מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتّقييم
מיסודן של האוניברסיטאות בישראל

Rate it Twice: Using the Wisdom of Many in One Mind to Improve Performance Evaluation

Report

RR 16-10

December 2016

Meir Barneron ❙ Avi Allalouf ❙ Ilan Yaniv

NITE REPORT RR-16-10

ISBN:978-965-502-203-2

Rate it Twice: Using the Wisdom of Many in One Mind to Improve Performance Evaluation

Meir Barneron

Hebrew University of Jerusalem and National Institute for Testing and Evaluation

Avi Allalouf

National Institute for Testing and Evaluation

Ilan Yaniv

Hebrew University of Jerusalem

December 2016

Author Note

Meir Barneron is a doctoral candidate at the Hebrew University of Jerusalem. Avi Allalouf is the director of scoring and equating at the National Institute for Testing and Evaluation, Jerusalem, Israel. Ilan Yaniv is a faculty member of the Department of Psychology and the Federmann Center for the Study of Rationality, Hebrew University of Jerusalem. The study was supported by a research grant from the National Institute for Testing and Evaluation (2016). Ilan Yaniv was additionally supported by grants from the Israeli Science Foundation (327/10) and the Chicago Wisdom Project (John Templeton Foundation, Subaward FP050019-B). Meir Barneron was supported by the Hoffman Leadership and Responsibility program and the Federmann Center for the Study of Rationality. Address correspondence to meir.barneron@gmail.com or Avi Allalouf at avi@nite.org.il or Ilan Yaniv at ilan.yaniv@mail.huji.ac.il.

Table of Contents

Abstract 2

Rate it Twice: Using the Wisdom of Many in One Mind to Improve Performance Evaluation 3

Method 5

Results 6

Discussion 10

References 13

List of Figures

Figure 1. Accuracy of evaluation: Mean squared errors. 10

Figure 2. Accuracy of evaluation: Mean correlation with criterion. 10

Abstract

The wisdom of crowds refers to the idea that judgmental estimation can be improved by averaging the estimates of different judges. Recent research has suggested that combining multiple estimates made by the same judge also yields accuracy gains, evidencing what is called "the wisdom of many in one mind." The present study extends the use of the wisdom of many in one mind to performance evaluation, specifically the evaluation of essays written as part of standardized college admissions tests. The participants in our field study were professional raters who were asked to evaluate a set of essays twice. The findings suggest that combining evaluations (within raters) is beneficial in that such combinations are more accurate than single evaluations in terms of squared errors and correlations. The within-rater combinations were also compared to combinations of pairs of independent raters. The independent-rater combinations were more accurate than the within-rater combinations. Notably, the within-rater combinations realized two thirds of the accuracy gains obtained from combining independent raters. That performance evaluation can benefit from the wisdom of many in one mind should be of interest to theorists as well as professionals in applied fields, such as human resources, education, and testing & assessment organizations.

Keywords: wisdom of crowd; wisdom of many in one mind; performance evaluation; standardized tests; essay writing task

Rate it Twice: Using the Wisdom of Many in One Mind to Improve Performance Evaluation

The wisdom of crowds refers to the idea that judgmental estimation can be improved by averaging the estimates of different judges (Budescu & Chen, 2014; Surowiecki, 2004; Yaniv, 2004). The reason for this improvement is that judges make different kinds of errors, some of which cancel out each other when the judgments are aggregated. Consequently, the error of the combined judgment tends to be smaller than the errors of the individual judgments (Larrick & Soll, 2006; Yaniv & Choshen-Hillel, 2012).

Recent research further suggests that accuracy gains could be obtained by combining multiple estimates made by the same judge, evidencing what might be called "the wisdom of many within one mind" (Herzog & Hertwig, 2009, 2014a; Vul & Pashler, 2008). Indeed, studies show that asking participants to estimate quantities twice (e.g., "What percentage of the world's airports are in the United States?") and then computing within-person averages yields estimates that are typically more accurate the individual estimates (Hourihan & Benjamin, 2010). Such findings suggest that the judgments produced by the same person are at least partially independent. Judges produced such estimates either because of the passage of time between the first and second estimations (Steegen, Dewitte, Tuerlinckx, & Vanpaemel, 2014) and/or because they were instructed to think differently about the questions (Herzog & Hertwig, 2009).

Our research investigated the wisdom of many in one mind in performance evaluation. In particular, we investigated whether gains could be accrued in evaluating essays. Essays written by college applicants are evaluated by professional raters. The raters' evaluations are important because they affect the applicants' test scores and thus their chances of admission. Our study tested the idea that the wisdom of many in one mind could be used to enhance the accuracy of essay evaluation.

The Psychometric Entrance Test is one such test, developed and administered by the National Institute for Testing & Evaluation (NITE) in Israel. The test has three parts: verbal reasoning, quantitative reasoning, and English as a foreign language (Oren, Kennet-Cohen, Turvall, & Allalouf, 2014). The essay-writing task is one component of the verbal reasoning part. The test takers are required to write an essay (up to 50 lines) within the 30 minutes allotted for the task. This task is used to evaluate the applicants' writing ability as well as their ability to develop a point of view on a given topic and support it with reasons and examples (Allalouf, Shmulevich & Nijem, 2014; Sofer, Pompian & Gafni, 2013).

Similar essay-writing tasks are used in the SAT and the GRE (Briihl & Wasieleski, 2007; Kobrin & Kimmel, 2006; Mattern, Camara & Kobrin, 2007).

The essays are evaluated by professional raters who receive comprehensive training designed to ensure that their evaluations are reliable and fair. The essays are evaluated on content and language. The raters grade each component on a 6-point scale anchored at 1 (very poor) and 6 (excellent). The final grade is the sum of the content and language grades.

The essay task is a major component (25%) of the verbal part of the test battery and affects the applicant's chances of admission to college. It is therefore important to seek the utmost accuracy in grading. A well-accepted method for increasing accuracy is to average the evaluations of two independent raters (Dunbar, Koretz & Hoover, 1991; Penny, Johnson & Gordon, 2000). Here is where the study of the wisdom of many in one mind could be of practical importance. At times there is a shortage of qualified raters, and hence it may be difficult to obtain evaluations from two raters. This is the case, for instance, when the essays are written in rare foreign languages (e.g., Amharic), where there is a shortage of well-trained professional raters. (NITE administers the college admission test in a dozen foreign languages to give a fair chance to applicants from various backgrounds.)

The goal of this study was to explore the benefits of eliciting performance evaluations twice from the same raters. We focused on two comparisons. First, we compared the accuracy of within-rater combinations (the average of two evaluations made by the same rater) to the accuracy of single evaluations. Second, we compared the within-rater combinations to the method of averaging the grades of pairs of independent raters.

The present study is an important extension of previous studies based on lab tasks. Such tasks often involve simple estimations of quantities and take little time. The evaluation of an essay, in contrast, requires greater effort and is more complex. Raters need to read the essays carefully and then evaluate them for language (e.g., choice of words and syntax) as well as content (e.g., line of reasoning, coherence). Indeed it takes about 10 minutes to evaluate a single essay. The extension of the wisdom of many in one mind to judgments involving an evaluative component should be of interest to researchers, as it goes beyond previous results based on the simple estimation of factual information.

Method

Participants

Thirty raters (79.3% women; mean age: 48.03, standard deviation: 15.37) were recruited for the study. All of them were well-trained professional raters and had participated in 12.28 evaluation sessions on average (standard deviation 3.46). They were paid 68 shekels per hour (1 shekel was worth about $0.28 at the time of the study). They worked approximately 9 hours each. The raters were blind to the purpose of the study.

Essay Materials

One hundred essays were randomly selected from a large pool of several thousand. The essays were written by 100 test takers who had taken the Psychometric Entrance Test and were given the same prompt on the same test date.

Procedure

In the first phase of the study, the raters participated in a training workshop conducted online that lasted three and half hours. Two workshops were held by the same instructor on two consecutive days. Fifteen raters participated in each session. In the workshop, raters reviewed the evaluation procedure, discussed different approaches, and received intense training using real essays.

After completing the workshop, the raters were instructed to evaluate 20 essays in two days. The essays were evaluated individually on a secure online platform, using two 6-point scales, for content and language. In evaluating the content of an essay, the raters were asked to judge how well the test taker addressed the topic, the logical connections and level of critical thinking (e.g., the ability to distinguish between opinion and fact, to suggest different perspectives and to contend with opposing viewpoints). The raters were also asked to take note of unnecessary repetitions and vague statements. In evaluating the language of the essay, the raters considered the clarity of the language, word usage, richness of vocabulary, grammar, and the use of linguistic tools to organize the writing (conjunctions, transition sentences and correct paragraphing). It took the raters approximately 10 minutes to evaluate each essay.

A week later the raters were summoned to grade the same set of essays once again. Each rater received the same set of 20 essays in a different order. They were instructed to complete this second evaluation in two days as well. We refer to the grading sessions as Time 1 and Time 2. The gap of one week between the two evaluation sessions was needed

to minimize the chances that the raters would remember their exact prior evaluations. They should still remember the workshop instructions, however.

Design

The 30 raters were randomly divided into five groups with six raters in each. The raters in each group evaluated 20 different essays. One rater dropped out after the first evaluation session. Thus, the database for analysis included 29 raters x 20 essays evaluated in each session x 2 sessions, that is, 1160 data points.

Results

This results section is organized as follows. First, we defined the accuracy measures (dependent variables). Second, we tested the hypothesis that eliciting and averaging two ratings from each rater (within-rater combinations) improves accuracy. Third, we compared the within-rater combinations to independent-rater combinations.

Dependent variables

For each rater and each essay a final grade was computed which equaled the sum of the grades given on language and content. Since the component grades were given on 6-point scales, the final grades ranged from 2 (poor) to 12 (excellent). To assess the accuracy of the grades, it was necessary to have a criterion (true score). The criterion was based on data from an earlier study in which a different sample of 15 independent raters each evaluated all 100 essays used in the present study (Cohen, 2015; Cohen & Allalouf, 2016). For each essay the mean of the 15 grades was computed to form the criterion against which rater accuracy was assessed in the present study. Two measures of accuracy were used, the squared errors of the final grades and the correlations between the final grades and the criterion.

Squared errors. We calculated the squared deviation of the final grade (per rater) from the criterion for each essay. Since the raters evaluated each essay twice, two separate squared errors were obtained, one for each of the final grades, at Times 1 and 2. Within-rater combinations were obtained by averaging the final grades at Time 1 and Time 2, so that the squared deviation of the within-rater combination could also be computed. In addition, we created pairs of independent raters (who had evaluated the same essays) and averaged their grades at Time 1. The squared errors of the independent-rater combinations were also computed.

The hypothesis of the wisdom of many in one mind stipulates that the within-rater combinations should be more accurate than either of the two component grades (at Times 1 and 2). We also compared the accuracies of the within- and different-rater combinations.

Correlations with the criterion. The second dependent variable was the correlation between the final grades and the criterion. Separate correlations were computed for each rater, first between the Time 1 evaluations and the criterion, second between the Time 2 evaluations and the criterion, and third between the within-rater combinations and the criterion. In addition, correlations were computed between the different-rater combinations and the criterion.

Our hypothesis states that the correlations between within-rater combinations and the criterion should be higher than that between the separate Time 1 / Time 2 evaluations and the criterion. Our design also allowed us to compare the accuracies of the within- and the different-rater combinations.

Testing the wisdom of many in one mind

In this section we investigate whether combining grades within a rater yields an improvement in accuracy in our two measures (squared errors and correlations).

Mean Squared Errors. We computed the mean squared error (MSE) for each rater across the 20 essays. We then submitted the MSE to a two-way analysis of variance, with grade type (three levels: Time 1, Time 2, within-rater combination) as a within-subject factor and essay group (five levels) as a between-subjects factor.

The ANOVA revealed a significant main effect of grade type, $F(2,48) = 8.01$, $p < .001$, $\eta_p^2 = .25$. Neither a main effect of the group, $F(4,24) = 1.14$, $p = .36$, $\eta_p^2 = .16$, nor an interaction between group and grade type were observed, $F(8,48) = 1.90$, $p = .08$, $\eta_p^2 = .24$.

We followed up the ANOVA with planned contrasts. In accordance with our hypothesis, the MSE of the within-rater combination (3.09) was significantly less than the MSE of the Time 1 evaluation (3.99), $F(1,24) = 24.14$, $p < .001$, $\eta_p^2 = .50$. Similarly, the MSE of the within-rater combination (3.09) was significantly less than the MSE of Time 2 evaluation (3.57), $F(1,24) = 8.93$, $p < .01$, $\eta_p^2 = .27$. None of the interactions with essay group was significant (*ps* > .13). The difference between the MSEs for Times 1 and 2 (3.57 vs 3.99) was not significant, $F(1,24) = 1.92$, $p = .18$, $\eta_p^2 = .07$.

Finally, following the recommendation of Judd, Westfall and Kenny (2012), we used Mixed Effect Models to analyze the data, with random intercepts for raters and essays; the squared error was the dependent variable. The model included two contrasts (within-rater combinations versus Time 1; within-rater combinations versus Time 2), dummy coded, as fixed effects. The analysis revealed a similar pattern: The within-rater combinations were more accurate than the Time 1 evaluations ($\beta = 0.9$, SE = 0.24, $t(1615.3) = 3.81$, $p < .001$) or the Time 2 evaluations ($\beta = 0.48$, SE = 0.24, $t(1615.3) = 2.02$, $p < .05$). Since random intercepts and slopes were included in the analyses, we can conclude that the within-rater combination fixed effects were not driven by specific essays or raters.

Correlations with the criterion. For each rater, we computed the correlation between the criterion and each of the following grades: Time 1 evaluation, Time 2 evaluation and within-rater combination. The correlations were then submitted to a two-way analysis of variance, with the grade type (three levels: Time 1, Time 2 and within-rater combination) as a within-subject factor and essay group as a between-subjects factor (five levels).

The ANOVA revealed a significant main effect of grade type $F(2, 48) = 6.66$, $p < .01$, $\eta_p^2 = .22$. A main effect of the group was observed, $F(4, 24) = 3.50$, $p < .05$, $\eta_p^2 = .37$, but no interaction between the two factors was found, $F(8,48) = .88$, $p = .54$, $\eta_p^2 = .13$.

In line with our predictions, planned contrasts revealed that the average correlation between the criterion and the within-rater combinations (.74) was higher than the average correlation between the criterion and the evaluations in Time 1 (.66), $F(1,24) = 19.74$, $p < .001$, $\eta_p^2 = .45$. Similarly, the average correlation between the criterion and the within-rater combinations (.74) was higher than the average correlation between the criterion and Time 2 evaluations in (.70), $F(1,24) = 4.76$, $p < .05$, $\eta_p^2 = .17$. Here too we did not find an interaction between essay group and grade type (ps > .41).

The mean correlation between the criterion and the Time 1 grades was not significantly different from the mean correlation between the criterion and the Time 2 grades, .70 vs .66, $F(1,24) = 2.49$, $p = .13$, $\eta_p^2 = .09$.

Comparing within- and independent-rater combinations

Next, we compared the two methods of combining grades, within raters and across raters, in terms of squared errors and correlations.

Mean Squared Errors. For each rater and essay, the data included the squared deviations of both the within-rater combinations and the different-rater combinations from the criterion. We computed the mean squared error for each rater, across the 20 essays. This dependent variable was submitted to a two-way analysis of variance, with grade type (two levels: within vs different-rater combination) as a within-subject factor and group as a between-subjects factor (five levels).

The MSE of the independent-rater combinations (2.58) was significantly less than the MSE of the within-rater combinations (3.09), $F(1,24) = 3.33$, $p < .05$, one tail, $\eta_p^2 = .12$. No interaction was found between essay group and grade type (p = .57).

This pattern was confirmed in a mixed-model analysis, with the squared errors as the dependent variable. The model included a comparison between the within-rater and independent-rater combinations as fixed effects, random intercepts by raters and essays. The analysis revealed that, on average, different-rater combinations were more accurate than within-rater combinations ($\beta = 0.51$, SE = 0.19, $t(1036.4) = 2.74$, $p < .01$).

Correlation with the criterion. For each rater we computed the correlation between the criterion and the within-rater combinations, and between the criterion and the different-rater combination. We then submitted the correlations to a two-way analysis of variance, with the grade type (two levels: within vs different-rater combination) as a within-subject factor and the group as a between-subjects factor (five levels). The analysis revealed that the average correlation for the different-rater combination (.742) was not significantly different from the average correlation for the within-rater combination (.735), $F(1,24) = 0.11$, $p = .37$, one tail, $\eta_p^2 = .005$. The results are plotted in Figure 1 (mean squared errors) and Figure 2 (mean correlations with the true scores).

Figure 1. Accuracy of evaluation: Mean squared errors.

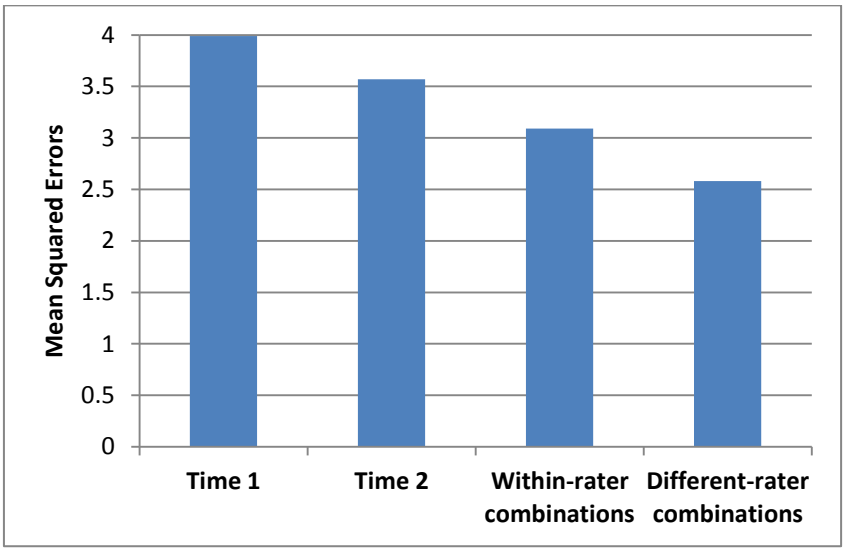
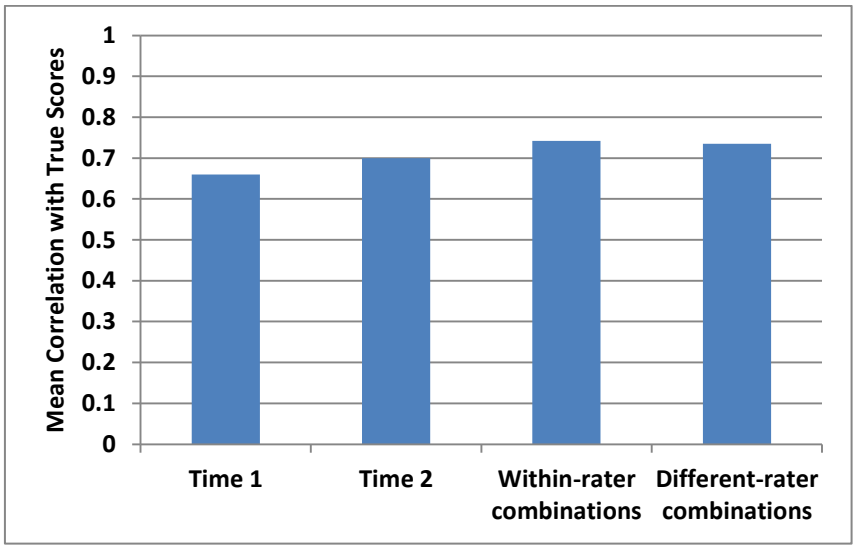


Figure 2. Accuracy of evaluation: Mean correlation with criterion.



Discussion

The present research was motivated by findings in the domain of judgment and decision-making suggesting that accuracy gains could be obtained by combining multiple judgments produced by the same judge (Fraundorf & Benjamin, 2012; Herzog & Hertwig, 2014b; Rauhut & Lorenz, 2010). We investigated the applicability of the wisdom of many in one mind to performance evaluation. In a field study, professional raters performed two separate evaluations of essays written by college applicants taking psychometric entrance tests. We investigated whether combining evaluations within raters would yield greater

accuracy, and found that the within-rater combinations were indeed more accurate than the single evaluations. This improvement was found in two accuracy measures, squared errors and correlations.

We also compared the within-rater and independent-rater combinations (averages of pairs of independent raters). We found that the latter were more accurate in terms of squared errors (though not in terms of correlation with the criterion). Importantly, however, the within-rater combinations realized 64% of the accuracy gains (in terms of squared errors) obtained from averaging pairs of independent raters. The accuracy of the independent-rater combinations provides an important benchmark, since testing centers use this method regularly with the goal of increasing accuracy. The disadvantage of the within-rater combination is appreciable, yet not exceedingly large.

Research contributions

This research provides both theoretical and practical contributions. Finding evidence for the wisdom of many in one mind in performance evaluation is theoretically important to students of judgment and decision-making. Research so far has used quantitative estimations of facts, such as "How many calories are there in a bowl of cooked rice?" or "In what year was the Suez Canal first opened for use?" (Soll & Larrick, 2009; Yaniv & Choshen-Hillel, 2012; Yaniv & Milyavsky, 2007). Such questions have objectively verifiable answers, making the truth criterion unambiguous.

In contrast, the evaluation of performance in the essay-writing task is less clear cut. For example, evaluating essays requires standard criteria. NITE has developed criteria for evaluating the quality of essays which are known to test takers and can be used to prepare for the test. NITE administers intensive training workshops regularly for the raters, with the goal of refreshing the evaluation criteria. Evaluating essays for language and content is far more complex than the estimation tasks used in the past, which involved evaluating the answers to factual knowledge questions. The advantage of combining judgments within one person has nevertheless been documented in the present study.

Our findings also make an important contribution to the domain of performance evaluation. Evidence for the wisdom of many in one mind in this domain should be of interest to researchers in varied applied fields. Evaluating performance accurately is important to professionals in various applied fields, such as human resources, education, testing and assessment, and similar areas.

Concluding remarks

Finding the wisdom of many in one mind is based on the assumption that the evaluations are made (at least partly) independently. In the present study, the second evaluation session took place one week after the first. The raters were blind to the purpose of the study and did not know that they would be required to evaluate the same set of essays again and thus did not attempt to remember their evaluations. Routine application of the within-rater procedure would require certain conditions to reduce the chances that the evaluations would be remembered or stored in some way, so as to maintain independence.

An interesting issue involving the use of the within-rater method is the possibility that the evaluations at Time 1 and Time 2 may differ in accuracy. The trend observed in terms of squared errors and correlations was not statistically significant, yet it is intriguing. The practical objective of performance evaluation is to obtain a rank ordering of the candidates rather than estimate their performance in absolute terms. Raters assigned to evaluate the same essays for a second time may develop expertise at ranking essays according to their relative quality. Consequently, the evaluations at Time 2 could be more accurate than those at Time 1. Future research should explore this issue further.

Finally, the key to the wisdom of crowds is access to opinions and judgments that are produced independently of one another. Future research should focus on alternative methods for maintaining a certain degree of independence between evaluations by the same person. Indeed, asking participants to "think differently" in the second round of evaluation appears to increase the independence among the judgments (Herzog & Hertwig, 2009, 2014b; Winkler & Clemen, 2004; but see White & Antonakis, 2013). The introduction of a larger time gap between the evaluations is another method. We speculate that there might be other methods, such as instructing raters to make lenient or strict evaluations at different times, so as to evoke different perspectives on the object of the evaluation. Future research should focus on the issue of independence to harness the power of the wisdom of many in one mind for performance evaluation.

References

Allalouf, A., Shmulevich, J., & Nijem, H. (2014). *The essay task in the psychometric entrance test. Descriptive and statistic report, 2012-2013* (Report No. 409). National Institute for Testing and Evaluation (In Hebrew).

Briihl, D. S., & Wasieleski, D. T. (2007). The GRE Analytical Writing Test: Description and Utilization. *Teaching of Psychology, 34(3),* 191-193.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. doi:10.1287/mnsc.2014.1909

Cohen, Y. (2015). The "Third Rater Fallacy" in Essay Rating: An Empirical Test. *Meeting of the National Council on Measurement in Education*, symposium on "Issues in Human Scoring of Constructed-Response Items", Chicago, IL.

Cohen, Y. & Allalouf, A. (2016) Scoring of essays by multiple raters- Procedure and descriptive statistic. NITE Technical Report No. TR-2, Jerusalem: NITE

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-303.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38. doi:10.1016/j.jml.2013.10.002

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2013). The crowd within and the benefits of dialectical bootstrapping: A reply to White and Antonakis (2013). *Psychological Science*, *24*(1), 117–119. doi:10.1177/0956797612457399

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. doi:10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 218–232. doi:10.1037/a0034054

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068–1074. doi:10.1037/a0019694

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi:10.1037/a0028347

Kobrin, J. L., & Kimmel, E. W. (2006). Test Development and Technical Information on the Writing Section of the SAT Reasoning Test™. Research Notes RN-25. *College Board*. Retrieved from http://eric.ed.gov/?id=ED562863

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. doi:10.1287/mnsc.1050.0459

Mattern, K. D., & Kobrin, J. L. (2007). SAT writing: An overview of research and psychometrics to date. Retrieved from http://research.collegeboard.org/publications/content/2012/05/sat-writing-overview-research-and-psychometrics-date

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema*, *26(1),* 117–126.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7(2),* 143-164.

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology, 55(2),* 191–197. doi:10.1016/j.jmp.2010.10.002

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. doi:10.1037/a0015145

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: a pre-registered replication study. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00786

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Garden City, NY: Doubleday.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science, 19(7)*, 645–647.

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvement in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science*, 24, 115–116.

Winkler, R. L., & Clemen, R. T. (2004). Multiple Experts vs. Multiple Methods: Combining Correlation Assessments. *Decision Analysis*, *1*(3), 167–176. doi.org/10.1287/deca.1030.0008

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78.

Yaniv, I., & Choshen-Hillel, S. (2012a). Exploiting the wisdom of others to make better decisions: Suspending judgment reduces egocentrism and increases accuracy: exploiting the wisdom of others. *Journal of Behavioral Decision Making*, *25*(5), 427–434. doi:10.1002/bdm.740

Yaniv, I., & Choshen-Hillel, S. (2012b). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology*, *48*(5), 1022–1028. doi:10.1016/j.jesp.2012.03.016

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. doi:10.1016/j.obhdp.2006.05.006