If both attributes are equally important, a competent HR employee should give equal overall ratings to the two, and explain the difference at the individual-skills level in a written review or by providing skills/attributes sub-ratings.

Alternatively, a less competent critic may discriminate on peripheral attributes that are not diagnostic of overall quality, but are salient or easy to process (Castellan, 1973; Chinander & Schweitzer, 2003; Tsay, 2014). To use the example above, a less competent HR person may rate candidate A lower than candidate B because A is from out of state or is younger, even though both individuals are equally good on important attributes such as interpersonal skills and relevant work experience. Such errors in judgment are not uncommon; evidence from multiple domains, including medical decision-making (Poses, Cebul, Collins, & Fager, 1985), financial forecasting (Yates, McDaniel, & Brown, 1991), sports results forecasting (Andersson, Edman, & Ekman, 2005), music performances (Tsay, 2014), and agriculture (Gaeth & Shanteau, 1984) reveals that even experienced professionals may discriminate based on evidence that is less diagnostic but more salient or easier to process. In this sense, how ratings are formed (based on diagnostic vs. non-diagnostic evidence) is more relevant than the final pattern (uniform vs. varied). However, because of people's strong tendency to rely on overall rating assessments, we predict the following:

Hypothesis 2a.. *The ratings pattern heuristic remains influential even when process-based information (individual attribute sub-ratings) reveals that both critics are equally capable of discriminating at the attribute level or that the varied critic relied on non-diagnostic attributes.*

Hypothesis 2b.. *The ratings pattern heuristic disappears only when people are forced to assess the critic's ability to discriminate at the attribute level prior to making a final assessment.*

We test this proposition in studies 4a, 4b, and 5. We find that sub-ratings of individual attributes, prominently shown before overall ratings, do little to reduce preference for a varied critic over a uniform one (study 4a). Only when participants are explicitly directed to evaluate the critics' discriminating ability at the attribute level, the bias against the uniform critic disappears (study 4b). In study 5, we pit uniformity at the summary ratings level against ability to consider diagnostic criteria. We show that unless explicitly prompted, participants judge a critic giving uniform overall ratings as less expert even when these ratings are based on an assessment of more relevant and important attributes.

Finally, we also examine whether critics correctly anticipate a uniform ratings bias and adjust their ratings accordingly. Unlike observers who only have access to critics' ratings, critics have access to more information to assess the appropriateness of their judgments. As a result, they may be less sensitive to the pattern of ratings that they produce when predicting how others would judge their expertise.

Hypothesis 3.. *The impact of uniform ratings on expertise will be larger from judges' perspective than from critics' perspective.*

We test this hypothesis in our final study, in which we compare people's perceptions of uniform critics with the critics' own intuition of the impact of their uniform ratings on others' judgements.

1.5. Pilot study: prevalence of phenomenon

Our research investigates the impact of providing equal (vs. varied) evaluations on perceptions of expertise. To establish the organizational relevance of our research, we first conducted a pilot study, designed to assess the prevalence of the studied phenomenon. We asked individuals with leadership experience how often they had to evaluate equally attractive options and how they typically behaved in such situations.

1.5.1. Method

One hundred and two participants recruited from the Prolific online research panel (51% female, M_{age} = 36.58, SD_{age} = 9.78) took part in this survey in exchange for a payment. Participants were from the US and the UK with full or part-time employment status, and with experience in a leadership position or position involving supervisory duties.

Participants were asked to indicate whether they had ever been in a situation where they had to evaluate (a) alternative strategies/courses of action, (b) alternative projects/contracts for their company to take on, (c) alternative products (such as software, equipment, supplies, etc.) for their company to purchase, and (d) alternative service providers (the four categories were presented in counterbalanced order). For each of these categories, if the participant answered "no," they were taken to the next category (or to the demographics section at the end of the survey). If they answered "yes," they were asked if the alternative options happened to be equally good sometimes. Again, if they answered "no," they were taken to the next category. If they answered "yes," they were asked four more questions: how often, in their experience, the alternatives were equally good (1 = very rarely; 7 = very often); how common such situations were (1 = not at all common; 7 = quite common); how often, when the alternatives were equally good, they gave the very same overall evaluation, instead of trying to differentiate them (1 = always gave the same evaluation; 2 = frequently gave the same evaluation; 3 = frequently gave at least slightly different evaluation; 4 = always gave at least slightly different evaluation). Participants were also asked to give one or two examples of a similar situation from their own experience. In the end, all participants were asked to indicate when, in general, they thought people were more likely to be perceived as experts: when they gave different evaluations to the alternatives, when they gave the same evaluation, or no difference. They also indicated how often they had to "make decisions for/on behalf of others in the organization," and how often they had to "evaluate strategic options, people, or products" (1 = "never," 7 = "all the time").

1.5.2. Results and discussion

Seventy-six percent of respondents indicated they often had to make decisions for or on behalf of others (M = 4.95) and evaluate strategic options, people, or products (M = 4.94).

Ninety-six percent of participants indicated they had experience evaluating alternatives in at least one of the four listed categories (75% for strategies, 52% for projects/contracts, 65% for products, and 54%

for service providers). Importantly, among these, 79% on average indicated the alternatives had sometimes been equal in quality (80% for strategies, 71% for projects/contracts, 86% for products, and 80% for service providers). Participants further thought that such situations happened frequently (M = 5.16) and were relatively common (M = 5.27). Forty-one percent stated that when presented with equal alternatives, they frequently gave them the same overall evaluation, 31% said they always gave the same evaluation, 20% said they frequently gave at least slightly different evaluations, and only 8% said they always gave slightly different evaluations. However, when asked when, in general, one was more likely to be perceived as an expert, 46% said when they gave different evaluations, 19% said when they gave the same evaluation, and, importantly, 35% thought it didn't matter either way. As we will see in our final study, people indeed underestimate the negative impact of uniform ratings.

In sum, results from our pilot study suggest that encountering alternatives that are of equal quality is relatively common in the workplace, and that in such situations most people tend to assign equal ratings. Having established this, we next explore the implications of this phenomenon for perceptions of expertise.

2. Study 1: ratings pattern vs. objective probability

Study 1 was designed to provide initial evidence consistent with the ratings pattern heuristic. We also wanted to test our first hypothesis, namely that the association between expertise and discrimination is so salient that strong accuracy cues are discounted in the presence of uniform ratings. We chose wine as the product category, since people often turn to wine critics for advice before making a decision. We gave participants information, in the form of base rates, indicating that a sample of three wines was much more likely to be of the same, rather than different, quality (study 1a) or type (study 1b). We then contrasted participants' judgments of the likelihood that the wines were of the same (vs. different) quality or type against participants' judgments of the expertise of two critics who had presumably sampled and judged the wines in a blind taste test.

2.1. Study 1a

2.1.1. Method

2.1.1.1. Sample and design. We initially ran two exploratory studies on Mechanical Turk with 60 participants each. The differences between conditions were large and consistent with our hypothesis, so we sought to replicate them in study 1a (and 1b), using similar design, but a larger sample size and a different population (college students). We aimed for a minimum of 80 participants (40 per condition) and ended data collection at the end of the day, which resulted in a few additional participants. Eighty-five undergraduate students from a large university (54% female, M_{age} = 20.35, SD_{age} = 1.37) completed this study for partial course credit. Participants were randomly assigned to one of two conditions: probability judgment condition and expertise judgment condition.

2.1.1.2. Procedure. All participants were told that three wines were randomly selected from a collection of 100 wines, 70% of which were 4-star wines (based on a 5-star scale), 15% were 3-star wines, and 15% were 5-star wines. Then, one half of respondents (probability judgment condition) were asked to indicate which of the following three outcomes was more likely: (a) all three wines were 4-star wines, (b) one wine was a 4-star wine, another was a 3-star wine, and one was a 5-star wine, or (c) "a" and "b" were equally likely. The other half of respondents (expertise judgment condition) were additionally told that two individuals, Jack and Paul, had tasted and rated the three wines in a blind test. Jack had given all three wines a 4-star rating, whereas Paul had given ratings of 4, 3, and 5 stars (we counterbalanced whether Jack or Paul was the uniform critic). Respondents in this condition were asked to indicate who was likely to be a better wine expert: (a) Jack, (b) Paul, or (c) "a" and "b" were equally likely. They were also asked to explain their reasoning.

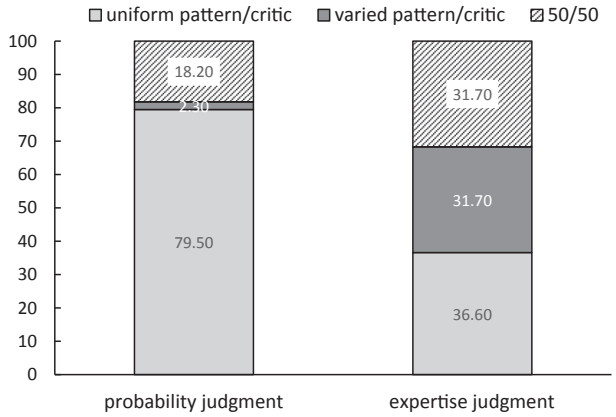


Fig. 1A. Study 1a: Percentage of participants choosing each of the three options.

2.1.2. Results and discussion

In the probability judgment condition, 79.50% of respondents thought that three 4-star wines was the most likely outcome. In contrast, in the expertise condition, only 36.60% of respondents thought that the critic who rated the three wines as 4 stars was more of an expert ($\chi^2(1) = 14.87$, $p < .001$; Please see Fig. 1A). These results provide initial evidence that a uniform pattern of ratings is taken as a signal of lack of discriminating ability and inferior expertise which can offset probability information about the likely distribution of the three wines. Although the vast majority of participants in the probability judgment condition acknowledged that it was more likely that all three wines were 4 stars, only a small group was willing to fully consider this piece of information when judging expertise.

The simple design of study 1a allowed us to directly contrast the use of probabilistic information with and without reference to critics' ratings variance. However, one could still argue that wine ratings are subjective and that the fact that 70% of the wines in the collection were described as 4-star does not mean that 4 stars was indeed the most appropriate rating. Perhaps a judge with a more discriminating palate could still discern qualitative differences. In study 1b we test whether this heuristic is used in a different type of judgment, which also reflects the ability to discriminate but is less subjective – the ability to identify the correct *type* of wine.

2.2. Study 1b

2.2.1. Method

2.2.1.1. Sample and design. Sample size was determined in a manner similar to that of study 1a. Participants were randomly assigned to one of two conditions: probability judgment and expertise judgment. Ninety-five undergraduate students from a large university (57.9% female, M_{age} = 20.51, SD_{age} = 1.46) completed this study for partial course credit.

2.2.1.2. Procedure. All participants were told that three wines were randomly selected from a collection of 100 wines, 70% of which were Chardonnay, 15% were Pinot Gris, and 15% were Sauvignon Blanc. One half of respondents (probability judgment condition) were then asked to indicate which of the following three outcomes was more likely: (a) the three wines selected from the wine collection were all Chardonnay, (b) one wine was a Chardonnay, another was a Pinot Gris, and one was a Sauvignon Blanc, or (c) "a" and "b" were equally likely. The other half of respondents (expertise judgment condition) was further told that two

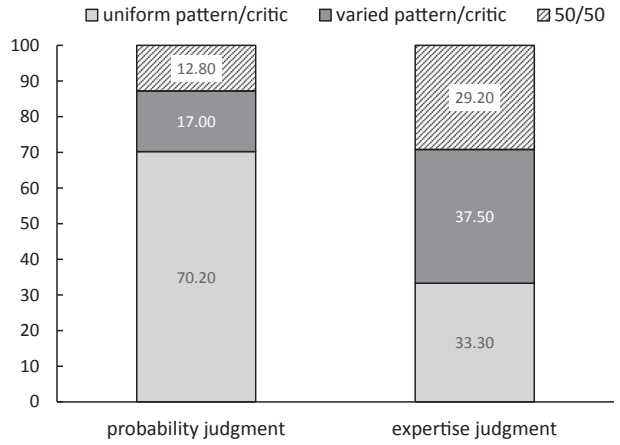


Fig. 1B. Study 1b: Percentage of participants choosing each of the three options.

individuals, Jack and Paul, had tasted and the classified wines in a blind test. Jack had identified all three wines as a Chardonnay, whereas Paul had identified them as a Chardonnay, a Pinot Gris, and a Sauvignon Blanc (Jack and Paul were counterbalanced). We asked respondents to indicate which of the two was likely to be a better wine expert: (a) Jack, (b) Paul, or(c) "a" and "b" were equally likely, and to briefly explain their reasoning.

2.2.2. Results and discussion

In the probability judgment condition, 70.20% of respondents thought that all Chardonnay was the most likely outcome. In contrast, in the expertise condition, only 33.30% of respondents thought that the critic who identified all three wines as Chardonnay was more of an expert ($\chi^2(1) = 12.30$, $p < .001$; Please see Fig. 1B).

Taken together, the results from studies 1a and 1b provide support for hypothesis 1. Participants overwhelmingly recognized that the uniform distribution was more likely. However, when judging expertise, only a minority considered the uniform critic as more knowledgeable, suggesting that uniformity completely offsets the probability information.

An analysis of the open-ended explanations that participants gave for their choice of the varied critic provides insight into the underlying mechanism. We asked four coders blind to the study hypotheses to go over the explanations given by participants in studies 1a and 1b and indicate whether they referred to ability to discriminate (yes/no). Pairwise agreement (average agreement between each pair of coders) was high (75%). Explanations were deemed to refer to ability to discriminate on 82% of occasions (e.g., "Jack couldn't tell the difference so he might not know much about wine," "A good wine expert would be able to taste the differences between the wines, whereas someone inexperienced would believe they were all the same," "Jack rated them all the same whereas Paul noticed differences between the three and was able to compare them against one another," etc.).

One could argue that the fact that individuals did not take probabilistic information into consideration when judging expertise is not particularly surprising, as individuals have been shown to neglect base rate information – a pattern referred to as the base rate fallacy (Kahneman & Tversky, 1973). Despite its popularity, considerable research suggests that the prevalence of base rate neglect has been largely overestimated (Ginosar & Trope, 1980; Koehler, 1996), and under many common circumstances, individuals do behave in line with what would be prescribed by the Bayesian model, especially when the individuating information has little diagnostic value (Davidson & Hirtle, 1990; Ginosar & Trope, 1980; Ofir, 1988). In this sense, our results indicate that ratings pattern is considered a strong individuating piece of information that may offset the value of probability information.

We should acknowledge that the greater dispersion in responses in the expertise-judgment conditions may be, at least in part, attributed to the greater complexity of the task in these conditions. Whereas participants in the probability-judgment conditions only had to consider which outcome was more likely, those in the expertise-judgment conditions had to integrate base rates with their own intuitions about discrimination ability and expertise. Further, because there was an objectively correct and relatively easy answer in the probability condition, but not in the expertise condition, it is reasonable to expect a greater concentration of answers in the former rather than the latter condition. In this sense, the choice data while consistent with our hypothesis cannot rule out task complexity as an alternative explanation. In study 2, and in all remaining studies, we designed the judgment tasks so they were of comparable complexity.

3. Study 2: ratings pattern vs. past performance

One may also wonder whether individuals would continue to disregard diagnostic information if their judgments had real implications for them and if the information was not presented in the form of base rates. Thus to provide a more robust test of Hypothesis 1, study 2 used a different context – financial decision making – in which the critic's past performance served as the diagnostic expertise information about the degree of difference among the options and actual monetary reward was at stake.

3.1. Method

3.1.1. Sample and design

We expected a moderate-to-large effect size ($d > .60$). Using G*power (Faul, Erdfelder, Buchner, & Lang, 2009), we estimated that for power of 80% and $d = .60$, 90 participants were required. We aimed to collect 100. One hundred and one participants from Mechanical Turk (44% female, $M_{age} = 32.30$, $SD_{age} = 11.39$) took part in this study for a baseline compensation and an opportunity to earn extra money based on their decision. The study employed a single-factor design with two conditions (ratings vs. no ratings). Participants were randomly assigned to one of these two conditions.

3.1.2. Procedure

The cover story introduced participants to a financial decision scenario in which they were presented with information about two analysts who they could later choose to consult in a subsequent financial decision task. All participants read that a year earlier the analysts had evaluated the attractiveness of three investment funds and had assigned each fund a star rating (out of 6 stars). All participants were also provided with information regarding the extent to which the analysts' ratings were consistent with the actual performance of the funds a year later.

In the no-ratings condition, participants were not shown the specific numerical ratings. They were only informed that one of the analysts' ratings were more consistent with the actual performance of the funds. Specifically, participants learned that one of the analysts had assessed the fundamentals of the three funds and had rated each one on a 6-star scale. The other analyst had assessed the fundamentals, as well as a number of additional indicators, and had also rated each fund using the same scale. Participants in this condition were also told that the actual performance of the funds one year later matched more closely the ratings of the first analyst (the one who had assessed only the fundamentals of the funds) than that of the second analyst (the one who had assessed the fundamentals of the funds as well as a number of additional indicators). Put simply, in the no-ratings condition, participants obtained the objective diagnostic information (i.e., past performance) and were *not* presented with the target non-diagnostic information (i.e., ratings pattern).

In the ratings condition, on the other hand, participants were shown

the numerical ratings. Specifically, participants read that according to one of the analysts, "the fundamentals of all three funds were solid" and he, therefore, expected them to perform equally well, as indicated by his ratings (the ratings were 5, 5, and 5 stars, on a 6-star scale). According to the second analyst, "even though the fundamentals of all three funds were solid, they differed on a number additional indicators," therefore he expected some degree of difference in performance, as indicated by his ratings (the ratings were 5, 4, and 6 stars, on a 6-star scale). Participants were then told that in reality, a year later the three funds had performed equally well—that is, closer to the predictions of the first analyst, who provided equal ratings. In doing so, in the ratings condition, we pitted objective diagnostic information about the difference in performance of the three funds (and thus, also, about the analysts' performance) against the target non-diagnostic information (i.e., ratings pattern) to assess the strength of the ratings pattern heuristic on expertise judgment.

We expected that reliance on past performance to judge expertise and choose a financial analyst (the right thing to do) would be observed in the no-ratings condition, but would be mitigated in the condition where the ratings pattern was also made available.

Participants evaluated the expertise of the two analysts by indicating who they thought was more knowledgeable, who they would be more likely to consult, and whose opinion they trusted more (expertise index; $\alpha = .96$), on seven-point bipolar scales from −3 (definitely [name of analyst A]) to 3 (definitely [name of analyst B]), where zero indicates indifference between the two analysts. We counterbalanced which analyst provided the uniform ratings. Next, participants were told that in the second part of the study they would be presented with a real financial decision task. Before reading a description of the task, participants learned that they could enlist the help of one of the two financial analysts and were asked to choose between them. This choice represented our main behavioral dependent measure. On the following screen, participants saw a list of five stocks and were asked to pick the one they thought would give them the best gain in the short term (three months). They were told that in three months they would be paid a bonus determined by the actual performance of the stock at that time.

Before deciding, participants saw the stock that the financial analyst they chose to help them had recommended, with a brief justification of the recommendation. In the end, participants indicated how much they knew about investing (1 = know very little, 7 = know a lot). Three months later, all participants received the same bonus payment. See detailed procedure in Appendix A.

3.2. Results

An ANCOVA on the Perceived Expertise index, with self-rated investment knowledge as a covariate, revealed a strong main effect of the ratings information ($F(1, 97) = 25.00$, $p < .001$). The covariate was also significant ($F(1, 97) = 6.82$, $p < .01$), as less knowledgeable participants were more likely to prefer the more accurate analyst. Apparently, more knowledgeable participants gave relatively more weight to the fact that one analyst used more information. We also tested for an interaction between knowledge and condition, but it was not significant ($p = .70$). Critically, participants' preference for the analyst with objectively superior past performance was clear when ratings patterns were *not* made available ($M = -1.60$, $SD = 1.23$), but preferences shifted away from him and toward the analyst with worse past performance but non-uniform ratings in the condition where the historically better performing analyst provided uniform ratings ($M = -.15$, $SD = 1.65$; $F(1, 97) = 25.00$, $p < .001$, $d = 1.00$). Note that in this and all subsequent studies responses were coded such that positive values indicate preference for the varied/less accurate critic, and negative values indicate preference for the uniform/more accurate critic.

A logistic regression on participants' consequential decision confirmed the robustness of the effect. In the no ratings condition, 84% of participants chose the analyst with superior past performance. However, consistent with the ratings pattern heuristic, this preference dropped to only 43% when the superior past performance information was accompanied by a uniform ratings pattern ($\chi^2(1) = 16.18$, $p < .001$). There were no effects for gender, age, or self-rated investment knowledge (all $p > .25$). Further analysis revealed that perceptions of expertise mediated the impact of ratings on choice of advisor. When condition and expertise are simultaneously used to predict choice, only the latter remains significant (condition: $b = .89$, $z = 1.33$, $p = .18$; expertise: $b = 1.71$, $z = 4.62$, $p < .001$). The 95% confidence interval of the indirect effect through expertise does not include zero (1.22, 4.97). Finally, in the end, 80% of participants followed the advice of their financial analyst.

3.3. Discussion

Study 2 provides further evidence for a ratings pattern heuristic in judgments of expertise. Participants judged a financial analyst significantly less favorably, as indicated in a subsequent choice, when they learned that he had given the same 5-star rating to three different funds. Importantly, this was observed even though participants knew that the funds had indeed performed equally well. This implies that for a significant share of participants, variance in ratings was a stronger signal of expertise than actual performance. The significance of this result is underscored by the fact that the choice had real financial consequences for participants in this study. Thus inattention or lack of engagement cannot explain the insufficient weight given to the accuracy criterion.

One may wonder whether correct prediction of one-year fund performance is a valid indication of expertise. After all, a correct prediction may be entirely due to luck. While it can be hard to disentangle the influence of luck and skills, for the purpose of our investigation what really matters is how participants interpreted this cue. Did they consider it to be a valid expertise signal? Our findings suggest that this is the case, since in the absence of ratings participants clearly perceived the match between predictions and performance to be a strong signal of expertise. Still, to test the robustness of the ratings pattern heuristic against a more unequivocal accuracy cue, study 3 relies on a context where skill is less likely to be confounded with luck.

4. Study 3: ratings pattern vs. past performance vs. confidence

We designed study 3 with two goals in mind. First, we wanted to provide a more conservative test of hypothesis 1, which poses that uniform ratings negatively impact perceptions of expertise even in the presence of accuracy cues that favor uniformity. Specifically, we wanted to use a less ambiguous accuracy cue that would be less attributable to luck. To that end, we replaced the financial scenario with an organizational one in which the accuracy cue was the performance of consulting projects. Second, we wanted to rule out confidence as an alternative explanation. Previous research has shown that highly confident judges are viewed as more knowledgeable, sometimes even when their accuracy is objectively lower (Price & Stone, 2004). It is possible that participants in our studies interpreted the uniform ratings as lack of confidence and thus judged the uniform critic to be less expert not because he was less discriminating but because he was less confident. To rule out this alternative explanation, we explicitly manipulated the critics' confidence. If confidence drives differences, the impact of ratings should disappear or be mitigated once confidence in judgment is explicitly manipulated. In contrast, if discrimination, and not confidence, is responsible for the impact of ratings, as we hypothesize, the effect of ratings pattern should remain significant regardless of expressed confidence. In other words, we expected a main for effect ratings (and possibility for expressed confidence), but no interaction.

4.1. Method

The design was similar to the one used in study 2, but with the added confidence manipulation. For that reason, we expected a medium-to-large effect size ($d > .25$, for a two-way ANOVA). This would require 158 participants for a comparison, which divided by 6 group (2 × 3) would lead to just 26. We decided to collect 45 participants per cell, which gives the analysis a power over 99%.

Two hundred and sixty nine respondents from Mechanical Turk (41% female, $M_{age} = 35.2$, $SD_{age} = 10.62$) completed this study in exchange for monetary compensation. Participants were randomly allocated to six conditions in a 2 (ratings vs. no ratings) × 3 (no confidence information, accurate rater more confident; accurate rater less confident) between-subjects design.

Participants were told that a business consulting company was considering new projects to take on and had two of their managers – John and Albert – evaluate the potential of three of these projects. In the no-ratings no-confidence information condition, participants were informed that the company decided to take on all three projects and that a year later the projects' performance was more in line with John's evaluations. In the no-ratings, accurate rater more confident condition, participants also learned that the rater whose evaluations were more aligned with actual performance had been 90% confident, whereas the other one had been 70% confident. In the no-ratings, accurate rater less confident condition the percentages were reversed.

In the ratings no-confidence information condition, prior to learning about the projects' performance, participants were shown the numerical ratings given by John and Albert (5-3-4 and 4-4-4, both out of 5). There was no information about confidence. In the other two ratings conditions, participants were shown the confidence percentages (90% vs. 70% and 70% vs. 90%). See Appendix B for details.

Next, all participants were asked to indicate who was more knowledgeable, who they would be more likely to consult, and whose opinion they trusted more (all on 7-point bipolar scales with the two raters at the end points; expertise index, $\alpha = .97$). In the end, everyone indicated who was more confident in his evaluations (definitely John vs. definitely Albert, on a 7-point bipolar scale).

4.2. Results and discussion

A two-way ANOVA on the expertise index yielded no significant interaction between ratings and expressed confidence ($F(2, 263) = .99$, $p > .30$) but two main effects: a main effect for the presence of ratings ($F(1, 263) = 36.56$, $p < .001$, $d = .75$) and a main effect for the confidence information ($F(2, 263) = 5.17$, $p = .006$, $d = .28$). Preference for the accurate rater decreased significantly in the presence of ratings ($M = -.38$, $SD = 1.66$), relative to the no-ratings conditions ($M = -1.50$, $SD = 1.37$). Preference for the accurate rater also decreased when he was less confident ($M = -.51$, $SD = 1.64$) relative to when he was more confident ($M = -1.22$, $SD = 1.53$; $F(2, 263) = 10.04$, $p = .002$), or relative to the no-confidence information conditions ($M = -1.03$, $SD = 1.63$; $F(2, 263) = 5.35$, $p = .02$). Replicating the results of study 2, in the ratings with no confidence information condition, participants were indifferent between the two raters ($M = -.33$, $SD = 1.79$, $t(44) = .22$). Please see Fig. 2.

A two-way ANOVA on the perceived confidence measure revealed main effects of ratings ($F(1, 263) = 4.16$, $p < .05$) and expressed confidence ($F(2, 263) = 228.09$, $p < .001$), but no interaction ($F(2, 263) = 1.50$, $p > .20$). The rater was also seen as more confident when he expressed high rather than low confidence (M = -2.28, $SD = 1.38$ vs. M = 2.24, $SD = 1.30$), or relative to the no-confidence information condition ($M = -.32$, $SD = 1.61$). Although the effect on perceived confidence was mostly driven by expressed confidence, ratings also had a small impact, as the more accurate rater was perceived as more confident in the absence of ratings information ($M = -.30$, $SD = 2.30$) than with ratings information ($M = .06$, $SD = 2.37$). Finally, an ANCOVA on expertise using perceived confidence as a covariate revealed a very similar pattern: main effects for ratings ($F(2, 262) = 32.13$, $p < .001$), confidence ($F(2, 262) = 8.12$, $p < .001$), and perceived confidence ($F(1, 262) = 58.63$, $p < .001$), but no interaction ($F(2, 262) = .65$, $p > .50$).

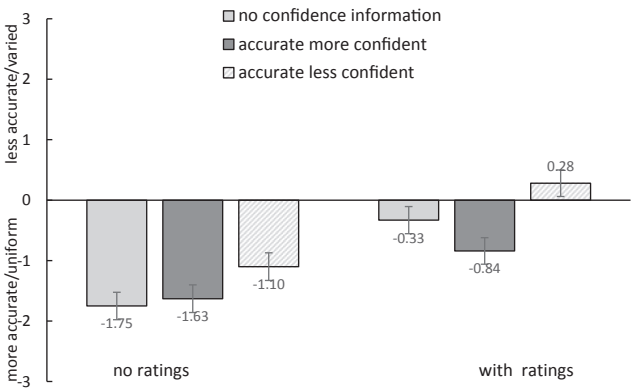


Fig. 2. Study 3: Perceptions of expertise as a function of ratings and confidence information. *Note*: Error bars represent confidence intervals.

In sum, study 3 replicated the pattern obtained in study 2, but using a different accuracy cue. The presence of ratings information significantly decreased preference for the accurate rater. In the absence of explicit confidence information, the ratings information made participants indifferent between the more and the less accurate raters. Expressed confidence had the predicted impact on preferences in that participants preferred the more confident rater. This effect, however, was independent of the ratings pattern and thus cannot explain the impact of ratings on perceived expertise.

We have argued that there are two reasons why the ratings pattern heuristic may be so resistant to other expertise diagnostic cues. The first reason involves a strong belief that uniform ratings signal inability to differentiate, which countervails the influence of more diagnostic cues. Consistent with hypothesis 1, preference for the more accurate critic significantly decreased (studies 1–3) and even flipped (studies 1a and 1b) when he provided uniform ratings. Participants' justifications (study 1) provided further evidence for the "expertise-discrimination" association and the negative inference people make when uniform ratings are presented.

The second reason for the robustness of the ratings pattern heuristic, formalized in hypothesis 2, refers to people's tendency to focus on the pattern of final ratings at the expense of information about the process of arriving at these ratings, although the latter serves to help justify the former. In our next studies, we seek evidence for hypotheses 2a and 2b.

5. Study 4a: summary ratings pattern vs. sub-ratings pattern 1

Because of people's strong tendency to rely on overall rating assessments, the ratings pattern heuristic should remain influential even when information indicates that both critics are equality capable of discriminating between options at the attribute level (hypothesis 2a). To further test the robustness of our findings, the current study also employed a procedure in which participants faced a consequential choice.

5.1. Method

5.1.1. Sample and design

Based on the choice results of study 2, we expected a strong preference for the varied rater in the control condition (around 85%, as in study 2) and a drop of approximately 30 percentage points, close to indifference, in the ratings conditions (again, as in study 2). Calculations using G*power (Faul et al., 2009) led to a minimum

sample of 82, so we rounded it up to 90.

We aimed to recruit a minimum of 90 participants from the available student pool. Ninety-six undergraduate students from a large university (57.3% female, M_{age} = 20.28, SD_{age} = 1.62) took part in this study in exchange for partial course credit. Participants were invited to lab session and were randomly assigned to one of two experimental conditions (sub-ratings vs. no sub-ratings).

5.1.2. Procedure

The cover story was that students taking part in a previous lab session had evaluated three math apps for kids (Wee Kids Math Path, Marble Math Junior, and Math Tales). Participants in the main study could see the score sheets presumably completed by two of the students in the previous session, and could choose one of the students to work with in a related future study (designing a science app for kids).

In the no sub-ratings condition, the score sheets featured only the overall ratings the students in the previous study had given to each app (on a 5-point scale from "very poor" to "excellent"). One of the students had given the same overall rating to all three apps (i.e., 4 = "good"), whereas the other had given varied overall ratings (3 = "average," 4 = "good," and 5 = "excellent").

In the sub-ratings condition, for each app, the score sheets also featured the ratings on four attributes (learning value, ease of use, entertainment values, and graphics), which were followed by overall ratings as in the no sub-ratings condition. Of importance, both raters displayed varied sub-ratings, thereby making it clear that both raters (uniform rating student and varied rating student) could and did discriminate at the attribute level. See Appendix C for details. After examining the score sheets, participants indicated which of the two students they would like to partner with, and briefly explained their choice.

5.2. Results and discussion

Nine students did not choose a partner (left the answer box blank or wrote something else, e.g., the name of one of the math apps) and were therefore excluded from the analysis. Participants' choices, coded as "1" (uniform rating student) or "0" (varied ratings student), were regressed on the presence of sub-ratings. As in the previous studies, there was a clear preference for the student who gave varied summary ratings; 84% chose this student over the one who gave a uniform rating. Further, the results from a binary logistic regression revealed that the presence of sub-ratings was not a significant factor ($\chi^2(1) = .12, p > .70$). In both conditions, the percentage of participants choosing the student who gave uniform ratings was very low (14.6% in the no sub-ratings condition, 17.4% in the sub-ratings condition).

These results are consistent with hypothesis 2a and provide further evidence for the strength of the ratings pattern heuristic, as it remained influential even when sub-ratings indicated that both critics were equality willing and capable of discriminating at the attribute level. Participants seemed to ignore the process through which the rater arrived at the summary ratings. This occurred despite the prominence of sub-ratings, which were presented before overall ratings and occupied three-quarters of the page.

6. Study 4b: summary ratings pattern vs. sub-ratings pattern 2

We proposed in hypothesis 2a that people spontaneously assess critics' ability to discriminate – and thus their expertise – from their summary ratings and do not give as much weight to discrimination at the sub-ratings level, unless explicitly prompted to do so (hypothesis 2b). In study 4b we seek evidence for this mechanism by asking participants to assess the critics' discrimination ability at the attribute level. We expect this prompt to eliminate the ratings pattern heuristic. Specifically, we expect it would lead participants to realize that the uniform critic is as discriminating – and, therefore, as expert – as the non-uniform one.

This study also addresses a potential alternative explanation. It is possible that uniform ratings may lead people to infer not only that the reviewer is less of an expert (i.e., for lack of skills or effort) but also that the ratings themselves are less "practical/helpful" to people who may want a more precise advice on what option to choose. In principle, then, one could argue that ratings pattern heuristic does not derive from direct inferences about the critic's expertise but from the fact that the critics, even if accurate, is less "helpful" (e.g., "I need the expert to tell me what to do!"). Study 4b attempts to rule out this possibility.

Finally, in this study we move away from the bipolar scales as they may lead to exaggerated differences between the two critics. Instead, in 4b we ask participants to evaluate the expertise of each critic separately. We also include a measure of perceived effort. This is done for exploratory purposes, even though we expect it to be highly correlated with the other expertise measures.

6.1. Method

6.1.1. Sample and design

Based on the results of our previous studies, we calculated the sample for an effect size of d = .60, which led to 90 participants. We ended up with a few more as presumably some participants completed the study in Qualtrics, but did not submit it in Mechanical Turk. Ninety-four participants from Mechanical Turk (30% female, M_{age} = 31.43, SD_{age} = 8.51) took part in this study for monetary compensation. The study employed a 2 (ratings pattern: uniform vs. varied; within) × 2 (discrimination prompt: present vs. absent; between) × 2 (category replicates: dishwasher vs. MBA program; within) mixed design.

6.2. Procedure

Participants were asked to share their opinion of critics who had evaluated options in two categories (dishwashers and MBA programs, in that order), using 5-point scales. In each category, one critic gave uniform summary ratings to the options, whereas the other gave varied summary ratings. Participants also saw how the critics had evaluated each of the options on two key attributes: cleaning power and water/energy efficiency for dishwashers; teaching quality and network opportunity for MBA programs (attribute sub-ratings). Both critics provided varied sub-ratings. The table below summarizes the procedure for the dishwasher category:

John's ratings:

	Brand A	Brand B	Brand C
Cleaning Power	3	5	4
Water/energy efficiency	3	1	2

Final ratings: Brand A: 3, Brand B: 3, Brand C: 3

Edward's ratings:

	Brand A	Brand B	Brand C
Cleaning Power	3	3	5
Water/energy efficiency	1	3	3

Final ratings: Brand A: 2, Brand B: 3, Brand C: 4

After that, a discrimination prompt manipulation took place. Participants in the prompt condition were asked to indicate which of the two critics seemed more capable of evaluating the cleaning power, the energy efficiency, and the overall quality of the dishwashers [teaching quality, network opportunity, and overall quality of the MBA programs] (in this order, on three separate 7-point bipolar scales, with the two critics at the two ends of the scale, and the mid-point indicating

indifference). The purpose of these questions was to make participants reflect on the fact that even though the uniform critic's final ratings were the same, he had given varied ratings to the specific attributes and was thus no less discriminating than the varied critic. Participants in the no-prompt condition did not see these questions and proceeded directly to the main dependent measures.

Next, all participants completed the main dependent measures. They indicated how capable and knowledgeable each critic was (1 = "not at all", 7 = "a lot") and how likely they were to consult each of the two critics in the future (1 = "not at all likely," 7 = "very likely"), if looking for advice on dishwashers [MBA programs] (expertise $\text{index}_{\text{dishwashers}}$, $\alpha = .85$; expertise $\text{index}_{\text{mba}}$, $\alpha = .92$). Participants also indicated how much effort each had invested in the review process (1 = "not at all," 7 = "a lot") and how useful and helpful the ratings of each of the critics would be if they were actually choosing among the three brands of dishwashers [MBA programs] (1 = "not at all," 7 = "a lot," usefulness $\text{index}_{\text{dishwashers}}$, $r = .82$; usefulness $\text{index}_{\text{mba}}$, $r = .89$).

6.3. Results

A repeated-measures ANOVA on the expertise index, with prompt as a between-subjects factor and ratings pattern (uniform vs. varied) and category replicate (dishwashers vs. MBAs) as within-subjects factors revealed a significant effect for ratings pattern ($F(1, 92) = 11.75$, $p = .001$), a significant effect for category replicate ($F(1, 92) = 7.52$, $p = .007$), and an interaction between presence of prompt and ratings pattern ($F(1, 92) = 6.74$, $p = .01$). When no prompt highlighting discriminating ability at the attribute level was available, findings were consistent with the ratings pattern heuristic (dishwashers: $M_{\text{varied}} = 5.41$, $SD = 1.07$ vs. $M_{\text{uniform}} = 4.09$, $SD = 1.26$, $F(1, 92) = 13.02$, $p < .001$, $d = 1.13$; MBAs: $M_{\text{varied}} = 5.43$, $SD = 1.16$ vs. $M_{\text{uniform}} = 4.49$, $SD = 1.38$, $F(1, 92) = 6.60$, $p = .01$, $d = 1.27$). However, when participants were prompted to evaluate the critics' ability for judging each attribute, the preference for the varied critic vanished (dishwashers: $M_{\text{varied}} = 4.90$, $SD = 1.46$ vs. $M_{\text{uniform}} = 4.74$, $SD = 1.43$, $F(1, 92) = .18$, *ns*; MBAs: $M_{\text{varied}} = 5.04$, $SD = 1.30$ vs. $M_{\text{uniform}} = 4.89$, $SD = 1.46$, $F(1, 92) = .15$, *ns*). Please see Fig. 3.

A repeated-measures ANOVA on the usefulness index revealed a similar pattern: a main effect for ratings pattern ($F(1, 91) = 16.11$, $p < .001$; there was one missing data point for this measure), a main effect for category replicate ($F(1, 91) = 10.18$, $p = .002$), and a significant interaction between ratings pattern and prompt ($F(1, 91) = 11.42$, $p = .001$). Without a prompt, participants perceived the uniform ratings as less useful than the varied ratings (dishwashers: $M_{\text{varied}} = 5.42$, $SD = 1.20$ vs. $M_{\text{uniform}} = 3.66$, $SD = 1.62$, $F(1, 91) = 18.13$, $p < .001$; MBAs: $M_{\text{varied}} = 5.45$, $SD = 1.35$ vs. $M_{\text{uniform}} = 4.10$, $SD = 1.55$, $F(1, 91) = 10.67$, $p < .001$). The presence of a prompt eliminated this difference (dishwashers: $M_{\text{varied}} = 4.79$, $SD = 1.53$ vs. $M_{\text{uniform}} = 4.61$, $SD = 1.65$, $F(1, 91) = .18$, *ns*; MBAs: $M_{\text{varied}} = 5.08$, $SD = 1.33$ vs. $M_{\text{uniform}} = 4.99$, $SD = 1.48$, $F(1, 91) = .05$, *ns*).

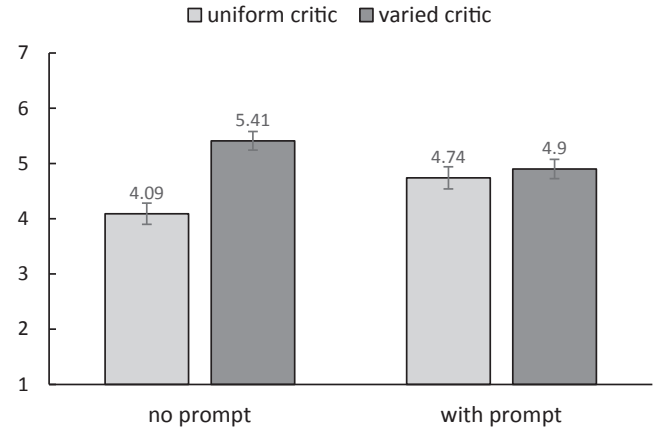


Fig. 3. Study 4b: Perceptions of critic expertise (dishwasher category). *Note*: Error bars represent confidence intervals.

Finally, a repeated-measures ANOVA on the effort measure yielded the same pattern: a main effect for ratings pattern ($F(1, 91) = 5.55$, $p = .02$), a main effect for category replicate ($F(1, 91) = 13.99$, $p < .001$), and a significant interaction between ratings pattern and prompt ($F(1, 91) = 4.25$, $p < .05$). Table 1 displays all means, standard deviations, the two critical simple main effects as well as the interaction term for each replicate on each measure.

6.4. Discussion

Study 4b confirmed hypothesis 2b. The ratings pattern heuristic vanished only when information about the critic's discriminant ability was available and participants were prompted to consider it. Otherwise, participants relied on the pattern (varied vs. uniform) of the summary ratings.

Also, the aversion to uniform ratings was not driven by the fact that the uniform critic was not helping the decision process by not providing a specific recommendation on which option to choose. If this were the case, then prompting participants to consider internal ratings should *not* have eliminated the effect. After all, even with the different internal ratings, it is still not clear which alternative, if any, the uniform critic would recommend. We note however that our manipulation increased also participants' perceptions of the usefulness of the uniform critic's ratings. This is not unreasonable. When the uniform critic is viewed as more knowledgeable, the information that he provides should also be perceived as more useful even if it still does not point out which option is superior. Still, to more conclusively rule out an alternative, or complementary, explanation based on the specificity of the recommendation, we ran a post-test in which we explicitly asked participants to evaluate the ability of the critics to provide specific recommendation.

6.4.1. Post-Test

Eighty-two participants from Mechanical Turk (54% female, $M_{\text{age}} = 35$) took part in this post-test in exchange for monetary compensation. Participants were presented with the same dishwasher and MBA program scenarios as those used in the main study. Similar to the main study, half of participants were assigned to a no-prompt condition, and the other half to a prompt condition. Participants in the prompt condition were first asked to indicate which of the two critics was more capable of evaluating the dishwashers (MBA programs) on the separate attributes, and then proceeded to the main dependent measure. Participants in the no-prompt condition did not see this question and proceeded directly to the main dependent measure. For the main dependent measure, all participants indicated which of the two critics had provided a more specific recommendation. As expected, participants viewed the varied critic as providing a more specific recommendation both for dishwasher (70.7%, $\chi^2 = 13.22$) and MBA programs (74.4%, $\chi^2 = 17.76$, $p < .001$). Critically, these perceptions were independent of the presence of a prompt (both $p > .70$). That is, in the prompt condition, where preference for the varied critic disappeared in study 4b, people still perceived the varied rating as a clearly more specific recommendation than the uniform rating. Thus, "helpfulness" of the rating, as an alternative account, cannot explain our findings.

7. Study 5: ratings pattern vs. appropriateness of criteria

We designed study 5 to further test the strength of the ratings pattern heuristic against a different type of expertise-diagnostic process information, namely the relevance of the attributes on which the critics

Table 1
Study 4b: Expertise, usefulness, and effort.

		No prompt		Prompt		p.
		Uniform (N = 49)	Varied (N = 49)	Uniform(N = 45)	Varied(N = 45)	
Expertise index	Dishwashers	4.09 (1.26)	5.41 (1.07)	4.74 (1.43)	4.90 (1.46)	.006
	MBAs	4.49 (1.38)	5.43 (1.16)	4.89 (1.46)	5.04 (1.30)	.064
	Dishwashers	3.66 (1.62)	5.42 (1.20)	4.61 (1.65)	4.79 (1.53)	.001
Usefulness index	MBAs	4.10 (1.55)	5.45 (1.35)	4.99 (1.48)	5.08 (1.34)	.007
Effort	Dishwashers	4.21 (1.58)	5.19 (1.23)	4.71 (1.38)	4.64 (1.37)	.017
	MBAs	4.71 (1.38)	5.40 (1.09)	4.98 (1.55)	5.16 (1.24)	.208

Note: Standard deviations are in parentheses. Shadowed areas indicate evidence of ratings pattern heuristic (i.e., significantly smaller mean for the uniform than the varied rating condition within a given row). "p." corresponds to the p value of the interaction term.

had based their overall ratings (hypothesis 2a). We provided participants with information suggesting that the uniform critic had taken mostly relevant attributes into consideration when arriving at the final ratings, whereas the varied critic had been strongly influenced by less relevant attributes. Consistent with hypotheses 2a and 2b, we expected that, unless explicitly prompted to consider this piece of information, participants would continue to favor the varied critic. Finally, we tested the use of the heuristic with yet another set of product categories.

7.1. Method

7.1.1. Sample and design

We expected a large difference between conditions because we anticipated a preference reversal in the prompt condition. We estimated an effect size of d = .70, which implies a required sample of 68. We asked for a few more. Seventy-eight participants from Mechanical Turk (51% female, M_{age} = 33.27, SD_{age} = 11.52) took part in this study in exchange for a payment. This study used a 2 (prompt: general impression vs. process prompt; between) × 2 (ratings pattern: uniform vs. non-uniform; within) × 3 (product category: body lotions vs. blenders vs. toothpaste; within) mixed design.

7.2. Procedure

Participants were asked to share their opinion of critics who had evaluated three products in three different categories (body lotion, blenders, and toothpaste), using 5-star scales. They learned that in each category, the critics rated the products on the same three attributes and provided an overall product evaluation. Then, they were presented with a table showing each critic's attribute ratings, followed by the overall rating. The final overall ratings of one of the critics were uniform (i.e., 4, 4, and 4), whereas those of the other critic were varied (e.g., 5, 3, and 2). The attribute sub-ratings of the two critics exhibited a comparable degree of variance. Furthermore, two of the attributes in each category were important, whereas the third attribute was peripheral. For example, body lotions were rated in terms of moisturizing properties and non-greasiness on skin – two diagnostic attributes – as well as bottle shape and design, a peripheral attribute. Attribute importance was established in a pre-test with 78 participants from the same subject pool as those in the main study. In each category, the peripheral attribute was evaluated as significantly less important than the other two attributes (lotions: M_{bottle_shape} = 2.33 vs. $M_{moisturizing\_property}$ = 6.26, $F(1, 77) = 267.51$, $p < .001$; M_{bottle_shape} = 1.94 vs. $M_{non\text{-}greasy}$ = 5.92, $F(1, 77) = 198.00$, $p < .001$; blenders: M_{color_range} = 2.69 vs. $M_{blending\_power}$ = 6.22, $F(1, 77) = 197.79$, $p < .001$; M_{color_range} = 2.69 vs. $M_{settings}$ = 5.46, $F(1, 77) = 132.41$, $p < .001$; toothpaste: $M_{flavors}$ = 3.48 vs. $M_{cavity\_protection}$ = 6.42, $F(1, 77) = 147.98$, $p < .001$; $M_{flavors}$ = 3.48 vs. $M_{breath\_freshening}$ = 6.68, $F(1, 77) = 116.45$, $p < .001$).

Importantly, the final ratings of the uniform critic were obtained by taking an average of the two important attributes, ignoring the peripheral attribute rating. In contrast, the final ratings for the varied critic were strongly influenced by the peripheral attribute. See the stimuli for toothpaste below. All materials are presented in Appendix D.

Mark's ratings:

	Brand A	Brand B	Brand C
Cavity protection	4	5	3
Breath freshening	5	4	4
Range of flavors	3	1	5

Final ratings: Brand A: 3, Brand B: 2, Brand C: 4.

James' ratings:

	Brand A	Brand B	Brand C
Cavity protection	4	5	3
Breath freshening	4	3	5
Range of flavors	3	2	5

Final ratings: Brand A: 4, Brand B: 4, Brand C: 4

After seeing the ratings in one of the categories, participants were asked, in an open-ended question, to either describe their overall impressions of each critic (general impression condition) or to describe how each critic had arrived at their overall ratings for each product (process prompt condition). The latter question aimed at drawing participants' attention to the weight the critics had given to the diagnostic and peripheral attributes. Next, participants indicated how knowledgeable and capable each critic was (1 – Not at all, 7 – Extremely), and how likely they were to consult him (1 – Not at all likely, 7 – Very likely; perceived expertise index, all α's > .86). This procedure was then repeated for the other two categories. The order of presentation of critics (uniform vs. non-uniform) and categories was fully randomized.

7.3. Results and discussion

A repeated-measures ANOVA on the perceived expertise index revealed a significant two-way interaction between prompt type and ratings pattern ($F(1, 76) = 21.92$, $p < .001$). Category had no impact. When asked to provide overall impressions of the critics, participants judged the varied critic to have higher expertise (M = 4.75, SD = 1.01) than the uniform critic (M = 3.88, SD = 1.22, $F(1, 37) = 9.51$, $p < .005$). In contrast, when prompted to think about how each critic formed their final ratings prior to judging expertise, the effect reversed (M_{varied} = 3.69, SD = 1.54 vs. $M_{uniform}$ = 4.82, SD = 1.22, $F(1, 39) = 12.57$, $p < .001$). Please see Fig. 4. Results for each of the three product categories are reported in Table 2.

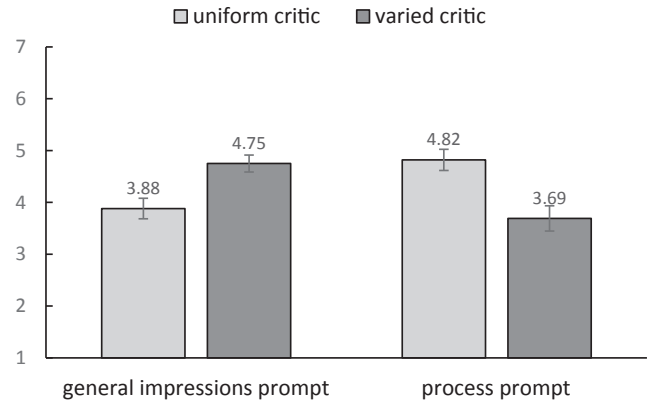


Fig. 4. Study 5: Perceptions of expertise (across all categories). *Note*: Error bars represent confidence intervals.

Table 2
Perceptions of expertise in study 5.

		General impression prompt		Process prompt	
		Uniform (N = 38)	Non-uniform (N = 38)	Uniform (N = 40)	Non-uniform (N = 40)
Expertise index	Lotions	3.80 (1.54)	4.68 (1.40)	4.65 (1.42)	3.81 (1.73)
	Blenders	4.03 (1.38)	4.76 (1.23)	5.01 (1.35)	3.48 (1.67)
	Toothpaste	3.83 (1.39)	4.81 (1.31)	4.80 (1.46)	3.80 (1.69)

Note: Standard deviations are in parentheses.

Results from the general-impression condition replicated those of studies 4a and 4b, indicating that the presence of sub-ratings is not enough to eliminate reliance on the ratings pattern heuristic. Even though both critics discriminated at the attribute level, participants continued to view the non-uniform critic as more knowledgeable. This pattern however was reversed when they were prompted to consider how each critic formed their final ratings. It is noteworthy that the same information was available in both conditions, but participants relied on the pattern of overall ratings to make their judgments, unless directed to think about how these ratings were formed.

8. Study 6: critics vs. judges

In studies 1–5, we presented participants with ratings from two critics and asked them to judge their expertise. In our final study, we decided to expand our investigation to also examine whether critics themselves are aware of the ratings pattern heuristic. We had proposed that the impact of uniform ratings on expertise is larger from the judges' perspective than from critic's perspective (hypothesis 3). To test this proposition, we manipulated whether the pattern generated by critics was uniform or varied and examined the reactions of others (who we refer to as "judges"), as well as the critic's own predictions about these reactions. At a methodological level, we also provide a test of our hypothesis in a fully between-subjects design, where judges were presented with ratings from just one critic.

In order to implement this study, we needed to take two issues into account. First, as in a real world situation, we wanted critics to be more knowledgeable than judges. To achieve this, we asked Indian Mturk participants to rate the prestige of a list of Indian universities, and we asked American Mturk participants to judge the Indian participants' expertise, based on their ratings. Thus the Indian participants served as critics, as they have superior knowledge about Indian universities than American participants who served as judges.

The second issue to consider was how to assign critics to different conditions. We wanted critics to provide their own ratings, but we also wanted to compare uniform to varied ratings. We accomplished this by creating a two-step procedure. First, we asked participants to rate several universities from a list. From this list, we dynamically selected a set of universities to which the critics had assigned equal ratings (uniform condition) or different ratings (varied condition) and presented them again to the critics, explaining that they would be judged based on how they rated these four universities. We explained to the critics that they could change their ratings. We then presented these sets of ratings to the judges and asked them to rate the expertise of the critics.

Following hypothesis 3, we expected judges to view a uniform critic as less knowledgeable than a varied one (the ratings pattern heuristic). Critics, however, could fail to fully take ratings pattern into account when predicting observers' reactions to their evaluations. This is because critics were knowledgeable about the universities and were thus likely to refer to the individual ratings rather than the pattern they formed when anticipating observers' reactions.

8.1. Method

8.1.1. Sample and design

Given the effect size observed in our previous studies, we used d = .70 to estimate sample size, resulting 34 participants per cell. Following our rule of at least 40 per cell, we asked for 80 participants in the critics group and 80 in the judges group. Although 81 participants from the critics group completed the study, 2 did not answer the key dependent variables resulting in usable sample at 159 (79 critics and 80 judges).

Participants were members of Mechanical Turk from India (critics) or the United States (judges). The data were collected in two stages: we first collected data from the critics and then used their responses to create the stimuli for the judges. The design was 2 (role: critic vs. judge) × 2 (pattern: uniform vs. varied) between-subjects. The study was run on Qualtrics with additional java script programming to select the set of universities in the critics condition as described below.

Participants in the critics condition (n = 79, 34% women, M_{age} = 32.01, SD_{age} = 7.87) were asked to rate the level of prestige of 16 Indian universities ("1-Not prestigious," "2-Slightly prestigious," "3-Prestigious," "4-Very prestigious"). They were then told that we were also interested in how others (American participants from Mechanical Turk) would judge them based on their ratings. We further explained that in order to keep this second task manageable, the American participants would only see how they had evaluated four of the universities. Next, the critics were presented with a subset of four of the universities and the ratings they assigned to them earlier, and informed that their expertise would be judged based on these ratings. They were also informed they could revise and change any of the ratings or leave them as they were. The ratings were presented using the same original multiple choice questions with the four prestige categories (see Appendix E for a screenshot). In the uniform critic condition, the program identified the most common rating and presented four universities that had been given this rating. Since there were originally 16 universities and four possible ratings, we were guaranteed to have a group with at least four universities that had the same rating. In the varied critic condition, the program looked for one rating from each category. If there were no four different ratings, then the most common rating was repeated. After reviewing and potentially changing their ratings for the set of four universities, participants were asked whether they had

changed their ratings (yes, no) and why. Next, they predicted how they expected the American participants to evaluate their expertise after seeing their ratings (1-Not knowledgeable at all/Not thoughtful ratings/No expertise at all, 7-Very knowledgeable/Very thoughtful ratings/A great deal of expertise, $\alpha = .87$).

We copied the four ratings from each critic into a separate question in our survey. The questions were grouped either in a uniform block or a varied block based on the condition assigned to the critic who generated those ratings. The uniform block had 40 questions (40 critics), while the varied block had 39. From each block, Qualtrics randomly and evenly selected a question for each participant in the judge condition (i.e., trying to keep an even number of participants per question). Each participant in the judge condition (n = 80, 35% women, $M_{age} = 34.11$, $SD_{age} = 10.93$) saw the ratings provided by one critic and evaluated the critic's expertise using the same seven-point scales as those used by the critics earlier ($\alpha = .95$).

8.2. Results and discussion

Although critics were given a chance to revise their ratings, very few of them did (14%) and this difference was not affected by whether they were in the uniform or varied condition ($\chi^2(1) = .14$, $p > .25$).

An ANOVA on expertise revealed a main effect for role, as judges evaluated the critics more negatively than critics anticipated ($M_{critic} = 5.47$, $SD = .93$ vs. $M_{judge} = 4.42$, $SD = 1.38$, $F(1, 155) = 32.98$, $p < .001$), as well as a main effect for pattern ($F(1, 155) = 6.25$, $p < .05$) and an interaction ($F(1, 155) = 4.46$, $p < .05$). Consistent with our previous results, judges viewed the varied critic was much more knowledgeable than the uniform one ($M_{varied} = 4.84$, $SD = 1.01$ vs. $M_{uniform} = 4.00$, $SD = 1.58$, $F(1, 155) = 10.71$, $p < .001$, $d = .64$). This discrepancy however was not anticipated by critics, who were largely insensitive to the potential impact of their ratings pattern on the judges' perceptions ($M_{varied} = 5.50$, $SD = .91$ vs. $M_{uniform} = 5.43$, $SD = .97$, $F(1, 163) = .07$, $p > .25$). Please see Fig. 5.

Results from study 6 provide converging evidence for a strong ratings pattern effect in judgments of expertise. We replicated the effect observed in our previous studies in a between-subjects design, which speaks to the robustness of the heuristic. The results from study 6 also reveal that, in spite of this robustness, critics seem to be unaware of the effect of their own ratings on others' perceptions.

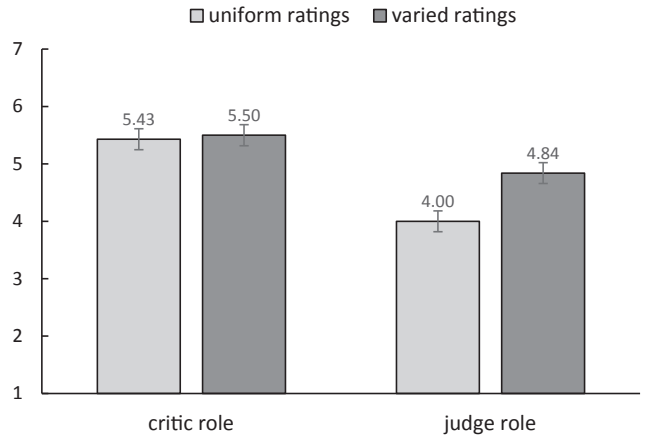


Fig. 5. Study 6: Perceptions of expertise. *Note*: Error bars represent confidence intervals.

9. General discussion and conclusions

Situations where individuals evaluate multiple alternatives are quite common. HR managers evaluate multiple job applicants, project managers assess different contracts to take on, investors rate various funds, and doctors and pharmacists may have alternative treatment options available to them. When the assessed alternatives are truly different in quality, it is important that the rater acknowledges these differences (Alba & Hutchinson, 1987; Weiss & Shanteau, 2003). It is equally important, however, to acknowledge when the alternatives are similar, or when they differ only on peripheral and non-diagnostic attributes but are equivalent on the important attributes (Gaeth & Shanteau, 1984; Weiss & Shanteau, 2003).

In this research, we demonstrate that whereas the first ability is well appreciated, the latter one is not, and is often taken as a sign of inferior knowledge. We propose that this happens for two reasons. First, people perceive a strong relationship between expertise and the ability to discriminate, and cues signaling discriminating ability are more salient than cues about the likely degree of difference among the rated items. Second, people focus predominantly on the summary pattern of ratings and ignore the pieces of information the critic relied on to arrive at these ratings. Results from eight studies provide converging evidence for these hypotheses.

In studies 1a and 1b, participants preferred a wine critic who had rated three wines differently or had identified them as being of different type (vs. the same), even though participants were given information suggesting that a uniform distribution was much more likely. Studies 2 and 3 replicated this effect in organizational contexts. In study 2, participants' preferences for a financial analyst with accurate past performance were substantially reduced when they learned that he had (correctly) predicted equal performance for three investment funds. Similarly, in study 3, a project manager was seen as significantly less competent when he assessed the potential of three projects to be comparable (rather than different), even though the projects did perform equally a year later. This study also ruled out confidence as an alternative explanation. In studies 4a and 4b, information indicating that the uniform critic was as discriminating as the non-uniform one at the attribute level attenuated the effect, but only when participants were explicitly asked to judge discrimination at the attribute level. Preferences were only reversed when the non-uniform critic had discriminated on non-diagnostic attributes *and* participant were prompted to consider the critics' criteria (study 5). Finally, study 6 revealed that when assuming the role of critics, individuals are largely unaware of the heuristic.

9.1. Theoretical and empirical contributions

The present research contributes to the literature on perceptions of expertise. Expertise is highly relevant in an organizational context. Organizations make decisions, based on the perceived expertise of sources, such as consultants, on a daily basis. Expertise is also a key determinant of the perceived value of an advisor and the degree of advice utilization (Bonaccio & Dalal, 2006). Understanding the factors that influence perceptions of expertise is thus crucial. Previous research has examined the role of source characteristics, such as past performance, experience, certifications, or similarity to the self (Brown & Reingen, 1987; Feick & Higie, 1992; Gershoff et al., 2001; Shanteau et al., 2002). Communication style can also serve as a source for judging expertise: Individual who express higher confidence (Price & Stone, 2004; Sniezek & Van Swol, 2001) or who communicate in more abstract (vs. concrete) terms (Reyt, Wiesenfeld, & Trope, 2016) are seen as more expert. And the valence of reviews can also impact perceptions of the reviewer (Amabile, 1983; Folkes & Sears, 1977). We extend this

research by showing that the degree of variance of evaluations can also significantly impact perceptions of expertise, and can even overshadow other more diagnostic expertise information. This finding is particularly relevant in an organizational context where people often have access to the individual ratings provided by a single employee, consultant, or adviser, and not just the aggregate average of the ratings of multiple sources.

Our findings also provide insight into people's perception of what it means to be an expert. Expertise has been conceptualized as the ability to respond differently to different stimuli, but also similarly to similar stimuli (Shanteau et al., 2002; Weiss & Shanteau, 2003). Relatedly, experts are said to be more capable of making fine distinctions among options in a category, but also better able to identify higher-level commonalities (Alba & Hutchinson, 1987). Our results suggest that people are quite sensitive to signals that indicate discriminating ability, but give little weight to the ability to identify commonalities at a higher level.

Our paper also adds to the literature on heuristics (Gigerenzer & Todd, 1999; Tversky & Kahneman, 1974) by identifying a new heuristic and demonstrating its resistance to the influence of diagnostic information. We draw parallels to research on the confidence heuristic which shows that people consider a more confident judge as more expert even when environmental information suggests that his confidence level is not well calibrated (Price & Stone, 2004; Sniezek & Van Swol, 2001; Van Swol & Sniezek, 2005). Van Swol and Sniezek (2005) proposed that this may be the case because confidence is a cue that is easier to process and thus receives greater weight, relative to other more diagnostic cues such as accuracy. This is broadly consistent with the heuristic and systematic model of information processing (Chen & Chaiken, 1999) according to which easily processed judgment cues overshadow more cognitively demanding pieces of information. Along the same lines, we find that discriminating ability, deduced from the variance in a critic's ratings, overshadows other, often more important expertise judgment criteria, such as accuracy or diagnosticity of the evaluation criteria.

The heuristic documented in this paper may be part of a broader class of "outcome biases" (Baron & Hershey, 1988) whereby people tend to neglect information about the process that leads to a specific outcome, and instead focus exclusively on the outcome itself. For example, participants in one study were given exactly the same information about the process through which a surgeon reached a decision to operate on a patient (Baron & Hershey, 1988). Yet, participants evaluated the decision as better and the surgeon himself as more competent when the surgery outcome was favorable (rather than unfavorable). Similarly, in an organizational context, employees often neglect to consider the fairness of a procedure; instead, their level of satisfaction is based on how favorable to them the procedure turns out to be (Brockner & Wiesenfeld, 1996). In a similar way, we have found that individuals focus on the final ratings (i.e., whether they are uniform or not), and do not consider the pieces of information the critic relied on to arrive at these ratings.

It should be noted that the ratings pattern heuristic is not inherently unreasonable. In fact, in most cases where no other diagnostic information is available, the use of the heuristic would most likely lead to accurate judgments. Our focus however is on cases where the spontaneous application of the "equal ratings mean incompetence" rule no longer yields correct judgments, that is, where enough information is available for people to realize that uniform ratings are actually more accurate. Our position in this sense is closer to the classic work on clinical inference (Brunswick, 1952; Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964) which emphasized the adaptiveness of employing intuitive probabilistic means when making judgments in uncertain environments. It is also in line with the more recent tempered views of heuristics (see Gigerenzer & Gaissmaier, 2011 for a review), according to which heuristics "are not inherently good or bad, or accurate or inaccurate" (pg. 10, Gigerenzer & Brighton, 2009). Instead, the match between the heuristic and the environment under which it is being used determines its appropriateness.

9.2. Limitations and future research

Although we have investigated the use of the ratings pattern heuristic across multiple contexts, manipulations and designs, a few elements were kept fixed. Varying some of these parameters may increase our understanding of this phenomenon, as well as identify potential boundary conditions. First, we note that we used sets of three or four alternatives. Naturally, we expect that reliance on the heuristic would be stronger for larger sets, since a larger set makes equal ratings less likely and further suggests inability to discriminate. It might be more interesting to test the lower limits of the heuristic, i.e., would the ratings pattern effect be there even for sets comprised of only two options?

Similarly, the granularity of the rating scale should also make a difference. Throughout our studies we used 5-point or 6-point rating scales, the 5-point scale being the standards scale used on most consumer websites such as amazon.com, tripadvisor.com, yelp.com, etc., as well as by HR staff when evaluating job applicants (Dattner, 2016). In some situations, however, a finer rating scale, such as a 7-point, or even a 100-point scale, might be more appropriate. In such cases of really fine ratings scales, we would expect reliance on the heuristic to be even stronger.

Finally, we have demonstrated an aversion to uniformity, but one may also wonder whether there is a more general relationship between rating variance and perceptions of expertise. Would someone who sees small differences between alternatives be considered less knowledgeable than someone who sees larger differences? And would large differences appear more representative of the parent population and also signal greater ability to discriminate? Indeed, outcomes characterized by high variance are more informative (Coombs, Dawes, & Tversky, 1970) and useful (West, 1996; West & Broniarczyk, 1998). In this sense, it is possible that in the absence of other expertise cues, individuals take degree of variation as a sign of expertise. On the other hand, smaller differences may be taken as a sign of precision, and research suggests that people are more likely to follow advice that is more precise (Jerez-Fernandez, Angulo, & Oppenheimer, 2013). Future research could further explore these competing possibilities.

9.3. Managerial implications

The finding that people have a strong aversion toward uniform evaluations has implications for those who provide evaluations, as well as those who consume them. Those concerned with their expertise reputation are advised to avoid making judgments indicating that alternatives are qualitatively similar. When advisors believe that this is the best judgment they can provide, they should weigh the negative consequences that a uniform judgment may have on their reputation. This is relevant particularly since many rating systems, such as Morningstar's (an investment research firm that maintains the "Morningstar Risk Rating"), the ones used on consumer review sites such as Amazon or Trip Advisor, as well as those used by HR staff to evaluate job applicants, use a 5-star system, which doesn't allow for fine differentiation in terms of ratings. In such situations, a knowledgeable critic who reviews alternatives that are qualitatively similar could signal expertise by drawing attention to the finer differences in the text reviews or by making more explicit his or her ratings of specific attributes of the options. For example, an HR employee who has given the same "Very good" overall evaluation to 3 interviewees, should stress

the importance of the evaluations of specific attributes such as leadership ability, communication skills, etc.

Finally, it is worth noting that we have used clear and strong cues indicating that the uniform critic was more accurate (studies 1 and 2) and as capable of making a discriminating judgment (studies 4 and 5) and still preferences were strongly in favor of the non-uniform critic. In most contexts, accuracy cues may be considerably less clear and even unavailable. In this sense, the impact of the ratings pattern heuristic is likely to be even stronger in real life

Appendix A

Procedure of Study 2

One year ago, we sent a survey to several financial analysts. Among other things, we asked them to rate some popular funds.

In this study today, you will be presented with information about two of these analysts, who you may choose to help you in a later task. Last year, we asked them to rate the following funds: T. Rowe Price New Horizons, Oakmark International, and Third Avenue Real Estate Value. Below are the ratings that Jack and Paul gave to the three funds:

No-ratings condition

Jack said that he assessed the "fundamentals" of the three funds and gave his ratings (in a 6 stars scale).

Paul said that he assessed the "fundamentals" and a number of indicators and gave his ratings (in a 6 stars scale).

One year later, an assessment of the funds actual performance indicates that:

- -Jack's ratings were correct for all three.
- -Paul's ratings were correct for one of the three funds and differed by one star for the other two.

Ratings condition

Jack:

T. Rowe Price New Horizons:	5 stars (out of 6)
Oakmark International:	5 stars (out of 6)
Third Avenue Real Estate Value:	5 stars (out of 6)

According to Jack, the fundamentals of all three funds were solid and he expected them to perform equally well, as indicated by his ratings.

Paul:

T. Rowe Price New Horizons:	5 stars (out of 6)
Oakmark International:	4 stars (out of 6)
Third Avenue Real Estate Value:	6 stars (out of 6)

According to Paul, even though the fundamentals of all three funds were solid, they differed on a number of indicators, hence he expected some degree of difference in performance, as indicated by his ratings.

A year later, the three funds have performed equally well.

Both conditions

Which of the two is likely to be more knowledgeable?

Which of the two would you be more likely to consult?

Which of the two would you be more likely to trust?

This week we asked the same financial analysts to pick a stock that they thought would give the best gain in the short term. We asked them to consider a short list of some of the most negotiated stocks. We will ask you the same question and the performance of the stock (to be known in 3 months) will determine your bonus. You can see the recommendation of one of the analysts. Who do you choose to provide you with advice? Jack/Paul

As we said, we asked the financial analysts to pick a stock that they thought would give the best gain in the short term. We asked them to consider a short list of some of the most negotiated stocks:

BAC: Bank of America Corporation

PBR: Petroleo Brasileiro SA Petrobras

TLM: Talisman Energy Inc

ORCL: Oracle Corporation

GE: General Electric Co

By short-term we mean end of March. [The study was run in late December].

Naturally, the question is not which company is better, more famous or wealthier, but which stock choice would give a greater gain in the short term. Stock performance is simply the variation in price during this period. If it increases, you will get a bonus. The more it increases, the higher your bonus. If it decreases, there is no bonus. Each percentage increase is an extra 10 cents. If price goes up by 1%, you get 10 cents, if it goes up by 10%, you get $1. The most we will pay is $2, so any increase above 20% will lead to a $2 bonus.

Here the recommendation from the analyst you selected:

"Stock markets are definitely a great option in the long run. In the short run, predictions are always more volatile. Still, these stocks have a significant volume and therefore a lot of data. My assessment is that Talisman Energy is undervalued at the moment and by the end of winter, its price should present a solid gain relative to its current position."

What is your decision?
BAC: Bank of America Corporation
PBR: Petroleo Brasileiro SA Petrobras
TLM: Talisman Energy Inc
ORCL: Oracle Corporation
GE: General Electric Co

Appendix B

Procedure of Study 3

Ratings condition:

The Pratt Company, specializing in business consulting, is considering several new consulting projects to take on over the next few months. Two of the company project managers – John Malcolm and Albert Stone – have been asked to evaluate the potential of three of the projects.

After studying the available information, John gave the following evaluation:
Project 1: 5 (out of 5)
Project 2: 3 (out of 5)
Project 3: 4 (out of 5)
John indicated he was 90% [70%] confident in his evaluations.
After studying the available information, Albert gave the following evaluation:
Project 1: 4 (out of 5)
Project 2: 4 (out of 5)
Project 3: 4 (out of 5)
Albert indicated he was 70% [90%] confident in his evaluations.
The company decided to take on all three projects and assign them to different teams.
A year later, the three projects have generated comparable profit.
Based on all the information above, who would you say is more knowledgeable?

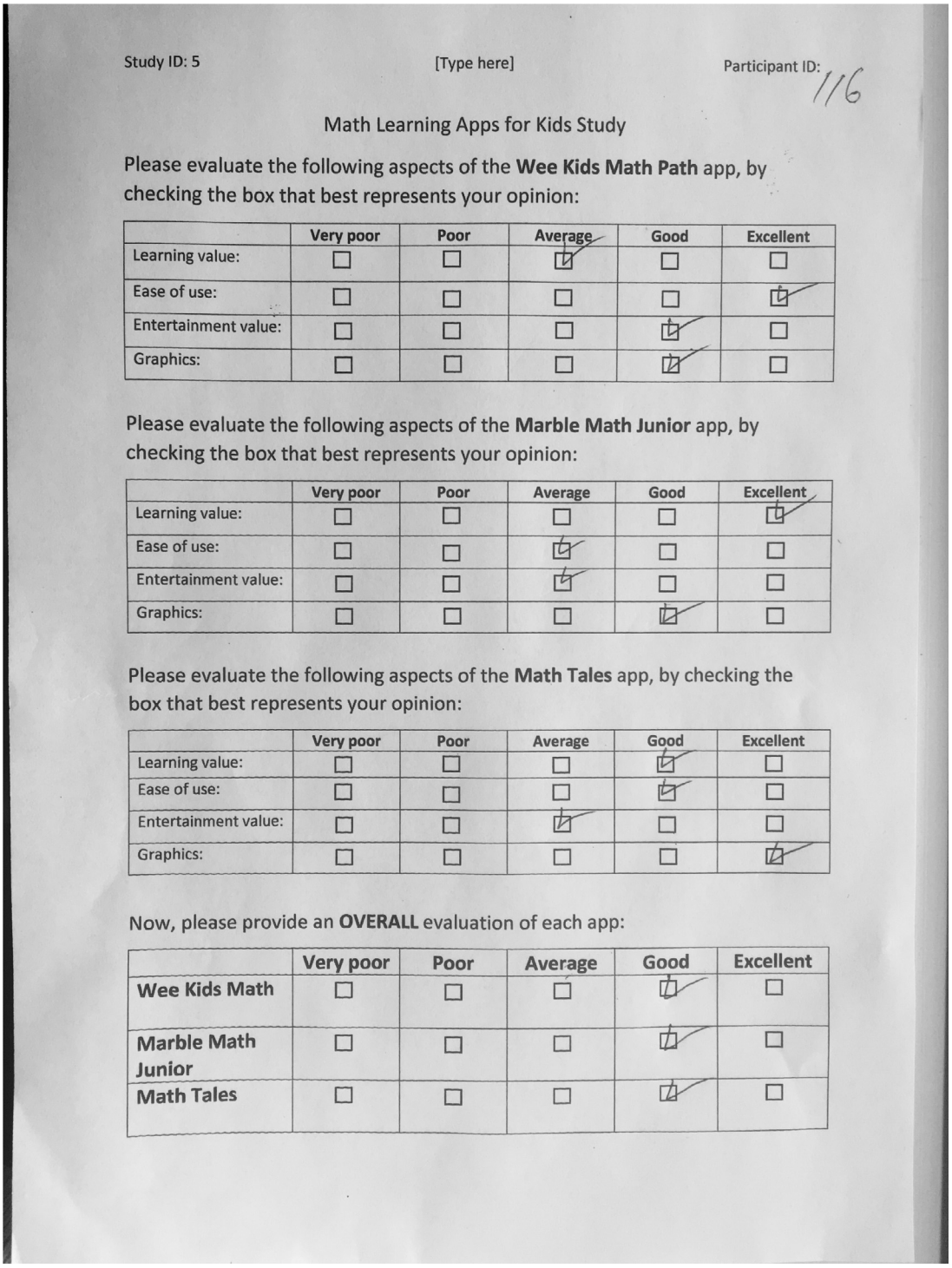
No ratings condition

The Pratt Company, specializing in business consulting, is considering several new consulting projects to take on over the next few months. Two of the company project managers – John Malcolm and Albert Stone – have been asked to evaluate the potential of three of the projects.

After studying the available information, John and Albert each provided their evaluations.
John indicated he was 90% [70%] confident in his evaluations.
Albert indicated he was 70% [90%] confident in his evaluations.
The company decided to take on all three projects and assign them to different teams.
A year later, the performance of the three projects is more in line with the evaluations given earlier by John.
Based on all the information above, who would you say is more knowledgeable?

Appendix C

Study 4a: Scoresheet with sub-ratings, uniform ratings student



Study ID: 5 [Type here] Participant ID: 116

Math Learning Apps for Kids Study

Please evaluate the following aspects of the **Wee Kids Math Path** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☑	☐	☐
Ease of use:	☐	☐	☐	☐	☑
Entertainment value:	☐	☐	☐	☑	☐
Graphics:	☐	☐	☐	☑	☐

Please evaluate the following aspects of the **Marble Math Junior** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☐	☐	☑
Ease of use:	☐	☐	☑	☐	☐
Entertainment value:	☐	☐	☑	☐	☐
Graphics:	☐	☐	☐	☑	☐

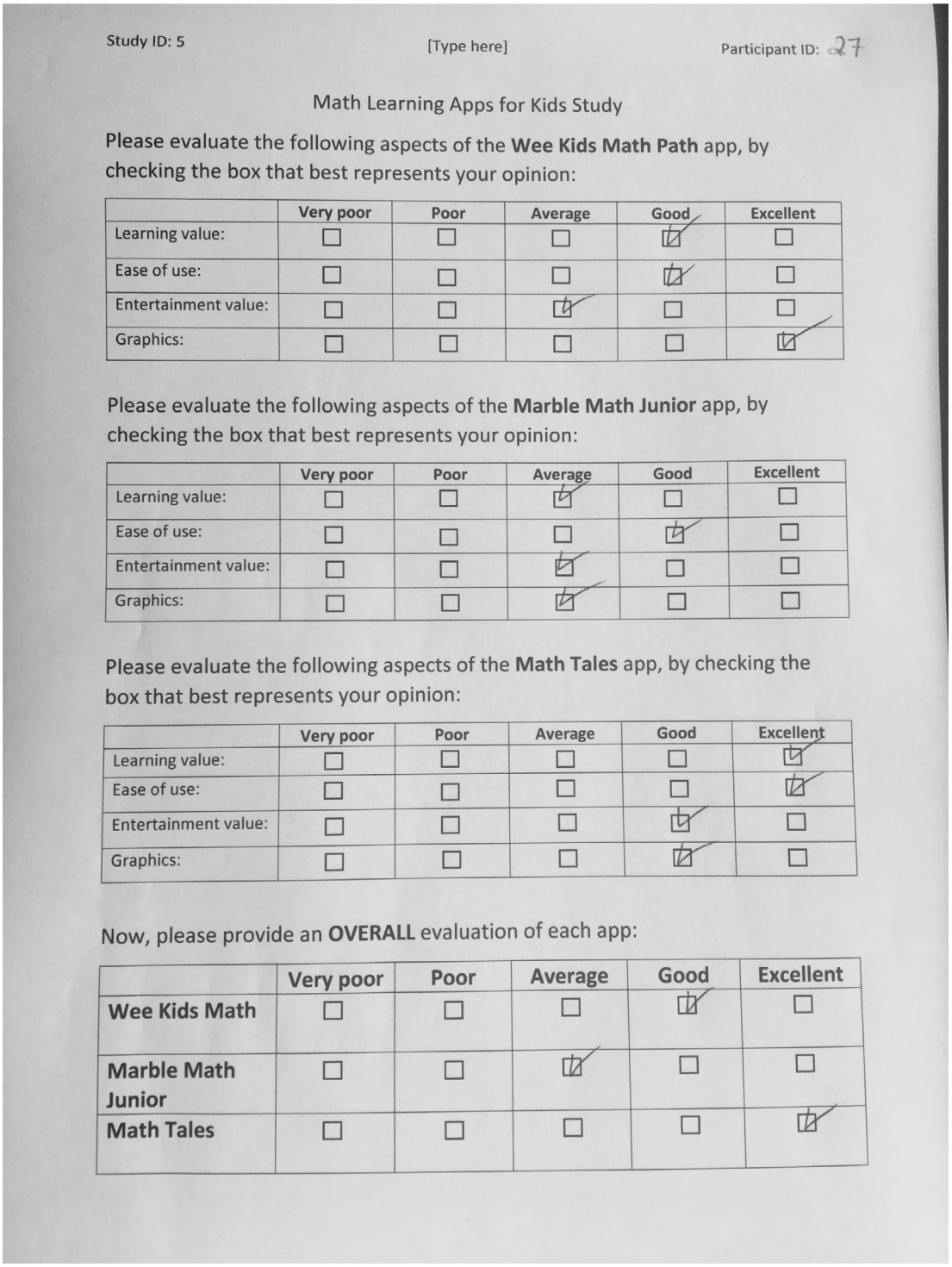
Please evaluate the following aspects of the **Math Tales** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☐	☑	☐
Ease of use:	☐	☐	☐	☑	☐
Entertainment value:	☐	☐	☑	☐	☐
Graphics:	☐	☐	☐	☐	☑

Now, please provide an **OVERALL** evaluation of each app:

	Very poor	Poor	Average	Good	Excellent
Wee Kids Math	☐	☐	☐	☑	☐
Marble Math Junior	☐	☐	☐	☑	☐
Math Tales	☐	☐	☐	☑	☐

Study 4a: Scoresheet with sub-ratings, varied ratings student



Study ID: 5 [Type here] Participant ID: 27

Math Learning Apps for Kids Study

Please evaluate the following aspects of the **Wee Kids Math Path** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☐	☑	☐
Ease of use:	☐	☐	☐	☑	☐
Entertainment value:	☐	☐	☑	☐	☐
Graphics:	☐	☐	☐	☐	☑

Please evaluate the following aspects of the **Marble Math Junior** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☑	☐	☐
Ease of use:	☐	☐	☐	☑	☐
Entertainment value:	☐	☐	☑	☐	☐
Graphics:	☐	☐	☑	☐	☐

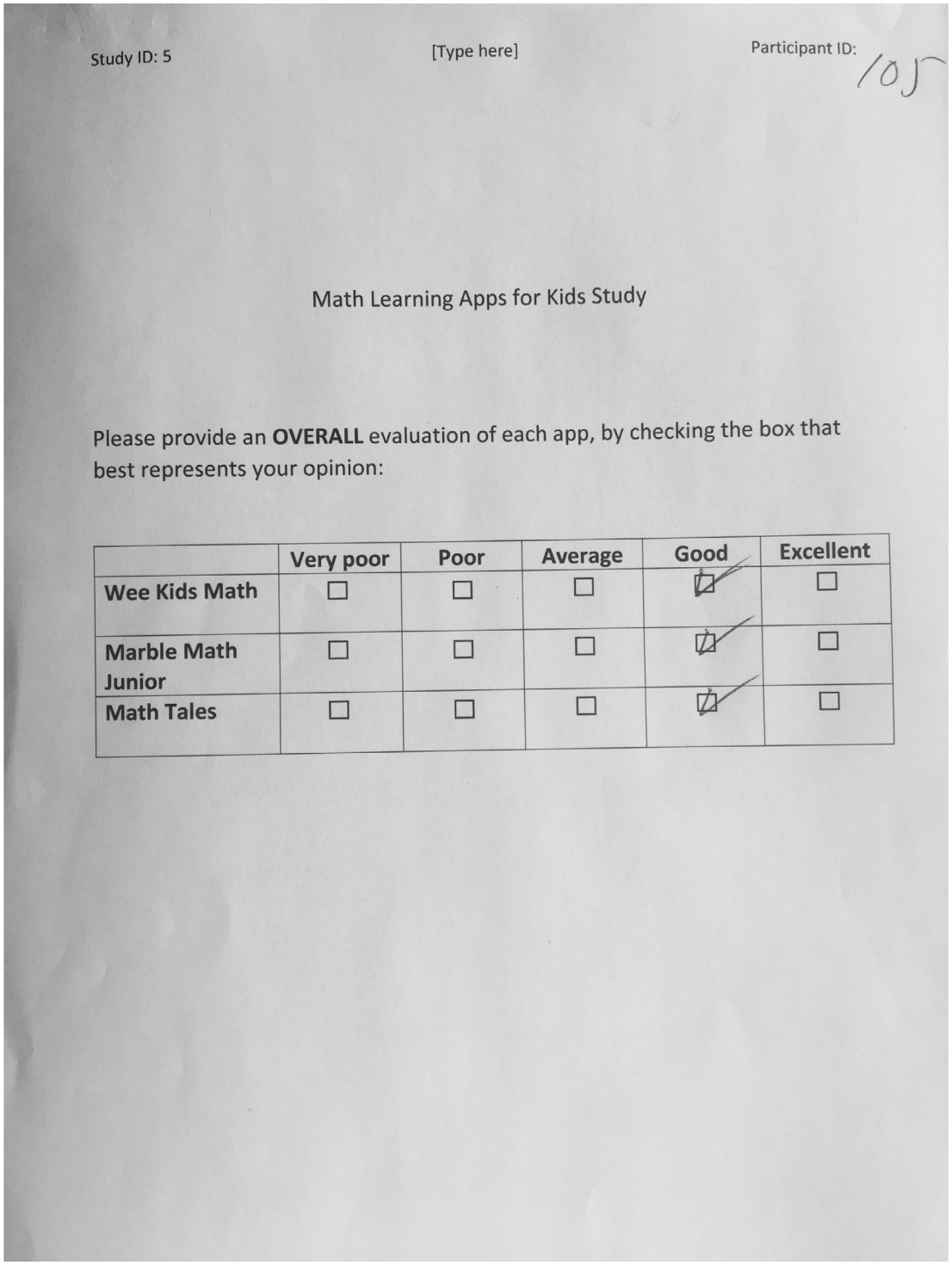
Please evaluate the following aspects of the **Math Tales** app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Learning value:	☐	☐	☐	☐	☑
Ease of use:	☐	☐	☐	☐	☑
Entertainment value:	☐	☐	☐	☑	☐
Graphics:	☐	☐	☐	☑	☐

Now, please provide an **OVERALL** evaluation of each app:

	Very poor	Poor	Average	Good	Excellent
Wee Kids Math	☐	☐	☐	☑	☐
Marble Math Junior	☐	☐	☑	☐	☐
Math Tales	☐	☐	☐	☐	☑

Study 4a: Scoresheet without sub-ratings, uniform ratings student



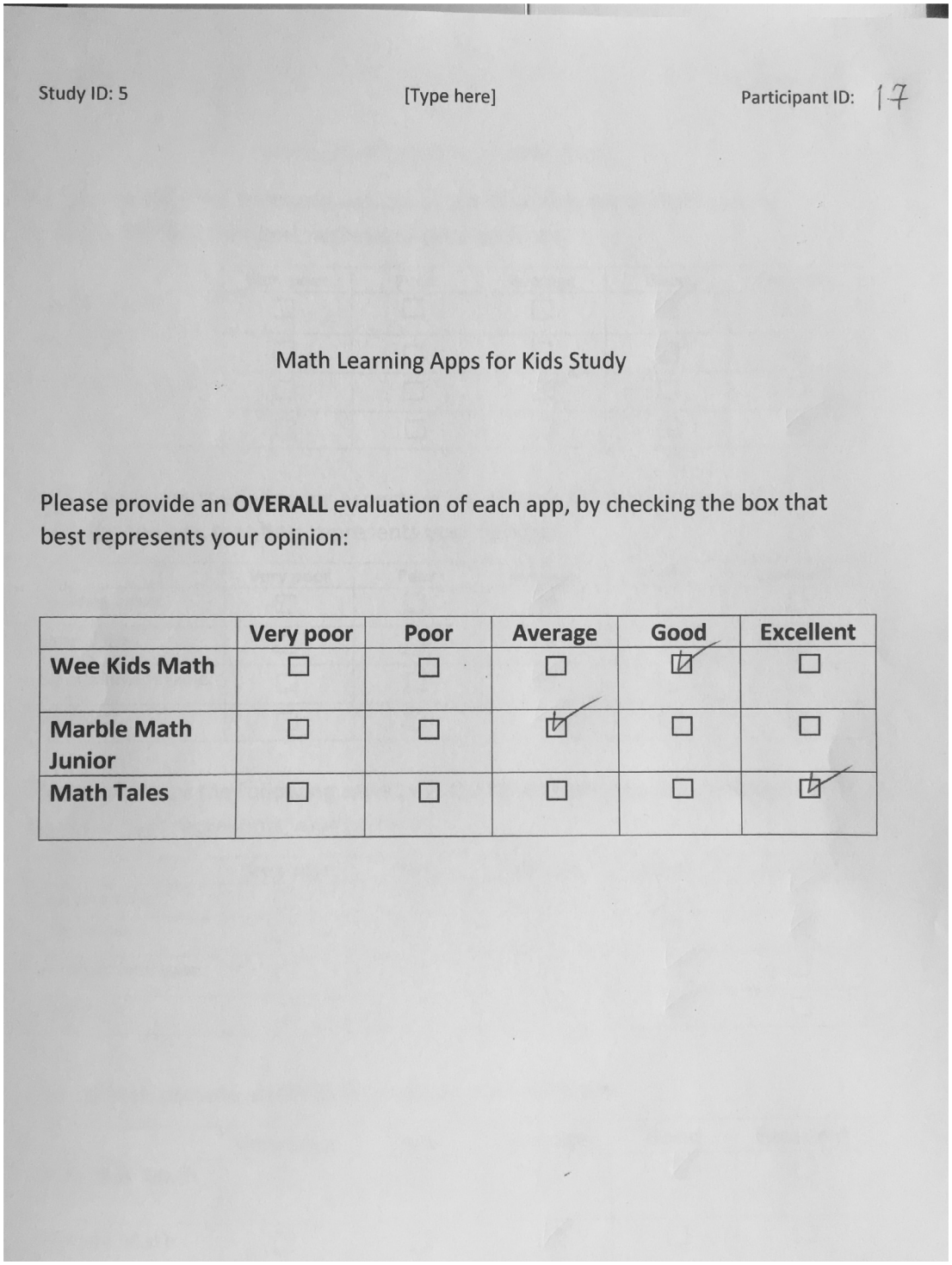
Study ID: 5 [Type here] Participant ID: 105

Math Learning Apps for Kids Study

Please provide an **OVERALL** evaluation of each app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Wee Kids Math	☐	☐	☐	☑	☐
Marble Math Junior	☐	☐	☐	☑	☐
Math Tales	☐	☐	☐	☑	☐

Study 4a: Scoresheet without sub-ratings, varied ratings student



Study ID: 5 [Type here] Participant ID: 17

Math Learning Apps for Kids Study

Please provide an **OVERALL** evaluation of each app, by checking the box that best represents your opinion:

	Very poor	Poor	Average	Good	Excellent
Wee Kids Math	☐	☐	☐	☑	☐
Marble Math Junior	☐	☐	☑	☐	☐
Math Tales	☐	☐	☐	☐	☑

Appendix D

Procedure of Study 5
In this study, we ask you to share your impressions of several individuals based on their product ratings.

Body lotion

Mary and Anne evaluated three brands of body lotion.
Here is how Mary evaluated the three brands of body lotion (5-point scales):

	Brand A	Brand B	Brand C
Moisturising property	3	4	4
Non-greasy on skin	3	2	2
Bottle shape/design	5	4	3

Final ratings: Brand A: 3; Brand B: 3; Brand C: 3

Here is how Anne evaluated the three brands of body lotion (5-point scales):

	Brand A	Brand B	Brand C
Moisturising property	3	4	5
Non-greasy on skin	4	3	3
Bottle shape/design	5	4	3

Final ratings: Brand A: 5; Brand B: 4; Brand C: 3

***No prompt condition*:**

Based on the ratings above, what is your overall impression of **Mary**?
Based on the ratings above, what is your overall impression of **Anne**?

Prompt condition:

Looking at how **Mary** rated each brand on the separate attributes, how do you think she arrived at the overall ratings of the brands?
Looking at how **Anne** rated each brand on the separate attributes, how do you think she arrived at the overall ratings of the brands?

Both conditions:

Based on these ratings, how likely are you to consult **Mary**, if you are looking for advice on body lotions in the future? 1-Not at all likely, 7-Very likely

Based on these ratings, how likely are you to consult **Anne**, if you are looking for advice on body lotions in the future? 1-Not at all likely, 7-Very likely

How capable is Mary of providing accurate ratings? 1-Not at all, 7-A lot
How knowledgeable is Mary about body lotions? 1-Not at all, 7-A lot
How capable is Anne of providing accurate ratings? 1-Not at all, 7-A lot
How knowledgeable is Anne about body lotions? 1-Not at all, 7-A lot

Blenders

John and Edward were asked to evaluate the quality of 3 brands of blenders.
Here is how John evaluated the three brands (on a 5-point scale)

	Blender A	Blender B	Blender C
Blending Power	3	5	4
Settings/speeds	5	3	4
Color range	4	3	2

Final ratings: Brand A: 4; Brand B: 4; Brand C: 4

Here is how Edward evaluated the three brands (on a 5-point scale)

	Blender A	Blender B	Blender C
Blending Power	3	5	5
Settings/speeds	4	3	4
Color range	5	3	2

Final ratings: Brand A: 5; Brand B: 3; Brand C: 2

Toothpaste

Mark and James were asked to evaluate the quality of three brands of toothpaste.

Here is how Mark evaluated the tooth paste brands (5-point scales).

	Brand A	Brand B	Brand C
Cavity protection	4	5	3
Breath freshening	4	3	5
Range of flavours	3	2	5

Final ratings: Brand A: 4; Brand B: 4; Brand C: 4

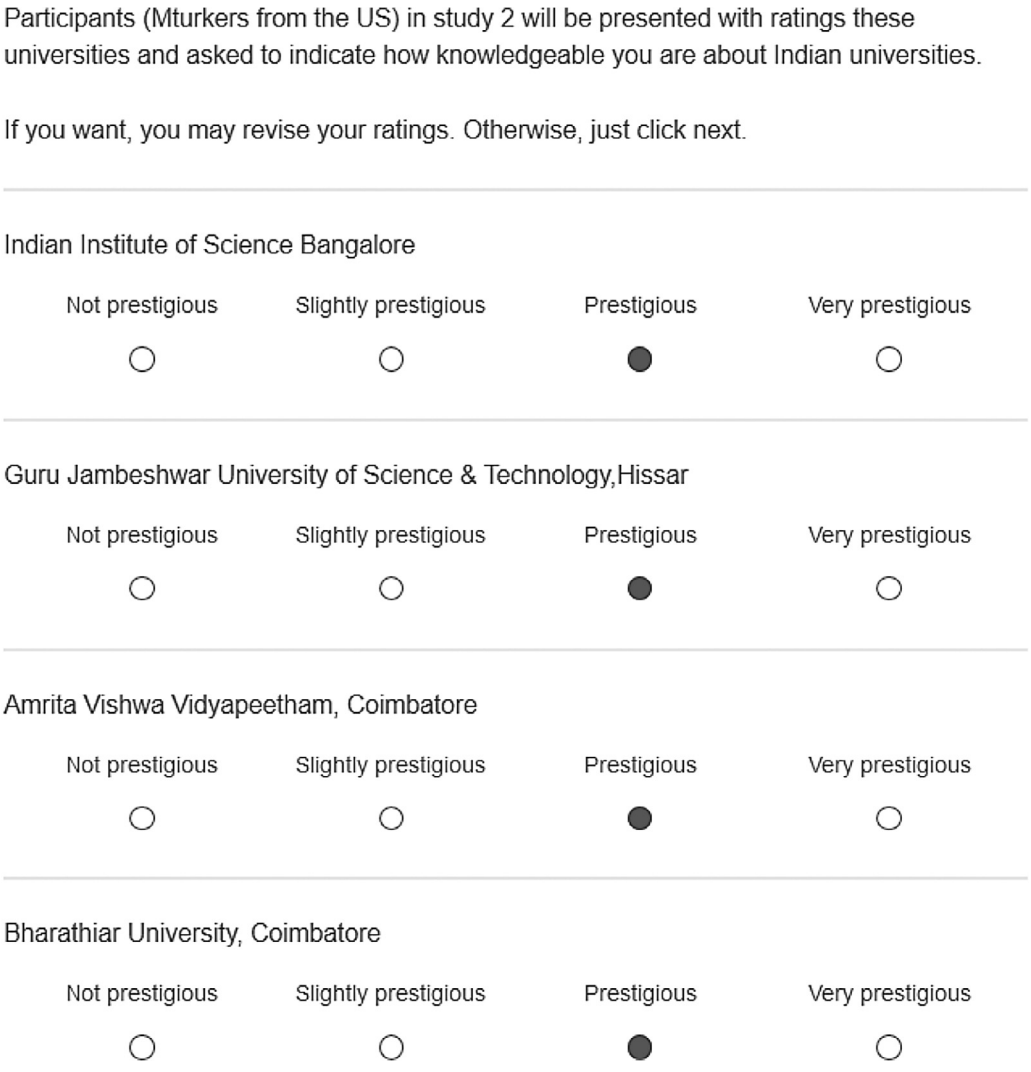
Here is how James evaluated the tooth paste brands (5-point scales).

	Brand A	**Brand B**	**Brand C**
Cavity protection	4	5	3
Breath freshening	5	4	4
Range of flavours	3	1	5

Final ratings: Brand A: 3; Brand B: 2; Brand C: 5

Appendix E

Screenshot of study 6



A Ratings Pattern Heuristic in Judgments of Expertise: When Being Right Looks Wrong.

References

Alba, J. W., & Hutchinson, W. J. (1987). Dimensions of consumer expertise. *Journal of Consumer Research, 13*(4), 411–454.

Amabile, T. M. (1983). Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology, 19*(March), 146–156.

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-Experts. *International Journal of Forecasting, 21*(3), 565–576.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*(4), 569–579.

Birnbaum, M., & Stegner, S. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37*(1), 48–74.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*, 127–151.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull, & R. S. Wyer (Vol. Eds.), *Advances in Social Cognition: Vol. 1*, (pp. 1–36). Hillsdale, NJ: Erlbaum.

Brockner, J., & Wiesenfeld, B. M. (1996). An integrative framework for explaining reactions to decisions: Interactive effects of outcomes and procedures. *Psychological Bulletin, 120*(2), 189–208.

Brown, J. J., & Reingen, P. H. (1987). Social ties and Word-of-Mouth referral behavior. *The Journal of Consumer Research, 14*(3), 350–362.

Brunswick, E. (1952). *The conceptual framework of psychology*. Chicago, IL: The University of Chicago Press.

Castellan, N. J. (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Performance, 9*(1), 16–29.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source vs. message cues in persuasion. *Journal of Personality and Social Psychology, 39*, 752–766.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken, & Y. Trope (Eds.). *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford.

Chen, P., Hong, Y., & Liu, Y. (2018). The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science* (in press).

Chinander, K. R., & Schweitzer, M. E. (2003). The input bias: The misuse of input information in judgments of outcomes. *Organizational Behavior and Human Decision Processes, 91*(2), 243–253.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science, 29*(5), 944–957.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Oxford, England: Prentice Hall.

Dattner, B. (2016). A scorecard for making better hiring decisions. Harvard Business Review; < https://hbr.org/2016/02/a-scorecard-for-making-better-hiring-decisions > .

Davidson, D., & Hirtle, S. C. (1990). Effects of non-discrepant and discrepant information on the use of base rates. *The American Journal of Psychology,* 343–357.

De Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research, 42*(6), 817–833.

Dougherty, J. (1978). Salience and relativity in classification. *American Ethnologist, 5*(1), 66–80.

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods, 41*(4), 1149–1160.

Feick, L., & Higie, R. A. (1992). The Effects of preference heterogeneity and source characteristics on ad processing and judgements about Endorsers. *Journal of Advertising, 21*(2), 9–24.

Fiske, S. T., & Neuberg, S. F. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Vol. Ed.), *Advances in Experimental Social Psychology: Vol. 23*, (pp. 1–74). San Diego, CA: Academic Press.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, P. T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing, 90*(June), 217–232.

Folkes, V. S., & Sears, D. O. (1977). Does everybody like a liker? *Journal of Expertimental Social Psychology, 13*, 505–519.

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behaviour and Human Performance, 33*(2), 263–282.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191–197.

Gershoff, A. D., Broniarczyk, S. M., & West, P. M. (2001). Recommendation or evaluation? Task sensitivity in information source selection. *The Journal of Consumer Research, 28*(3), 418–438.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143.

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

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* USA: Oxford Univ Press.

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology, 16*(3), 228–242.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*, 438–456.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*(2), 117–133.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-probability studies. *Psychological Review, 71*(1), 42–60.

Jerez-Fernandez, A., Angulo, A. N., & Oppenheimer, D. M. (2013). Show me the numbers: Precision as a cue to others' confidence. *Psychological Science, 25*(2), 633–635.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The Psychology of Intuitive Judgment* (pp. 49–81). New York: Cambridge Univ. Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Keren, G., & Teigen, K. H. (2001). Why is p = . 90 better than p = . 70? Preference for definitive predictions by lay consumers of probability judgments. *Psychonomic Bulletin & Review, 8*(2), 191–202.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1–17.

Kostyra, D. S., Reiner, J., Natter, M., & Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing, 33*, 11–26.

Lim, Y., & Van Der Heide, B. (2015). Evaluating the wisdom of strangers: The perceived credibility of online consumer e-views on Yelp. *Journal of Computer-Mediated Communication, 20*, 67–82.

MacKie, D. M., Gastardo-Conaco, M. C., & Skelly, J. J. (1992). Knowledge of the advocated position and the processing of in-group and out-group persuasive messages. *Personality and Social Psychology Bulletin, 18*(2), 145–151.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology, 13*(4), 330–334.

Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes, 42*(3), 343–363.

Petty, R. E., & Cacioppo, J. A. (1981). *Attitudes and persuasion: Classic and contemporary approaches.* Dubuque, IA: William C. Brown.

Poses, R. M., Cebul, R. D., Collins, M., & Fager, S. S. (1985). The accuracy of experienced physicians' probability estimates for patients with sore throats: Implications for decision making. *JAMA, 254*(7), 925–929.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*(1), 39–57.

Reyt, J., Wiesenfeld, B. M., & Trope, Y. (2016). Big picture is better: The social implications of construal level for advice taking. *Organizational Behavior and Human Decision Processes, 135*, 22–31.

Rosario, A. B., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The effect of electronic Word of Mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing, 53*(June), 297–318.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(July), 382–439.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research, 136*(2), 253–263.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes, 84*(2), 288–307.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Erlbaum.

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs non-musicians: An event-related potential and behavioral study. *Experimental Brain Research, 161*(1), 1–10.

Tsay, C. (2014). The vision heuristic: Judging music ensembles by sight alone. *Organizational Behavior and Human Decision Processes, 124*, 24–33.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(1124–1130).

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology, 44*, 443–461.

Wang, S., Cunningham, N. R., & Eastin, M. S. (2015). The impact of eWOM message characteristics on the perceived effectiveness of online consumer reviews. *Journal of Interactive Advertising, 15*(2), 151–159.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(1), 104–116.

West, P. M. (1996). Predicting preferences: An examination of agent learning. *The Journal of Consumer Research, 23*(1), 68–80.

West, P. M., & Broniarczyk, S. M. (1998). Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *The Journal of Consumer Research, 25*(1), 38–51.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes, 103*(1), 104–120.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes, 49*(1), 60–79.

Yeung, Catherine W. M., & Soman, D. (2007). The Duration Heuristic. *The Journal of Consumer Research, 34*(3), 315–326.