https://doi.org/10.22364/hssl.31.1.05
pp. 65–79

HUMANITIES AND
SOCIAL SCIENCES
LATVIA

31(1) 2023

PERFORMANCE: DIFFERENCES IN MEASURING PERFORMANCE

Daniel Philipp Schettler[1]
University of Latvia

Abstract. An important issue in human resources is the procedure of the employees' performance evaluation. The appraisal is essential in the sense of employee appreciation and motivation. Most employers use a subjective performance evaluation of a single superior or a group of persons involved in the employee's working processes. The subjective evaluation of a group or one person is often questioned about being appropriate. An often-named solution for a more objective criteria could be data driven performance measures. Professional sport provides a unique opportunity to compare objective and subjective performance evaluation measures. A data set of the German Bundesliga was used to test if the two different performance measure come to equal results. It is shown that differences in means exist but equivalence tests support the hypothesis that both measures could be treated as equal. In toto, it seems that in an environment where performance is relatively good to measure objective and subjective performance evaluations lead to equivalent results.

Keywords: performance appraisal, subjective and objective performance.

[1] **Contact:** Daniel Philipp Schettler: setlers.daniels@inbox.lv; Faculty of Economics and Management, University of Latvia, Aspazijas bulvaris 5, Riga, LV-1050, Latvia.

Introduction

> *“There is perhaps not a more important human resources system in organizations than performance evaluation.”* (Judge and Ferris 1993, 80)

A huge majority of employers use performance appraisals as a part of their human resources management (Cappelli and Conyon 2018, p. 89). The procedure is important for many decisions in human resources management (HRM) and due to its impact, an important tool for the companies' performance. “Performance appraisal is a key tool in companies that provides information about employees performance in order to make important decisions, such as salary adjustments, promotions, identification of training and development needs, documentation of performance levels or behaviours that may cause firing or sanctions.” (Espinilla et al. 2013, p. 459)

In a standard performance appraisal supervisors evaluate the performance of their supervisees, requesting employees to act in the interests of the employer (Cappelli and Conyon 2018, p. 88). Typically the process follows a yearly routine, beginning with the definition of the performance goals which can be assessed or redefined until the final performance appraisal by the superior (Frederiksen et al. 2017, p. 411). “Traditional conceptualizations of the performance-rating process imply that performance is a knowable and observable objective reality and that performance ratings are reasonable reflections of that reality.” (Judge and Ferris 1993, 97)

Because of its important role in HRM many researchers promoted a better understanding of the employee evaluation processes. Organisational justice is an often-cited theory in the appraisal context. It can be distinguished into four dimensions, distributive, procedural, interpersonal, and informational justice (Colquitt 2001, p. 386). Whereas Colquitt (2001) demonstrated that the justice dimensions influence important corporate outcomes as commitment. Important for this article is the procedural justice, which is defined as the perceived fairness of the procedures the decision is based on (Roberson and Stewart 2006, p. 284). Procedural justice is fostered by a fair decision making process (Colquitt 2001, p. 386). Roberson and Stewart suggested that perceived accuracy in the employee feedback process might motivate to improve performance (Roberson and Stewart 2006, p. 293).

In the study of Taylor et al. employees perceived a greater accuracy and fairness in the appraisal system when evaluated with a due-process appraisal (Taylor et al. 1995, 518). One of the favourable rules of this evaluation method is the accuracy rule, accordingly, managers and employees record performance accurately and use these records for the justification of performance evaluations. Albright and Levy concluded that the more credible the sources and their feedback were the more favourable was the evaluation of the feedback (Albright and Levy 1995).

Additionally, many other researchers focus on the biases that can occur during the performance evaluation of the supervisor. The process of performance appraisals is often perceived as unfair because employees often only depend on the single opinion of the supervisor (Selvarajan and Cloninger 2012, p. 3067). Subjectivity in performance appraisals are prone to different biases such as social relations (Choon and Embi 2012, p. 190) or demographic similarities that can influence the evaluation process (Judge and Ferris 1993, 87). "Supervisors can (ab) use their discretion in determining subjective performance evaluations by directing subordinates toward activities that are not valued by the organization. Biased supervision is costly and reduces the optimal strength of subjective performance pay, as in case of incongruent verifiable performance measurement." (Delfgaauw and Souverijn 2016, p. 120) A brief digression: Objective criteria for quantitative and qualitative employee performance could be a turnover rate whereas subjective indicators could be the perceived satisfaction with the employees work (Wesche and Sonderegger 2019, pp. 200–201).

The arguments for the use of algorithms in the decision-making process are often focusing on such biases and the discretion of the supervisor. With automatization of the decision making processes human discretion and biases can be limited (Danaher 2016, pp. 262–263). The use of data mining in human resource management (HRM) is also a prospering research field (Strohmeier and Piazza 2013, p. 2410). Taking this into account, a reduction of human limitations with the help of an algorithm should lead to a fairer perceived process.

This article uses the definition of Lee for algorithms, who defined it as "[..] a computational formula that autonomously makes decisions based on statistical models or decision rules without explicit human intervention. This reflects the recent advancement of the autonomous decision-making capabilities of algorithms from artificial intelligence and machine learning, and current usage of the term in popular media." (Lee 2018, p. 3)

Nevertheless, it is not said, that an algorithmic decision leads per se to a higher perceived fairness of the decision-making process. Nagtegaal (2021) focused on the effect of the inclusion of algorithms in managerial decisions on procedural justice perceptions. Her research results suggest that adding an algorithm to a manager's decision-making process can increase the perception of procedural justice for high complexity practice (Nagtegaal 2021, p. 1). Whereas decisions in high complexity made only by computers would be perceived as lower in procedural justice than decisions made by a manager.

Lee (2018) arguments that the perceived fairness of an algorithmic decision would depend on task characteristics. With tasks that require human skills, like hiring and work evaluation, the human decisions were perceived as fairer than the algorithmic decisions (Lee 2018, p. 1).

The phenomenon of negative tendencies toward algorithmic decision-makers of people, knowing about the algorithmic presence, is called algorithm aversion (Köbis and Mossink 2021, p. 2). These reservations about automated processes also occur when it is obvious that an algorithm can achieve better results than an expert (Filiz et al. 2021, p. 1).

> *"If verifiable performance measures are imperfect, subjective performance evaluation may provide a more accurate assessment of employees' performance, thereby providing better incentives for employees."*
> (Delfgaauw and Souverijn 2016, p. 107)

This article uses a sports data set to test if subjective and objective performance measures come to equal performance evaluation results. Professional soccer offers a unique chance to undertake this analysis, to the author's knowledge, this is the first article, which uses this opportunity to gain new helpful insights regarding performance evaluation processes.

Research results

> *"Sport* [..] *provides opportunities to observe, accurately measure, and compare variables of interest over time and to test hypotheses with highly motivated respondents in quasi-laboratory conditions."* (Wolfe et al. 2005, p. 185)

The research setting of this article is professional football and uses the unique research opportunities that come with it. "There is no research setting other than sports where we know the name, face, and life history of every production worker and supervisor in the industry. Total compensation packages and performance statistics for each individual are widely available, and we have a complete data set of worker-employer matches over the career of each production worker and supervisor in the industry." (Kahn 2000, p. 75)

As performance is relatively good to measure in sports like professional football, many statistical measures exist. Even if these football statistics are not (yet) as distinctive as they are in the big North American sports like Baseball or Basketball. There are some differences between football and the popular North American sports. Work interdependencies between the players are often named as a factor that makes evaluating football players more complicated than in other sports, where isolated actions can be assessed with statistics (Della Torre et al. 2018, p. 126). Nevertheless, in recent years, more and more statistical analyses appear also in professional football.

From the literature three streams in measuring performance can be distinguished. Firstly, some authors use composite performance measures or simple indexes. An index is a numerical result of several individual indicators. An important criterion for the quality of an index is the indicator selection and the indicator weight. Experts can, for example, rate the weights of the indicators or be estimated empirically with statistical analyses. However, the simplest form of an unweighted indicator is to use an additive index. Therefore, the numerical indicator results are summed up and averaged (Bortz and Döring 2006, pp. 143–149).

The second possibility to measure performance in football stems from the media coverage of the industry. Several sports magazines or newspapers in Europe rate football players (e.g., the German Kicker, the Spanish Marca or the Italian Gazzetta dello Sport). The rating person could be a group of many or a single journalist which elaborate on each player's game performance. In the case of the German sports magazine kicker, the ratings correspond to German school grades ranging from 1 (exceptional) to 6 (very poor) (Frick 2011, p. 102). According to Della Torre et al., the specialized journalists' performance perception of a player captures two performance dimensions, the quantitative (e.g., goals scored) and the qualitative dimension (e.g., effectiveness) (Della Torre et al. 2018, p. 125). Nevertheless, these ratings are subjective performance measures (Frick 2011, p. 102), which makes them prone to individual biases.

Thirdly, there are more and more algorithmic performance measures in professional sports. Such as the LigaInsider Performance index for the players of the first German division in professional football. Their algorithms take more than 250 variables into account and calculates depending on the players position a school rating as the journalists from the kicker magazine do (LigaInsider). The authors of the webpage claim to have the fairest performance evaluation because it would not depend on a journalist's subjective perception. The biggest difference towards the first category is the sheer amount of data which is used for the calculation of the index. In the first stream researchers only focus on a few indicators like assists, tackling or goals.

The data set of this article consists of two different sources. The performance measures of all 1. Bundesliga players were collected for the seasons 2019–2020 and 2020–2021. On the one hand, from the homepage of the German sports magazine kicker, the player's average evaluation the kicker Note. On the other hand, the same was done for all average performance index from LigaInsider of all players which were evaluated during the mentioned campaigns. Both measures are good to compare because both follow the abovementioned German school grade system, the grades can range from 1 (exceptional) to 6 (very poor). A school grade would be technically speaking classified as an ordinal variable. Nevertheless, if a variable has at least five increments, it can be treated as a metric variable (Berry 1993, p. 47). Furthermore both data sources treat the variable as metric, as they offer an average and additionally, researchers as well, used the kicker grade as dependent variable of their regression analysis (Frick 2011, p. 104).

Table 1 demonstrates important statistical measures of both performance measures. In total 794 evaluations were made by kicker and 435 were made by LigaInsider. From the mentioned quartiles it can be already seen that both evaluations have 50 percent of the observations in a pretty similar range. Which is also graphically demonstrated with the following Boxplot analysis.

Table 1. Descriptive Statistics

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max	N
Kicker	1.500	3.283	3.605	3.636	3.970	6.000	794
LigaInsider	2.350	3.480	3.740	3.722	3.980	4.680	435

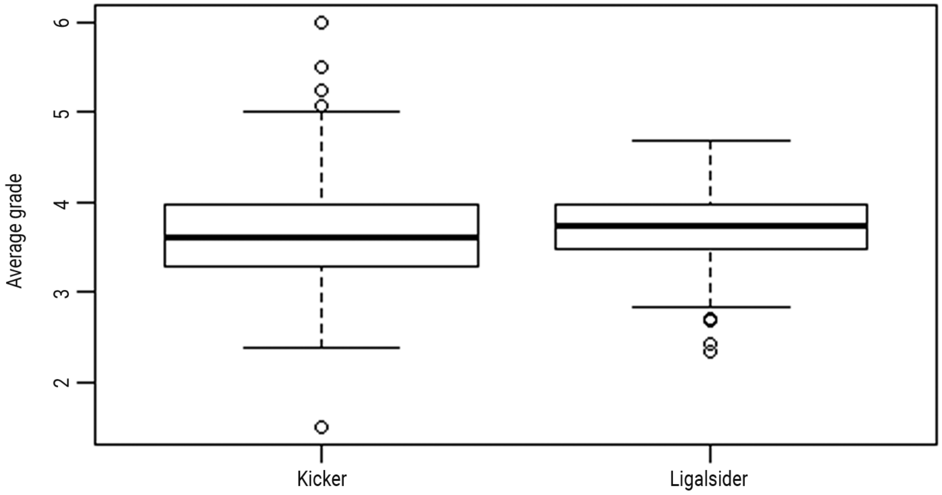

Figure 1. Boxplot Kicker and Liga Insider Ratings

Figure 1 displays the data for all ratings which were allotted during the two seasons. The analysis shows that the kicker has outliers on both ends whereas the LigaInsider has only outliers on the lower end of the rating system. The minimum and maximum kicker grade show that almost the full spectrum from 1 to 6 was used. Indeed, both mean and median seem to be very similar, as well as the 3rd quartile which is symbolized with the upper end of the boxes.

For 432 player / season data points a kicker as well as a LigaInsider rating is given. In Figure 2 the data was plotted against each other.

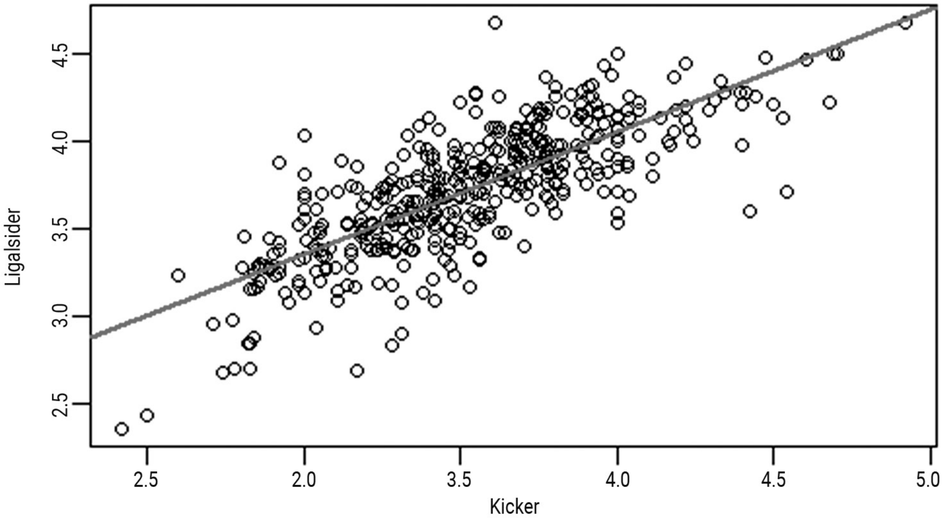

Figure 2. Plot

From the data points and the regression line a positive correlation between the two measures can be assumed. This impression is supported by the Pearson and Spearman correlation test, the test results are displayed in Table 2. Normally Spearman's rank correlation test is used for ordinal data and Pearson correlation for continuous data. Spearman's rho is with 0.78 very high, one would indicate a perfect association of ranks. The same strong positive relationship between the kicker and the LigaInsider grades supports the assumption drawn from the Pearson correlation. The higher the kicker grade, the higher the LigaInsider grade, and vice versa. Not only the high rho-value but the low p-value below 0.05 allows to rejects the null hypothesis, which assumes that there exists no correlation between the two variables.

Table 2. Correlation Test

Correlation Test / Variable	**Kicker- LigaInsider**	*p-value*
Spearman	**0.782**	*< 2.2e-16*
Pearson	**0.775**	*< 2.2e-16*

For further investigations of the differences between the average grades of kicker and LigaInsider, a normality test was performed. The results of the performed Shapiro-Wilk normality test suggest no normal distribution. The null hypothesis, that the variables are from a normal distribution, is rejected because the p-value is lower than 0.05. Therefore, the data seems not to be sampled from a normal distribution. Therefore, the test results advocate to proceed with non-parametric test for the analysis of the differences between the two grades.

Table 3. Normality Test

Variable / Normality Test	**Shapiro-Wilk**
Kicker and LigaInsider	p-value = 0.001748

To compare the two means of the different performance measures a paired t test will be used. T tests are especially in clinical research a widely used research method (Kim 2015, p. 540). As the sample data does not seem to meet standard distributional assumptions the bootstrap method is considered as an alternative approach (Frey 2018, p. 218). To test for differences between the paired samples of the kicker and the LigaInsider performance evaluation, a bootstrapped paired t tests was performed. The null hypothesis tests if the difference in means is zero and the alternative hypothesis is therefore that the difference in means is different from zero. If there is no difference between the two performance ratings the results would be close to zero and the differences in the means would be zero (Kim 2015, p. 544). Table 3 displays the test results. The low p-value indicates that there exit differences in means between the two performance measures. Nevertheless, the goal of this paper is to test if the two measures can be considered as equivalent and not if the difference in means equals zero. Therefore, the next step is to test for equivalence of the two performance ratings.

Table 4. Bootstrapped t test

estimate	**statistic**	**p. value**	**parameter**	**conf. low**	**conf. high**	**method**	**Alter-native**
−0.193	−15.309	1.47E-42	431	−0.2180	−0.1684	Bootstrapped Paired t-test	two.sided

To conclude statistical equivalence, the difference between groups is smaller than what is considered meaningful and statistically falls within a previously defined interval, the equivalence bounds (Lakens et al. 2018, p. 260). With two one-sided tests (TOST) equivalence is tested against the smallest effect size of interest (SESOI). Lakens et al. described three subjective approaches to justify the SESOI (Lakens et al. 2018, p. 262). The author divides between objective or subjective justification of the SESOI. The former variant would be based on quantifiable theoretical predictions. Whereas, three categories are described for the subjective justification. Firstly, with benchmarks, where the SESOI is set to a standardized effect size. Secondly, it can be based on related studies and thirdly, it can be based on a resource question.

As there is – to the authors knowledge – no previous research, the second approach is not applicable. As a consequence, the third approach subjective justification of raw differences will be used as first attempt. Remembering that both performance measures are using school grades from one to six. The allocation of quarter grades is a commonly used basic approach in student evaluation. That means, that the grades are commonly summarized in 0.25-steps, quarter-grades (e.g., a grade of 3.40 and a grade of 3.60 are commonly both reported as a 3.5). One of the data sources, the kicker magazine, allocates only half-grades. Nevertheless, the common use of quarter grades makes it plausible to take these steps as the borders for a significant difference. Accordingly, in the first test case, quarter grades −0.25 and +0.25 will be used as raw differences.

The results are displayed in Figure 3 and Table 5. Graphically it is shown in Figure 3 that the 90 % confidence interval lies in between the raw differences. If the confidence interval lies in between the equivalence margin [−0.25; 0,25]. If

Table 5. TOST Results with Raw Bounds

TOST Results						
			t	df	p	
Kicker	LigaInsider	t-test	−7.72	431	<.001	
		TOST Upper	−18.5	431	<.001	
		TOST Lower	3.11	431	0.001	
Equivalence Bounds						
					90% Confidence interval	
			Low	**High**	**Lower**	**Upper**
Kicker	LigaInsider	Cohen's d	−0.521	0.521		
		Raw	−0.250	0.250	−0.216	−0.140

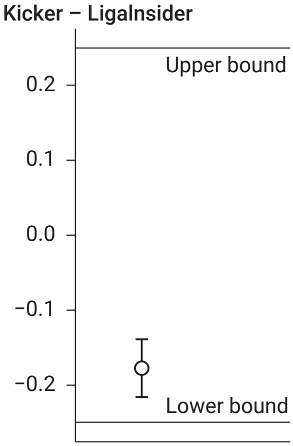

Figure 3. TOST Results with Raw Bounds

the difference between the measures falls in these boundaries, statistical equivalence can be concluded or that the difference is too small to care about. With this result, statistical equality can be expected between the two performance measures. Table 5 displays besides the earlier discussed t-test, which assumes differences in means, the significance of the upper and lower TOST, the calculated verification that the two measures are both falling in the raw boundaries.

As a robustness check, the approach with a standardized effect size will be tested additionally. For that, the SESOI is set to a medium effect size of d = 0.5. According to the literature review of Lakens this would make it possible to reject only effects in the upper 25 %–33 % of the distribution effect sizes (Lakens et al. 2018, p. 262). Comparing Cohen's d in Table 5 and Table 6, it can be seen that both values are pretty similar (0.521 and 0.5). This shows that the boundaries do not differ much if the standardized or the raw effect size will be used. Whereas the standardized effect size is a little bit stricter than the raw effect size approach, as Cohen's d is 0.5 and therefore smaller. Nonetheless, the result does also not change, also with the stricter, standardizes effect size, statistical equivalence can be proved (see Figure 4 and Table 6).

In summary, with the bootstrapped t test differences in means where found. However, both equivalence test results proof statistical equivalence between the two performance measures. Which means that the objective and the subjective performance evaluation come to equivalent results. In the research setting of the German football Bundesliga, the evaluation of an expert, the kicker rating, seems to come to the same conclusion about performance than the data driven approach of LigaInsider does.

Table 6. TOST Results with Standardized Bounds

TOST Results						
			t	**df**	**p**	
Kicker	LigaInsider	t-test	−7.72	431	<.001	
		TOST Upper	−18.1	431	<.001	
		TOST Lower	2.67	431	0.004	
Equivalence Bounds						
					90% Confidence interval	
			Low	**High**	**Lower**	**Upper**
Kicker	LigaInsider	Cohen's d	−0.500	0.500		
		Raw	−0.240	0.240	−0.216	−0.140

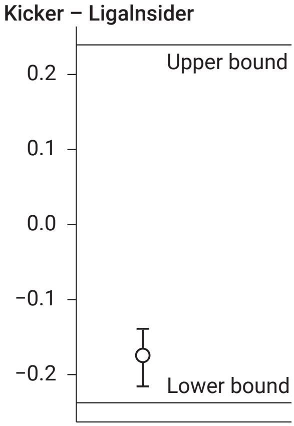

Figure 4. TOST Results with Standardized Bounds

Discussion and Summary

> *"Objective measures are generally unavailable for workers who perform many different tasks in frequently changing environments or work in teams or in administrative and cross-divisional functions such as HR, legal, accounting, or finance."* (Frederiksen et al. 2017, p. 409)

Despite conceptual arguments suggesting how functional performance appraisals should be, in practice they are one of the most unpopular and criticized aspects of the modern workplace (Cappelli and Conyon 2018, p. 96). Roberson and Stewart conclude that motivating properties of feedback may depend on the perceived correctness of the feedback itself (Roberson and Stewart 2006, p. 283). The acceptance of a decision made by a computer or an algorithm is important when for example employee motivation should be addressed with it. Research results of several studies showed that humans face computer-made decisions with scepticism (Wesche and Sonderegger 2019, p. 204).

The results of this article demonstrate that it could be possible that an expert can evaluate the performance of employees as good as an algorithm can when performance is relatively easy to measure. As suggested by Nagtegaal, algorithms can be expected to become a great decision-making partner in highly complex practice, instead of becoming a substitute for manager (Nagtegaal 2021, p. 6). This describes also a transformation of the relationship between human users and computers. The clear master-slave paradigms changes to a more equal level of hierarchy between operators and computers (Wesche and Sonderegger 2019, p. 197). Delfgauw and Souverijn supposed that the combination of incongruent performance measures and biased supervision could moderate wrong incentives on employee's effort (Delfgaauw and Souverijn 2016, p. 107). The 360-degree appraisal overcomes some of the disadvantages of the traditional single source appraisal (Espinilla et al. 2013, p. 459). In this case the algorithmic evaluation can be used as another evaluating source in the process and help enhance procedural fairness.

Nevertheless, for the interpretation the specialities of the data set should be kept in mind. In total, it is important to remark that the perfect measure for the overall performance evaluation of a player is not available (Della Torre et al. 2018, p. 127). Nevertheless, performance in sport is relatively easy to measure the results of such a research setting might be different in a setting in which performance is not as easy to observe (Harder 1992, p. 332). Hall, Szymanski, and Zimbalist even state that the usually hidden information actions are not plausible in a sports context because the players regularly perform in front of large audiences (Hall et al. 2002, p. 157). However, other researchers make a direct link to the corporate world. Della Torre et al. state that a group of workers, having high skills and salaries like executive or senior management are to a certain extend comparable (Della Torre et al. 2014).

Regardless, this article demonstrated statistically that subjective and objective performance evaluation is coming to results that can be considered as equal. Further research could investigate the influence of the combination of human and algorithmic multi-source feedback on perceived procedural fairness. As the acceptance of multi-source feedback also depends as well on the perceived procedural fairness (McCarthy and Garavan 2007, p. 912). Whether the combination of personal and data-based performance evaluation can influence the perceived fairness positively would be one of many possible research fields.

REFERENCES

Albright, M. D., Levy, P. E. (1995). The Effects of Source Credibility and Perfromance Discrepancy on Reactions to Multiple Raters. *Journal of Applied Social Psychology*, 25 (7), 577–600. https://doi.org/10.1111/j.1559-1816.1995.tb01600.x

Berry, W. D. (1993). *Understanding Regression Assumptions*. Thousand Oaks: SAGE Publications.

Bortz, J., Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4th ed. Heidelberg: Springer Medizin Verlag.

Cappelli, P., Conyon, M. J. (2018). What Do Performance Appraisals Do? *ILR Review*, 71 (1), 88–116. https://doi.org/10.1177/0019793917698649

Choon, L. K., Embi, M. A. (2012). Subjectivity, Organizational Justice and Performance Appraisal: Understanding the Concept of Subjectivity in Leading Towards Employees' Perception of Fairness in the Performance Appraisal. *Procedia – Social and Behavioral Sciences* 62, 189–193. https://doi.org/10.1016/j.sbspro.2012.09.030

Colquitt, J. A. (2001). On the Dimensionality of Organizational Justice: A Construct Validation of a Measure. *Journal of Applied Psychology*, 86 (3), 386–400.

Danaher, J. (2016). The Threat of Algocracy – Reality, Resistance and Accommodation. *Philosophy and Technology*, 29 (3), 245–268.

Delfgaauw, J., Souverijn, M. (2016). Biased supervision. *Journal of Economic Behavior & Organization*, 130 (1), 107–125. https://doi.org/10.1016/j.jebo.2016.06.012

Della Torre, E., Giangreco, A., Legeais, W., Vakkayil, J. (2018). Do Italians Really Do It Better? Evidence of Migrant Pay Disparities in the Top Italian Football League. *European Management Review* 15 (1), 121–136. https://doi.org/10.1111/emre.12136

Della Torre, E., Giangreco, A., Maes, J. (2014). Show Me the Money! Pay Structure and Individual Performance in Golden Teams. *European Management Review*, 11 (1), 85–100. https://doi.org/10.1111/emre.12025

Espinilla, M., Andrés, R. de, Martínez, F. J., Martínez, L. (2013). A 360-degree performance appraisal model dealing with heterogeneous information and dependent criteria. *Information Sciences*, 222 (1), 459–471. https://doi.org/10.1016/j.ins.2012.08.015

Filiz, I., Judek, J. R., Lorenz, M., Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, 31 (100524), 1–8. https://doi.org/10.1016/j.jbef.2021.100524

Frederiksen, A., Lange, F., Kriechel, B. (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization*, 134 (2–3), 408–429. https://doi.org/10.1016/j.jebo.2016.12.016

Frey, B. B. (2018). The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation. https://doi.org/10.4135/9781506326139

Frick, B. (2011). Performance, Salaries, and Contract Length: Empirical Evidence from German Soccer. *International Journal of Sport Finance*, (6), 87–118.

Judge, T. A., Ferris, G. R. (1993). Social Context of Performance Evaluation Decisions. *The Academy of Management Journal*, 36 (1), 80–105.

Kahn, L. M. (2000). The Sports Business as a Labor Market Laboratory. *The Journal of Economic Perspectives*, 14 (3), 75–94.

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68 (6), 540–546. https://doi.org/10.4097/kjae.2015.68.6.540

Köbis, N., Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114 (2), 1–13. https://doi.org/10.1016/j.chb.2020.106553

Lakens, D., Scheel, A. M., Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1 (2), 259–269.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5 (1), 1–16. https://doi.org/10.1177/2053951718756684

LigaInsider: Deutschlands fairste Fussballnote. Available online at: https://www.ligainsider.de/deutschlands-fairste-fussballnote/, checked on 2/11/2021.

McCarthy, A. M., Garavan, T. N. (2007). Understanding acceptance of multisource feedback for management development. *Personnel Review*, 36 (6), 903–917. https://doi.org/10.1108/00483480710822427

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, 38 (1), 1–10. https://doi.org/10.1016/j.giq.2020.101536

Roberson, Q. M., Stewart, M. M. (2006). Understanding the motivational effects of procedural and informational justice in feedback processes. *British Journal of Psychology* (*London, England: 1953*), 97 (Pt 3), 281–298. https://doi.org/10.1348/000712605X80146

Selvarajan, T. T., Cloninger, P. A. (2012). Can performance appraisals motivate employees to improve performance? A Mexican study. *The International Journal of Human Resource Management*, 23 (15), 3063–3084. https://doi.org/10.1080/09585192.2011.637069

Strohmeier, S., Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, 40, 2410–2420.

Taylor, S. M., Tracy, K. B., Renard, M. K., Harrison, K., Carroll, S. J. (1995). Due Process in Performance Appraisal: A Quasi-Experiment in Procedural Justice. *Administrative Science Quarterly*, 40 (3), 495–523.

Wesche, J. S., Sonderegger, A. (2019). When computers take the lead – The automation of leadership. *Computers in Human Behavior*, 101, 197–209.

Wolfe, R. A., Weick, K. E., Usher, J. M., Terborg, J. R.; Poppo, L., Murrel, A. J. et al. (2005). Sport and Organizational Studies. Exploring Synergy. *Journal of Management Inquiry*, 14 (2), 182–210.