The Discrepancy Between Talented Football Players' Self-Perceived Performance and Their Coach's Evaluation

Tim van der Kooi

s4655796

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 06

Supervisor: Niklas Neumann

Second evaluator: Dr. Ben Gützkow

In collaboration with: Minsoe Veenstra, Aaron Connemann, Lucas Reijnhoudt, Sietse Witteveen and Thomas Klunder.

June, 2025

Abstract

This longitudinal study investigates whether young talented football players from a Dutch premier league club rate their own performances differently than their coach. Existing literature about this mismatch suggests that players probably rate themselves higher than their coach. Getting insights into this potential mismatch can allow sports psychologists to develop targeted interventions aimed at improving communication strategies between coaches and players, thus enhancing player development. Performance was measured by using a single-item questionnaire for every day the 39 players were at the club, across two seasons. Differences were measured by performing a paired samples t-test. Results show that there is a significant difference between players' own performance ratings and their coaches' ratings of their performances with $t(749) = -3.306$, $p < .001$, $d = -0.121$. A possible explanation for the results is that the present research was conducted at the group level, however, previous research suggests that group-level results are not always applicable to individual-level results, which is known as the nonergodicity problem. This study contributes to science by providing an insight into players' and coaches' evaluations of performance over a long period. Gaining knowledge about performance mismatches between players and coaches can contribute to player development.

Keywords: performance, mismatch, football players, coaches, nonergodicity

The Discrepancy Between Talented Football Players' Self-Perceived Performance and Their Coach's Evaluation

During the summer transfer window of 2021 Manchester United legend Cristiano Ronaldo signed for the club for a second time, after playing for Real Madrid and Juventus. Ronaldo became the club's top scorer in his first year back at the club. After this season, Erik ten Hag was appointed as Manchester United's new head coach. Ten Hag wanted to revolutionize Manchester United's way of playing. He often favored other players over Ronaldo in the starting eleven, which irritated the 37-year-old superstar. After a period of limited playing time, Ten Hag wanted to bring Ronaldo on for the final few minutes of a game against Tottenham Hotspur. Ronaldo refused to come on and left the stadium early. Ronaldo later explained that he felt disrespected by Ten Hag. In an interview with Piers Morgan Ronaldo said: "A coach putting me on for three minutes in a game is not acceptable to me. Sorry, I'm not that kind of player. I know what I can give to teams." (Reuters, 2022). It became clear that Ronaldo still perceived himself as the world-class player he used to be, expecting to play a key role for his team. However, ten Hag was not convinced by Ronaldo's performances, which is why he often benched him. The relationship between player and coach did not improve over the next few weeks, ultimately leading to Ronaldo leaving the club in November 2022.

While Ronaldo's case is an extreme example of a player rating his own performance higher than his coach does, discrepancies between self-assessed and coach-assessed performance likely occur at all levels of the sport. Importantly, such mismatches are not necessarily unidirectional. In some cases, coaches may rate a player's performance more favorably than the player does himself. Regardless of the direction, these discrepancies could negatively impact the player-coach relationship, potentially affecting things such as communication, trust, and player development.

Harris & Schaubroeck (1988) conducted a meta-analysis on the discrepancies in how people rate themselves versus how others rate them. They found that across multiple studies, the average correlation between self-ratings (individuals' own assessments of their performance) and supervisor-ratings (evaluations given by their supervisors) was 0.35. This means that while there is a positive relationship, the agreement between self- and supervisor ratings is only moderate, indicating a considerable degree of mismatch. Furthermore, people on average rated themselves 0.70 standard deviations higher than their supervisors' assessments. This suggests that people generally tend to rate themselves more favorably than external evaluators do.

To explain this phenomenon, the researchers identified three types of egocentric bias. The first is defensive bias, which occurs when individuals rate themselves more positively than others do in order to maintain a positive self-image (Holzbach, 1978; Steel & Ovalle, 1984). The second is moderated defensive bias, which suggests that the extent to which a person rates themselves more highly than others do is influenced by a third variable, such as self-esteem (Baird, 1977; Kay, Meyer, & French, 1965). Higher self-esteem leads to more inflated self-ratings, whereas lower self-esteem reduces this effect. Finally, attribution bias, as proposed by attribution theory (DeVader, Bateson, & Lord, 1986; Jones & Nisbett, 1972), suggests that people tend to attribute their own good performances to internal factors and their failures to external factors. Conversely, they tend to attribute others' successes to external factors and others' failures to personal shortcomings. Another bias that supports the idea that people tend to rate themselves more highly than others do is illusory superiority. This well-documented cognitive bias refers to the tendency of individuals to perceive themselves as above average in a given skill, despite the statistical reality that only 50% of people can be above the average.

Hofseth et al. (2017) provide empirical support for these biases. The researchers examined a sample of 338 young Norwegian football players, with an average age of 17.63 years, from elite youth academies, comparing their self-perceived performance ratings to those given by their coaches. Performance ratings were assessed using a structured questionnaire, in which both players and coaches rated the player's overall contribution, tactical execution, and technical skills. The study found that 63% of players rated themselves higher than their coach did, with discrepancies being largest in offensive positions.

Looking back at the study by Harris & Schaubroeck (1988), in addition to biases, they also identified the nature of the evaluated job as a factor influencing differences in correlations between different raters. They found that in more complex jobs, there was lower agreement between different raters. Blue-collar and service jobs are mostly routine-based, and performance is relatively well-defined compared to managerial or professional roles (Harris & Schaubroeck, 1988). Performance in football is also harder to define, as it is highly subjective. Therefore, it can be expected that the correlation between a player's self-perceived performance and their coach's evaluation will be low.

Doeven et al., (2017) investigated whether there is a mismatch between professional basketball players' perceived exertion, perceived recovery, and their coaches' observed exertion and observed recovery. They found that coaches systematically overestimated both players' perceived exertion and perceived recovery. The correlations between player and coach ratings were weak: $r = .25$ for exertion and $r = .21$ for recovery. This suggests that players and coaches often judge their own efforts in different ways.

Furthermore, more research found mismatches between players' perceived exertion and coaches' intended and observed exertion. Brink et al. (2017) studied this among young talented football players and found that coaches generally underestimated players' perceived exertion. Building on this, Brink & Frencken (2018) examined whether giving the coach feedback about how hard players actually trained could reduce this mismatch. They found that, after feedback was provided, the average difference between coaches' observed exertion and players' perceived exertion reduced from 1.0 to 0.7 points, with $p < .003$, on a 1 to 10 scale. This significant reduction in the mismatch shows that, with feedback, coaches can better understand players. It could also be the case that, when there is a mismatch between players' and coaches' ratings on performance such as in Hofseth et al. (2017), the mismatch can be reduced via feedback.

Taking the discussed theories and empirical findings into account, it seems plausible that young talented football players tend to rate their own performance higher than their coach's evaluation. However, most research on self-rated performance and other-rated performance is conducted outside of football. Hofseth et al. (2017) found that most players rate themselves higher than their coach. But they only measured performance ratings once. No research has been done on this discrepancy with multiple data points per player. Therefore, it remains unclear whether the mismatch between self- and coach ratings is a consistent pattern across many observations. To address this gap, the present study examines whether there is a systematic discrepancy between players' self-rated performance and their coach's evaluations, using a dataset containing many individual ratings per player.

Researching the potential discrepancy between players' self-rated performance and their coaches' evaluations is important for various reasons. Firstly, when a player rates himself significantly higher than his coach does, this can lead to miscommunication, frustration on both sides, and a decline in trust. Secondly, if a player's self-assessment is lower than the coach's evaluation, it may indicate a lack of self-confidence, which can negatively impact the player's development, performance, and well-being. Thirdly, identifying these discrepancies allows sports psychologists to develop targeted interventions aimed at improving communication strategies between coaches and players. Finally, examining these mismatches contributes to a better scientific understanding of how sports performance should be assessed and interpreted.

Building on these considerations, this study aims to answer the following question. To what extent do young football players' self-perceived performance ratings differ from their coach's evaluation, and in which direction does this mismatch occur? It is expected that young talented football players will, on average, rate their own performance higher than their coach do.

Methods

Participants

This study was conducted in cooperation with a Dutch premier league (Eredivisie) club. In total, 94 players from their male U-21, U-18, and U-16 teams (15 to 20 years old) participated in the study. After preprocessing, data of 39 players were used (inclusion criteria will be explained later in the methods). The teams all played in the highest national league for their age category. Data was collected for two seasons, with usual weeks containing five days of data collection, consisting of four training days and one matchday. All players were informed about the data collection at the start of the data collection, or when players started playing for the club during the data collection process. By signing an informed consent, players could decide whether they agreed with their data being used for research purposes. All players used in this study have agreed. Their coaches also agreed to their data being collected and used for research purposes. Due to personal data protection, further details regarding the demographics of players and coaches cannot be provided. This study was conducted according to the requirements of the Declaration of Helsinki and was approved by the ethics committee of the Faculty of Behavioral and Social Sciences of the University of Groningen (research code: PSY-2425-S-0016).

Materials and Procedures

A single-item questionnaire was used to measure self-perceived performance. Song et al. (2023) demonstrated that single-item measures can achieve levels of predictive validity comparable to those of multi-item questionnaires, and in some cases may even exceed them. In addition, single-item questionnaires reduce participant burden and fatigue, which may enhance response quality and compliance in repeated-measures designs (Song et al., 2023). Players had to use a tablet provided by the club to fill in the single-item questionnaire. They did so at the end of every training day, after usually two sessions, and after every matchday, maximally 30 minutes after the training/match. The question "How well did you perform today?" was asked to measure performance. It was based on existing literature and adjusted to the current context (e.g., Cohen et al., 2006; Totterdell, 2000). Answers were given on a 0-100 visual analog scale, with the annotation that "0" means "very bad (far below my capabilities) and a "100" means "maximally (to the best of my capabilities)". To answer on the scale, players had to drag a slider across a horizontal line. A number was only shown when they let go of the slider. Answers were given by all players in the locker room one after another. The coaches gave players a score for their performance on the same scale as on which the players scored themselves. The coaches did this once per week, at the start of a week, regarding the average performance of a player during the week before. Coaches did not know about the ratings the player gave himself. From the age of 15 years old, players were familiarized with daily data collection, or since when they first joined the club during the study. Monitoring players is important to the clubs' philosophy and is integral to the development of the players. At multiple times throughout the seasons, the coaches emphasized the importance of the data collection: to enhance performance and to foster individual development. This approach addresses common limitations of self-reports, including social desirability, response fatigue, and compliance issues (Saw et al., 2015).

Statistical Analysis

In order to preprocess the data, average weekscores for the players' self-perceived performance ratings were calculated, and the ratings by the coaches were aligned with the corresponding week of player ratings, since coaches gave their evaluation in the week after the player. These adaptations form the variables "PlayerRating" and "CoachRating". Preprocessing led to the exclusion of 55 players due to the following inclusion criteria: Firstly, a week of data collection for a specific player was deemed usable when the player had a response rate of at least 80% during that week, while the coach also scored the performance of the player for that week. This comes down to 4 out of 5 answers in a standard week. When a week has only 4 or less days of measurements, answers should have been given every day to meet the at least 80% response rate. Thereafter, out of all the weeks a player has participated in the experiment, they should have at least 70% eligible weeks out of the weeks where the coach has also responded, in order to have a good and fair balance between the individual players contributing to the analysis. Finally, players had to have at least 13 eligible weeks, which is approximately 65 days of data collection, to be included in the analysis. This relatively strict border has been set, because looking at the amount of eligible weeks per player revealed that most other options (i.e. 12, 11, or 10 eligible weeks) only added a few additional players to the analysis. All inclusion criteria have been implemented to have a good and fair balance between the individual players contributing to the analysis. These inclusion criteria lead to the omission of 55 out of 94 players, which brings to total of participants used for analysis to 39. The average of data points per player/coach pair is 19.231, with *SD* = 6.623, minimum = 13 and maximum = 36. To test the hypothesis, a paired samples t-test will be performed to see whether PlayerRating is significantly lower than CoachRating across the dataset. The software of JASP 0.19.3 is used for analyzing.

Results

Descriptive Statistics

The descriptive statistics for the values of the variable PlayerRating, measured on a visual analog scale with possible values ranging from 0 to 100, based on the total number of responses for PlayerRating are: valid = 750, *M* = 72.541, *SD* = 9.003, minimum = 32 and maximum = 94.667. The descriptive statistics for the values of the variable CoachRating, measured on a visual analog scale with possible values ranging from 0 to 100, based on the total number of responses for CoachRating are: valid = 750, mean = 74.186, *SD* = 9.846, minimum = 44 and maximum = 100.

Assumption Check

Before a paired samples t-test can be performed, statistical assumptions have to be checked. Firstly, the variables that are going to be compared are dependent of each other. Each PlayerRating must be meaningfully paired with a CoachRating. This assumption is met as this is inherent to the dataset. Secondly, there should be no spurious outliers. This assumption was initially not met, as there was one week of data for one player with a PlayerRating of 81.227 and a CoachRating of 0. This extreme outlier is removed from the dataset, as it is more likely to be an error by the coach than his actual rating, because the CoachRating minimum apart from that score was 44. After omitting this week, the assumption has been met. Finally, the difference scores between the two variables should be distributed approximately normally. A histogram (presented in the appendix) of the differences showed that this assumption has been met.

Main Analysis

This study aims to answer the following question: To what extent do young football players' self-perceived performance ratings differ from their coach's evaluation, and in which direction does this mismatch occur? It was expected that players would rate their own performances higher than their coach would. To test whether players rate their own performances higher than their coach, a paired samples t-test was conducted. The result of this t-test is $t(749) = -3.306$, $p < .001$, $d = -0.121$. So, the result shows that PlayerRating is lower than CoachRating, and that this result is significant, but with a very small effect size. This means that, contrary to the expectation formulated in the hypothesis, the players' perceived performance ratings are not higher than the coaches' perceived performance ratings.

Discussion

Main Discussion

This study investigated whether, and in what direction, there is a mismatch between young talented football players' self-perceived performance ratings and their coaches' evaluations. The expectation was that the players would overestimate their performances in comparison to their coach. This expectation was not met, as there was a significant difference between PlayerRating and CoachRating, with CoachRatings being higher on average than PlayerRatings. However, the effect size of this difference was very small. So, because the means are very close to each other, it would make more sense to treat them as more or less equal.

The results contradict the findings from a similar study conducted by Hofseth et al. (2017), who found that players systematically rate their own performances higher than their coach. The main difference between their study and the present study is that they only measured performance once, whereas this study tried to get a good image of overall performance by measuring repeatedly. Measuring repeatedly can provide a more reliable representation of performance. So, it could be the case that Hofseth et al. (2017) would have found similar results if they also measured repeatedly.

The results can also be explained with the help of Brink & Frencken (2018). They showed that differences between coaches' and players' observed and perceived exertion decreased after the coach received feedback about players' actual exertion. So, when the coach acquired more knowledge about the player, his estimations grew closer to the player's evaluation. In the context of the present study, a similar thing may occur in the opposite direction, namely, feedback coming from the coach to the players. Although it cannot be stated with full certainty, it seems plausible that professional coaches give regular feedback to their players about their performance. This gives players more insight into when their coach thinks they perform well. Thus, it seems logical that players, as they learn what their coach values in terms of performance, will gradually align their own opinion with their coaches' opinion. In addition to this, Harris & Schaubroeck's (1988) statement that tasks become harder to judge when complexity of the task increases may play a role in this. Performing as a football player can be seen as complex, but through feedback from the coach, the task can become less complex because it becomes clearer for a player what the coach expects. The task has thus become less complex, hence easier to judge. This can reduce differences between PlayerRating and CoachRating.

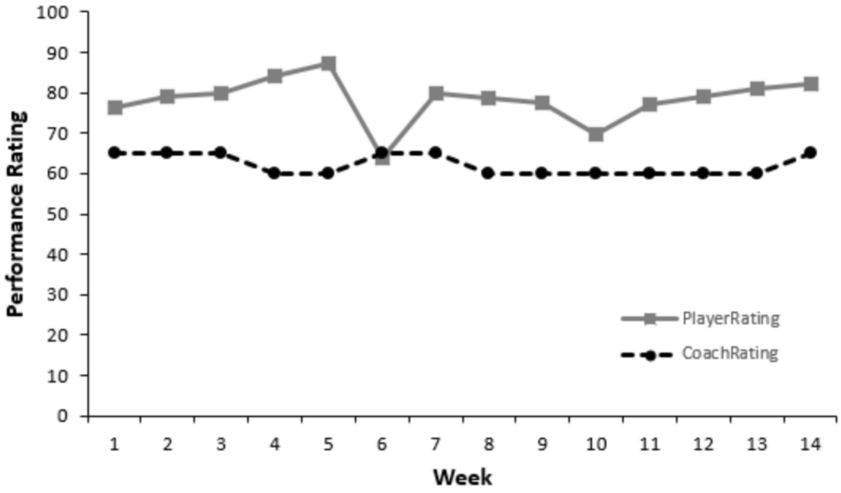
Similar to this study, Doeven et al. (2017) also found that coaches gave higher ratings than players, but for different constructs. They found that coaches systematically overestimated both players' perceived exertion and perceived recovery. However, because these constructs differ a lot from performance in terms of definition, no sensible generalizations can be made from their findings to the findings of the present study.

Furthermore, it is important to note that while this study used a group-level approach, there is a possibility of nonergodicity. This means that for group-based results to apply to individuals, there should be homogeneity (the same statistical models apply to all individuals) and stationarity (statistical variables remain stable over time) (Neumann et al., 2021). Neumann et al. (2021) demonstrated that these requirements are usually not met in sports. In the present study, large individual differences can be seen in terms of differences between the

PlayerRatings and CoachRatings. For instance, some players consistently rated themselves higher than their coach, as you can see for a specific player in Figure 1 below.

Figure 1

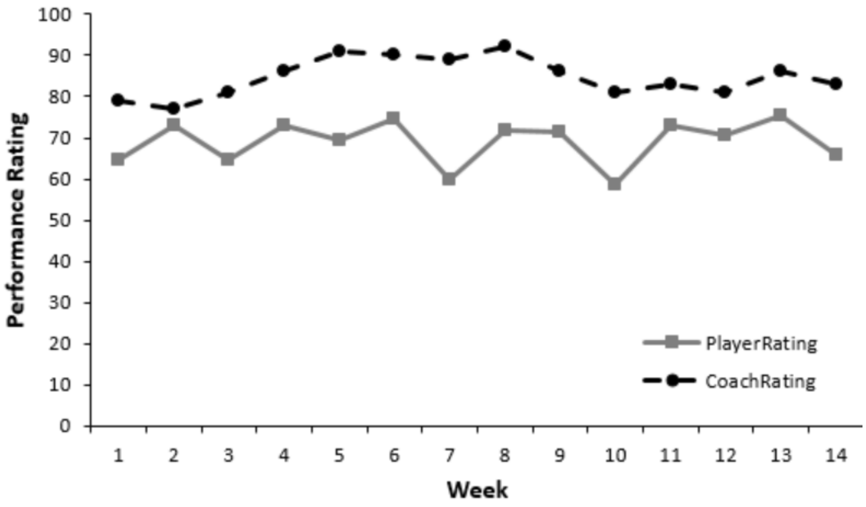
Data of a Single Player with Constant PlayerRating > CoachRating



This trend is not unique to this player. A detailed look at the dataset reveals that plenty of players tend to rate themselves higher than their coach. However, there are also players who consistently rated themselves lower than their coach, as you can see for a specific player in Figure 2.

Figure 2

Data of a Single Player with Constant PlayerRating < CoachRating



These plots show that the nonergodicity problem is present in this study, because the group level results are do not apply to these two players.

Strenghts, Limitations and Future Research

A strength of this study is that it measures performance by young football players repeatedly, which helps to create a more reliable picture of the comparison between player ratings and coach ratings than when performance is measured only once. Another strength is the use of a single-item questionnaire to combat participant burden and fatigue (Song et al., 2023). This study also has a high level of ecological validity, because it has been conducted in a real-life setting.

A first limitation of this study is the nonergodicity problem. The group-level results do not always translate into individual results, as can be seen in the figures (Neumann et al., 2021). Future research should find a way to predict differences between the coach rating and player rating on an individual level. Another limitation is that there is a lot of missing data, which is normal in research in sports. For the present study, inclusion criteria were relatively strict. Future research could set more lenient inclusion criteria in order to investigate a larger dataset. The final limitation is that coaches and players rated performance on a different time scale. Players rated themselves every day, whereas coaches only rated players once a week. It is possible that coaches had a less accurate picture of a player's day to day performance, also because the player only has to remember his own performance from one day at a time, while the coach has to remember performances from an entire week, with usually the first day of the rated week being exactly one week before the moment where the coach gives his rating. This is a problem, because memories get less accurate over time (Murre & Dros, 2015). Future research could combat this problem by making players and coaches judge performance on the same time scale, preferably once a day, because recent memories tend to be more accurate (Murre & Dros, 2015). Finally, future research should also investigate whether strong alignment between CoachRating and PlayerRating is actually a consequence of feedback from the coach for the players, in combination with the task of performing being made easier to define (Harris & Schaubroeck, 1988)

Conclusion

The expectation that players would rate their performance higher than their coach was not confirmed, at least not at the group level. It is possible that the, on group level, small difference between players and coaches is a consequence of adequate feedback by the coach, making it easier for the player to understand when the coach believes the player performs well, although it cannot be stated with full certainty that such feedback was regularly provided. Due to the nonergodicity problem, the results are not always applicable to individuals, as there are instances of players systematically rating themselves either higher or lower than their coach. Future research should focus on finding a solution to the problem of nonergodicity in this context, as well as to determine the nature of the results of this study.

References

Baird, L. S. (1977). Self and Superior Ratings of Performance: As Related to Self-esteem and Satisfaction with Supervision. *Academy of Management Journal*, *20*(2), 291–300. https://doi.org/10.2307/255402

Brink, M. S., & Frencken, W. G. P. (2018). Formative feedback for the coach reduces mismatch between coach and players' perceptions of exertion. *Science and Medicine in Football*, *2*(4), 255–260. https://doi.org/10.1080/24733938.2018.1451651

Brink, M. S., Nederhof, E., Visscher, C., Schmikli, S. L., & Lemmink, K. a. P. M. (2010). Monitoring load, recovery, and performance in young elite soccer players. *The Journal of Strength and Conditioning Research*, *24*(3), 597–603. https://doi.org/10.1519/jsc.0b013e3181c4d38b

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*(4), 385. https://doi.org/10.2307/2136404

Doeven, S. H., Brink, M. S., Frencken, W. G., & Lemmink, K. A. (2017). Impaired Player–Coach perceptions of exertion and recovery during match congestion. *International Journal of Sports Physiology and Performance*, *12*(9), 1151–1156. https://doi.org/10.1123/ijspp.2016-0363

Harris, M. M., & Schaubroeck, J. (1988). A META-ANALYSIS OF SELF-SUPERVISOR, SELF-PEER, AND PEER-SUPERVISOR RATINGS. *Personnel Psychology*, *41*(1), 43–62. https://doi.org/10.1111/j.1744-6570.1988.tb00631.x

Hofseth, E., Toering, T., Jordet, G., & Ivarsson, A. (2017). Self-evaluation of skills and performance level in youth elite soccer: Are positive self-evaluations always positive? *Sport Exercise and Performance Psychology*, *6*(4), 370–383. https://doi.org/10.1037/spy0000094

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579–588. https://doi.org/10.1037/0021-9010.63.5.579

Jones, E. E. (1972). The actor and the observer : Divergent perceptions of the causes of behavior. *Attribution: Perceiving the Causes of Behavior*, 79–94. https://ci.nii.ac.jp/naid/10030006009

Kay, E., & Meyer, H. H. (1965). Effects of threat in a performance appraisal interview. *Journal of Applied Psychology*, *49*(5), 311–317. https://doi.org/10.1037/h0022522 Murre, J. M. J., & Dros, J. (2015). Replication and Analysis of Ebbinghaus' Forgetting Curve. *PLoS ONE*, *10*(7), e0120644. https://doi.org/10.1371/journal.pone.0120644

Neumann, N. D., Van Yperen, N. W., Brauers, J. J., Frencken, W., Brink, M. S., Lemmink, K. A., Meerhoff, L. A., & Hartigh, R. J. D. (2021). Nonergodicity in load and recovery: group results do not generalize to individuals. *International Journal of Sports Physiology and Performance*, *17*(3), 391–399. https://doi.org/10.1123/ijspp.2021-0126

Reuters. (2022, November 18). *Ronaldo felt "provoked" by Man Utd boss Ten Hag in Spurs win*. https://www.reuters.com/lifestyle/sports/ronaldo-felt-provoked-by-man-utd-boss-ten-hag-spurs-win-2022-11-18/.

Saw, A. E., Main, L. C., & Gastin, P. B. (2015). Monitoring athletes through self-report: factors influencing implementation. *PubMed*, *14*(1), 137–146. https://pubmed.ncbi.nlm.nih.gov/25729301

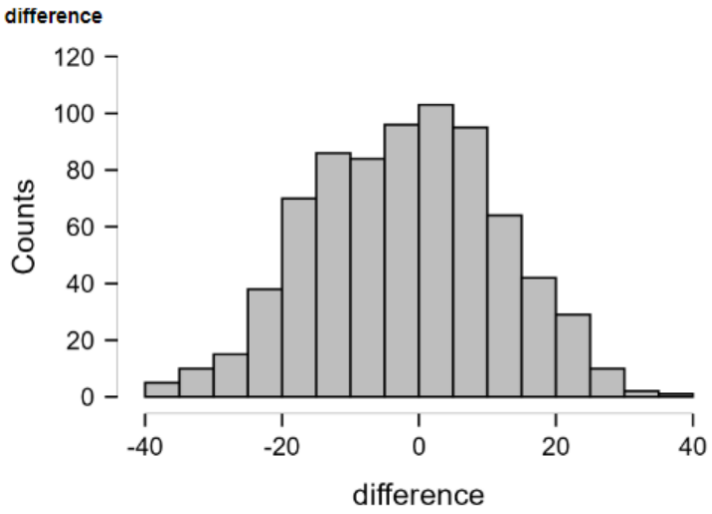
Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2022). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662–1671. https://doi.org/10.1177/10731911221113563

Steel, R. P., & Ovalle, N. K. (1984). SELF-APPRAISAL BASED UPON SUPERVISORY FEEDBACK. *Personnel Psychology*, *37*(4), 667–685. https://doi.org/10.1111/j.1744-6570.1984.tb00532.x

Totterdell, P. (2000). Catching moods and hitting runs: Mood linkage and subjective performance in professional sport teams. *Journal of Applied Psychology*, *85*(6), 848–859. https://doi.org/10.1037/0021-9010.85.6.848

Appendix

Histogram of Difference Scores



Note. This is the histogram of the difference scores between PlayerRating and Coachrating, as discussed in the "Assumption Checks" section.