ONLINE APPENDIX

APPENDIX A: Study 1: Video Collection and Selection Pretest

We collected the videos of circus acts included in Study 1 by hosting an online circus video contest. Through announcements on social media, we invited circus creators to submit online videos of their acts to the contest. A total of 324 videos were submitted by 266 creators (42 creators submitted multiple videos, and no one submitted more than five videos). To ensure the number of forecasts would not be spread too thin among videos, we selected a subset of these videos to serve as the performances on which participants made forecasts in the forecasting survey. To select this subset of videos, we administered a pretest survey to the 266 creators who submitted videos. Each was asked to watch and rate ten randomly selected videos, and 181 creators/hybrids completed this pretest survey (68.05 percent response rate). Of these 181 respondents, 152 then completed the forecasting survey (83.98 percent response rate; 57.14 percent effective response rate).

In the pretest, we trimmed the pool of videos to a subset of 100 videos using four selection criteria. First, we omitted videos submitted by creators who did not complete the pretest survey and thus could not provide informed consent to use their videos in the study. Second, because some videos would not display properly in certain countries due to copyright laws on the music, we omitted videos that did not play properly for more than half of the creators who tried to view them. This ensured that every qualifying video was rated by at least three creators in the pretest.

Third, to ensure that only videos of stage acts were included, we omitted videos with means lower than 5.00 on the item “This video showed an act that can be performed on stage” (1 = “Strongly Disagree” to 7 = “Strongly Agree”); the cutoff of five implied at least moderate agreement. This omitted acts that could exist only in video form. Fourth, to simulate managers’ typical experience of evaluating and selecting among acts that could realistically be included in future shows, we set a minimum threshold on the quality of ideas included. To set a quality threshold but still capture a fairly wide range of quality, we omitted videos that scored below the mean of 4.56 (S.D. = 1.20) on the item “I liked this video,” which creators rated on the same Likert-type scale. These four criteria narrowed the subset to 109 videos, but nine of these videos were taken offline by their owners before the study concluded, bringing the final subset included in the forecasting survey to 100 videos. A total of 61 videos were rated below the mean in liking but met the other three criteria. Because these 61 videos were rated by the creators/hybrids who submitted them, they were included in the audience survey to help test H3–H5, making a total of 161 videos included in the study.

APPENDIX B: Study 1: Supplementary Analyses

Hypothesis 1. As a further test of H1 in terms of placement accuracy, I tested whether the results for H1 replicate when the raw ratings were used, as opposed to the rankings implied by the ratings. I calculated the correlation between each participant's raw forecasts (the three forecasting items averaged together) and the raw audience results (the z-scores of the three dimensions of audience success averaged together) for the set of videos he or she rated. I then calculated the average correlation for participants in each role. In terms of Pearson correlations, creators (mean = .32, S.D. = .31) were significantly higher on average than managers (mean = .22, S.D. = .32), $t(295) = 2.43$, $p < .05$, but not significantly different from laypeople (mean = .28, S.D. = .31), $t(325) = 1.12$, $p = .26$, or hybrids (mean = .31, S.D. = .34), $t(217) = 0.16$, $p = .87$. But in terms of Kendall's tau-b correlations, which may be more appropriate given the small sample size for each participant (10 or fewer videos rated), creators (mean = .23, S.D. = .25) were significantly higher on average than managers (mean = .15, S.D. = .25), $t(295) = 2.96$, $p < .01$, and laypeople (mean = .18, S.D. = .23), $t(325) = 2.96$, $p < .05$, but not significantly different from hybrids (mean = .21, S.D. = .26), $t(217) = 0.55$, $p = .59$. These results provide additional support for H1.

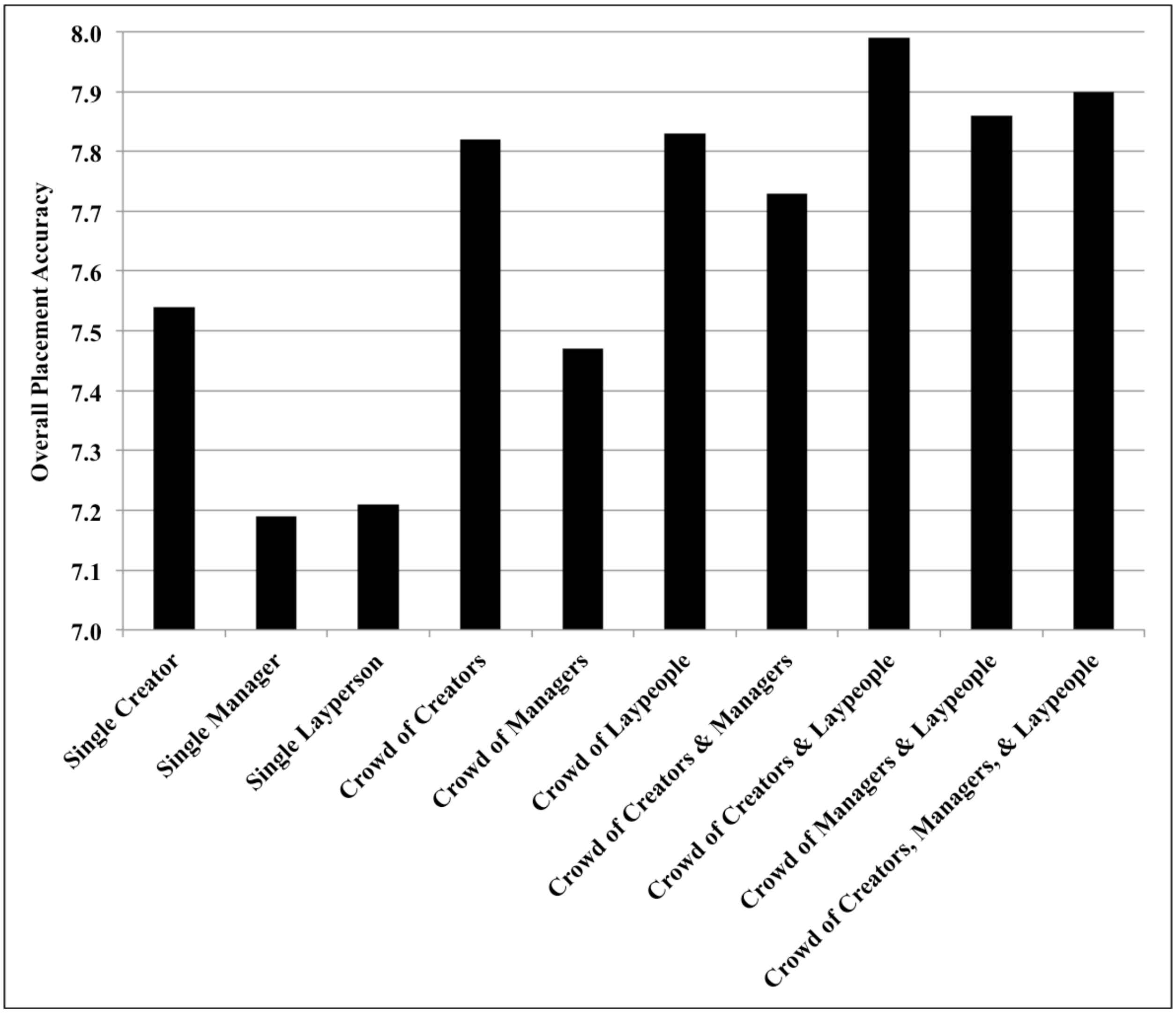
It is possible that creators have an advantage over managers only in avoiding relatively small mistakes and that the two roles are equal in avoiding big mistakes. To test this possibility, I scored big mistakes as overall accuracy scores more than one standard deviation below the mean (mean = 7.34, S.D. = 2.08), which meant scores of 5 or less. I conducted a binary logistic regression model with big mistakes as the dependent variable, role as the independent variable, and number of videos scored as a control. Across all roles, big mistakes were made 17 percent of the time [$b = –1.57$, Wald χ^2 (1) = 507.34, $p < .001$]. Managers were 56.46 percent more likely to make big mistakes than creators [$b = 0.26$, Wald χ^2 (1) = 6.61, $p < .05$], and laypeople were 54.73 percent more likely to make big mistakes than creators, although this effect was marginal [$b = 0.19$, Wald χ^2 (1) = 3.74, $p = .06$]. No other comparisons between roles were significant. Consistent with H1, this suggests that creators had an advantage over managers in avoiding severe false positives and false negatives, which are likely to have greater practical significance than more minor mistakes.

I also tested whether creators had an advantage in identifying the single best idea from the set they evaluated, using the inaccuracy scores for the top-ranked idea (according to the audience) in each participant's set of ideas. As expected, creators (mean = –2.23, S.D. = 2.36) ranked the best idea significantly closer to first than managers (mean = –2.98, S.D. = 2.64), $t(295) = 2.58$, $p < .05$, and laypeople (mean = –3.18, S.D. = 2.49), $t(325) = 3.55$, $p < .001$, but not significantly different from hybrids (mean = –2.55, S.D. = 2.44), $t(217) = 0.79$, $p = .43$. Thus creators had an advantage over managers in identifying the single best idea from the set they evaluated.

Comparing crowds of creators, managers, and laypeople. To help enrich and test the robustness of the results, I conducted supplementary analyses on the crowd-level forecasts of creators, managers, and laypeople. According to theory and research on the "wisdom of crowds"

(Clemen, 1989; Surowiecki, 2005), it is possible that creators no longer have an advantage over managers and laypeople in placement accuracy when individuals' predictions are averaged together (the results on estimation accuracy would be the same at the crowd level). For each of the 100 videos in the forecasting survey, I calculated the crowd average on the three forecasting items for creators, managers, and laypeople. To compare the overall placement accuracy of each individual participant with each crowd, I used these crowd averages to formulate crowd rankings for each subset of videos watched by individual participants. I calculated overall accuracy scores for each crowd (creator, manager, and layperson) in the same fashion as individual participants. Thus I scored the crowds on the same subsets of videos that were randomly assigned to each individual; see figure B1.

Figure B1. Study 1: Overall placement accuracy by role and crowd.



Paired-samples *t*-tests comparing the overall accuracy scores of individuals and crowds showed that all three crowds were significantly more accurate than individuals (mean = 7.56, S.D. = 2.08), including crowds of creators (mean = 7.82, S.D. = 1.84), *t*(4668) = 13.43, *p* < .001, managers (mean = 7.47, S.D. = 1.88), *t*(4668) = 3.67, *p* < .001, and laypeople (mean = 7.83, S.D. = 1.79), *t*(4641) = 12.72, *p* < .001. But the manager crowd was significantly less accurate than the creator crowd, *t*(4668) = 11.66, *p* < .001, and the layperson crowd, *t*(4641) = 9.81, *p* < .001, while the creator and layperson crowds were approximately equal, *t*(4642) = 0.06, *p* = .95. Individual creators (mean = 7.54, S.D. = 1.98) were significantly less accurate than the creator crowd (mean = 7.92, S.D. = 1.82), *t*(1614) = 6.89, *p* < .001, and layperson crowd (mean = 7.95, S.D. = 1.73), *t*(1602) = 6.19, *p* < .001, but approximately equal to the manager crowd (mean = 7.54, S.D. = 1.85), *t*(1614) = –0.06, *p* =.95. Thus an average creator was as accurate as the entire crowd of managers.

To test the possibility that the manager crowd (N = 120) was at a disadvantage because it contained fewer participants than the creator (N = 177) and layperson (N = 150) crowds, I calculated overall accuracy scores for four combined crowds using the forecasts of all the participants in the combined crowd: creator–manager (mean = 7.73, S.D. = 1.80), creator–layperson (mean = 7.99, S.D. = 1.75), manager–layperson (mean = 7.86, S.D. = 1.78), and creator–manager–layperson (mean = 7.90, S.D. = 1.74). The crowd of just creators was significantly more accurate than the creator–manager crowd, *t*(4668) = 4.46, *p* < .001. Also, the manager–layperson crowd was not significantly more accurate than the crowd of just laypeople, *t*(4641) = 1.88, *p* = .06. Thus adding managers made the creator crowd worse and did not significantly improve the layperson crowd. The most accurate forecasts were from the creator–layperson crowd, which was significantly more accurate than the creator crowd, *t*(4668) = 9.08, *p* < .001, the layperson crowd, *t*(4641) = 7.29, *p* < .001, and the creator–manager–layperson crowd, *t*(4668) = 5.78, *p* < .001.

I also scored the crowds in predicting the 1–100 rankings of the 100 videos included in the forecasting survey. Consistent with the other results, the creator crowd (mean = 77.76, S.D. = 18.48) was significantly more accurate than the manager crowd (mean = 73.26, S.D. = 17.97), *t*(99) = 2.28, *p* < .05, but the layperson crowd (mean = 77.41, S.D. = 16.47) did not significantly differ from the creator or manager crowds. Taken together with the other results, these crowd-level results suggest that creators have an advantage over managers regardless of whether individuals or crowds are making the forecasts, but creators may have an advantage over laypeople only at the individual level—crowds of laypeople may be just as accurate as crowds of creators, though a crowd of both creators and laypeople may be most accurate. In general, these crowd-level results support the value of examining the effect that role designs have on creative forecasting accuracy at the individual level, as some roles may provide better raw material for crowd-level forecasts than others.

REFERENCES

Clemen, R. T.
1989 "Combining forecasts: A review and annotated bibliography." International Journal of Forecasting, 5: 559–583.

Surowiecki, J.
2005 The Wisdom of Crowds. New York: Anchor.