Examining the reliability and predictive validity of performance assessments by soccer coaches and scouts

Bergkamp, Tom L. G.; Meijer, Rob R.; den Hartigh, Ruud J. R.; Frencken, Wouter G. P.; Niessen, A. Susan M.

Published in:
Psychology of sport and exercise

DOI:
10.1016/j.psychsport.2022.102257

Psychology of Sport & Exercise 63 (2022) 102257

Contents lists available at ScienceDirect

Psychology of Sport & Exercise

journal homepage: www.elsevier.com/locate/psychsport

Examining the reliability and predictive validity of performance assessments by soccer coaches and scouts: The influence of structured collection and mechanical combination of information☆

Tom L.G. Bergkamp [a,*], Rob R. Meijer [a], Ruud. J.R. den Hartigh [b], Wouter G.P. Frencken [c,d], A. Susan M. Niessen [a,**]

[a] Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712TS, Groningen, the Netherlands
[b] Department of Developmental Psychology, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712TS, Groningen, the Netherlands
[c] Center for Human Movement Sciences, University of Groningen, University Medical Center Antonius Deusinglaan 1, 9713 AV, Groningen, the Netherlands
[d] Football Club Groningen, Groningen, the Netherlands

ARTICLE INFO

Keywords:
Assessment
Prediction
Structure
Mechanical combination
Soccer
Coaches and scouts

ABSTRACT

Soccer coaches and scouts typically assess in-game soccer performance to predict players' future performance. However, there is hardly any research on the reliability and predictive validity of coaches' and scouts' performance assessments, or on strategies they can use to optimize their predictions. In the current study, we examined whether robust principles from psychological research on selection – namely structured information collection and mechanical combination of predictor information through a decision-rule – improve soccer coaches' and scouts' performance assessments. A total of $n = 96$ soccer coaches and scouts participated in an elaborate within-subjects experiment. Participants watched soccer players' performance on video, rated their performance in both a structured and unstructured manner, and combined their ratings in a holistic and mechanical way. We examined the inter-rater reliability of the ratings and assessed the predictive validity by relating the ratings to players' future market values. Contrary to our expectations, we did not find that ratings based on structured assessment paired with mechanical combination of the ratings showed higher inter-rater reliability and predictive validity. In contrast, unstructured-holistic ratings yielded the highest reliability and predictive validity, although differences were marginal. Overall, reliability was poor and predictive validities small-to-moderate, regardless of the approach used to rate players' performance. The findings provide insights into the difficulty of predicting future performance in soccer.

1. Introduction

Talented soccer players are typically identified by soccer coaches and scouts, who aim to predict players' future performance on the basis of a number of indicators, often through assessing in-game soccer performance (Bergkamp et al., 2019; Larkin & O'Connor, 2017). Because selecting players who will excel in the future can yield significant financial and competitive advantages for clubs, it is important that these performance predictions are reliable and valid (Den Hartigh et al., 2018; A. H. Roberts et al., 2020; Till & Baker, 2020). However, there is hardly any research on how coaches and scouts should retrieve and use information on performance indicators to optimize predictions (Den Hartigh et al., 2018). Therefore, we examine this topic in the present study. In particular, we introduce and apply a number of robust principles from psychological research on selection which are relevant for assessing in-game soccer performance. These principles relate to the way

☆ This study was preregistered on the Open Science Framework: https://osf.io/qfbc7/?view_only=31560d776b5147ccadf7b4939373d500

* Corresponding author. Heymans Institute for Psychological Research, Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712TS, Groningen, the Netherlands.
** Corresponding author. Heymans Institute for Psychological Research, Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712TS, Groningen, the Netherlands.
E-mail addresses: t.l.g.bergkamp@rug.nl (T.L.G. Bergkamp), a.s.m.niessen@rug.nl (A.S.M. Niessen).

https://doi.org/10.1016/j.psychsport.2022.102257
Received 17 January 2022; Received in revised form 17 July 2022; Accepted 22 July 2022
Available online 31 July 2022

information on performance indicators is *collected* and *combined* into a final assessment by decision-makers such as coaches and scouts (Meehl, 1954; Nolan & Highhouse, 2014; Sawyer, 1966).

1.1. Structured information collection

The information collection method of a scout or coach can be defined by the degree of structure in their assessment strategy. Huffcutt and Arthur (1994) and Chapman and Zweig (2005) described two facets of structure that are relevant for scouting soccer players, namely indicator structure and rating structure. Indicator structure refers to the degree to which decision-makers assess different individuals (e.g., players) on the same indicators, whereas rating structure refers to the level of standardization in rating these indicators (Chapman & Zweig, 2005; Huffcutt & Arthur, 1994). Thus, these principles imply whether coaches and scouts observe and score different performance indicators separately and consistently (i.e., indicator structure), and on the same scale (i.e., rating structure). For example, a soccer coach who does not assess performance indicators separately, but rather assesses players with a single rating based on the player's overall performance, applies a relatively unstructured approach. In contrast, a soccer coach who always evaluates players on passing, dribbling, and sprinting ability separately, and rates each of those predefined indicators on an anchored rating scale, uses a highly structured approach to assess performance.

Research from selection psychology has repeatedly shown that structured information collection outperforms unstructured information collection in terms of reliability and predictive validity (Conway et al., 1995; Huffcutt et al., 2013, 2014). The main reason for this finding is that information is collected more consistently when assessed in a structured manner. Accordingly, unstructured information collection usually results in suboptimal predictive validity, because it leads to inconsistent (and thus, unreliable) assessments within and between decision-makers (Kahneman et al., 2016; Karelaia & Hogarth, 2008; A. H. Roberts et al., 2020). For example, it is likely that different scouts or coaches who assess the same player through an unstructured approach differ in the performance indicators they take into account (i.e., indicator structure) and how they score them (i.e., rating structure).

A systematic review of different qualitative studies showed that most soccer coaches did not use of a set of separate, explicit performance indicators on which they based their assessment (A. H. Roberts, Greenwood, et al., 2019). Instead, they used an unstructured approach and primarily predicted performance by using their expertise intuitively (Christensen, 2009; Johansson & Fahlén, 2017). Coaches constructed an image of the ideal player in their head and recognized a future professional player in a way that 'they knew it when they saw it.' However, they had difficulty verbalizing what the performance indicators looked like exactly and did not score them (A. H. Roberts, Greenwood, et al., 2019). In contrast, a recent study showed that soccer scouts used a somewhat structured assessment approach, as most scouts always or very frequently assessed different players – of the same position and age – on the same indicators (Bergkamp et al., 2022).

1.2. Holistic vs. mechanical information combination

In performance prediction, multiple performance indicators are often considered. Decision-makers can combine the information they have collected on those indicators in either a holistic or a mechanical way to form their final assessment. In holistic combination, information is combined 'in the head' of the decision-maker (Dawes et al., 1989). For example, a coach who assesses players with a single, overall rating based on their overall impression uses holistic combination to form their final assessment. A coach who rates passing, dribbling, and sprinting ability separately (i.e., structured assessment), but integrates these ratings 'intuitively' in their head to form a final assessment also uses holistic combination. Thus, it is possible for decision-makers to use a structured assessment approach paired with holistic information combination. Indeed, a recent study among soccer scouts indicated that they often used this approach to scout players: most scouts used a structured assessment approach, but still relied on their intuition to form their final assessment (Bergkamp et al., 2022).

In contrast, mechanical combination means that information is combined according to a pre-determined decision-rule (Meijer et al., 2020). This decision-rule can be relatively simple. For instance, coaches use mechanical combination when they rate each indicator separately, and base their final assessment on the mean or sum of their separate ratings (Den Hartigh et al., 2018). Such mechanical combination typically outperforms holistic combination of information, because information is weighted more consistently when combined mechanically (Ægisdóttir et al., 2006; Grove & Meehl, 1996).

Nevertheless, decision-makers in many domains prefer to use unstructured holistic assessment approaches to make predictions. The primary reason for this seems to be that they experience autonomy and control over their predictions when they make them holistically (Nolan & Highhouse, 2014), and feel they can accurately 'make sense' of important information (Dana et al., 2013). Consequently, holistic combination is often used in practice to make predictions across a spectrum of contexts, such as clinical psychiatry, criminal justice decisions, and hiring interviews (Bishop & Trout, 2002; Lilienfeld et al., 2013; Neumann et al., 2021).

1.3. Structured-mechanical assessment

Few studies have explicitly examined the benefit of structured assessment based on observations *paired* with mechanical combination of those assessments. So far, the benefits of a structured assessment approach have been most evident in the literature on hiring interviews (Huffcutt et al., 2013; 2014, McDaniel et al., 1994), but it is relatively unclear whether scores on the indicators were also combined mechanically, and how that may have influenced the findings (see Conway et al., 1995, for an exception, who found a moderating effect of mechanical combination). At the same time, evidence for the benefit of mechanical combination is mostly based on studies in which different performance indicators were already quantitative in nature (e.g., test scores) and were combined in a data-driven linear model (Ægisdóttir et al., 2006; Grove & Meehl, 1996). That is, the indicators did not have to be quantified by the decision-maker based on their observations.

Notable exceptions are the studies by Arkes et al. (2006) and Dana and Rick (2006). Arkes et al. (2006) examined a structured-mechanically combined assessment approach based on raters' observations. They asked participants to rate scientific convention sessions and posters by either giving a single overall rating or a structured procedure in which one rating was given to each of five indicators. The authors found that the mean of the structured ratings yielded higher inter-rater reliabilities than the holistic procedure in which one overall rating was given. Moreover, Dana and Rick (2006) asked participants to predict final semester GPA either holistically, or by predicting the grade for different courses and taking the mean of those grades as the GPA prediction. They found that this structured-mechanical combination of the predicted course grades was a better predictor of actual final GPA than the holistically derived predicted GPA.

1.4. The current study

The potential benefit of a structured assessment approach paired with mechanical combination of information is particularly relevant for soccer coaches and scouts, who typically use their own observations of performance to make predictions. In this study, we experimentally examined the reliability and predictive validity of coaches' and scouts' assessments of soccer performance, based on structured vs. unstructured information collection and holistic vs. mechanical combination of information. Coaches and scouts assessed players' performance on video,

which resulted in a 1) structured-mechanical, 2) structured-holistic, and 3) unstructured-holistic performance rating. Additionally, the study included a condition without video observation. With this additional condition, we aimed to explore whether the observation of players' in-game performance, a key component of talent identification in practice, contributes to or hurts coaches' and scouts' performance predictions. Therefore, in the 'no-observation' condition, participants did not view a player's performance on video, but made a performance prediction based on simple background information of the player. Finally, we asked participants to indicate their confidence in their predictions and intentions to use each approach to predict performance. We formulated the following hypotheses:

H1. Structured-mechanical performance ratings yield the highest *inter-rater reliability*, followed by structured-holistic ratings, followed by unstructured-holistic ratings.

H2. Structured-mechanical performance ratings yield the highest *predictive validity*, followed by structured-holistic ratings, followed by unstructured-holistic ratings.

We expected to find the largest differences between the structured-mechanical and unstructured-holistic performance ratings, for which we hypothesized to find observed reliabilities of $ICC_{structured\text{-}mechanical} = 0.37$ and $ICC_{unstructured\text{-}holsitic} = 0.15$ and predictive validities of $r_{structured\text{-}mechanical} = 0.3$ and $r_{unstructured\text{-}holsitic} = 0.1$ (Arkes et al., 2006; McDaniel et al., 1994).

2. Methods

The study was preregistered on the Open Science Framework (OSF). To keep the method section concise, we refer to the preregistration (https://osf.io/qfbc7/?view_only=31560d776b5147ccadf7b4939373d500) for more details on specific subsections of the methodology.

2.1. Participants

We recruited soccer coaches and scouts who were associated with the Royal Dutch Football Association (KNVB) and professional soccer clubs in the Netherlands (see OSF preregistration, section 3.3, 'Data collection procedures'). A total of $n = 117$ coaches and scouts ultimately participated in the experiment (48% were associated with the KNVB), of which $n = 94$ fully completed and $n = 2$ completed at least one condition. $N = 25$ responses were removed because participants did not complete at least one condition or did not meet the eligibility criteria (see OSF preregistration, section 5.4, 'data exclusion'). $N = 91$ (95%) participants identified themselves as male and $n = 5$ (5%) as female. Participants were on average 50.71 ($SD = 14.74$) years old and had 10.21 ($SD = 9.92$) years of experience as a scout or coach.

Power analysis for the validity analyses indicated that a sample size of $n = 147$ participants was necessary to detect the expected validity differences (See section 3.5 – 'sample size rationale' – of our preregistration for a more elaborate explanation of the required sample size for the primary analyses). Thus, we did not obtain the required sample size, meaning that our analyses were underpowered (a power analysis with $n = 96$ for the same effect size specified in the pre-registration yielded 64% power). Ethical approval was granted by the Ethical Committee of Psychology of the University of Groningen (code PSY-2021-S-0142) and informed consent was obtained for all participants prior to the experiment.

2.2. Materials and measures

2.2.1. Stimulus material

Participants were presented with videos of adult, male, professional soccer players in competitive 11-vs-11 soccer games in the 2015–2016 soccer season (video duration was 15–20 min per game). These videos showed all successful and unsuccessful events and actions of the player in that game, including passes forward, running actions, dribbles, shots, and duels. We selected soccer players from the following international competitions: Super League 1 (Greece), Bundesliga (Austria), Super League (Switzerland), Fortuna Liga (Czech Republic), Eliteserien (Norway), Superliga (Denmark), and Allsvenkan (Sweden). The combination of historic videos and foreign leagues limited Dutch participants' recognition of players or potential recollection of players' performance.

We controlled for players' playing position and age by selecting a random sample of $k = 25$ players who 1) were all *full backs* 2) were younger than 23 years old at the time and 3) had played at least 10 full 90-min games during the 2015–2016 season. We selected compilation videos of two games in which each player was not substituted, against opponents of similar strength (see OSF Section 3.2, 'Explanation of existing data'). Videos were obtained from the online scouting platform Wyscout (www.wyscout.com). Finally, we retrieved players' age, games played, and market value (from www.transfermarkt.com) at the end of the 2015–2016 soccer season.

2.2.2. Criterion

We used players' market value at the end of the 2018–2019 season as the criterion measure. These market values were estimated by users from the forum www.transfermarkt.com and can be considered 'wisdom of the crowd' judgments (Herm et al., 2014). Bergkamp et al. (2019) argued for the use of an in-game soccer performance criterion that can differentiate between individual players, to study more meaningful predictor-criterion relationships in talent identification settings. In addition to in-game performance, estimated market values are related to by a multitude of factors, such as player popularity attributes, age, injuries, total club market values, and league the player performs in (Herm et al., 2014; Müller et al., 2017; Rodríguez et al., 2019). Still, these studies found that in-game performance or expert ratings of performance were the most important contributors in predicting market values. Finally, estimated market values are highly correlated with actual transfer fees (Torgler & Schmidt, 2007). Given these results and that market values are able to differentiate between individual players (Bergkamp et al., 2019), we considered these estimates an adequate proxy for players' performance. These market values are publicly available. We chose a predictive interval of three seasons between the compilation videos and the market values so that there was some time for the values to reflect players' performance over the years.

2.2.3. Structured-mechanical rating

We created a list of eight soccer performance indicators that are deemed important for the full back position. These indicators were determined based on prior research (c.f. Bergkamp et al., 2022; Larkin & O'Connor, 2017; S. J. Roberts, Greenwood, et al., 2019) and in collaboration with the KNVB (see Table 1).

In the structured condition, players' performance was measured by asking participants to "rate each of the eight performance indicators on a 7-point scale (1 = very poor; 7 = excellent)". Because we had no reason to assume that some indicators should be considered more important than others, we took the mean of these ratings and used this composite rating as the structured-mechanical performance rating.

2.2.4. Structured-holistic rating

After participants rated the player on the eight criteria in the structured condition, they were asked to "rate the player's *overall* soccer performance on the eight criteria with a single rating, on a 7-point scale (1 = very poor; 7 = excellent)." This was used as the structured-holistic rating.

2.2.5. Unstructured-holistic rating

In the unstructured condition, participants did not rate each of the eight performance criteria. Instead, they were solely asked to "rate the player's *overall* soccer performance on the eight criteria with a single rating, on a 7-point scale (1 = very poor; 7 = excellent)" to obtain the

Table 1
Performance indicators deemed relevant for the full-back position.

Team function	Task	Examples of skills, actions, and abilities:
Defending	Retains compactness	Cuts off space between ball and goal, sprints back, contains vertical and horizontal spaces together with teammates, intercepts ball.
	Disrupts the offensive build up	Applies pressure on the ball; keeps opponent in front of him or provides coverage; forces opponent to play ball backwards; enters duels; applies coverage for center backs when ball is on the other side.
	Preventing goal scoring opportunities around the 18-yd box	Plays man to man, marks man, fights back in duels without fouling opponent, blocks shots, clears ball from penalty area.
Transitioning – defense to attack	Positions himself so that he can obtain the ball – make a progressive dribble or pass	Goes deep, away from the ball, between the lines, dribbles in, deep pass, guards distances with teammates, creates scoring opportunities.
Attacking	Widening space	Positions himself at the right moment, vertically and horizontally, goes deep, does not move towards ball (dependent on the situation)
	Building up offensively	Attacks space, deep, is available for the pass, creates overload with central defender, dribbles, passes.
	Creating goal scoring opportunities	Through combination with teammates or individual action creates early cross, dribbles, passes, sprints deep.
Transitioning – attack to defense	Is available to stop the counter, apply pressure, and retain compactness.	Applies pressure, sprints back, tackles, does not lose challenges, blocks passing lanes.

Note: performance indicators are phrased as tasks (i.e., middle column), which are categorized under four team functions: defending, attacking, and transitioning (from attack to defense and vice versa, i.e., left column). Each task includes a number of corresponding actions, skills, and abilities as examples (i.e., right column).

unstructured-holistic rating.

2.2.6. Prediction of market value

In all three conditions, we measured the prediction of players' market value by asking participant to "make a prediction of the player's market value at the end of the 2018/2019 soccer season." This prediction was made on a continuous scale in millions of euros with 1 decimal (e.g. 0.4 million = 400,000). To provide participants with a reference point, we included the range from the lowest to the highest market value for the group of full backs in the background information.

2.2.7. Confidence and use intentions

Confidence was measured in each condition, after they made their predictions, by asking participants how confident they were that their assessment and/or prediction were accurate (1 = *no trust*, 5 = *a lot of trust*). Participants' intention to use the assessment approaches was measured through a three-item scale that was used in previous personnel selection research (Nolan & Highhouse, 2014) that we translated into Dutch and adapted to this context by replacing "hiring decisions" with a Dutch translation of "future talent selection decisions". Internal consistencies of the use intentions scale based on our data were acceptable-to-good (Unstructured-holistic $\alpha = 0.68$; structured-mechanical $\alpha = 0.83$; Structured-holistic $\alpha = 0.84$; No-observation $\alpha = 0.81$).

2.3. Procedure

The digital experiment was distributed via Qualtrics (Qualtrics, Provo, Utah). Before distribution, the questions in the experiment were reviewed by a KNVB scouting coordinator and two coaches and two scouts of a professional soccer club to improve terminology, consistency, and clarity. Participants were randomly allocated to a version of the questionnaire that contained either the structured or unstructured condition as the first condition (See OSF preregistration, section 2.4, 'randomization'). The no-observation condition was the final condition in both versions. Ethical approval was granted by the Ethical Committee of Psychology of the University of Groningen (code PSY-2021-S-0142) and informed consent was obtained for all participants prior to the experiment.

After they provided consent and answered five questions on demographics, participants were shown a description that stated to imagine a situation in which they were a scout for a sub-top (i.e., positions 4–9 out of 18) Eredivisie club. The club was interested in finding a new full back and wanted participants to assess the current performance of several players. Participants were given the list with the eight performance indicators that the club deemed important for the full back position (see Table 1). In each condition, a different player was randomly drawn from the sample of 25 players. We aimed to evenly distribute the players shown to participants across conditions, so that each player was rated (approximately) an equal number of times.

In the structured condition, participants were presented with the player's compilation video and were asked to watch the full video. Afterwards, participants were asked to rate each of the eight indicators. We took the mean of these ratings to obtain the structured-mechanical rating. Participants then provided their structured-holistic rating. Next, participants were shown the ratings for each indicator they just provided, their structured-holistic rating, and the player's background information: the player's age, number of competition games played, and market value in the 2015–2016 season. They were then asked to make a prediction of the player's market value in the 2018–2019 season. Finally, participants were asked to indicate the confidence they had in their prediction and their intention to use this method for talent selection decisions. Use intentions and confidence were measured for both structured-mechanical and structured-holistic assessment approaches.

The unstructured condition was similar to the structured condition, but participants were not asked to rate each performance indicator separately. Instead, they were asked to provide their unstructured-holistic rating. They were also asked to predict this player's market value, based on their unstructured-holistic rating and the same background information as provided in the structured condition. Furthermore, they were asked to indicate their use intentions and confidence.

Finally, participants predicted a third player's market value solely based on the aforementioned background information, without any video material. We also measured participant's confidence and use intentions in this condition.

2.4. Statistical analysis

2.4.1. Reliability

The reliability of the performance ratings in each assessment condition was assessed by computing the intraclass correlation coefficient (ICC, one-way random effects, single measures, (Koo & Li, 2016). We used a bootstrap procedure to compare the different ICC values between the three ratings (1 = structured-mechanical vs. unstructured-holistic, 2 = structured-mechanical vs. structured-holistic, 3 = structured-holistic vs. unstructured-holistic). For each comparison, we resampled with replacement the existing data 5000 times and computed the difference

between two ICC's each iteration. We then computed a 95% confidence interval around this estimate.

The number of observations per player was not perfectly evenly distributed, as some observations were removed because the participant did not meet the eligibility criteria. In short, most players had four observations, whereas a few had five or three (see Appendix A for full overview). We used a player's four most recent observations in case that player had 5 observations. Moreover, we used the 'iccNA' from the 'irrNA' R package (v0.2.2 (Brueckl & Heuer, 2021), to compute the ICC's, which can handle randomly missing data for players who had three observations.

2.4.2. Predictive validity

The distribution of players' market values was highly right-skewed and the relationship with participants' performance ratings could not be described as linear. Therefore, we computed Spearman's correlations (r_s) between the performance ratings from each assessment condition and players' market value in the 2018–2019 season.[1] We assessed whether the difference between two coefficients was statistically significant using the method for dependent correlation coefficients – common index - described by (Steiger, 1980) .[2]

2.4.3. Contribution of observing in-game performance

To explore if observing players' in-game performance helps or hurts predictive validity, we computed Spearman's correlations between participants' prediction of market value and players' actual market value in the 2018–2019 season in the three conditions.[1] We compared the correlation in the no-observation condition against the unstructured and structured assessment condition, using the method for dependent correlations – common index – by Steiger (1980) described above.

2.4.4. Model of participants' structured-holistic assessment approach

In the structured condition, we constructed a linear model regressing participants' prediction of the 2018–2019 market value on their ratings of the separate performance indicators, the players' age, number of games played, and market value at the end of the 2015–2016 season. Because we had relatively many performance predictors compared to the number of observations, we reduced the data by computing for each participant an average attacking and defending rating, by taking the mean of the three attacking and three defending ratings, respectively. Based on Q-Q and fitted vs. residuals plots, the assumptions of linearity, homoscedasticity, and normality or errors for this model were violated. Therefore, we took the natural logarithm of participants' market value prediction and the 2015–2016 market value predictor, which improved these assumptions.[3] For this model with transformed variables, we computed the relative weights of each predictor in explaining the R^2 by using the 'relaimpo' R package (Grömping, 2006).

2.4.5. Confidence and use intentions

We constructed a mixed model for the confidence question (i.e., "how confident are you that your assessment and/or prediction is accurate") and the mean score of the use intention scale (e.g., "how likely are you to use this assessment and/or prediction approach in future talent identification practices"), with observations nested within individuals and the four conditions as a fixed within-subjects factor. We compared the estimated marginal means in a post-hoc analysis.[4]

3. Results

3.1. Inter-rater reliability

The inter-rater reliabilities were very small for all performance ratings. The ICC of the unstructured-holistic rating was the largest (ICC = 0.14, 95% CI = −0.04; 0.39), followed by the structured-holistic rating (ICC = 0.07, 95% CI = −0.09; 0.31) and the structured-mechanical rating (ICC = 0.04, 95% CI = −0.11; 0.27). Because the differences were not in the expected direction, we did not test the ICC differences for statistical significance.

3.2. Predictive validity of performance ratings

The validities of the different performance ratings in predicting players' market values were small-to-moderate and statistically significant (Cohen, 1988). The unstructured-holistic rating yielded the largest predictive validity ($r_s = 0.31$, 95% CI = 0.11; 0.48, $p < 0.01$), followed by the structured-mechanical rating ($r_s = 0.25$, 95% CI = 0.06; 0.43, $p = 0.01$) and the structured-holistic rating ($r_s = 0.22$, 95% CI = 0.02; 0.40, $p = 0.03$). Except for the difference between the structured-mechanical and the structured-holistic rating, differences in correlation coefficients were not in the expected direction. The difference between the structured-mechanical and structured-holistic rating was small and not statistically significant ($r_{s\ \text{difference}} = 0.03$, $p = 0.38$).

3.3. Correlation of participants' market value prediction

Correlations between participants' prediction of players' market value and players' actual market value were moderate and statistically significant. Validity for participants' predictions in the structured condition was the largest ($r_s = 0.41$, 95% CI = 0.22; 0.56, $p < 0.01$), followed by predictions from the unstructured condition ($r_s = 0.38$, 95% CI = 0.19; 0.54, $p < 0.01$) and the no-observation condition ($r_s = 0.25$, 95% CI = 0.05; 0.43, $p < 0.01$). Differences in correlation coefficients between the no-observation condition and the two other assessment conditions were small and not statistically significant (see Table B1, appendix B). Hence, we found no evidence that observing soccer players in games hurt or helped validity, but the differences point more towards 'helps' than 'hurts.'

3.4. Model of participants' structured assessment

Participants' structured ratings on the indicators and the players' background information explained 53% of the variance in participants' predictions of market value ($R^2 = 0.53$, $R^2_{adj} = 0.49$, $F(7, 88) = 14.26$, $p < 0.01$; see Table B2 and B3 in appendix B for the regression results and correlation matrix, respectively). Figure 1 presents the relative importance of each predictor in explaining the variance in participants'

[1] Our pre-registration specified that we would compute Pearson's correlations for these analyses. However, given the skewness of player' market values and participants' prediction of market value, we opted for a non-parametric alternative (i.e., Spearman's correlation).

[2] The test for differences in dependent correlations requires the correlation between the predictor measures (e.g., correlation between unstructured-holistic and structured-mechanical rating). However, this correlation is dependent on the indexing of the observations within a player. Therefore, we computed the correlation coefficient between each pair of columns with the 4 most recent ratings (4 x 4 = 16 correlations) and averaged these coefficients through a meta-analysis with the Fisher r-to-Z transformation. This average correlation was used as the estimate of the dependent correlation between ratings in each condition. This procedure is not described in our pre-registration.

[3] Our pre-registration did not specify any transformations of the variables. The violation of the assumptions is likely due to the skewness of the market value variables. As there is no straightforward non-parametric regression variant, we opted to transform these variables by taking the logarithm.

[4] Our pre-registration specified that we would conduct a repeated measures ANOVA (RMA) to assess confidence and use intentions in each condition. However, we opted to conduct this analysis in the mixed model framework, as our design was not fully balanced (i.e., unstructured-holistic $n = 95$, structured-mechanical $n = 96$, structured-holistic n = 96, no-observation $n = 94$) and this approach tends to be more flexible than RMA's with regard to missing values.

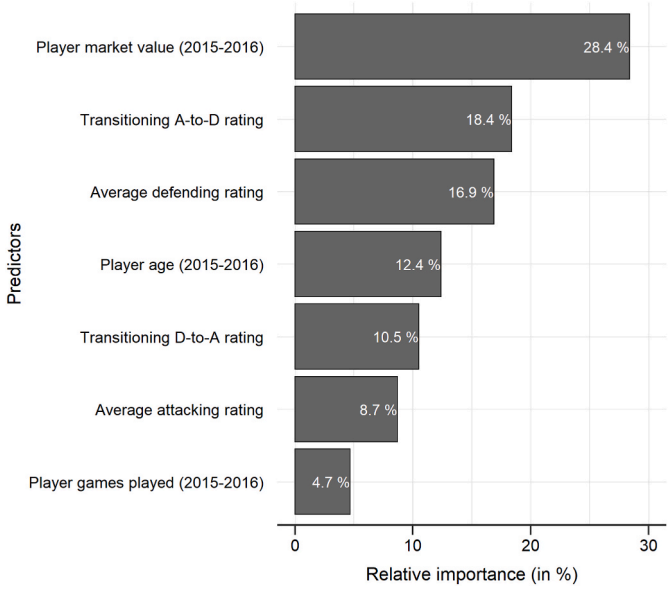

Figure 1. Relative importance of each predictor in predicting the logarithm of participants' 2018–2019 market value prediction. Note: Relative importance is scaled to sum to 100%.

predictions of players' market value. Player's market value in the 2015–2016 season had the largest contribution of the individual predictors in determining participants' prediction of market value (relative contribution to $R^2 = 28.4\%$). When combined, the performance ratings contributed 54.5%, with the transitioning A-to-D rating (contribution = 18.4%) and average defending rating (contribution = 16.9%) having the largest contribution.

3.5. Use intentions and confidence

The mixed model for the mean use intention score with assessment approach as a factor and a random intercept for participants was statistically significant ($F(3, 283.06) = 44.87$, $p < 0.01$). Post-hoc comparisons of the marginal means of the fitted model showed that the mean use intention of the no-observation approach was significantly lower ($M = 2.62$, $SD = 0.62$) than the mean of the unstructured-holistic ($M = 3.23$, $SD = 0.55$), structured-mechanical ($M = 3.16$, $SD = 0.51$), and structured-holistic approach ($M = 3.29$, $SD = 0.45$). Comparisons among the other assessment approaches did not differ significantly (see Table B4 in appendix B).

The mixed model with the confidence score as the dependent variable and the three prediction approaches was also statistically significant ($F(3, 282) = 82.68$, $p < 0.01$). Post-hoc comparisons of the marginal means also showed that the mean confidence in the no-observation approach ($M = 1.99$, $SD = 1.02$), was substantially lower than the mean confidence in the unstructured-holistic ($M = 3.21$, $SD = 0.83$), structured-mechanical ($M = 3.11$, $SD = 0.81$), and structured-holistic approach ($M = 3.30$, $SD = 0.68$). Comparisons among the latter three assessment approaches also did not differ significantly (see Table B5 in appendix B).

4. Discussion

The aim of the present study was to examine whether a structured observational assessment approach paired with mechanical combination of information improves the reliability and predictive validity of soccer coaches' and scouts' performance ratings. Moreover, the exploratory section of this study examined (a) whether observing soccer players in-game performance helps or hurts predictive validity, (b) how different sources of information contribute to coaches' and scouts' predictions, and (c) how different assessment approaches affect participants' use intentions and confidence.

4.1. Reliability and validity of performance ratings

Our hypotheses were that the structured-mechanical ratings yielded the highest inter-rater reliability and predictive validity, followed by structured-holistic ratings, and the unstructured-holistic ratings. Contrary to our expectations, the unstructured-holistic performance ratings were the most reliable and predictively valid, although the differences were marginal. Moreover, the reliabilities of the ratings overall were very poor, which likely decreased the chance of finding high predictive validities in general. Accordingly, the predictive validities of the ratings overall were small-to-moderate.

The absence of systematic differences in reliability was not in accordance with prior research on structured collection and mechanical combination of information. For example, while the ICC estimate of the unstructured-holistic rating was similar to the estimate found in the study by Arkes et al. (2006) on rating scientific presentations (ICC = 0.14 compared to ICC = 0.15 by Arkes), the ICC of the structured-mechanical rating was much smaller (ICC = 0.04 compared to ICC = 0.37). Given that reliability is typically a necessary, but not sufficient condition for validity, these results make finding large validities, as well as the expected differences in validity highly unlikely.

Interestingly, the reliability of the structured-mechanical rating exceeded the theoretical limit of the square root of its reliability. This is possibly due to correlated errors in the ratings and market values (Nimon, Reichwein Zientek, & Henson, 2012). This can likely explain why we did find predictive validities that approximated the range of expected values ($0.1 < r_s < 0.3$), even though we found poorer reliabilities than expected.

Given these ambiguous reliabilities, we found no evidence that mechanical combination of the ratings substantially improved its predictive validity, which disagrees with the findings by Dana et al. (2013) on predicting GPA scores or findings on the benefit of mechanical combination when using already quantified predictors (Ægisdóttir et al., 2006; Kuncel et al., 2013). Interestingly, the reliability and predictive validity estimates of the structured-holistic ratings were also smaller than those of the unstructured-holistic ratings' estimates. Thus, we did not find evidence of a benefit of structure – independent from mechanical combination of information (Huffcutt & Arthur, 1994).

The current findings could suggest that the structured assessment approach implemented in this study was not structured enough. Compared to rating multiple pre-established indicators (i.e., as in the current study), an even higher level of rating structure is established when observations are evaluated against pre-established benchmark answers (e.g., anchored rating scale) and on more narrowly defined tasks. Establishing this level of rating structure also requires structuring the tasks that candidates (i.e., players) have to demonstrate. However, task structure is low in soccer when observing player's in-game performance, because the tasks that each player encounters are not standardized and thus not consistent across games or players. For example, an interviewer can ask each candidate the exact same questions, which can subsequently be checked against benchmark answers. In contrast, the dynamic nature of a soccer game implies that some 'tasks' may show up more or less often (or not at all) and may vary in difficulty or complexity. This makes assessing in-game performance on a narrower task level and developing broadly applicable, explicit benchmarks very difficult. Moreover, participants in our study at least observed the same game of each player, but task consistency is even lower in practice, because scouts and coaches typically observe the same player in different games. Thus, the level of structure implemented in the current study is realistically near the highest possible level when assessing in-game soccer performance.

Possible explanations for the poor reliability and predictive validity

in the structured condition are that participants' interpretation of the eight performance indicators and the rating system differed based on their backgrounds. The current sample included coaches and scouts of (many) different soccer organizations. This may have attenuated the consistency across participants in their assessment of the eight indicators, yielding a lower reliability for the structured-mechanical rating. However, overcoming this issue by using anchored rating scales is very difficult in the absence of task structure, as explained above. Moreover, it is likely that the typical scouting approach within each soccer organization differs in terms of structure. This would imply that the level of familiarity and experience with applying a structured assessment approach differed across participants prior to the start of the experiment, which may have also affected their ability to assess each performance indicator separately. As a future avenue, the different interpretation of performance indicators may be addressed by letting coaches and scouts define the indicators collectively or through training (Roch et al., 2012). This creates a shared agreement and definition of each performance indicator among participants (Kahneman et al., 2016). Although this was impossible in the current experiment, it is an important first step in practice when a soccer club wants to implement a structured assessment approach.

Finally, it can be argued that the current performance indicators did not cover the most important performance facets for scouts and coaches. For instance, previous studies have shown that coaches and scouts had difficulty formulating specific performance indicators, but instead assessed more general performance categories, such as 'technique' or 'physical attributes' (Bergkamp et al., 2021; A. H. Roberts, Greenwood, et al., 2019). It is possible that the specific list of indicators used in the current study did not allow participants to assess such performance categories. However, note that including these 'broadly-defined' categories also leaves more room for interpretation among participants, making it doubtful whether this practice will improve reliability estimates.

Taken together, the current study did not find support for hypotheses H_1 and H_2. Future studies should examine whether the reliability and predictive validity of coaches' and scouts' structured-mechanical ratings are, as suggested by the outcomes of the study, not superior to structured-holistic and unstructured-holistic ratings, or whether they *are* superior when accounting for the design-related arguments mentioned above.

4.2. Contribution of observing performance, use intentions, and confidence

Correlations between participants' prediction of market values and players' actual market values were larger after observing the player on video (i.e., in the structured and unstructured conditions) than after not observing a player (i.e., in the no-observation condition), although the differences were not statistically significant. This suggests that participants extracted valid information from the videos. Relatedly, there was no strong evidence that participants predictions were hurt by being exposed to irrelevant information such as psychical appearance. This finding differed from the literature on unstructured hiring interviews, which have been shown to hurt the predictive validity of decision-makers predictions (Dana et al., 2013).

Nevertheless, it is difficult to assess which valid cues participants extracted from the videos. According to the linear model on participants' prediction of market value, participants based their prediction mostly on players' prior market value (28.4%) and their ratings of performance (combined 54.5%). The prior market value was a strong predictor of future market value ($r_s = 0.42$), which participants correctly took into account. Furthermore, approximately half of the variance was unexplained. It is possible that this half consists of valid observations in the video that were not captured by the list of specific performance indicators in this study.

However, if participants were to consistently observe, assess, and integrate the same valid indicators, then this should also be reflected in the inter-rater reliability of the unstructured-holistic or structured-holistic ratings. Yet, the reliability of these ratings was poor. This makes it unlikely that participants were consistent in which (valid) indicators they used, and in how they assessed and integrated them. In sum, future studies should investigate further which valid cues soccer coaches and scouts observe in games and how they integrate them in their performance predictions.

Finally, participants indicated that they had substantially less intentions to use and confidence in an assessment approach that did not involve observing a player's in-game performance. This suggests that participants feel they can more adequately 'make sense' of their assessments and predictions when based on their own observations of players' performance (Dana et al., 2013). Moreover, we did not find significant differences in mean confidence and use intentions between the unstructured-holistic, structured-mechanical and structured-holistic assessment approaches. This finding also differed from the literature on hiring interviews, where structured-mechanical assessment approaches have been found to yield lower use intentions and confidence among participants (Nolan & Highhouse, 2014). Taken together, it suggests that participants may be open for using either an unstructured or structured assessment approach, granted that they can observe the player's in-game performance.

4.3. Limitations

The present study's limitations may lie in its ambition to mimic a soccer scouting context. For example, to accurately portray each player's skills and abilities, we included two different soccer games in each compilation video. However, this made the videos relatively long (i.e. approximately 30 min), and it took participants' approximately 1.5–2 h to complete the entire experiment. Therefore, fatigue could have affected how serious participants' assessed players' performance. Moreover, most scouts and coaches did not regularly assess players' performance *on video* and could have been relatively unfamiliar with this approach. However, video observations were necessary to make sure that participants based their assessment on the same information.

Furthermore, a limitation of this study is that the main analyses were underpowered. We aimed to include soccer coaches and scouts who worked at the highest competitive levels. Unfortunately, it was simply impossible to include more participants who met our inclusion criteria. However, given that high-level coaches and scouts are a very specific population, the current number of participants included can be considered relatively large for the field of sports sciences.

Another limitation was that not every player was observed an exactly equal number of times, meaning that we had missing data for the reliability analyses. While the analysis technique was able to account for this limitation, a balanced design would have been more robust and powerful. Finally, a methodological limitation is that we had to take the average of the attacking and defending ratings for the regression analysis, due to the number predictors relative to the number of observations. This prevented us from assessing the relative contribution at the level of the independent performance indicators.

4.4. Concluding remarks

It is important that soccer coaches' and scouts' assessment of soccer performance are reliable and predictively valid. While previous studies have shown that assessment approaches based on structured information collection and mechanical combination of information typically yield stronger reliability and predictive validity than unstructured holistic assessment approaches, the present study did not find evidence for this hypothesis in the context of scouting soccer players. Inter-rater reliabilities of participants' ratings were poor, and predictive validities small-to-moderate. Moreover, the exploratory findings tentatively suggest that observing players' performance does not hurt, but may help

predict performance, and participants indicated that they had more confidence and intention to use an assessment approach that involved observing players.

The ambiguous findings make it difficult to formulate clear implications for scouting soccer players on the basis of this study. Nevertheless, the current study is the first to examine the potential benefit of structured information collection and mechanical combination information in a soccer context. Given the strong evidence on the benefit of structured information collection and mechanical combination of information in other domains, we consider it worthwhile for future research to investigate how these principles can contribute to improve soccer scouting. For example, future research may consider whether structured assessment of a (smaller) list of indicators defined collectively by a group of coaches and scouts with the same organizational background improves predictive validity and reliability. The current study has laid the groundwork for research examining structured and mechanical information collection and combination in soccer, and opened up fruitful avenues for future research to consider.

Funding

This research was partially funded by the Royal Dutch Football Association (Koninklijke Nederlandse Voetbalbond, KNVB, www.knvb.com). The KNVB did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

Conflict of Interest: The authors declare that no conflict of interest exists.

Data availability

Data will be made available on request.

Acknowledgement

We would like to thank Casper Albers for his helpful suggestions regarding the data analysis and Sil Piek for assisting in the conceptualization of the performance indicators. Furthermore, we would like to thank Jan Verbeek, Maurice Hagebeuk, Talent Performance Coaches of the KNVB's Jeugdplan Nederland, and heads of scouting of participating clubs for assisting in recruiting scouts and coaches.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.psychsport.2022.102257.

Appendix A

Table A1
Number of observations per player in each assessment condition

	Condition		
Player number	Unstructured	Structured	No-observation
1	5	3	3
2	4	4	3
3	4	4	4
4	4	3	4
5	5	3	3
6	4	4	4
7	4	4	4
8	3	4	2
9	4	4	4
10	4	4	4
11	4	3	4
12	4	4	3
13	5	4	4
14	5	5	4
15	4	4	3
16	4	3	3
17	4	4	4
18	3	4	5
19	2	4	4
20	3	3	4
21	4	4	5
22	4	4	4
23	3	5	4
24	3	4	4
25	2	4	4

Table B1
Correlational differences between participants' market value predictions across assessment conditions.

Comparison	r_{s12}	r_{s13}	r_{s23}	r_s difference	*t*	*df*	*p*
Unstructured vs. No-observation	0.38	0.25	0.19	0.13	1.06	93	0.29
Structured vs. No-observation	0.41	0.25	0.32	0.16	1.40	91	0.17

Note: r_{s12} = Spearman's correlation between participants' market value predictions and first condition in 'comparison' column (e.g., 'Unstructured'), r_{s13} = Spearman's correlation between participants' market value predictions and second condition in 'comparison' column (e.g., No-observation), r_{s23} = Spearman's correlation between first and second condition in 'comparison' column, r_s difference = difference in Spearman's correlations between participants' market value prediction and first and second condition in comparison column, respectively (i.e., $r_{s12} - r_{s13}$).

Table B2
Results from regression model predicting the logarithm of players' market value in the 2019–2020 season

Predictor	β	SE	*t*	*p*	*Relative importance (in %)*
(Intercept)	8.36	1.22	6.84	<0.01	–
Player market value[a,b]	0.44	0.09	4.88	<0.01	28.4
Transition A-to-D rating	0.15	0.08	1.86	0.07	18.4
Average defending rating[c]	0.10	0.11	0.88	0.38	16.9
Player age[b]	−0.25	0.07	−3.77	<0.01	12.4
Transition D-to-A rating	0.09	0.08	1.03	0.30	10.5
Average attacking rating[c]	0.04	0.08	0.47	0.64	8.7
Player games played[b]	0.01	0.01	0.88	0.38	4.7

$R^2 = 0.53$, $R^2_{\text{adj}} = 0.49$, $F(7, 88) = 14.26$, $p < 0.001$
Note: All predictors, with the exception of 2015–2016 player market value, were mean centered before the analysis. Relative importance is scaled to sum to 100%; a = natural logarithm of player market value; b = in the 2015–2016 soccer season; c = Average of three attacking and defending ratings, respectively

Table B3
Correlations between different predictors in regression model for participants' market value prediction in structured condition

	Market value pred.	Att. rating 1	Att. rating 2	Att. rating 3	Avg att. rating	Trans. A-to-D rating	Def. rating 1	Def. rating 2	Def. rating 3	Avg def. rating	Trans. D-to-A rating	Market value (2015–2016)[a]	Games played (2015–2016)[a]	Age (2015–2016)[a]
Market value pred.	1	–	–	–	–	–	–	–	–	–	–	–	–	–
Att. rating 1	0.45	1	–	–	–	–	–	–	–	–	–	–	–	–
Att. rating 2	0.48	0.57	1	–	–	–	–	–	–	–	–	–	–	–
Att. rating 3	0.38	0.58	0.47	1	–	–	–	–	–	–	–	–	–	–
Avg att. rating	0.42	0.5	0.43	0.5	1	–	–	–	–	–	–	–	–	–
Trans. A-to-D rating	0.53	0.71	0.72	0.57	0.57	1	–	–	–	–	–	–	–	–
Def. rating 1	0.33	0.42	0.44	0.45	0.77	0.47	1	–	–	–	–	–	–	–
Def. rating 2	0.35	0.44	0.35	0.42	0.91	0.47	0.66	1	–	–	–	–	–	–
Def. rating 3	0.37	0.4	0.32	0.4	0.84	0.51	0.4	0.63	1	–	–	–	–	–
Avg def. rating	0.53	0.87	0.82	0.81	0.57	0.8	0.52	0.49	0.45	1	–	–	–	–
Trans. D-to-A rating	0.44	0.46	0.52	0.41	0.68	0.56	0.72	0.64	0.41	0.55	1	–	–	–
Market value (2015–2016)[a]	0.46	0.13	0.24	0.11	0.11	0.2	0.07	0.06	0.14	0.19	0.13	1	–	–
Games played (2015–2016)[a]	0.24	−0.06	0.2	−0.04	0.13	0.14	0.1	0.11	0.11	0.04	0.17	0.37	1	–
Age (2015–2016)[a]	−0.26	−0.15	−0.17	−0.09	−0.14	−0.06	−0.19	−0.05	−0.13	−0.17	−0.13	0.2	0.14	1

Note: att. = attacking, avg. = average, trans. = transitioning, def. = defending, Att rating 1–3 = (1) widening space, (2) building up, (3) creating scoring opportunities, Def rating 1–3 = (1) Retaining compactness, (2) disrupting build up, (3) preventing scoring opportunities, a = denotes background information of the player in the 2015–2016 soccer season

Table B4
Difference in mean use intentions between different assessment approaches

Comparison	Mean difference	SE	df	*t* ratio	p^a
Structured-mechanical vs. unstructured-holistic	−0.07	0.06	282.65	−1.04	0.72
Structured-mechanical vs. structured-holistic	−0.13	0.06	282.08	−2.05	0.17
Structured-mechanical vs. No-observation	0.54	0.06	283.03	8.33	<0.01
Unstructured-holistic vs. structured-holistic	−0.06	0.06	282.65	−1.00	0.75
Unstructured-holistic vs. No-observation	0.61	0.06	282.46	9.35	<0.01
Structured-holistic vs. No-observation	0.67	0.06	283.03	10.37	<0.01

a = Controlling for multiple comparison with Tukey's post hoc test

Table B5
Difference in mean confidence between different assessment approaches

Comparison	Mean difference	SE	df	*t* ratio	p^a
Structured-mechanical vs. unstructured-holistic	−0.10	0.095	282.56	−1.05	0.72
Structured-mechanical vs. structured-holistic	−0.19	0.095	282.05	−1.98	0.20
Structured-mechanical vs. No-observation	1.12	0.095	282.87	11.79	<0.01
Unstructured-holistic vs. structured-holistic	−0.09	0.095	282.56	−0.93	0.79
Unstructured-holistic vs. No-observation	1.22	0.095	282.36	12.81	<0.01
Structured-holistic vs. No-observation	1.31	0.095	282.87	13.76	<0.01

a = Controlling for multiple comparison with Tukey's post hoc test

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Arkes, H. R., Schaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making, 19*, 429–439. https://doi.org/10.1002/bdm

Bergkamp, T. L. G., Frencken, W. G. P., Niessen, A. S. M., Meijer, R. R., & den Hartigh, R. J. R. (2021). How soccer scouts identify talented players. *European Journal of Sport Science*, 1–39. https://doi.org/10.1080/17461391.2021.1916081

Bergkamp, T. L. G., Frencken, W. G. P., Niessen, A. S. M., Meijer, R. R., & den Hartigh, R. J. R. (2022). How soccer scouts identify talented players. *European Journal of Sport Science, 22*(7), 994–1004. https://doi.org/10.1080/17461391.2021.1916081

Bergkamp, T. L. G., Niessen, A. S. M., den Hartigh, R. J. R., Frencken, W. G. P., & Meijer, R. R. (2019). Methodological issues in soccer talent identification research. *Sports Medicine, 49*(9), 1317–1335. https://doi.org/10.1007/s40279-019-01113-w

Bishop, M. A., & Trout, J. D. (2002). 50 years of successful predictive modeling should be enough: Lessons for philosophy of science. *Philosophy of Science, 69*(S3), S197–S208. https://doi.org/10.1086/341846

Brueckl, M., & Heuer, F. (2021). *irrNA: Coefficients of interrater reliability—generalized for randomly incomplete datasets*. R package, 0.2.2.

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology, 58*(3), 673–702. https://doi.org/10.1111/j.1744-6570.2005.00516.x

Christensen, M. K. (2009). An eye for talent": Talent identification and the "practical sense" of top-level soccer coaches. *Sociology of Sport Journal, 26*(3), 365–382. https://doi.org/10.1123/ssj.26.3.365

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. In *Statistical power analysis for the behavioral sciences* (2nd dr.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*(5), 565–579. https://doi.org/10.1037/0021-9010.80.5.565

Dana, J., Dawes, R., & Peterson, N. (2013). Belief in the unstructured interview: The persistence of an illusion. *Judgment and Decision Making, 8*(5), 512–520. Retrieved from http://journal.sjdm.org/vol8.5.html.

Dana, J., & Rick, T. (2006). In defense of clinical judgment … And mechanical prediction. *Journal of Behavioral Decision Making, 19*, 413–428. https://doi.org/10.1002/bdm

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Den Hartigh, R. J. R., Niessen, A. S. M., Frencken, W. G. P., & Meijer, R. R. (2018). Selection procedures in sports: Improving predictions of athletes' future performance. *European Journal of Sport Science, 18*(9), 1191–1198. https://doi.org/10.1080/17461391.2018.1480662

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Herm, S., Callsen-Bracker, H. M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006

Huffcutt, A. I., & Arthur, W. (1994). Hunter and hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*(2), 184–190. https://doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21*(3), 264–276. https://doi.org/10.1111/ijsa.12036

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment, 22*(3), 297–309. https://doi.org/10.1111/ijsa.12078

Johansson, A., & Fahlén, J. (2017). Simply the best, better than all the rest? Validity issues in selections in elite sport. *International Journal of Sports Science & Coaching, 12*(4), 470–480. https://doi.org/10.1177/1747954117718020

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise. *Harvard Business Review, 2017-Janua*(128), 52–53.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Larkin, P., & O'Connor, D. (2017). Talent identification and recruitment in youth soccer: Recruiter's perceptions of the key attributes for player recruitment. *PLoS One, 12*(4), Article e0175716. https://doi.org/10.1371/journal.pone.0175716

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review, 33*(7), 883–900. https://doi.org/10.1016/j.cpr.2012.09.008

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599–616.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Meijer, R. R., Neumann, M., Hemker, B. T., & Niessen, A. S. M. (2020). A tutorial on mechanical decision-making for personnel and educational selection. *Frontiers in Psychology, 10*(January). https://doi.org/10.3389/fpsyg.2019.03002

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005

Neumann, M., Niessen, A. S. M., & Meijer, R. R. (2021). Implementing evidence-based assessment and selection in organizations: A review and an agenda for future research. *Organizational Psychology Review, 11*(3), 205–239. https://doi.org/10.1177/2041386620983419

Nimon, Kim, Reichwein Zientek, Linda, & Henson, K. Robin (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00102

Nolan, K. P., & Highhouse, S. (2014). Need for autonomy and resistance to standardized employee selection practices. *Human Performance, 27*(4), 328–346. https://doi.org/10.1080/08959285.2014.929691

Roberts, A. H., Greenwood, D., Humberstone, C., & Raynor, A. J. (2020). Pilot study on the reliability of the coach's eye: Identifying talent throughout a 4-day cadet judo camp. *Frontiers in Sports and Active Living, 2*(December), 1–8. https://doi.org/10.3389/fspor.2020.596369

Roberts, A. H., Greenwood, D. A., Stanley, M., Humberstone, C., Iredale, F., & Raynor, A. (2019). Coach knowledge in talent identification: A systematic review and meta-synthesis. *Journal of Science and Medicine in Sport, 22*(10), 1163–1172. https://doi.org/10.1016/j.jsams.2019.05.008

Roberts, S. J., McRobert, A. P., Lewis, C. J., & Reeves, M. J. (2019). Establishing consensus of position-specific predictors for elite youth soccer in England. *Science and Medicine in Football, 3*(3), 205–213. https://doi.org/10.1080/24733938.2019.1581369

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rodríguez, S. M., Ramírez Hassan, A., & Coad, A. (2019). Uncovering value drivers of high performance soccer players. *Journal of Sports Economics, 20*(6), 819–849. https://doi.org/10.1177/1527002518808344

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*(3), 178–200.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245. Nummer 2.

Till, K., & Baker, J. (2020). Challenges and [possible] solutions to optimizing talent identification and development in sport. *Frontiers in Psychology, 11*(April), 1–14. https://doi.org/10.3389/fpsyg.2020.00664

Torgler, B., & Schmidt, S. L. (2007). What shapes player performance in soccer? Empirical findings from a panel analysis. *Applied Economics, 39*(18), 2355–2369. https://doi.org/10.1080/00036840600660739