Journal of Experimental Psychology:
Learning, Memory, and Cognition
2017, Vol. 43, No. 2, 271–288

© 2016 American Psychological Association
0278-7393/17/$12.00 http://dx.doi.org/10.1037/xlm0000304

Adaptation and Fallibility in Experts' Judgments of Novice Performers

Jeffrey S. Larson and Darron M. Billeter
Brigham Young University

Competition judges are often selected for their expertise, under the belief that a high level of performance expertise should enable accurate judgments of the competitors. Contrary to this assumption, we find evidence that expertise can reduce judgment accuracy. Adaptation level theory proposes that discriminatory capacity decreases with greater distance from one's adaptation level. Because experts' learning has produced an adaptation level close to ideal performance standards, they may be less able to discriminate among lower-level competitors. As a result, expertise increases judgment accuracy of high-level competitions but decreases judgment accuracy of low-level competitions. Additionally, we demonstrate that, consistent with an adaptation level theory account of expert judgment, experts systematically give more critical ratings than intermediates or novices. In summary, this work demonstrates a systematic change in human perception that occurs as task learning increases.

Keywords: expertise, judgment and decision making, adaptation level, learning, judgment accuracy

Judges play an important role in society, because they evaluate human behavior during influential events. Judges evaluate skill-based performance in talent shows (e.g., *American Idol*), in the classroom, and in sporting events (e.g., the Olympics). Judges of human competitions are generally selected for their expertise at performing the competition task (Meyer & Booker, 2001) based on a belief that expertise in performing a task enables superior judgment of others' performance on the task. Many large competitions employ former competitors as judges because they are viewed as the most expert in the field. For example, 10 of the 11 horse dressage competition judges at the 2012 London Olympics were formerly champion riders. Similarly, the judges of major talent shows, such as TV's *The Voice* and *Dancing with the Stars*, are successful performers in that arena.

In this article, we investigate the effect of expertise on judgments of competitor performance. Although a large stream of prior research has investigated the effect of expertise on judgment (Camerer, 1981; Camerer & Johnson, 1991; Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Einhorn & Hogarth, 1978; Ericsson & Charness, 1994; Fiske, 1975; Kahneman & Klein, 2009; Kaufman, Baer, Cole, & Sexton, 2008; Kaufman, Gentile, & Baer, 2005), judgment in the domain of competitor performance is unique in that judgments of competitors are typically made at the same time as the judged performance, necessitating a short, intense judgment period. Also unique in our research is that instead of examining the effect of expertise on a uniform set of judgments, we examine how expertise affects judgments of both beginner-level competitors as well as advanced-level competitors. Where prior research (e.g., Kinney, 2009) and lay theory suggest higher levels of expertise should improve judgments of competitors at all performance levels, we propose expertise improves judgment accuracy at some competitive performance levels but harms it in others. Specifically, we hypothesize that greater expertise improves judgments of high-level competitive performers, but can harm judgments of low-level competitive performers.

In the next section, we review relevant research on the learning process that produces expertise and its known impact on judgment, and then lay out our proposal for how expertise both increases and decreases judgment accuracy. Although research with actual experts has been limited due to the difficulty of obtaining actual experts to participate in research studies, our studies employ internationally recognized experts as participants. We then present the results of three studies that provide evidence in support of our proposal and against several alternate explanations. Finally, we conclude with a discussion of the implications of this proposal.

Conceptual Development

Researchers have proposed many frameworks of skill acquisition to explain the nature and process of expertise development. In one of the most influential frameworks, Fitts (1964) defines three stages of skill acquisition. In the first stage (the cognitive stage), the learner is focused on gathering the facts needed to understand and perform the task. In the second, or trial phase, the learner begins to develop the movements associated with performing the task. In the experience phase, the learner achieves a high level of performance. In this last stage, the performance of many rudimentary tasks becomes fast, accurate, effortless, and largely removed from the learner's awareness as the learner slowly adapts and focuses on the execution of more complex aspects of performance.

Anderson (1982) provides a second widely used framework of skill acquisition, in which he describes skill acquisition in terms of changes in knowledge structure, from Stage 1 (the declarative stage), to Stage 2 (knowledge compilation), and finally to Stage 3 (procedural knowledge). In a review of this literature, Ackerman

This article was published Online First October 6, 2016.

Jeffrey S. Larson and Darron M. Billeter, Department of Marketing and Global Supply Chain, Marriott School of Management, Brigham Young University.

Correspondence concerning this article should be addressed to Jeffrey S. Larson, Department of Marketing and Global Supply Chain, Marriott School of Management, Brigham Young University, 672 TNRB, Provo, UT 84602. E-mail: jeff_larson@byu.edu

(1987) views the main insights of these two frameworks to be complementary. Expertise can be characterized as a process of automation, in which the automatic execution of low-level skills enables increasing cognitive focus on improved execution of higher-level skills. Similarly, the stages of both frameworks map onto the novice-intermediate-expert designation of performance expertise often used in the literature (e.g., Hinds, 1999; Plucker, Kaufman, Temple, & Qian, 2009).

Decades of research on expert performance supports the view that both physiological and cognitive changes occur as individuals develop expertise in a particular performance domain (Ericsson, Krampe, & Tesch-Romer, 1993; Hill & Schneider, 2005; Kanfer & Ackerman, 1989). For example, after years of practice, elite athletes experience changes in anatomical characteristics such as enlarged hearts (Sundberg & Elovainio, 1982), increased muscle fiber (Ericsson, 1990), reduction in body fat (Reilly & Secher, 1990), and flexibility in joint movement (Tucker & Collins, 2012). Expertise also changes one's cognitive structures (Alba & Hutchinson, 1987), increasing one's ability to recall (Chase & Ericsson, 1981; Chase & Simon, 1973), to multitask (Beilock, Carr, MacMahon, & Starkes, 2002), to perform the task easily (Evans, 2008), and even the prediction accuracy of one's own learning (Billeter, Kalra, & Loewenstein, 2011). A hallmark of expertise is a decrease in the cognitive effort required to perform the practiced task, as many aspects of the task become automated through practice (Anderson, 1982; Fitts, 1964; Haier et al., 1992). These cognitive changes enable the expert to perform a complex skill at a high level. These changes also affect how experts judge the performances of other competitors in that arena.

Competitor judgment is a complex assessment of the extent to which participant performance matches a predetermined ideal. Because the assessment includes some subjectivity, most competitive domains attempt to establish universal standards to increase the reliability and objectivity of performance evaluation. An accurate judgment in such competitions is one in which a judge properly assesses the extent to which each competitor's performance matches the objective standard for ideal performance. Generally, research supports the ability of judges to reliably identify superior performances in subjective domains (Hinds, Patterson, & Pfeffer, 2001; Smith, 2004). Still, other research has highlighted the low level of reliability attained by expert judges (Bergee, 2007; Thompson & Williamon, 2003). Indeed, because of the subjectivity inherent in competitive judgment, we contribute to this literature by utilizing multiple methods (alignment of judges' marks, identification of performance quality through Bayesian MCMC techniques) to identify the underlying accuracy of judgments.

The primary question we address in this article is how performance expertise affects competition judgments, especially the accuracy of these judgments. We examine the nature of the limitations of *perceptual* expertise that accompany *performance* expertise. Extensive research has demonstrated perceptual expertise develops alongside performance expertise (Kaufman, Gentile, & Baer, 2005; Kaufman, Baer, Cole, & Sexton, 2008; for a review, see Mann, Williams, Ward, & Janelle, 2007). This stream of research generally shows performance experts demonstrate enhanced perceptual skill in their domain of expertise. Although some of this research has found high levels of correlation between the perceptions of experts and novices, this work also found empirical evidence that experts were more reliable than novices in their judgments (Kaufman, Gentile, & Baer, 2005, p. 262). Our findings indicate a limitation in the enhanced perceptual expertise accompanying performance expertise—that expert performers are often less accurate than intermediate performers in their judgments of beginner-level performances. (Throughout this article, the terms *expert*, *intermediate*, and *novice* refer to the level of *performance* expertise attained by the judge, not his or her *perceptual* expertise. We avoid use of the term *expert judge* to avoid the potential confusion that the judge is an *expert at judging*. Our studies never measured expertise in judgment—only expertise in performance).

Past research on judgment accuracy utilizing Brunswik's (1952) lens model has established several necessary conditions for accurate judgment (Karelaia & Hogarth, 2008). In a prototypical study in this area, participants attempt to predict a specified outcome by observing the available cues and constructing their prediction based on those cues. Then, the predicted value is compared to the actual value to determine judgment accuracy. In competitive judgment, the cues are the characteristics of the observed performance, whereas a global judgment of the quality of the performance replaces the predicted value (Mills, 1991). For example, an Olympic diving judge gives a diver a score from 0 to 10 (i.e., the outcome) based on the characteristics of the observed performance, including the diver's body position midflight, alignment of legs, verticality upon entry, size of splash, and any other categories of performance that are accepted as aspects of ideal performance (i.e., the cues).

Three main factors that the lens model identifies as necessary for judgment accuracy are (for a review, see Karelaia & Hogarth, 2008): (a) the judge utilizes the correct cues (i.e., does not ignore valid cues and does not utilize invalid cues); (b) the judge properly weights the cues (i.e., does not credit an item as being more or less diagnostic than it is); and (c) the judge accurately evaluates each cue (i.e., detects the level, from very high to very low, of each cue). We examine the effect of expertise in facilitating or hindering each of these conditions for accurate judgment.

The first requirement for accurate judgment of competitor performance is knowledge of all pertinent cues. For example, if an Olympic diving judge does not know to assess a cue such as the verticality of the diver upon entry into the water, the judge's marks will have more error and thus be less accurate. The literature on judgment consistently finds expertise facilitates a superior knowledge of cues. For example, experts can see cues in x-rays that novices cannot (Lesgold et al., 1988). A study of expert and novice consumers found novices' information search over product attribute cues to be influenced by spurious background image primes, whereas experts' information search was unaffected (Mandel & Johnson, 2002). In over 50 expertise-related judgments, Garb (1989) found clinical psychologists had a greater ability than less experienced people to identify the factors that contributed to the correct medical diagnosis. A review of auditing research shows an increase in the expertise of the auditor is associated with an ability to correctly identify errors (Tubbs, 1992). An examination of chess players found experts could more easily and accurately identify the cues on the chessboard that were influential in determining the ideal move than nonexperts (De Groot, 1965). Because chess players' expertise develops with the express goal of being able to perform a task in an ideal fashion, knowledge of the dimensions that contribute to ideal performance increases directly with expertise. Thus, judges who gained expertise through past performance

in a competition domain should have an increased knowledge of pertinent cues relative to those with less expertise.

The second requirement for accurate judgment is a proper weighting of the cues. Constructing a global judgment of performance through a weighted combination of multiple cue values is a complex mental activity. Prior research on judgment finds expertise increases a judge's capacity to properly weight the performance cues. For example, research into clinical judgment demonstrates experts are more capable of determining the relative importance of cues (Meehl, 1954). To the extent that ideal performance can be codified, the development of expertise should aid in the cognitive processes needed to properly weight the various cues of a competitor's performance. By contrast, a judge's ability to accurately assess the level of each cue may not systematically increase with expertise.

The third requirement for accurate judgment is an ability to accurately assess the level of each cue (Einhorn, 1972). If an Olympic diving judge knows all the pertinent cues and correctly weights them, her judgment will still be inaccurate if she cannot accurately evaluate whether the diver's verticality upon entry was *perfect* or just *very good*. As expertise increases, an expert's focus is increasingly centered on ideal performance. As a result, expertise increases an expert's discriminative capacity in this focal area, near ideal performance (Weiss & Shanteau, 2003). For example, Frederick and Libby (1986) demonstrated that auditors, who had additional expertise, were significantly superior to students in identifying the deviations from ideal accounting practices. We propose that the increased discriminative capacity near ideal performance has a downside—it decreases an expert's ability to discriminate among lower-level performances. Adaptation level theory states that discriminative capacity is greatest at the area of focus and decreases with greater distance from this focal area (Helson, 1948; Dinnerstein, 1965). In other words, experts will be unable to discriminate which of two beginner divers had a worse body position during the dive, because their adaptation level is far removed from this low performance level. A novice has not developed as keen a focus on the ideal as the expert and has an adaptation level closer to the beginner-level performance. The novice will thus be better able to discriminate the differences in quality of two beginners' body positions.

In summary, judgment accuracy is dependent on three capacities: (a) knowledge of the pertinent cues, (b) proper weighting of the cues, and (c) accurate assessment of the level of each cue. Whereas expertise universally improves the first two capacities, it decreases the accuracy of cue assessment under certain conditions. Thus, we propose the accuracy of judgments of competitor performance depends on the expertise of the judge and the performance level of the competitors, in the following manner. Expertise improves a judge's knowledge of the ideal, which encompasses both the knowledge of pertinent cues and the ability to properly weight the cues. But expertise reduces a judge's ability to discriminate among low-level performances, and thus can lead to a reduction in judgment accuracy when evaluating these low-level performers. For illustrative purposes, we assume judgment accuracy is an additive function of knowledge of the ideal and discriminative capacity. In the top panel of Figure 1, we depict the relationship between expertise and knowledge of the ideal, with knowledge of the ideal increasing with expertise. In the second panel, we illustrate discriminative capacity as a function of both expertise level and performance level. According to adaptation level theory (Helson, 1948), for experts, discriminative capacity is highest near the ideal; for novices, discriminative capacity is highest at low performance levels; for intermediates, discriminative capacity is highest at intermediate performance levels. Finally, the last panel of the figure depicts judgment accuracy as an additive function of discriminative capacity and knowledge of the ideal. As the figure depicts, in the lower levels of performance, intermediates make better judgments than experts. Intermediates' superior discrimination ability at lower performance levels overcomes their inferior knowledge of the ideal relative to experts. If adaptation level theory holds, expertise would assist a judge in the assessment of expert performance, but in some instances may hinder a judge's ability to assess lower-level performances. In this article, we test whether expertise can hinder assessment of lower-level performances at the same time that it improves assessment of high-level performance.

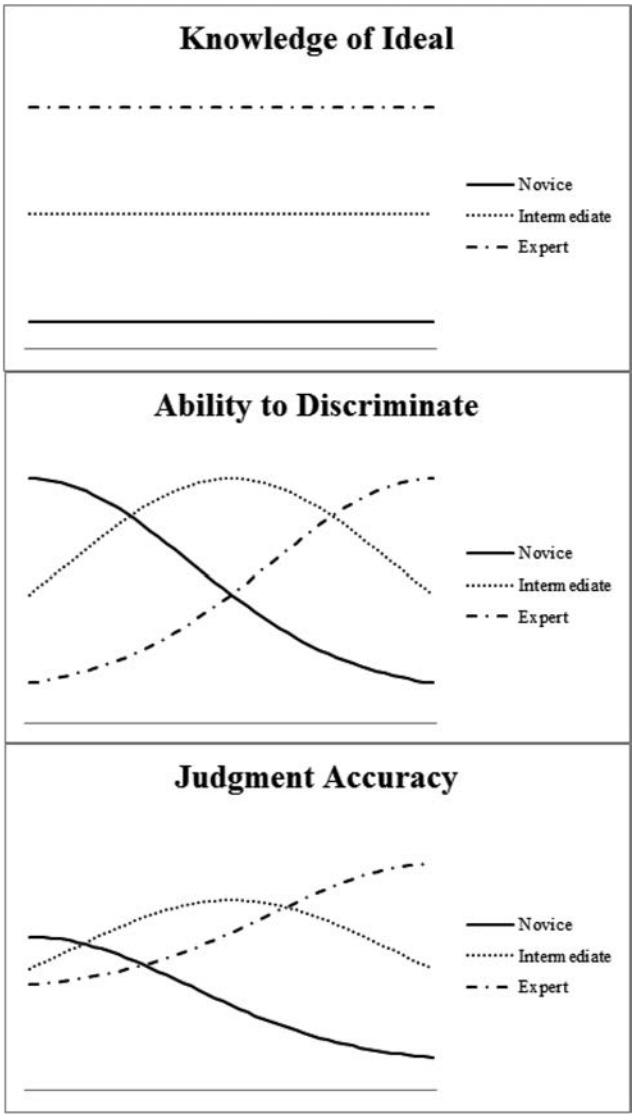


Figure 1. The effect of expertise on knowledge of ideal, discrimination ability, and judgment accuracy.

Our contribution to the literature on expert judgment extends beyond the demonstration that experts have limitations in their judgments of beginner performance. We also hypothesize experts' judgments are systematically more critical than novices' judgments. Stated more precisely, if both experts and novices evaluate the same performance (at any skill level), experts will evaluate that performance to be worse than will novices. Two aspects of expert judgment cause this severe judgment. First, because experts have larger knowledge structures (Alba & Hutchinson, 1987), they are better able to recognize all the pertinent cues for ideal performance, and they will observe more mistakes from a performance than will a novice. Second, because experts' adaptation level is closer to ideal performance, an expert will judge any single error in performance to be more egregious than will a novice. Thus, when both experts and novices judge the same set of performances on the same scale, we should observe the following: (a) experts' judgments are lower on the scale than novices' judgments, (b) experts notice more errors than do novices, and (c) experts rate the same performance error to be more egregious than do novices.

In three studies, we examine the role of expertise in competitive judgment. In Study 1, we examine the judgment accuracy in the domain of singing. Study 2 makes a similar analysis in the domain of ballroom dance. Study 3 returns to the domain of singing, but examines judgments of choral singing rather than solo singing. Results of the three studies in these varying domains provide evidence that the effects we hypothesize are robust across multiple domains of skill learning.

Study 1: Accuracy of Judgment by Expertise Level and Performance Level

The purpose of Study 1 is to examine the effect of singing expertise on judgments of vocal performance. Specifically, we seek to determine whether expert singers make less accurate judgments of low-quality singers than do intermediate singers. To ensure the expert singers were truly expert, we employed singers from a world-famous choir.

Method

Participants. First, we recruited 114 participants from an undergraduate subject pool to participate in a study on singing judgment. The participants from the undergraduate subject pool exhibited a high degree of variability in vocal expertise, including a few participants who could be considered true experts (e.g., membership in highly selective audition-only choirs). To ensure sufficient representation of true experts in the study, we recruited 11 participants from a world-famous Emmy- and Grammy-winning choir. These participants also had extensive musical education and training.

Materials. To study judgment of vocal performance, we first had to collect recordings of vocal performance. We recruited 20 women to participate in a study on music in exchange for $5. When participants showed up for the study, they were informed they would be recorded singing the chorus from the song "Let it Go," from the Disney movie musical *Frozen*. The recording took place in a small, enclosed room. Participants stood in a corner of the room in front of a recording microphone, which was attached to a laptop computer. The female research assistant sat in front of the laptop to control recording and playback.

Before singing, participants listened to a short recording of the same chorus being sung by a female research assistant unaccompanied by instrumentation. Participants were asked to sing in the same key and same tempo they had just heard in the recording. All participants were familiar with the song. After singing the chorus, participants were paid and dismissed.

We edited each of the 20 audio recordings to play only the first two lines of the chorus. The duration of each recording was about 15 s.

Procedure. Participants in Study 1 were asked to listen to seven recordings of the song "Let it Go," which were randomly sampled from the 20 music clips, and to rank order the seven clips from best to worst. After rank ordering the seven music clips, participants were then asked to rank order another seven music clips, which were composed of an additional random sample of seven of the 20 recorded vocalists. Thus, participants made two different rank orderings of two random samples (with replacement) of seven of 20 vocalists.

After rank ordering the two sets of vocalists, participants were asked "Please describe the extent of your singing experience" and were provided an open-ended response box to type their response. Participants also made a self-assessment of their vocal expertise using a three-item scale (Park, Mothersbaugh, & Feick, 1994) that included the questions, "How much do you feel you know about singing?" "Compared to your friends and acquaintances, how much do you feel you know about singing?" and "Compared to a singing expert, how much do you feel you know about singing?" Participants responded on a 5-point semantic differential scale anchored by *very little* and *very much* (Cronbach's alpha = .88).

Results

The central hypothesis of this research is that experts are less accurate than intermediates in their judgments of beginner-level performers. To investigate whether expert vocalists were less accurate in their judgments of the beginner-level singers, we examine the accuracy of the judgments of experts, intermediates, and novices in judging lower-quality singers. This analysis requires us to determine both the expertise level of the judge as well as the quality level of the vocalists. We first describe our method for assessing judges' expertise.

Categorization of judges' expertise. At the end of the judgment survey, judges described their singing experience. A research assistant read each participant's description and rated that expertise on a 5-point scale (1 = *very little singing experience*, 5 = *the highest level of expertise*). The research assistant gave a 0 rating to participants who entered "none" or "0" or who failed to respond to this question. We thus considered novices to be those receiving a rating of 1 or 2; intermediates were those with a rating of 3 or 4; experts were those with a rating of 5. Under this classification system, 19 participants qualified as experts, and 41 qualified as intermediates.

Assessment of singer quality. We hypothesized that expert singers would be less accurate than intermediate singers in rank ordering low-quality vocalists. To examine whether this hypothesis was correct, we had to determine the "true" quality of the 20 vocalists. Because the central hypothesis of this research is that experts do not provide reliable judgments of low-quality singers, we could not rely on the judgment of an expert to report the true quality of the 20 vocalists. Instead, we modeled the data from all 125 participants to find the true quality of the 20 vocalists. Spe-

cifically, we applied a Bayesian MCMC model to the participants' ranking data to extract the latent singing quality demonstrated by the 20 vocalists (Appendix A provides details of the model).

To test the hypothesis that experts would be less accurate than intermediates in their assessments of lower-quality singers, we needed to select the performers (singers) that qualified as "low quality." Our specific hypothesis was that the judges with the most expertise would have difficulty correctly determining the relative quality of two low-quality singers. To classify the quality of the singers being judged, we utilized the average latent-quality score of the two singers. We used three different cut-offs for classifying low-quality singer pairs: (a) the bottom tercile, (b) the bottom quartile, or (c) the bottom quintile.

Alignment of judgments within expertise class. We use interjudge alignment as a measure of judgment quality, which researchers have compellingly argued as a necessary condition for high judgment quality (Einhorn, 1972; Einhorn, 1974; Thompson & Williamon, 2003). If judges exhibited no skill at all and randomly assigned rankings, we would expect alignment to be near 50%. If judges exhibited perfect judgment, we would expect alignment to be 100%. Thus, we begin our examination of judgment accuracy by comparing interjudge alignment within each level of expertise by performance-quality level. Figure 2 shows the percentage of judgments that coincided within each class of judges at each performance-quality level.

To measure alignment within a class of judges, we calculated the percentage of pairwise rankings that coincided within that class of judges. For example, perhaps 10 of the 19 experts performed a ranking that included both Singer 1 and Singer 2. If seven of these 10 judges ranked Singer 1 higher than Singer 2, the alignment for this pair would be 70%. (If three of these 10 judges ranked Singer 1 higher than Singer 2, the alignment for this pair would still be 70%.) We calculated the alignment within each class of judges for each of the 190 pairwise-rankings comparisons.

Intermediates were significantly more aligned in their judgments of low-quality singers ($M_{\text{Intermediates}} = 87.8\%$) than were experts ($M_{\text{Experts}} = 82.1\%$; $\chi^2(1) = 4.36$, $p < .05$, Cohen's $d = .15$, LCL = .01, UCL = .30). Intermediates were also significantly more aligned in their judgments of low-quality singers than were novices ($M_{\text{Novices}} = 80.5\%$; $\chi^2(1) = 13.31$, $p < .01$, Cohen's $d = .22$, LCL = .10, UCL = .33). Thus, our central hypothesis regarding the limitation of expertise was confirmed: experts were less accurate in their judgments of low-quality performers than were intermediates.

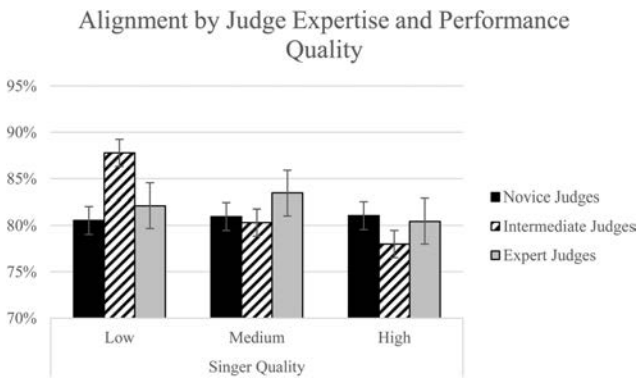


Figure 2. Study 1 within-group judgment alignment by expertise level and performance level.

If we limit our attention to singers in the bottom quartile, the differences become more dramatic. Intermediates' judgments of singers in the bottom quartile were more aligned with one another ($M_{\text{Intermediates}} = 87.2\%$) than were experts' judgments ($M_{\text{Experts}} = 77.4\%$; $\chi^2(1) = 8.60$, $p < .01$, Cohen's $d = .25$, LCL = .08, UCL = .42). We also observed this pattern in judgments of singers in the bottom quintile ($M_{\text{Intermediates}} = 87.7\%$ vs. $M_{\text{Experts}} = 77.1\%$; $\chi^2(1) = 8.17$, $p < .01$, Cohen's $d = .27$, LCL = .09, UCL = .46). This finding is consistent with our theory. Experts have difficulty differentiating among low-quality performers because their adaptation level is far removed from the observed performance level. The lower the quality of the performance, the more difficulty experts have differentiating them. As a result, experts' accuracy decreased as the singer quality decreased, which the alignment numbers reflected.

Judgment accuracy. An alternate explanation of these results is that intermediates use similar heuristics in rank ordering low-quality performers, but do not necessarily provide more accurate judgments. In other words, perhaps intermediates' judgments are more aligned simply because they are more similar to one another and not because they are more accurate. To address this alternate explanation, we conducted another analysis based not on judges' alignment with one another, but on judges' concordance with the "true" rankings of the singers as indicated by the Bayesian model described earlier.

The Bayesian model provides an estimate of the quality of each of the 20 singing performances. This method provides several advantages. First, the quality estimates are gained without imposing a narrowed view of the factors that constitute superior performance. By using judgments from all participants, we do not impose the experts', intermediates', or novices' particular judgment criteria. Second, this method provides a way to directly measure judgment accuracy, which is the main purpose of this research. Because the model provides an estimate of the quality of each performer, we can compare the "true" ranking of performers with the ranking provided by each judge. In other words, we can directly measure each judge's accuracy. Although judgment alignment is a reliable proxy for judgment accuracy, it is still a proxy; thus, a direct measurement of accuracy is desirable. Third, the model provides measures of uncertainty about the quality estimates, which allows us to run robustness checks for our findings. We provide these checks in Appendix B. Additional follow-up analyses provided evidence for the validity of the quality estimates. Details of these analyses are in Appendix C.

A direct measure of judgment accuracy is the concordance between the "true" rankings and the rankings given by the judge. We utilized the same method to measure concordance as that used in the Kendall rank correlation coefficient, namely, by calculating the percentage of ranked pairs that are ordered the same by each set of rankings. If the true quality ranking of a group of seven vocalists, A-B-C-D-E-F-G, is 1-2-3-4-5-6-7, and the judge rank orders them 1-3-2-4-5-7-6, the level of concordance between the judge's rankings and the "true" rankings is 90.4%, or 19 of 21 ranked pairs being concordant between the judge's rankings and the "true" rankings. (The judge incorrectly evaluated C to be better

than B, and G to be better than F, but was still correct in all other pairwise assessments.)

Each rank ordering participants performed resulted in 21 binary judgments, and each of the 125 participants performed two rank orderings each, yielding a dataset of 5,250 judgments. However, one participant did not perform either rank ordering, and nine other respondents skipped one of the two rank orderings, yielding a dataset of 5,019 judgments.

Figure 3 shows the accuracy of judges at all three levels of expertise when judging singers classified by quality level into terciles. As the figure shows, experts exhibited a lower accuracy (M = 82.6%) than intermediates (M = 88.0%) when judging low-quality singers ($\chi^2(1) = 4.20$, $p < .05$, Cohen's $d = .15$, LCL = .007, UCL = .29). Consistent with our theory, the difference in judgment accuracy between experts and intermediates became more dramatic as we narrowed our focus to even lower-quality singers. When examining singing pairs only in the lowest quartile, accuracy of experts decreased (M = 77.4%) while intermediate accuracy decreased only slightly (M = 87.2%; $\chi^2(1) = 8.60$, $p < .01$, Cohen's $d = .25$, LCL = .08, UCL = .42), reinforcing the finding that experts provide less accurate evaluations of beginner performance than intermediates. When examining the judgment accuracy of singing pairs in the bottom quintile, we find the results are similarly strong. Experts' judgment accuracy (M = 77.3%) is lower than intermediates' accuracy (M = 88.0%; $\chi^2(1) = 8.49$, $p < .01$, Cohen's $d = .28$, LCL = .09, UCL = .46).

One potential objection to this analysis is that a point estimate cannot summarize the quality of an artistic performance, because equally good performances can be perceived differently due to differences in style, interpretation, expression, and so on (Bergee, 2007; Thompson & Williamon, 2003). For example, although Renee Fleming is objectively a superior opera singer to Madonna, determining whether she is a better singer than Luciano Pavarotti may be difficult, because the measurement of their quality cannot be separated from preference on nonquality-determining attributes such as style and musicality. For this reason, we conducted additional analyses in which we removed data on singers whose relative superiority could not be determined with a high level of confidence. These analyses also consistently showed intermediates were more accurate than experts when rank ordering low-quality singers. For readability, we do not describe these nine additional significance tests here but report them in Appendix B.

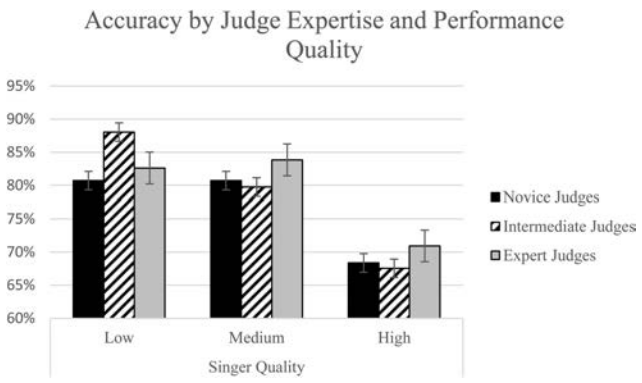


Figure 3. Study 1 within-group judgment accuracy by expertise level and performance level.

Discussion

Two separate analyses confirmed the central hypothesis of this article. Interjudge alignment is a reliable indicator of judgment quality, and based on this metric, intermediates provided better judgments of low-quality performers than did experts. A separate analysis, based on the "true" quality of the singers as estimated by a model, also found intermediates made more accurate judgments of low-quality singers than did experts. The simplest and most satisfying explanation for these corroborating analyses is the one we propose: experts are less capable than intermediates at judging low-level performers, because experts are less able to discriminate among low-level performances. The consistency of the empirical results across two different analyses is compelling, but we conducted two additional analyses related to the Bayesian analysis that provide further evidence in favor of the validity of the model results and the robustness of the analysis. These analyses are found in Appendix C and Appendix D.

As expertise in a performance domain increases, the ability to accurately and reliably distinguish differences in low-level performance diminishes, which results in experts providing less accurate judgments of low-level performers than intermediates. We found support for this effect in expert and intermediate singers' judgments of low-quality singers. A variety of analyses showed that when rank ordering singers, intermediates were more likely than experts to put low-quality singers in the correct order. To provide evidence of the general nature of the effect, we conducted Study 2 in a completely different performance domain—ballroom dance.

Study 2: Accurate Judgment of Ballroom Dance Competitors

In Study 2, we investigate the impact of expertise on judgment accuracy in an additional domain. We propose this effect occurs in complex subjective judgments of competitive performance. It occurs in domains in which expertise is developed through study and practice of ideal performance standards. Champion ballroom dancers work for years at developing their ability to perform a variety of dance movements according to the accepted ideal performance standards. The judges of these competitions are consistently former ballroom dance champions; thus, we expect to extend our findings from the domain of vocal performance to the complex motor skill of competitive ballroom dance.

Method

Participants. The five experts in our study were all former ballroom dance champions. We recruited intermediates and novices from ballroom dance classes being offered at the university that hosted the competition. The six novices were nearing the end of their first ballroom dance class. The seven intermediates were all enrolled in their third semester (advanced level) of a particular series of ballroom dance class. The novices had 2 and a half months of ballroom dance experience. The intermediates had between 1 and 3 years of ballroom dance experience. The experts had at least 15 years of ballroom dance experience.

Procedure. We conducted this field study at the National Ballroom DanceSport Championship, which takes place every year in the arena of a private university in the western United States. The judges evaluated all rounds of two different competitions: a beginner-level quickstep and an advanced-level waltz. While the experts stood around the dance floor to judge the couples, the intermediates and novices sat in the first two rows of seating. Because we were unable to obtain the same judging forms available to the experts, the intermediates and novices had to write down the number of each couple they wished to call back, rather than circle the number listed on their form. Thus, the intermediates and novices were at a disadvantage for providing good judgments relative to the experts.

The beginner-level competition was danced in six rounds, starting with 76 couples in Round 1 and decreasing to 54, then 40, then 25, then 14, then six couples over the subsequent rounds. The advanced-level waltz was danced in five rounds, starting with 55 couples in Round 1 and decreasing to 35, then 25, then 11, then seven couples over the subsequent rounds. In each round, judges received specific instructions on the number of couples to call back. We examine the judgments of experts, intermediates, and novices in Rounds 1 through 5 of the beginner-level competition and in Rounds 1 through 4 of the advanced-level competition. Because the last round of each competition utilized a new judgment (ranking rather than call-backs), we do not examine this last round of judgments.

Results

Consistent with Study 1, we begin by examining interjudge alignment as a measure of judgment quality. We calculated interjudge alignment using all possible pairwise comparisons between judges. For example, the first round of the beginner-level competition included 76 couples. Each pairing of the five experts could be compared across the 76 couples, leading to 760 marks of alignment (10 pairings for all 76 couples). Table 1 shows the interjudge alignment for every round of competition for each level of judging. In Appendix E, we also provide an alternate analysis of accuracy based on judges' alignment with latent dancer quality extracted from a model, similar to the analysis we presented in Study 1. This alternate analysis offers convergent evidence of the hypothesized effect.

We focus first on the beginner-level competition, in which we predicted intermediates would provide better judgments than both experts and novices. In both of the first two rounds, the difference in alignment between the experts' and intermediates' evaluations was not significant, $\chi^2(1) = 1.05$, $p = .30$; $\chi^2(1) = .56$, $p = .45$. In addition, intermediate alignment in these first two rounds is actually lower than that of experts, which is inconsistent with our hypothesis. Post hoc, it seems reasonable to conclude intermediates' unfamiliarity with the judgment procedure hindered their judgments in these first two rounds. The difficulty of judging a ballroom dance competition comes not only from the difficulty inherent in judging the quality of the dancing, but also in the logistical challenge of allocating one's attention to completing the required judgments in the short timeframe provided. The experts had practiced this logistical challenge hundreds of times in previous competitions. The intermediates and novices, on the other hand, had never performed such judgments. Their compliance with the judging requirements reflects this difficulty. In Round 1, judges were asked to recall 50 of the 76 couples. Four of the five experts recalled 50 couples and one recalled 49. Intermediates and novices recalled between 31 and 47 couples. In Round 2, judges were asked to recall 36 of the 54 couples. Experts recalled between 35 and 37 couples, and intermediates and novices recalled 24 to 35 couples. By Round 3 of the beginning-level competition, the intermediates and novices found their way. When asked to recall 24 of 40 couples, intermediates and novices all recalled between 21 and 24 couples and they continued to comply with judging requests in the remaining rounds. (The order of the rounds of dancing was as follows: Beginners Round 1, Beginners Round 2, Advanced Round 1, Beginners Round 3, Advanced Round 2, Beginners Round 4, Advanced Round 3, Beginners Round 5, Advanced Round 4).

Figure 4 shows the pooled alignment from Rounds 3 through 5 for each class of judges. Consistent with our hypothesis, the pooled alignment from Rounds 3 through 5 was significantly higher for intermediates ($M = 64.6\%$) than for experts ($M = 60.1\%$); $\chi^2(1) = 4.82$, $p < .05$, or novices ($M = 60.8\%$); $\chi^2(1) = 4.19$, $p < .05$, Cohen's $d = .08$, LCL = .003, UCL = .15. The difference between experts and novices was not significant, $\chi^2(1) = .12$, $p = .73$. Examining the results of each round of dancing individually, we found that in Round 3, intermediates' judgments were signif-

Table 1
Judgment Alignment by Round, Competition Level, and Level of Judges' Expertise

	Percent alignment				
Beginner-level competition	Round 1 ($n = 76$)	Round 2 ($n = 54$)	Round 3 ($n = 40$)	Round 4 ($n = 25$)	Round 5 ($n = 14$)
Experts	72.5	64.7	62.6	58.4	55.7
Intermediates	70.4	62.4	69.1	63.0	55.1
Novices	67.5	61.6	63.6	59.7	55.2

	Percent alignment			
Advanced-level competition	Round 1 ($n = 55$)	Round 2 ($n = 35$)	Round 3 ($n = 25$)	Round 4 ($n = 11$)
Experts	74.5	84.6	64.8	73.3
Intermediates	70.0	66.8	67.0	65.0
Novices	67.9	61.7	54.1	54.9

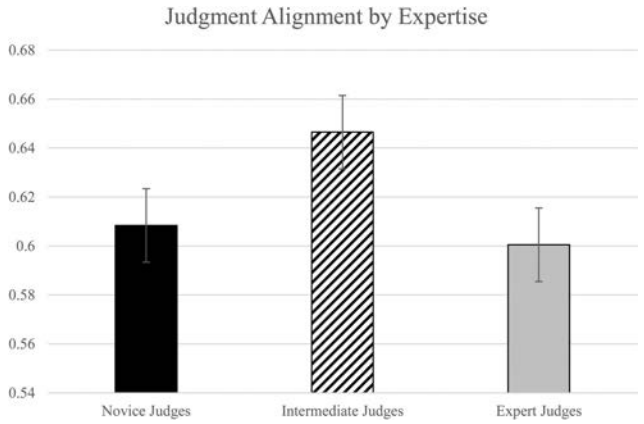


Figure 4. Study 2 within-group judgment alignment by expertise level for beginner-level dancers.

icantly better than both experts' judgments, $\chi^2(1) = 4.32$, $p < .05$, Cohen's $d = .12$, LCL = .007, UCL = .23, and novices' judgments, $\chi^2(1) = 3.90$, $p < .05$, Cohen's $d = .11$, LCL = .001, UCL = .21. This pattern continued into Round 4, though due to small sample sizes (only 25 couples danced in this round), these differences were not significant (intermediates vs. experts: $\chi^2(1) = 1.55$, $p = .21$; intermediates vs. novices: $\chi^2(1) = 1.02$, $p = .31$). Finally, in Round 5, judges at all three levels performed no better than chance, indicating that at this point of the competition, the performance ability was so similar that none of the three groups of judges was able to reliably distinguish their performance.

We hypothesized that experts would be less accurate than intermediates when judging beginner-level competitions, but more accurate when judging advanced-level competitions. Next, we investigate judgments of advanced-level performance. We focus on Rounds 2 through 4 of this competition, in keeping with our prior observations about the intermediates' and novices' learning on the logistical dimension of judgment. In contrast to the beginner-level competition, in this advanced-level competition, experts' judgments were more aligned ($M = 75.8\%$) than intermediates' judgments ($M = 66.6\%$); $\chi^2(1) = 19.4$, $p < .01$, Cohen's $d = .19$, LCL = .10, UCL = .27, and novice judgments were the least aligned ($M = 57.9\%$); experts vs. novices: $\chi^2(1) = 60.9$, $p < .01$, Cohen's $d = .37$, LCL = .28, UCL = .47; intermediates vs. novices: $\chi^2(1) = 20.5$, $p < .01$, Cohen's $d = .18$, LCL = .10, UCL = .26. This finding is consistent with our proposal. At higher performance levels, expertise provides the judge with both greater knowledge of the ideal and greater ability to discriminate between two advanced performers, leading to higher judgment accuracy. Examining the results of each round individually, we find that in all four rounds of advanced-level competition, experts were significantly better judges than novices (all $p < .01$). Furthermore, in two of four rounds (Rounds 2 and 4), experts were significantly better judges than intermediates, $\chi^2(1) = 37.4$, $p < .01$; $\chi^2(1) = 8.3$, $p < .01$; marginally significantly better judges in one round, Round 1, $\chi^2(1) = 3.80$, $p = .05$; and not significantly different in Round 3, $\chi^2(1) = .40$, $p = .52$. Intermediates were significantly better judges than novices in two of four rounds, Rounds 3 and 4, $\chi^2(1) = 15.9$, $p < .01$; $\chi^2(1) = 4.4$, $p = .03$); and marginally significantly better judges in Round 2, $\chi^2(1) = 3.4$, $p = .07$. These results are consistent with the notion that increased expertise leads to greater accuracy in the judgment of advanced-level competition.

An alternative explanation for these results is that intermediates do not exhibit superior judgments; rather, they simply use more similar heuristics than either experts or novices in making their judgments, as was discussed in Study 1. If the degree of agreement in utilized heuristics (as opposed to accuracy) drove judgment alignment, we would expect to see the same pattern of alignment in both the beginner and advanced dance events. Instead, the alignment of experts, intermediates, and novices changed from the beginner to the advanced event, with intermediates providing superior judgments in the beginner event but experts providing superior judgments in the advanced competition. This pattern is precisely the one our theory predicts, but is inconsistent with a heuristic-driven explanation. In addition, a heuristic-based explanation of these results would predict novices show more alignment than intermediates or experts. Novices have the least knowledge of ideal dance standards, and would thus be the most likely to utilize the same heuristics in judging dance quality (because they have a less diverse set of heuristics from which to choose). Finally, additional evidence from a model-based analysis in which we model the latent quality of each couple using the judgments of all 18 judges weakens this alternative explanation. We describe the model and the results in Appendix E.

Discussion

As hypothesized, we found evidence that intermediates can provide better judgments of lower-level competitions than experts. We propose experts' inability to reliably discriminate among lower-level competitors causes the decreased accuracy of experts. Consistent with this proposition, intermediates' superiority disappeared in the advanced-level event. In this higher-level competition, experts were better able to discriminate among competitors, so their judgments were superior to intermediates' judgments, and intermediates' judgments were superior to those of novices.

Whereas Studies 1 and 2 demonstrate the effect of expertise on judgment accuracy, Study 3 seeks to make a more direct examination of how expertise affects perceptions of performance. Specifically, we investigate how experts' adaptation level affects the mean ratings level they assign to performances as compared to novices' mean ratings level for the same performances. If experts are more sensitive to small deviations from ideal performance, as adaptation level theory proposes, the mean ratings they give to performances should reflect this greater sensitivity.

Study 3: Adaptation Level and Criticality in Expert Judgment

Study 3 seeks to provide evidence of the perceptual difference between experts and novices and thus validate our application of adaptation level theory to the area of competition judgment. Experts' greater discrimination ability near the ideal performance level should lead them to give lower evaluations than novices. Because experts can more easily detect small deviations from the ideal performance standard, if a novice and an expert notice the same performance error, the expert would be more likely to categorize this error as egregious, whereas the novice would be more

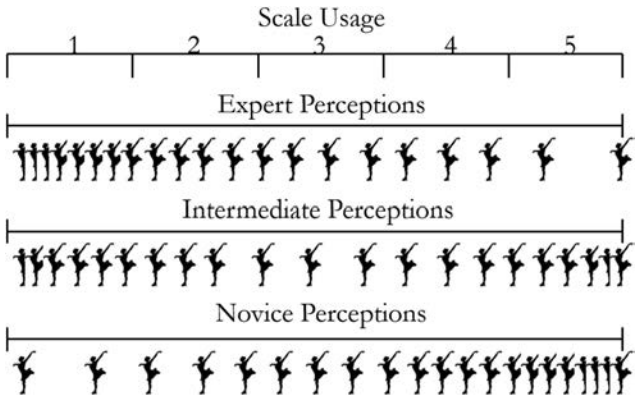


Figure 5. Adaptation level determines scale usage in performance rating.

likely to categorize it as minor. This hypothesized difference in mean ratings is illustrated in Figure 5, in which we illustrate experts', intermediates', and novices' perceptions of a set of dancers on a quality scale. Greater sensitivity to deviations from the ideal causes experts to perceive high-quality performances to be medium quality and medium-quality performances to be lower quality. As a result, experts perceive very few performances as meriting the top score on the scale. On the other hand, greater discrimination ability at lower levels for the novices causes them to use higher-scale ratings. In addition, lower evaluations from experts are also the result of experts' larger knowledge structures, which cause them to observe a greater number of errors than novices. Thus, this study investigates the nature of expert versus novice judgment, both in terms of the absolute number of errors noticed as well as their perceptual differences in judging those errors, or in other words, their evaluations of the level of the performance cues. We hypothesize that experts will notice more errors than intermediates and novices and that they will evaluate the same error more harshly.

Method

Participants. We recruited 57 judges to critique 10 short clips of choral music. To ensure the judges represented a range of expertise levels from novice to expert, we recruited judges both from an undergraduate subject pool and from the same world-famous choir used in Study 1. Further, the judges recruited from this award-winning choir all had received at least a bachelor's degree in music.

Materials. We collected recordings of 10 different choirs singing 30 s of the same section of the well-known Christmas tune "Carol of the Bells."

Procedure. After listening to each music clip, judges provided a rating of the choir on a 5-point scale. After providing this judgment of the overall quality of the choir, judges then listened to the same choir a second time, and provided comments on the choir's deficiencies. Judges were asked to list all the mistakes they heard and rate the severity of each mistake. For example, one judge commented on the deficiencies of Choir 1, "Balance - 2; Consistency of Tone - 1," indicating balance was a moderately severe problem and tone consistency was a severe problem. After rating all 10 music clips on both overall quality and individual mistakes, judges reported their choral-music expertise on the same three-item scale described in Study 1 (Cronbach's alpha = .87).

Results

According to adaptation level theory, experts are more attuned to small deviations from the ideal. As a result, a very good choir that receives a 5 out of 5 rating from a novice would likely receive a lower rating from an expert, because the expert would (a) perceive mistakes the novice did not notice and (b) perceive each mistake to be more severe than would the novice.

To test whether experts are more critical than novices or intermediates in their evaluation of performances, we ran an ordered logistic regression on choir rating with choir fixed effects and expertise as an independent variable. As hypothesized, expertise significantly predicted rating, such that greater expertise led to lower evaluations of choral quality, $\chi^2(1) = 28.2$, $p < .01$. As an illustration, Figure 6 shows the scale usage of novices, intermediates, and experts. (Novices were those with a rating of 2.33 or lower on the three-item expertise scale, intermediates, 3.67 or lower, and experts, 4.0 and above, to yield groupings of size 18, 23, and 16, respectively.) As expected, usage of the lower rating values increased with expertise.

We proposed two explanations for these lower judgments by experts. First, because experts have greater knowledge of ideal performance standards, they should notice more mistakes than less expert singers. All judges were asked to identify mistakes they noticed from each choir and to rate the severity of those mistakes. We found experts noticed more mistakes than intermediates and novices. The average number of mistakes novices, intermediates, and experts noticed was 14, 18, and 29, respectively (novices vs. intermediates: $t(54) = 1.98$, $p = .05$; novices vs. experts: $t(54) = 4.63$, $p < .01$; intermediates versus experts: $t(54) = 2.97$, $p < .01$). Thus, the greater number of mistakes experts noticed can partially explain experts' lower overall evaluations.

A second possible explanation for the experts' lower ratings is their increased discrimination capacity near the ideal. This capacity should cause experts to rate each mistake as more severe than the same mistake noticed by an intermediate or novice. To test whether experts rated individual mistakes more severely than intermediates or novices, we ran another logistic regression on the ratings of the individual mistakes (setting aside the overall ratings

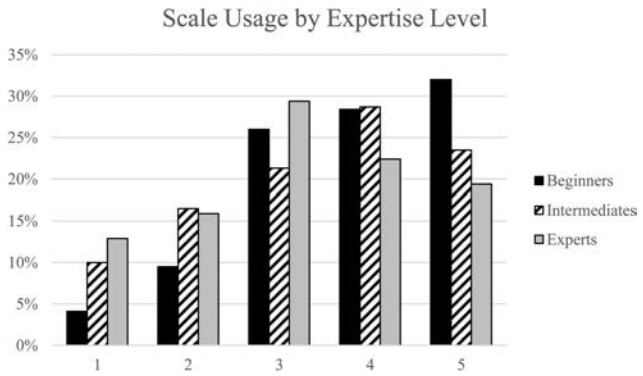


Figure 6. Study 3 share of scale usage for experts, intermediates, and novices.

of the choirs). We employed a choral-music expert to categorize each error so that we could include category fixed effects in this logistic regression. Even when controlling for the type (category) of error on the same music clip (the model also included choir fixed effects), experts gave significantly lower ratings than novices, $\chi^2(1) = 5.04$, $p = .02$. This finding indicates experts' noticing a greater number of errors did not solely cause the lower evaluations. Experts also evaluated each individual error to be more severe than did novices. This finding is consistent with the application of adaptation level theory to competition judgment. Expertise increases the ability to discriminate among performance differences near the ideal, but decreases the ability to discriminate among lower-level performances.

Further analysis of the comment data from judges across the expertise continuum provides additional evidence of the role of expert knowledge structures in judgment differences. We modeled the category comment data using a Dirichlet-multinomial model with the expertise measure as a covariate on each category parameter (Wagner & Taudes, 1986; see Appendix F for details). Table 2 shows the expected category usage for a novice (a score of 1 on the expertise scale) and an expert (a score of 5). Consistent with our expectations, experts' broader knowledge structures led them to notice mistakes from a broader set of categories. The standard deviation of the 15 usage probabilities in the novice column is .060 versus .048 for the expert column. Because experts have a broader set of knowledge about choral music, their comments were less concentrated into individual categories.

This model also provides some insights into the types of choral skills that can be considered low level versus high level. Low-level choral skills include balance, dynamics, and energy. Novices frequently noted these skills, but experts mentioned them less often. (The numbers in Table 2 indicate the relative frequency of each category of mistake, not the absolute frequency. But even after adjusting for the greater number of mistakes noticed by experts, novices notice these categories more frequently than experts.) High-level choral skills include vocal production and vowels, which novices almost never noted but experts mentioned frequently. These data are consistent with prior expertise frameworks (Anderson, 1982; Fitts, 1964), which propose low-level skills are automated by experts and thus less likely to be noticed consciously, whereas novices do not have a knowledge of high-level skills, but experts readily note them.

Table 2
Study 3 Model-Based Analysis of Performance Cues

	Category usage	
Category	Novice	Expert
Intonation	.112	.162
Tempo	.145	.070
Blend	.016	.106
Balance	.253	.067
Articulation	.037	.015
Vowels	.001	.176
Dynamics	.118	.041
Attacks	.010	.038
Unity	.026	.047
Energy	.065	.010
Tone quality	.093	.116
Diction	.067	.054
Style	.027	.029
Vocal production	.009	.033
Phrasing	.021	.035

Discussion

We propose that experts less reliably discriminate among lower-level performances as a natural consequence of their adaptation to higher levels of performance. Their expertise shifts their adaptation level nearer to the ideal, and thus their discriminative capacities diminish when perceiving performances of lower quality (and therefore far from the ideal performance levels). Study 3 provided evidence in support of this proposition. Experts gave lower ratings of performances, indicating a greater sensitivity to deviations near the ideal. Consistent with an ability to recognize additional performance cues, experts noted a greater number of performance errors than intermediates or novices. Furthermore, consistent with adaptation level theory, experts gave lower ratings than novices for the same mistake made by the same choir, giving further evidence of this perceptual shift.

General Discussion

Judges determine the winners of many kinds of competitions. In competitions with a set of objective standards, judges determine the extent to which each competitor matches the ideal standard. The typical assumption is that a person's expertise in the domain determines her ability to accurately make these judgments. Reliable judgment of competitive performance requires an ability to identify, weight, and assess performance cues. Although expertise enhances one's ability to identify and weight performance cues, we show it can hinder one's ability to assess performance cues of novice performance. Thus, experts make poorer judgments of low-level competitions. Intermediates, having a sufficient knowledge of the ideal and a superior ability to discriminate low-level competitors, are better judges of low-level performance than experts. Two studies provided evidence for this phenomenon. Study 1 found intermediates were more likely than experts to accurately detect the relative quality of two low-level singers. Study 2 found intermediates to be more accurate than experts in calling back beginner-level ballroom dancers.

Experts' inability to reliably discriminate performance differences at a lower level is caused by perceptual changes that occur as their expertise increases. The ability to discriminate performance differences at levels near their level of expertise is accompanied by a hindrance in the ability to discriminate performance differences far from their level of expertise. We propose this hindrance follows from adaptation level theory, which states the ability to discriminate is greatest near the adaptation level and decreases further away from it (Helson, 1948). Experts' study and development of ideal performance shifts their adaptation level toward ideal performance, decreasing their ability to discriminate at lower levels of performance. Although previous work has investigated the fallibility of recall (Baird, 2003; Mehta, Hoegg, & Chakravarti, 2011) and prediction (Chi, 1978; Glenberg & Epstein, 1987; Hinds, 1999) in experts, competitive judgment is distinct from these findings in that the cognitive task is simultaneous with the performance—it does not require recall of the past or prediction of the future. Study 3 provided evidence of adaptation level

theory (Helson, 1948) in the ratings expert choral judges provided about choral performance. The experts' ratings of the choirs were significantly lower than the ratings of those with less expertise. In addition, we found evidence that experts' ratings of individual mistakes were also lower than novices' ratings of those same mistakes. This finding supports our proposal that expertise increases discriminative capacity at higher performance levels (and thereby decreases discriminative capacity at lower performance levels), consistent with adaptation level theory.

Addressing a common criticism of expertise studies, namely, an overreliance on artificial, lab-induced experts (Camerer & Johnson, 1991), we studied judgments made by nationally and internationally recognized experts whose expertise was developed through years of training and performance. The use of actual experts is crucial to our investigation, because experts' reduced discriminative capacity in assessing beginner-level performance cues results from the natural development of expertise, and this perceptual shift may require a great deal of time and practice to occur. Our studies rely on expertise that results from the development of cognitive mechanisms acquired through extended periods of practice and not from inferences about expertise made from the study of novices (Ericsson, 2014). Whereas previous studies have investigated the amount of time required to develop exceptional expertise in a field (Ericsson & Lehmann, 1996; Ericsson, 2014), future research could investigate the minimum amount of time required for experts to develop new cognitive mechanisms that lead to adaptation and hence changes in judgment.

The use of absolute experts rather than merely relative experts was also critical to our investigation. The term "expert" can be applied in a relative sense to any group with knowledge that exceeds that of another group, but the effects we observed are unlikely to obtain in relative experts who have not fully adapted to a higher performance standard. Although running field studies with absolute experts poses logistical challenges, because experts are generally a very small subset of those in the field (e.g., ballroom dance), absolute experts are a group that have most likely adapted to high performance standards.

The use of absolute experts in our studies is simultaneously an important contribution and a limitation of this research. Because absolute expertise of the variety we study cannot be induced within a short lab session, we did not manipulate expertise in our studies. As a result, proving the causal role of expertise in the effects we studied is more difficult. We encourage future research to address this limitation.

Future research is also warranted on the boundary conditions of experts' poor judgment of beginners. Several open questions remain as to how and under what training conditions expertise produces a sufficient upward shift in adaptation level to produce the demonstrated effect. Is there a difference between an "expert" and an "expert judge"? We predict an "expert judge"—someone who receives focused training on judging competitors, independent of his or her ability to perform at a high level—would be less likely to show lower accuracy in the judgment of beginner-level performers. Expertise decreases accuracy in the judgment of low-level competitors by diminishing the expert's ability to discriminate performance quality at these low levels. Focused training on judging a specific performance level would likely reverse this poor discriminatory ability, but other unintended consequences may exist, such as less sensitivity to their current performance level. Future research should explore what factors overcome the limitations produced by adaptation levels.

Finally, in a domain as complex as competition judgment, alternate explanations exist as to how expertise decreases judgment accuracy in the domain of beginner-level performance. One possible alternate explanation is that experts' greater knowledge of the many factors that constitute an ideal performance leads them to give more variant judgments than novices or intermediates. In other words, perhaps experts are not less able to discriminate among low-level performers at all but are simply using more varied judgment criteria. While usage of more varied judgment criteria may be a contributing factor to the effects we demonstrate, our results provide several arguments against a more varied judgment criteria being the sole explanation for the results. First, analyses of the internal consistency of expert judges in Study 1 showed experts were less reliable than intermediates or novices at discriminating among beginner-level competitors (see Appendix D). Second, this alternate explanation would expect to find a lower accuracy from experts at all levels of competition, and not just when judging beginners. Instead, our results consistently showed experts made less accurate judgments only when judging beginners. Third, this alternate explanation would also predict the highest level of accuracy for novices, because they have the narrowest knowledge on which to base their judgments. Instead, we found intermediates produce the most accurate judgments of low-level performers, consistent with our theory. Despite these arguments against this alternate explanation, future research into the role of experts' expansive knowledge structures on their (in)ability to provide accurate judgment could prove fruitful.

Our research differs from investigations into expertise in the domains of accounting, clinical medical judgment, and chess in that we investigate judgments of competitive performance that are simultaneous with the observed performance. This scenario is generally the case in competitive performance, where judges provide their marks immediately following a performance evaluated in real time. The need to evaluate in real time likely accentuates the differences in novice, intermediate, and expert judgments. Experts are superior in their ability to detect performance flaws and weight various cues to construct a global judgment, and the time pressure inherent in judging competitor performance likely accentuates this superiority in experts. But where experts are inferior, namely, in their ability to accurately detect the severity of flaws from beginner-level performers, the time pressure inherent in competition judgment might also accentuate their inferiority on this dimension.

In this article, we focus on the benefits and limitations of expertise on judgments of varying levels of competitor performance. A natural extension of this research would be to question whether novices should choose to receive instruction from expert or intermediate instructors. Although our results would indicate expert instructors might have greater difficulty evaluating the novice's performance, the expert instructor's knowledge structure might compensate for their judgmental limitations by aiding the instructor with problem solving and assisting with pattern development. Initial work in this area by Hinds, Patterson, and Pfeffer (2001) indicates that in many beginner tasks, novices make better initial instructors than experts. Further research into the impact of expert, intermediate, and novice instructors on novice learning could provide an impactful solution to a decision millions

of consumers frequently encounter (e.g., finding a tutor or piano teacher for a child).

This article demonstrates expertise can hinder judgment of novice performance. Just as world-renowned experts such as Itzak Perlman, Greg Louganis, George Foreman, Nadia Comaneci, Yo Ma, Alessandra Ferri, and Renee Fleming would likely be reluctant to trust the evaluation of their performance to novices, novice performers should be reluctant to entrust their performance evaluation to experts.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27. http://dx.doi.org/10.1037/0033-2909.102.1.3

Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of consumer expertise. *Journal of Consumer Research, 13,* 411–454. http://dx.doi.org/10.1086/209080

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406. http://dx.doi.org/10.1037/0033-295X.89.4.369

Baird, R. R. (2003). Experts sometimes show more false recall than novices: A cost of knowing too much. *Learning and Individual Differences, 13,* 349–355. http://dx.doi.org/10.1016/S1041-6080(03)00018-9

Beilock, S. L., Carr, T. H., MacMahon, C., & Starkes, J. L. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied, 8,* 6–16. http://dx.doi.org/10.1037/1076-898X.8.1.6

Bergee, M. J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55,* 344–358. http://dx.doi.org/10.1177/0022429408317515

Billeter, D. M., Kalra, A., & Loewenstein, G. (2011). Underpredicting learning after initial experience with a product. *Journal of Consumer Research, 37,* 723–736. http://dx.doi.org/10.1086/655862

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

Camerer, C. F. (1981). *The validity and utility of expert judgment* (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). New York, NY: Cambridge University Press.

Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141–189). Hillsdale, NJ: Erlbaum.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81. http://dx.doi.org/10.1016/0010-0285(73)90004-2

Chi, M. T. (1978). Knowledge structures and memory development. *Children's thinking: What develops, 1,* 75–96.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152. http://dx.doi.org/10.1207/s15516709cog0502_2

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague, the Netherlands: Mouton.

Dinnerstein, D. (1965). Intermanual effects of anchors on zones of maximal sensitivity in weight-discrimination. *The American Journal of Psychology, 78,* 66–74. http://dx.doi.org/10.2307/1421082

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7,* 86–106. http://dx.doi.org/10.1016/0030-5073(72)90009-8

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59,* 562–571. http://dx.doi.org/10.1037/h0037164

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416. http://dx.doi.org/10.1037/0033-295X.85.5.395

Ericsson, K. A. (1990). Peak performance and age: An examination of peak performance in sports. In P. B. Baltes & M. M. Baltes (Eds.), *Successful aging: Perspectives from the behavioral sciences* (pp. 164–195). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511665684.008

Ericsson, K. A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population. *Intelligence, 45,* 81–103. http://dx.doi.org/10.1016/j.intell.2013.12.001

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49,* 725–747. http://dx.doi.org/10.1037/0003-066X.49.8.725

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305. http://dx.doi.org/10.1146/annurev.psych.47.1.273

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278. http://dx.doi.org/10.1146/annurev.psych.59.103006.093629

Fiske, H. E., Jr. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education, 23,* 186–196. http://dx.doi.org/10.2307/3344643

Fitts, P. M. (1964). Perceptual-motor skill learning. *Categories of human learning, 47,* 381–391.

Frederick, D. M., & Libby, R. (1986). Expertise and auditors' judgments of conjunctive Events. *Journal of Accounting Research, 24,* 270–290. http://dx.doi.org/10.2307/2491134

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 387–396. http://dx.doi.org/10.1037/0033-2909.105.3.387

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15,* 84–93. http://dx.doi.org/10.3758/BF03197714

Haier, R. J., Siegel, B. V., Jr., MacLachlan, A., Soderling, E., Lottenberg, S., & Buchsbaum, M. S. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task: A positron emission tomographic study. *Brain Research, 570,* 134–143. http://dx.doi.org/10.1016/0006-8993(92)90573-R

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika, 57,* 97–109. http://dx.doi.org/10.1093/biomet/57.1.97

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review, 55,* 297–313. http://dx.doi.org/10.1037/h0056721

Hill, N. M., & Schneider, W. (2005). Brain changes in the development of expertise: Neuroanatomical and neurophysiological evidence about skill-based adaptations. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 653–682). New York, NY: Cambridge University Press.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied, 5,* 205–221. http://dx.doi.org/10.1037/1076-898X.5.2.205

Hinds, P. J., Patterson, M., & Pfeffer, J. (2001). Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *Journal of Applied Psychology, 86,* 1232–1243. http://dx.doi.org/10.1037/0021-9010.86.6.1232

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

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74,* 657–690. http://dx.doi.org/10.1037/0021-9010.74.4.657

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

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20,* 171–178. http://dx.doi.org/10.1080/10400410802059929

Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly, 49,* 260–265. http://dx.doi.org/10.1177/001698620504900307

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education, 56,* 322–337. http://dx.doi.org/10.1177/0022429408328934

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ: Erlbaum.

Mandel, N., & Johnson, E. J. (2002). When web pp. influence choice: Effects of visual primes on experts and novices. *Journal of Consumer Research, 29,* 235–245. http://dx.doi.org/10.1086/341573

Mann, D. T., Williams, A. M., Ward, P., & Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: A meta-analysis. *Journal of Sport & Exercise Psychology, 29,* 457–478.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press. http://dx.doi.org/10.1037/11281-000

Mehta, R., Hoegg, J., & Chakravarti, A. (2011). Knowing too much: Expertise-induced false recall effects in product comparison. *Journal of Consumer Research, 38,* 535–554. http://dx.doi.org/10.1086/659380

Meyer, M. A., & Booker, J. M. (2001). *Eliciting and analyzing expert judgment: A practical guide* (Vol. 7). Philadelphia, PA: SIAM. http://dx.doi.org/10.1137/1.9780898718485

Mills, J. (1991). Assessing musical performance musically. *Educational Studies, 17,* 173–181. http://dx.doi.org/10.1080/0305569910170206

Park, C. W., Mothersbaugh, D. L., & Feick, L. (1994). Consumer knowledge assessment. *Journal of Consumer Research, 21,* 71–82. http://dx.doi.org/10.1086/209383

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology & Marketing, 26,* 470–478.

Reilly, T., & Secher, N. (1990). Physiology of sports: An overview. In T. Reilly, N. Secher, P. Snell, & C. Williams (Eds.), *Physiology of sports* (pp. 465–485). London, UK: E. & F. N. Spon.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian marketing and statistics.* West Sussex, UK: Wiley. http://dx.doi.org/10.1002/0470863692

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 484–494. http://dx.doi.org/10.1037/0278-7393.8.5.484

Smith, B. R. (2004). Five judges' evaluation of audiotaped string performance in international competition. *Bulletin of the Council for Research in Music Education, 160,* 61–69.

Sundberg, S., & Elovainio, R. (1982). Resting ECG in athletic and non-athletic adolescent boys: Correlations with heart volume and cardiorespiratory fitness. *Clinical Physiology, 2,* 419–426. http://dx.doi.org/10.1111/j.1475-097X.1982.tb00048.x

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception: An Interdisciplinary Journal, 21,* 21–41. http://dx.doi.org/10.1525/mp.2003.21.1.21

Tubbs, R. M. (1992). The effect of experience on the auditor's organization and amount of knowledge. *Accounting Review, 67,* 783–801.

Tucker, R., & Collins, M. (2012). What makes champions? A review of the relative contribution of genes and training to sporting success. *British Journal of Sports Medicine, 46,* 555–561.

Wagner, U., & Taudes, A. (1986). A multivariate polya model of brand choice and purchase incidence. *Marketing Science, 5,* 219–244. http://dx.doi.org/10.1287/mksc.5.3.219

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human factors: The journal of human factors and ergonomics society, 45,* 104–111.

(*Appendices follow*)

Appendix A

Study 1 Estimation Procedure

This model of judges' rank orderings utilizes data augmentation, similar to procedures used on threshold models (Rossi, Allenby, & McCulloch, 2005). The observed ranking data are R_{ij}, where i indexes singer and j indexes judge. The model augments each judge's ranking data with that judge's perceived quality of each ranked singer, z_{ij}. The augmented data, z_{ij}, are generated through random draws from a normal distribution that is truncated from below (in the case of the top-ranked singer) or above (in the case of the lowest-ranked singer) or both above and below (in the case of singers ranked 2 through 6). We then use the augmented data to estimate each singer's latent quality, λ_i. Let $z_{r(k)j}$ represent the perceived quality of the singer who received a ranking of k from judge j. So $z_{r(2)j}$ is the perceived quality of the second-ranked singer by judge j. Estimation proceeds by iterating between two steps:

- Step 1: Draw $z \mid R, \lambda$. For all i and j, we draw z_{ij} (judge j's perceived quality of singer i) from a truncated $N(\lambda_i, 1)$. If $R_{ij} = 1$, then $N(\lambda_i, 1)$ is truncated from below at $z_{r(2)j}$. If $R_{ij} = 7$, then $N(\lambda_i, 1)$ is truncated from above at $z_{r(6)j}$. If $R_{ij} = 2, 3, 4, 5$, or 6, then $N(\lambda_i, 1)$ is truncated from above and below by the better- and worse-ranked singer, respectively.
- Step 2: Find $\lambda \mid z$. For all i, draw λ_i (latent quality of singer i). For each i, a new candidate λ_i is drawn by adding a random draw from a $N(0, .5)$ distribution to the current value of λ_i. The likelihood of both the current λ_i and the candidate λ_i are given by $\prod_j N(z_{ij},1)$. The candidate λ_i is accepted if the ratio of the likelihood evaluations is greater than a draw from a random uniform distribution (Hastings, 1970). Because both the current and the candidate receive the same likelihood evaluation from the uniform prior, the prior does not affect the ratio.

Appendix B

Alternate Analyses of Study 1

Because the latent or "true" quality of the 20 singers comes from a model, the latent quality of each singer is known with error. For example, the model gives the latent quality of Singer 19 and 20 to be −1.74 and −1.76, respectively, thus determining singer 19 to be superior to singer 20. However, the model is only 53% confident in this determination. Thus, as a robustness check on the validity of the results reported in the main text, we repeat those analyses after removing data from singing pairs for which the model is less than 95% confident in the relative superiority of one singer over the other. Analysis on this data supported the robustness of the effect. When examining the bottom tercile of singer pairs in this truncated dataset, experts continued to exhibit a lower level of accuracy (M = 86.5%) relative to intermediates (M = 90.9%), though due to a lower sample size and the decreased difficulty of the task (i.e., the relative superiority of singers was known with greater confidence precisely because these singers exhibited a larger difference in quality, so determining their relative quality was easier), this difference was only marginally significant, $\chi^2(1) = 3.13$, $p = .08$, Cohen's $d = .13$, LCL = −.01, UCL = .28. However, if we limit attention to singing pairs in the bottom quartile, the difference in accuracy between experts (M = 82.6%) and intermediates (M = 90.7%) was highly significant, $\chi^2(1) = 6.64$, $p < .01$, Cohen's $d = .23$, LCL = .06, UCL = .41. Similarly, for singing pairs in the bottom quintile, experts show significantly lower accuracy (M = 82.9%) than intermediates (M = 91.4%; $\chi^2(1) = 6.16$, $p = .01$, Cohen's $d = .25$, LCL = .05, UCL = .45).

At the end of the judgment survey, judges were asked to report their singing expertise on a three-item scale. We classified as novices those with a rating of 2.33 or lower on the three-item expertise scale, intermediates, 3.67 or lower, and experts, 4.0 and above. This division of the scale points is roughly equal, with novices being composed of the lowest five possible scores (1.0, 1.33, 1.67, 2.0, 2.33), intermediates being composed of the next four possible scores (2.67, 3.0, 3.33, 3.67), and experts being composed of the top four scores (4.0, 4.33, 4.67, 5.0). This division yielded 16 experts, 54 intermediates, and 44 novices. Conducting the same statistical tests using this alternate measure of expertise provides yet another robustness check. (Recall that the categorization of experts, intermediates, and novices previously was based on participants' open-ended descriptions of their singing expertise).

(*Appendices continue*)

Based on this alternate system of expert qualification (and using the full data set), experts evaluated low-quality singers (pairs of singers in the bottom tercile) with less accuracy (M = 81.8%) than did intermediates (M = 87.1%; $\chi^2(1) = 3.86$, $p < .05$, Cohen's $d = .13$, LCL = .0003, UCL = .26). When we limit the focus to singing pairs in the bottom quartile, experts still demonstrated lower accuracy (M = 77.2%) than intermediates (M = 86.3%); $\chi^2(1) = 7.59$, $p < .01$, Cohen's $d = .22$ LCL = .01 UCL = .38. This comparison between intermediates and experts also yielded a significant difference when we considered only singing pairs in the bottom quintile, $\chi^2(1) = 11.49$, $p < .01$, Cohen's $d = .30$, LCL = .13, UCL = .47. Experts were correct in 75.7% of these judgments, whereas intermediates were correct in 87.9% of these judgments.

Finally, we also conducted these same analyses (with this self-reported measure of expertise) using only data from singing pairs for which the model calculated greater than 95% confidence in the relative superiority of the singers. Again, experts attained lower accuracy (M = 85.9%) than did intermediates (M = 90.1%) in rank ordering singers in the bottom quartile, $\chi^2(1) = 2.84$, $p = .09$, Cohen's $d = .12$, LCL = −.02, UCL = .26. Experts' accuracy level was even lower (M = 82.5%) when rank ordering singers in the bottom quartile, whereas intermediates' accuracy level increased slightly (M = 90.3%); $\chi^2(1) = 6.25$, $p = .01$, Cohen's $d = .21$, LCL = .05, UCL = .38. Finally, accuracy when rank ordering singers in the bottom quintile showed the same pattern. Experts' accuracy level (M = 81.5%) was significantly lower than intermediates' accuracy (M = 92.1%); $\chi^2(1) = 10.5$, $p < .01$, Cohen's $d = .31$, LCL = .12, UCL = .49.

Appendix C

Judgment Accuracy Based on Class-Specific Models

To demonstrate the validity of the model results from Study 1, we estimated the "true" quality of each singer using three different models: one based only on rankings from novices (Model I), another based only on rankings from intermediates (Model II), and another based only on rankings from experts (Model III). Figure C1 shows the accuracy of each class of judges when judging low-quality singers as calibrated by each model. If a biased model drove the results of Study 1, we would expect the results to change dramatically across the three models. Specifically, if bias drove the results, Model I would show novices to have the highest accuracy, Model II would show intermediates to have the highest accuracy, and Model III would show experts to have the highest accuracy. Instead, intermediates showed higher accuracy than experts in their judgments of low-quality singers across all three models. Note that in Model III, the "true" quality was calibrated only on judgments from a relatively small sample of experts, so the posterior estimates strongly favor the experts' judgments. Nonetheless, experts' judgments are still less accurate than intermediates'.

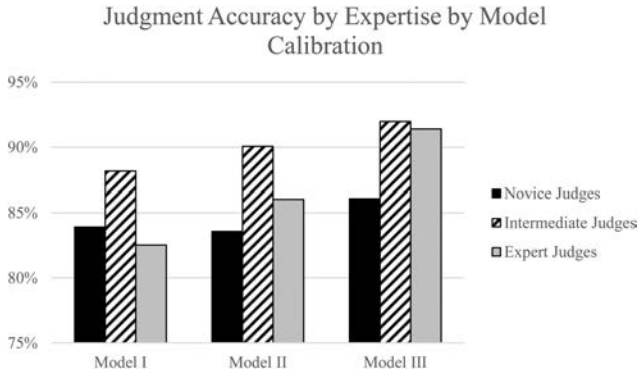


Figure C1. Judgment accuracy on low-quality performers calibrated on three different models.

(*Appendices continue*)

Appendix D

Analysis of Internal Consistency

Of the 5,019 judgments the participants in Study 1 made, 258 were repeated judgments. That is, judgments of the same two singers in both rankings by a participant occurred 258 times. These repeated judgments allowed us to analyze the internal consistency of judges by expertise and by performance level. Internal consistency is a reliable indicator of discriminatory ability. If a judge can reliably discriminate performance from a given performance level, in repeated judgments of the same two performers, she will reliably give a superior rank to one over the other. If she cannot reliably discriminate, her judgments will be more likely to change from one occasion to the next. We hypothesized that experts' discriminatory abilities would be lower for low-level performances and higher for high-level performances. We further hypothesized novices' discriminatory abilities would be higher for low-level performances and lower for high-level performances. To verify this hypothesis, we conducted a logistic regression using as the dependent variable a binary indicator of whether each of the 258 judgments was consistent across the two judgment occasions. As independent variables, we included (a) Qdiff: a measure of the difference in quality between the two singers (a control variable, because a large difference makes the judgment easier and more likely to be consistent); (b) AvgQual: a measure of the average quality of the two singers; (c) AvgQual2: AvgQual squared, to account for nonlinearity; (d) Exp: the expertise of the judge; (e) Exp2: Exp squared, also to account for nonlinearity; (f) AvgQual by Exp: interaction term (see Table D1).

The interaction term is the key test of internal consistency. Our conceptual model predicts discriminability is greatest at the adaptation level of the judge. Thus, internal consistency should be highest among novices when judging low-quality singers, and experts should show the highest internal consistency when judging high-quality singers. This pattern would produce an interaction in the logistic regression. The logistic regression model is shown below (see Figure D1).

The interaction term is significant, as our conceptual model predicts. To summarize the meaning of this model, Figure 6 shows the consistency of each judge level at each singer quality level according to the model. Internal consistency fits our conceptual model perfectly. Importantly, experts' consistency in judging low-quality singers is lower than novices' consistency. This result provides further corroboration of our conceptual model. (Internal consistency is a measure of judgment discrimination, not a measure of judgment accuracy. Thus, we expected novices to have higher consistency than intermediates and experts when judging low-quality singers. That is, we expected internal consistency to fit the pattern indicated in panel 2 of Figure 1, not panel 3 of Figure 1).

Table D1
Variables in the Equation

Variable	B	*SE*	Wald	*df*	Sig.	Exp (B)
Step 1[a]						
Qdiff	.177	.181	.957	1	.328	1.194
AvgQual	−1.073	.472	5.174	1	.023	.342
AvgQual2	−.982	.275	12.789	1	.000	.375
Exp	.470	.336	1.957	1	.162	1.600
Exp2	−.097	.060	2.649	1	.104	.907
AvgQual by Exp	.449	.167	7.231	1	.007	1.567
Constant	−.210	.462	.207	1	.649	.811

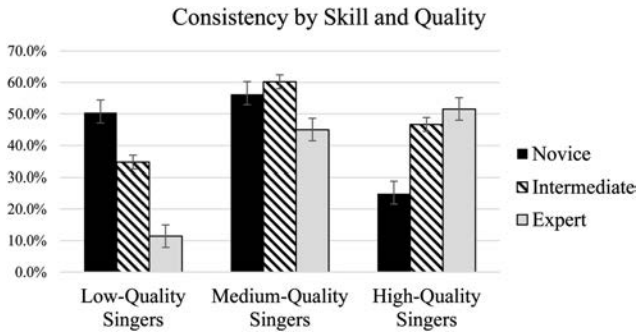


Figure D1. Internal consistency of judges by expertise and performance quality.

(*Appendices continue*)

Appendix E

Model-Based Analysis of Study 2 Judgments

In this model-based analysis, we first model each couple's latent dance quality. We can then use this latent dance quality to determine the accuracy of each judge in calling back the dance couples that deserved a callback. The model of judges' callbacks is a threshold model using data augmentation, similar to procedures outlined elsewhere (Rossi, Allenby, & McCulloch, 2005). The augmented data in this model is each judge's (j) perceived quality of each couple (i) in each round (r), z_{ijr}, which is related to the judges' observed callbacks, M_{ijr}, and each couple's latent quality, λ_i, in the following manner:

$$M_{ijr} = 1 \text{ if } \text{rank}(z_{ijr}) \leq \text{N}_{\text{jr}}, M_{ijr} = 0 \text{ otherwise},$$

$$z_{ijr} = \lambda_i + \varepsilon_{ijr},\ \varepsilon \sim \text{Normal}(0, \sigma_\varepsilon = 1),$$

where N_{jr} is the number of couples called back by judge j in round r, and rank(z_{ijr}) is the rank of couple i in round r for judge j when all z_{kjr} are ordered from highest to lowest. We used a noninformative uniform prior on the λ_i. Estimation proceeded by iterating between the following two steps:

- Step 1: Draw $z \mid M, \lambda$. For all i and r, we draw z_{ijr} (judge j's perceived quality of couple i in round r) from a truncated $N(\lambda_i, 1)$. If $M_{ijr} = 1$, then $N(\lambda_i, 1)$ is truncated from below at z_{hjr}, where z_{hjr} is the maximum z_{kjr} for all z_{kjr} such that $M_{kjr} = 0$. If $M_{ijr} = 0$, then $\text{N}(\lambda_i, 1)$ is truncated from above at z_{hjr}, where z_{hjr} is the minimum z_{kjr} for all z_{kjr} such that $M_{kjr} = 1$.
- Step 2: Find $\lambda \mid z$. For all i, draw λ_i (latent quality of couple i). For each i, a new candidate λ_i is drawn by adding a random draw from a $N(0, .5)$ distribution to the current value of λ_i. The likelihood of both the current λ_i and the candidate λ_i are given by $\prod_j \prod_r N(z_{ijr},1)$. The candidate λ_i is accepted if the ratio of the likelihood evaluations is greater than a draw from a random uniform distribution (Hastings, 1970). Because both the current and the candidate receive the same likelihood evaluation from the uniform prior, the prior does not affect the ratio. (The likelihoods are weighted to give each class of judges equal strength in determining the λ_i. Because we had five experts, seven intermediates, and six novices, likelihoods of the experts were weighted by 7/5, and likelihoods of the novices were weighted by 7/6).

After Step 2, we recorded the alignment of the callbacks of each class of judges in each round with the latent true dancer quality. After a 1,000-draw burn-in period, we simulated 100,000 draws from the posterior distribution.

Table E1 provides the posterior mean alignment with truth of each set of judges for each round of the two competitions according to this model. It also provides the 95% credible interval of each alignment number. First, let us examine the beginner-level competition. The pattern of alignment generally coincides with the analysis provided in the main body of the article. Experts' judgments were superior in Rounds 1 and 2, when the novices and intermediates struggled to manage the logistics of the judgment procedure. However, also similar to the prior analysis, experts' judgments were not significantly better than intermediates' judgments, but they were significantly better than novices' judgments in Round 1. Also similar to the prior analysis, intermediates' superior judgment capacity became evident in Round 3. Intermediates' judgments showed a higher mean accuracy in Rounds 3, 4, and 5. Interestingly, the higher level of accuracy intermediates showed was only significant in Round 4, as opposed to Round 3 in the prior analysis.

Table E1
Study 2 Model-Based Analysis of Judgment Accuracy by Round, Competition Level, and Level of Judges' Expertise

	Percent alignment with true quality				
Beginner-level competition	Round 1 (n = 76)	Round 2 (n = 54)	Round 3 (n = 40)	Round 4 (n = 25)	Round 5 (n = 14)
Experts	78.1 (75.8, 80.0)	70.8 (67.4, 74.1)	68.6 (64.0, 72.0)	62.5 (58.4, 66.4)	59.3 (51.4, 68.6)
Intermediates	74.7 (72.2, 76.7)	67.2 (64.0, 70.4)	71.8 (67.9, 75.0)	71.9 (66.9, 73.7)	60.3 (55.1, 65.3)
Novices	68.3 (66.2, 70.2)	66.2 (63.6, 68.5)	65.5 (61.7, 69.2)	65.4 (60.0, 69.3)	57.4 (52.4, 64.3)
	Percent alignment with true quality			
Advanced-level competition	Round 1 (n = 55)	Round 2 (n = 35)	Round 3 (n = 25)	Round 4 (n = 11)	
Experts	75.5 (72.4, 78.9)	85.3 (79.4, 89.7)	70.5 (64.8, 74.4)	55.4 (41.8, 67.3)	
Intermediates	72.6 (69.9, 74.6)	72.2 (69.0, 74.7)	72.9 (65.7, 78.3)	67.1 (61.0, 71.4)	
Novices	65.8 (63.0, 68.5)	67.0 (63.8, 69.5)	57.7 (53.3, 61.3)	58.3 (48.5, 63.6)	

(*Appendices continue*)

As a formal test of the superiority of intermediate judgment relative to expert judgment, we pooled the accuracy of Rounds 3 through 5, consistent with the analysis provided in the body of the article. The percentage of the 100,000 draws in which intermediates attained a higher accuracy than experts reflects the statistical confidence we can place in intermediates' judgment superiority. Intermediates attained a higher accuracy than experts in 98,968 of the 100,000 draws, for a confidence of almost 99%. Intermediates attained a higher accuracy than novices in 99,978 of 100,000 draws, for a confidence well above 99%. This analysis provides additional evidence that intermediates made more accurate judgments of the beginner-level dancers than did either experts or novices.

Appendix F

Model-Based Analysis of Judges' Evaluations

Consistent with prior literature, we propose experts' extensive knowledge structures cause experts to notice a greater variety of error types than novices (Schoenfeld & Herrmann, 1982). From the judges' comments on the choral mistakes, we were able to make several investigations into the nature of expert knowledge structures. First, we examined how expertise affected the types of choral mistakes noticed. Each judge's set of comments provided a profile of their knowledge of choral-music proficiency by indicating the breadth of choral-music categories noticed. We modeled the comment data using a Dirichlet-multinomial model with our expertise measure as a covariate on each alpha value to lend insight into the effect of expertise on judgment (Wagner & Taudes, 1986). We specified the model as follows:

$$P(\mathbf{Z}|\boldsymbol{\alpha}) = \frac{\Gamma(A)}{\Gamma(A+N)}\prod_{k=1}^{15}\frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)},$$

where $\boldsymbol{Z}$ is the vector $z_1, \ldots z_{15}$ of the number of comments made in each of the 15 categories, Γ is the gamma function, $A = \Sigma\alpha_k$, and $N = \Sigma n_k$. Further, we specified the covariates on each α_k:

$$\alpha_k = e^{\beta_0+\beta_{1k}*X},$$

where X is the self-reported measurement of expertise.

Received September 3, 2015
Revision received May 26, 2016
Accepted May 31, 2016 ■