Talent Assessment in Soccer

Bergkamp, Tom

DOI:
10.33612/diss.236732064

Publication date:
2022

Talent Assessment in Soccer

Predicting Performance Through the Lens of Selection Psychology

Thomas Lodewijk Gijsbertus Bergkamp

This PhD project was a collaboration between the University of Groningen, the Royal Dutch Football Association (Koninklijke Nederlandse Voetbalbond, KNVB), and FC Groningen. Research presented in this thesis was partially funded by the KNVB.

Cover design and illustrations: Job Kind (Instagram @jobkind www.instagram.com/jobkind)

Printed by: Ipskamp Printing

Talent Assessment in Soccer

Predicting Performance Through the Lens of Selection Psychology

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 15 september 2022 om 14.30 uur

door

Thomas Lodewijk Gijsbertus Bergkamp

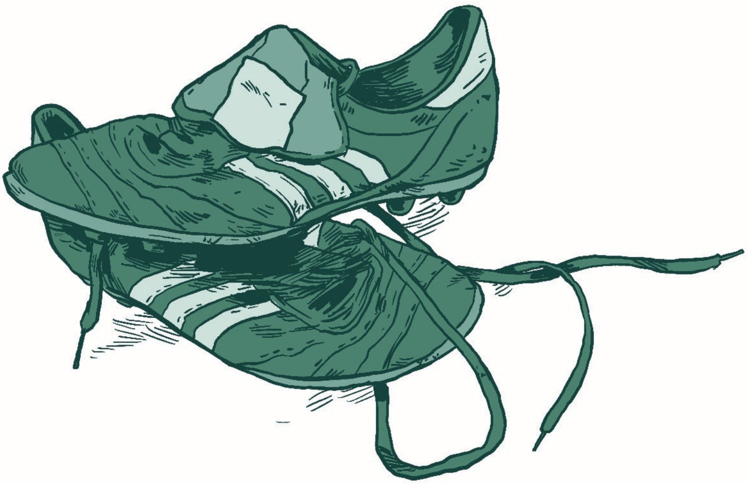
geboren op 17 november 1992
te Almelo

Promotor
Prof. dr. R.R. Meijer

Copromotores
Dr. J.R. den Hartigh
Dr. W.G.P. Frencken
Dr. A.S.M. Niessen

Beoordelingscommissie
Prof. dr. N. van Yperen
Prof. dr. M. Born
Prof. dr. G.J.P. Savelsbergh

CONTENTS



CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In 2010, Dutch defender Virgil van Dijk played in the under-19 team of Willem II's professional youth soccer academy. He was considered an 'average' talent by the club's scouts and coaches; while Van Dijk was strong and worked hard, he was deemed to lack certain technical and tactical skills. Therefore, he did not make the club's first team and was almost deselected from the academy (Visser, 2018). When Grads Fühler – a professional scout for FC Groningen – coincidentally observed Van Dijk in a soccer game, he was immediately sold on the young player. In contrast to the staff at Willem II, Fühler observed "an explosive player who was unbeatable in duels and had great passing instincts." Van Dijk reminded Fühler of Frank Rijkaard in his role as central defender at Ajax and the Dutch national squad. FC Groningen quickly moved in to sign Van Dijk and his career progressed rapidly: Van Dijk performed successfully at FC Groningen, Celtic, and Southampton, before being signed by Liverpool for a record fee of 85 million euros. Since then, he has often been described as one of the best defenders in the premier league. In 2019, he was even rated as the best premier league player by his peers.

Soccer clubs constantly have to decide which players are most likely to excel in the (near) future. Therefore, selection decisions are inherently tied to predicting future performance. For example, the coaches, scouts and staff at Willem II did not predict that Van Dijk would become an international superstar, or even a serviceable player in the Dutch Eredivisie, when they agreed to let FC Groningen sign him. Many similar examples exist. Although we do not know if Willem II could have known better, the case of Van Dijk begs the question what methods coaches, scouts, and clubs can use to make more accurate soccer performance predictions. In other words, how can soccer performance be predicted reliably and validly? This is the central question of the present thesis.

1.1.1 PREDICTING SOCCER PERFORMANCE

Throughout this thesis, I define soccer performance as "all observable and measurable actions, behaviors, and outcomes that soccer players engage in and which contribute to the team's tasks within a soccer game." This relatively broad definition emphasizes individual players' performance within competitive soccer games (i.e., in-game performance). Examples of soccer performance incorporated in this definition include in-game passing performance or high-intensity meters run (Goes et al., 2018; Pappalardo et al., 2019; Stevens et al., 2016). Of course, there are other definitions of soccer performance that are more abstract and include player accomplishments or performance levels, such as the status of being an elite player in an academy (Baker et al., 2015; Swann et al., 2015). The latter operationalizations of soccer performance have traditionally been used in the sports science literature.

The interest of the sports science literature in predicting soccer performance levels has been evident in the past decades. In two exemplary studies, Williams and Reilly (2000) and Reilly et al. (2000) proposed a conceptual model which highlighted potential predictors of soccer performance. The papers became two of the most cited papers in the history of the Journal of Sports Sciences (Williams et al., 2020). They sparked a plethora of studies that examined whether these predictors – including physical and physiological (e.g., sprinting speed), psychological (e.g., motivation), and technical (e.g., dribbling skills) performance indicators1 – could be used to discriminate between soccer performance levels (e.g., elite vs. non-elite players) and determine who would excel in the future (Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018; Sarmento, Anguera, et al., 2018). At the same time, researchers carefully explored whether tests assessing these indicators could be used to assist in talent identification procedures (Güllich & Cobley, 2017; Lidor et al., 2009). However, although different studies had various levels of success in discriminating between players in different soccer performance levels, they have not identified a consistent set of indicators which validly predicts future performance (Bergkamp et al., 2018; Breitbach et al., 2014; Johnston et al., 2018).

Recent discussions related the inability of the literature to find consistent predictors of soccer performance to the dynamic nature of sports talent: valid predictors for excellence may not be identifiable at the young age at which many players are selected (Davids, Araújo, Vilar, et al., 2013; Den Hartigh, Hill, et al., 2018). Researchers have also explained the mixed findings in light of the in the literature's 'reductionist' approach (Breitbach et al., 2014; Phillips et al., 2010; Renshaw et al., 2019). Specifically, by focusing on performance predictors tested in an isolated setting, the soccer literature has largely ignored in-game constraints that may be essential to understanding team-sports performance, such as the interaction with moving opponents or teammates (Den Hartigh, Niessen, et al., 2018; Pinder et al., 2011; Vilar et al., 2012). Given the complexity of soccer performance and development, some researchers even questioned whether studies that aim to predict soccer performance are worthwhile (Abbott et al., 2005; Breitbach et al., 2014; Güllich & Cobley, 2017; Phillips et al., 2010).

[1] In this thesis I use the terms 'performance indicators' and 'attributes' interchangeably to refer to any potential predictors of soccer performance.

I would argue that this conclusion is premature. I believe that there are promising opportunities for the field of sports sciences to optimize soccer performance predictions. However, in order to be most effective, this line of research needs to account for certain limitations in the current soccer performance prediction methodology. More specifically, alternative methods and approaches to predict soccer performance should cater to the complexity of soccer performance and development. This implies – among other things – that these methods should take the player, task, and environment interaction into account (Araújo et al., 2006). At the same time, such methods should ideally be tailored to the decision-making process of different stakeholders (e.g., coaches, scouts, and staff) in the soccer selection process. That is, they should aim to optimize the way in which these stakeholders make performance predictions and talent selection decisions in practice.

Although selecting talented players is difficult, the reality is that most sports organizations simply have to make selection decisions at some point, due to limited resources (e.g., financial, personnel, and facilities) or places available (Till & Baker, 2020). Accordingly, these selection decisions do not have to be based on near-perfect performance predictions: predictions that are more accurate than current procedures, but also yield imperfect reliability and validity, can contribute to making more accurate selection decisions.

1.1.2 OPTIMIZING PERFORMANCE PREDICTIONS THROUGH THE LENS OF SELECTION PSYCHOLOGY

Interestingly, psychological research on selection (further referred to as *selection psychology*, i.e., the field concerned with how to best select candidates for different achievement domains; Bergkamp et al. 2019) offers a framework that addresses these issues. Although concepts and principles from this framework are highly relevant for soccer performance predictions, they have rarely been considered in the field of sport sciences (Den Hartigh, Niessen, et al., 2018). Therefore, the primary aim of this thesis is to demonstrate – theoretically and empirically – how different assessment principles from selection psychology may improve the quality of research practices, as well as our understanding of predicting soccer performance. I will highlight two areas in which these principles offer valuable insights for the prediction of soccer performance, namely 1) to identify predictors and 2) to design procedures to collect and combine information on those predictors.

Predictors

Selection psychology offers various insights on what type of predictor information is effective and why (Mol et al., 2005; Schmidt & Hunter, 1998; Wernimont & Campbell, 1968). Besides the usefulness of the measurement of traits, different studies showed that high-fidelity, sample-based assessments that mimic the criterion performance are often good predictors of future performance, particularly in relatively homogeneous (i.e., preselected) samples (Lievens & De Soete, 2012; Niessen et al., 2018; Sackett et al., 2017). An example of such a sample-based assessment in the context of soccer is a small-sided version of an official 11-vs-11 game. Compared to official games, small-sided games (SSGs) are typically played on a smaller pitch, include less players, and are shorter in duration (Olthof et al., 2019; Van Maarseveen et al., 2017).

Performance in SSGs can be seen as a sample-based predictor because it 'samples' relevant soccer task- and performance constraints (Pinder et al., 2011). For example, this format includes the team's tasks to score goals or challenges an individual's ability to play a through ball between the defensive line. Accordingly, sample-based predictors in soccer closely align with recent insights in the field of sports sciences on how soccer performance emerges through the dynamic person-environment interactions (Davids, Araújo, Correia, et al., 2013). The ecological dynamics literature posits that this interaction – and thereby the coupling between perception and action – should remain intact in the predictor context and content, resulting in a design that is representative of the criterion performance (Araújo et al., 2006; Davids, Araújo, Correia, et al., 2013; Pinder et al., 2013). So far, hardly any research in soccer has examined SSG performance as a predictor of future soccer performance (Unnithan et al., 2012; Wilson et al., 2021). Therefore, the first aim of this thesis is to examine the predictive validity of small-sided game performance.

Collecting and combining information

Selection psychology also offers valuable insights on the way predictor information can be collected and combined to improve performance predictions. These insights relate to predictor information that is used in a quantitative form (e.g., test scores), but also to information that has to be judged and quantified by decision-makers, such as assessment of observations of performance. Given that soccer scouts and coaches typically use their own assessments of in-game performance to predict players' future performance (Jokuschies et al., 2017; Roberts et al., 2019a), insights on collecting and combining information to improve assessments are particularly relevant for these decision-makers.

Previous research in selection psychology showed that unstructured information collection and holistic combination of information based on intuitive judgments can yield inconsistent or biased predictions (Conway et al., 1995; Dana & Rick, 2006; Kahneman & Klein, 2009; Kuncel et al., 2013). To give an example in soccer, coaches collect information in an unstructured way when they assess players on indicators that happen to stand out to them, instead of on an explicit, pre-defined list of performance indicators. Moreover, when they integrate their impressions (through unstructured or structured collection) in their mind to form their overall impression, they use holistic combination to form their final assessment (Meehl, 1954). This approach is suboptimal, as information is weighted and combined inconsistently across coaches. In contrast, information is weighted and combined more consistently when decision-makers assess performance in a structured manner and combine scores 'mechanically' through a decision rule (Arkes et al., 2006; Dawes et al., 1989). This decision-rule can be relatively simple. For example, coaches who rate different performance indicators separately, and base their final assessment on the mean or sum of their separate ratings, use structured information collection paired with a decision rule (Den Hartigh, Niessen, et al., 2018; Meijer et al., 2020).

Structured information collection and combination through a decision rule are valid ways to improve performance assessments. Yet, it is unclear to what extent soccer decision-makers use these approaches, and to what extent it improves their performance assessments in terms of reliability and predictive validity. The second aim of this thesis is, therefore, to extend insights on the collection and combination of information for making performance assessments in soccer.

1.2 OUTLINE

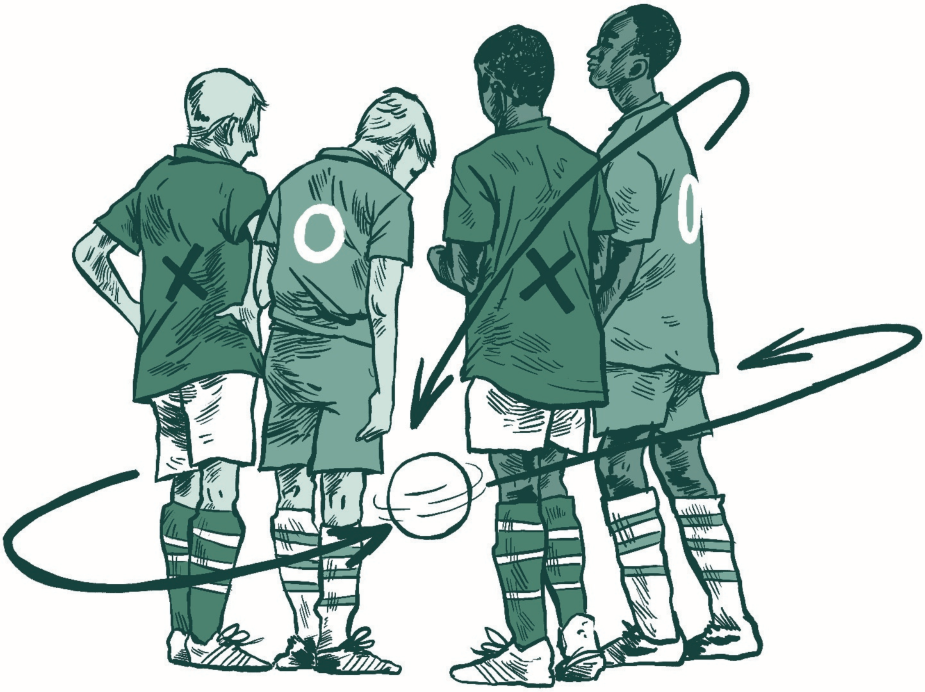
This thesis is outlined as follows. Chapter 2 provides a review of the talent identification literature in soccer from the perspective of selection psychology and performance prediction. It comprises a large set of empirical studies that have explored the relationship between performance indicators and (future) soccer performance (levels). We[2] critically discuss this literature and highlight principles from selection psychology that are relevant to the design, validity, and utility of talent identification research, but which are rarely considered in this field.

[2] Throughout this thesis, I will use 'we' when I refer to work that resulted from the collective efforts of me and my supervisors (e.g., conceptualization, design, analysis, and findings of studies in the different chapters). I will use 'I' when I refer to specific personal aims, thoughts, and reflections on the findings (i.e., mainly in the introduction and discussion of this thesis).

In chapter 3, we apply some of the suggestions discussed in chapter 2. We examine the validity of small-sided game performance in predicting 11-vs-11 soccer performance. In contrast to previous talent identification studies, we use a measure of in-game soccer performance to differentiate between individual soccer players at the predictor and criterion level.

In chapters 4 through 6 we study the decision-making process of soccer scouts and coaches. Research on how soccer scouts identify talented soccer players is scarce. Therefore, in chapter 4 we examine which soccer performance indicators scouts consider important predictors, and to what extent they assess players in a structured manner. Accordingly, in chapters 5 and 6 the reliability and predictive validity of scouts' and coaches' actual performance assessments is examined. In an experimental design and practical setting, respectively, we examine the influence of structured information collection and mechanical combination of information on their performance assessments.

Finally, I provide a summary, reflect on the findings in this thesis, and discuss some limitations and suggestions for future research in chapter 7.



CHAPTER 2

METHODOLOGICAL ISSUES IN SOCCER TALENT IDENTIFICATION RESEARCH

This chapter was published as:

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

ABSTRACT

Talent identification research in soccer comprises the prediction of elite soccer performance. While many studies in this field have aimed to empirically relate performance indicators to subsequent soccer success, a critical evaluation of the methodology of these studies has mostly been absent in the literature. In this position paper we discuss advantages and limitations of the design, validity, and utility of current soccer talent identification research. Specifically, we draw on principles from selection psychology that can contribute to best practices in the context of making selection decisions across domains. Based on an extensive search of the soccer literature, we identify four methodological issues from this framework that are relevant for talent identification research. These are (1) the operationalization of criterion variables (the performance to be predicted) as performance levels, (2) the focus on isolated performance indicators as predictors of soccer performance, (3) the effects of range restriction on the predictive validity of predictors used in talent identification, and (4) the effect of the base rate on the utility of talent identification procedures. Based on these four issues, we highlight opportunities and challenges for future soccer talent identification studies that may contribute to developing evidence-based selection procedures. We suggest for future research to consider the use of individual soccer criterion measures, to adopt representative, high-fidelity sample-based predictors of soccer performance, and to take restriction of range and the base rate into account.

2.1 INTRODUCTION

Sports organizations invest substantial resources in the search for players who have the potential to excel. These identification programs are aimed at detecting talented players who demonstrate strong performance in sport-specific abilities that are predictive of future career success (Lidor et al., 2009; Reilly et al., 2000; Williams & Reilly, 2000). Typically, these players are selected and recruited for specialized development programs that provide the appropriate learning conditions, facilities, equipment, and staff to realize the players' potential (Burgess & Naughton, 2010; Martindale et al., 2005).

Historically, talent identification programs are associated with the subjective evaluation of players' potential by coaches and scouts, who base their criteria primarily on personal taste, knowledge, and experience (Christensen, 2009; Meylan et al., 2010). In the last few decades, however, there has been an increasing interest in complementing these subjective assessments with evidence-based talent identification procedures, in order to increase the probability of selecting successful players. As a result, talent research has seen the integration of multidimensional and comprehensive models that detail performance indicators as potential predictors of successful adult performance (Unnithan et al., 2012; Vaeyens et al., 2008; Williams & Reilly, 2000), as well as a plethora of studies that have aimed to estimate the empirical relationships between these predictors and performance criteria in different sports.

Predicting future sports performance is inherently multifaceted and complex. Players' developmental trajectories are rarely linear, because cognitive and motor skills are intertwined and develop through dynamic interactions with the individual athlete's performance environment (Davids et al., 2008; Den Hartigh, Hill, et al., 2018; Den Hartigh, Niessen, et al., 2018; Gulbin et al., 2013; Güllich & Emrich, 2014). Several recently published systematic reviews have aimed to summarize the empirical evidence for performance indicators that may determine elite sports performance in general (Johnston et al., 2018; Rees et al., 2016), and in specific domains such as soccer (Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018; Sarmento, Anguera, et al., 2018). Results from these studies suggest that various physical, technical, tactical, and psychological indicators contribute to determining individual sport-specific success. However, due to the considerable variation in study designs, findings across individual talent identification studies are inconsistent and difficult to compare (Bergkamp et al., 2018; Höner & Feichtinger, 2016; Johnston et al., 2018; Murr, Raabe, et al., 2018), and therefore there is no clear set of variables that uniformly predict skill level (Breitbach et al., 2014; Johnston et al., 2018).

Still, a major aim in the field of sport sciences is to apply best-practice talent identification methods, that is, methods that allow for valid predictions of players' future performance. So far, various articles have been published discussing scientific or ethical challenges that hinder the possibilities of identifying talents (Baker et al., 2018; Breitbach et al., 2014; Pankhurst & Collins, 2013; Rees et al., 2016), such as the definition of the concept of talent (Baker et al., 2018), the influence of maturation on performance (Meylan et al., 2010), and the difficulties of early selection and early prediction of adult performance based on knowledge of how (physical) performance characteristics develop (Abbott et al., 2005; Den Hartigh, Hill, et al., 2018; Lidor et al., 2009; Pearson et al., 2006). Furthermore, several papers have discussed methodological and design features of talent identification studies (Breitbach et al., 2014; Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018). However, we observed that reflections of methodological issues specifically relevant for research on predictors and criteria used for selection purposes are scarce in the talent identification literature. Critical reflections on these issues are important for providing insight into how research results should be interpreted, and to provide guidelines for researchers in employing best-practices from a methodological point of view.

The aim of this position paper is to provide an overview of the talent identification literature and discuss some methodological issues that we consider particularly relevant in the context of selection. More specifically, we discuss methodological considerations commonly addressed in psychological research on selection (further referred to as selection psychology) regarding determinants of predictive validity, utility, and interpretability of assessment and selection procedures. Selection psychology is concerned with how to best select candidates for different achievement domains (Den Hartigh, Niessen, et al., 2018; Robertson & Smith, 2001; Vinchur & Bryan, 2012). It provides psychometric and statistical tools for measuring human traits, skills, abilities, and performance, and defines theoretical principles that affect the relationship between a (set of) predictor(s) and a criterion. While research in selection psychology has mostly focused on selecting candidates for jobs, its psychometric and statistical considerations are relevant for a wide range of performance and expertise contexts that involve selection, including higher education (Kuncel et al., 2013; Niessen & Meijer, 2017) and sports (Den Hartigh, Niessen, et al., 2018; Lyons et al., 2011).

Based on the selection psychology framework, we discuss four methodological topics that are relevant for talent identification research in soccer. Furthermore, we offer suggestions based on these topics that can improve the design of future talent identification studies and can contribute to the development of evidence-based talent identification practices. The topics are (1) the operationalization of criterion variables (the performance to be predicted), (2) the fidelity of the performance indicators used as predictors, (3) the effects of range restriction on the predictive validity of predictors used in talent identification, and (4) the effect of the base rate on the utility of talent identification procedures. Some of these issues have been briefly touched upon previously in the context of talent identification in sports (Baker et al., 2018; Breitbach et al., 2014; Güllich & Cobley, 2017; Vaeyens et al., 2008), but they are rarely thoroughly addressed (for an exception on some issues, see Ackerman (2014). Moreover, since these issues are not explicitly and specifically accounted for, we consider an in-depth evaluation valuable for advancing the field.

Because the aim of this article is to relate some specific methodological principles that are relevant in research on selection, and thereby for talent identification in soccer, we do not discuss analytic and design-related issues that have been discussed previously. Examples are the use of stepwise model selection methods (Henderson & Denison, 1989; Thompson, 1995), presenting exploratory results as confirmatory findings (Kerr, 1998; Tukey, 1980), the absence of cross-validation, issues related to multiple testing (Bender & Lange, 2001), and the use of small sample sizes, which are issues that are relevant across various scientific disciplines.

2.2 METHODOLOGICAL ISSUES

2.2.1 OPERATIONALIZING THE CRITERION

Talent identification in soccer involves the measurement of skills and abilities which are related to an indicator of soccer performance (i.e., the criterion, Breitbach et al., 2014; Lidor et al., 2009; Williams & Reilly, 2000). This criterion is ideally measured in the future (predictive validity), but is sometimes measured at the same time (concurrent validity). In our view, the talent identification literature has largely neglected to pay attention to the operationalization of criterion variables that provide information about the differences between players in terms of soccer performance after selection (Wilson et al., 2017). More specifically, an explicit measure of soccer performance is rarely used as a criterion. Instead, the criterion used in most studies is the selection decision itself, which is usually a categorical variable indicating performance or skill level. Examples of performance level indicators that have been used in studies are elite-, sub-elite-, and non-elite level (Huijgen et al., 2015; Kavussanu et al., 2011; Waldron & Worsfold, 2010), professional-, semi-professional-, or non-professional level (Haugaasen et al., 2014; Höner et al., 2017; Höner & Votteler, 2016), first team or reserves (Gravina et al., 2008), elite-, club level, or dropouts (Deprez, Fransen, et al., 2015; Figueiredo et al., 2009), national- or regional level (Zibung et al., 2016; Zuber et al., 2015, 2016), selected- and non-selected players (Den Hartigh et al., 2017; Gil et al., 2014; Goto et al., 2015; Huijgen et al., 2014), and nationally drafted or non-drafted players (Gonaus & Müller, 2012; see Table 2.1).

While using performance level as a criterion measure is understandable from a pragmatic point of view, it also carries some problems. First, this approach provides limited information on the individual differences between players on the actual outcome of interest (Phillips et al., 2010; Piggott et al., 2019), which is soccer performance in 11-a-side games (Unnithan et al., 2012). We believe that the ultimate aim of soccer talent identification research is to predict individual soccer performance as a function of performance in talent identification procedures, not selection as a function of performance in talent identification procedures (Wilson et al., 2016, 2017). Thus, talent identification procedures should strive to predict how players will perform, relative to others, but research designs that adopt a performance level criterion implicitly assume that all players within a performance level perform equally well. As a result of this operationalization, the predictive value of talent predictors is often investigated using statistical analyses based on mean differences between the selected and non-selected players (mostly through the use of t-tests or [multivariate] analysis of variance; see Figueiredo et al., 2009; Lago-Penas et al., 2014; le Gall et al., 2010). Although these statistical analyses can contribute to discovering relevant predictors for talent identification research to some extent, these designs cannot determine the value of different combinations of performance indicators in predicting an outcome variable indicative of individual soccer ability (Breitbach et al., 2014; Höner & Votteler, 2016; Wilson et al., 2017).

Secondly, determining indicators that predict individual soccer performance allows for successful selection of players on the basis of those variables. However, the use of a selection decision as the criterion can hinder this aim, because the judgment of a player's performance level might not be an accurate representation of individual soccer performance. This approach strongly depends on the validity of the coach's or scout's judgment in distinguishing between successful and 'non-successful' players. Yet, the validity of these judgements is not well established, and is often even biased (Den Hartigh, Niessen, et al., 2018). For example, judges are easily influenced by factors unrelated to a player's talent or performance, such as the player's skin color or reputation (Findlay & Ste-marie, 2004; Stone et al., 1997). In addition, the bias of judges to systematically select more mature players or players born earlier in the year has been well reported in the talent identification literature (Helsen et al., 2012; Musch & Hay, 1999).

Table 2.1 Design and methodological characteristics of soccer talent identification studies

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Reilly et al. (2000)	Cross-sectional	U17	16 15	Elite Sub-elite	*Low-fidelity:* Height, weight, body composition (*physical* - 7 variables) Speed, endurance, agility, strength (*physiological*- 10 variables) Dribbling and shooting (*soccer-specific* – 2 variables) Anxiety intension- and direction, anticipation, motivation (*psychological* – 11 variables)	Partially – authors briefly consider if findings will replicate in highly selected players, who are exposed to more systematic training.
Vaeyens et al. (2006)	Cross-sectional	U13 – U16	490[a]	Elite Sub-elite Non-elite	*Low-fidelity:* Height, weight (*physical* - 3 variables) Speed, endurance, agility, strength (*physiological*- 10 variables) Dribbling, shooting, passing, juggling (*soccer-specific* – 4 variables)	Yes – authors consider that differentiating ability of performance indicators might be dependent on competitive age class and relate findings to homogeneity of sample due to pre-selection.
Toering et al. (2009)	Cross-sectional	U12 – U18	159 285	Elite Non-elite	*Low fidelity:* Self-regulation (*psychological* – 6 variables)	No, but authors did control for effects of age.

2

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Coelho e Silva et al. (2010)	Cross-sectional	U14	69 45	Elite Local	*Low fidelity:* Maturity (3 variables) Height, weight, body composition (*physical* - 3 variables) Speed, endurance, agility, and power (*physiological* - 5 variables) Dribbling, shooting, passing (4 variables) Task- and ego orientation (*psychological* - 2 variables) *Other:* Soccer experience (1 variable)	No
Waldron and Worsfold (2010)	Cross-sectional	U14	69 32	Elite Sub-elite	*High fidelity:* Attempted, successful and unsuccessful skill involvements in a game, such as passing, shooting, tackling (18 variables)	No
Kavussanu et al. (2011)	Cross-sectional	U13 – U17	69 49	Elite Non-elite	*Low fidelity:* Task- and ego orientation, perceived parental environment (*psychological* – 11 variables)	No

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Waldron and Murphy (2013)	Cross-sectional	U15	15 16	Elite Sub-elite	*Low fidelity:* Speed, strength, agility (*physiological* – 5 variables) Dribbling (*soccer-specific* – 2 variables) *High fidelity:* Attempted, successful and unsuccessful skill involvements in a game, such as passing, shooting, tackling (6 variables) Physiological performance during games, such as intensity movements and distance covered (9 variables) *Other:* Heart rate and perceived exertion (2 variables)	No
Haugaasen et al. (2014)	Cross-sectional	U14 – U22	615 81	Non-professional Professional	*Other:* Engagement in soccer-specific activities (sociological – 4 variables)	Partially - authors specifically examine participation in soccer-specific activities in different age categories, but do not relate their findings to the homogeneity of the sample, due to preselection.

Table 2.1 (continued)

Study	**Prognostic period (follow up)**	**Age at assessment**	***N***	**Criterion**	**Predictors**	**Considers restriction of range**
Verburgh et al. (2014)	Cross-sectional	U9 – U17	84 42	Highly-talented Amateur	*Low fidelity:* Executive functions (*psychological* – 8 variables)	Partially - authors briefly state that findings can only be considered in context of samples, but authors do not examine differentiating ability of predictors per age category, and did not control for age.
Baláková et al. (2015)	Cross-sectional	U14	91[a]	Talented Less-talented	*Low fidelity:* Cognitive functions (*psychological* – 16 variables)	No
Goto et al. (2015)	Cross-sectional	U9 – U10	14 20	Retained Released	*Low fidelity:* Maturity (1 variable) *High fidelity:* Physiological performance during games, such as intensity movements and distance covered (6 variables)	No
Huijgen et al. (2015)	Cross-sectional	U14 – U18	47 41	Elite Sub-elite	*Low fidelity:* Lower and higher cognitive functions (*psychological – 6 variables)*	No

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Fenner et al. (2016)	Cross-sectional	U10	16	Rating of technical performance in SSGs[b]	*Low fidelity:* Speed, strength (physiological – 3 variables) *High fidelity:* Individual performance in SSGs, time-motion characteristics (5 variables)	Yes - authors compare findings to similar study with older players, and state findings did not replicate there due to homogeneity of technical skills in older, highly selected players.
Bennett et al. (2017)	Cross-sectional	U12 – U16	36 37	High-level Low-level	*High fidelity:* Attempted, successful and unsuccessful skill involvements in a game, such as passing, shooting, dribbling (13 variables)	No
Den Hartigh et al. (2017)	Cross-sectional	U11	49 39	Selected Non-selected	*Low fidelity:* Game reading based on video images (1 variable)	No
Rowat et al. (2017)	Cross-sectional	U18	27	Rating of technical performance in SSGs[b]	*Low fidelity:* Maturity (1 variable) Speed, endurance (*physiological* – 2 variables) Dribbling, passing, shooting (*soccer-specific* – 4 variables)	No

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Wilson et al. (2017)	Cross-sectional	NA	32	Individual performance in 1-vs-1 and 11-a-side games[b]	*Low fidelity:* Height, weight, body composition (*physical* – 7 variables, 2 latent variables) Speed, strength, balance (*physiological* – 7 variables, 3 latent variables) Dribbling, juggling, shooting, passing (*soccer-specific* – 5 variables, 2 latent variables)	No
Gil et al. (2007)	<1 year	U15 – U18	126 68	Selected Non-selected	*Low-fidelity:* Height, weight, body composition (*physical* - 22 variables) Speed, endurance, agility, power (*physiological*- 10 variables)	Partially – authors briefly consider that technical, tactical and psychological skills may have more discriminative power for selected players at later ages, when growth differences are less important.
Gravina et al. (2008)	< 1 year	U11 – U14	44 22	First team Reserves	*Low fidelity:* Height, weight, body composition (*physical* – 13 variables) Speed, strength (*physiological* – 10 variables)	Partially - authors very briefly relate findings to extended population, but do not discuss homogeneity of sample due to preselection.

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Huijgen et al. (2014)	< 1 year	U17 – U19	76 47	Selected Deselected	*Low fidelity:* Speed, endurance (*physiological* - 4 variables) Dribbling (*soccer-specific* - 4 variables) Tactical characteristic questionnaire (4 – variables) Task- and ego orientation, anxiety, concentration, motivation (*psychological* – 8 variables)	No, but authors did control for effects of age.
Lago-Penas et al. (2014)	< 1 year	U15/U17/U20	156[a]	Selected Non-selected	*Low fidelity:* Height, weight, body composition (*physical* – 6 variables) Speed, endurance, strength (physiological – 3 variables)	No
Zuber and Conzelmann (2014)	< 1 year	U13	140	Overall soccer performance rating[b]	*Low fidelity:* Achievement motive (*psychological* – 2 latent variables) Speed, endurance, strength, agility (*physiological* - 4 variables, 1 latent variable) Dribbling, juggling and ball control (*soccer-specific* – 3 variables, 1 latent variable)	Yes - authors relate findings to homogeneity of sample due to preselection.

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Aquino et al. (2017)	< 1 year	U17	28 38	Selected Non-selected	*Low fidelity:* Maturity (1 variable) Height, body composition (*physical* – 3 variables) Speed, endurance, strength (*physiological* – 7 variables) Shooting, ball control, dribbling, tactical skills questionnaire (*soccer-specific* – 4 variables)	No
Gil et al. (2014)	1 year	U10 – U11	21 43	Selected Non-selected	*Low fidelity:* Maturity (3 variable) Height, weight, body composition (*physical* – 9 variables) Speed, endurance, strength (*physiological* – 7 variables) *Other:* Soccer experience (1 variable)	No
Vestberg et al. (2012)	< 2 years	Adult	29 28 All	High division Low division Goals scored and assists[b]	*Low fidelity:* Executive functions (psychological – 3)	Yes - authors also have results for non-soccer players, and are therefore able to compare results to the general population.

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Vestberg et al. (2017)	< 2 years	U13 – U20	30	Goals scored and assists[b]	*Low fidelity:* Executive functions (*psychological* – 4 variables)	Yes - authors also have results for non-soccer players, and are therefore able to compare results to the general population.
Figueiredo et al. (2009)	2 years	U12 – U15	36 90 33	Drop-out Club Elite	*Low fidelity:* Height, weight, body composition (*physical* - 6 variables) Speed, endurance, agility, and power (*physiological*- 6 variables) Dribbling, shooting, passing (*soccer-specific* - 4 variables) Task- and ego orientation (*psychological* - 2 variables) *Other:* Soccer experience (1 variable) Rating of player's potential (1 – variable)	No

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Deprez et al. (2015)	2 years	U10 – U17	633 231 29 29 All	Club Dropout Contract No contract Total minutes played in first team[b]	*Low fidelity:* Maturity (2 variables) Height, weight, body composition (*physical* – 3 variables) Speed, power, endurance, motor coordination (*physiological* – 8 variables) Dribbling (*soccer specific* – 2 variables)	Yes - authors examine discriminatory power of variables per age group and discuss these results in relation to the homogeneity of each age group, in terms of physical abilities. Also briefly relate their findings to the extended, unselected population.
Zuber et al. (2015)	2 years	U13	10 82	National team Elite - Not selected	*Low fidelity:* Achievement motivation, achievement goal orientation, self-determination (psychological – 5 variables)	Yes - authors investigate distinct clusters formed of the different variables, for each age category. Also briefly consider homogeneity of the sample on examined variables.
Zuber et al. (2016)	3 years	U12	12 39 68	National Regional No talent card	*Low fidelity:* Maturity (1 variable) Net hope (*psychological* - 2 variables) Speed, endurance, strength (*physiological* – 3 variables) Dribbling, passing, juggling (*soccer-specific* – 3 variables)	Yes - authors investigate distinct clusters formed of the different variables, for each age category. Also note that results should only be considered in the context of their homogenous sample, and cannot directly be translated to the general population.

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Zibung et al. (2016)	3 years	U13	10 30 64	National Regional No talent card	*Low fidelity:* Speed, endurance, agility (*physiological* – 3 variables) Dribbling, passing, juggling (*soccer-specific* – 3 variables)	Yes - authors briefly discuss decrease of variance in performance over time, as a result of increasing homogeneity of sample due to preselection.
Huijgen et al. (2013)	1 – 3 years	U12 – U19	269 50	Selected De-selected	*Low fidelity:* Passing: Loughborough Soccer Passing Test (*soccer-specific* – 2 variables)	Partially - authors take the development of skills into account and relate results to different age categories, but only very briefly consider homogeneity of the sample, due to preselection.
Höner and Feichtinger (2016)	4 years	U12	308 2369	Youth Academy No youth academy	*Low fidelity:* Achievement motive, ego orientation, sport orientation, volition, self-concept, self-efficacy, anxiety (*psychological* – 17 variables)	Yes - authors relate their findings to the homogeneity of the sample due to preselection.
Kannekens et al. (2011)	3 – 5 years	U17 – U19	52 53	Professional Amateur	*Low fidelity:* Tactical skills questionnaire (soccer-specific – 4 variables) *Other:* Soccer experience, practice per week, non-specific sport practice	No

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Le Gall et al. (2010)	4 – 6 years	U14 – U16	48 167 235	International Professional Amateur	*Low fidelity:* Maturity (3 variables) Height, weight, body composition (*physical* - 3 variables) Speed, endurance, agility, and power (*physiological* - 14 variables)	Partially – authors examine discriminative power of performance characteristics per age group, but only very briefly consider how homogeneity of their sample, due to preselection, may affect findings.
Gonaus and Müller (2012)	1 – 6 years	U14 – U17	821 3912	Drafted Non-drafted	*Low fidelity:* Speed, endurance, strength, agility (*physiological* - 12 variables)	Yes - authors consider the homogeneity of the sample and relate discriminating power of variables to specific age group.
Höner and Votteler (2016)	4 – 7 years	U12	195 731 1025 20,892	National Regional Academy Not selected	*Low fidelity:* Sprinting, agility (*physiological* - 2 variables) Dribbling, ball control, shooting (*soccer-specific* – 3 variables)	Yes - authors mention restriction of range, relate findings to homogeneity of the sample due to preselection, and consider that discriminatory power may vary according to age group and homogeneity of sample.

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Höner et al.(2017)	8 – 10 years	U12	89 913 13,176	Professional Semi-prof. Non-prof.	*Low fidelity:* Relative age (1 variable) Height, weight (*physical* – 2 variables) Speed, agility (*physiological* – 2 variables) Dribbling, shooting, ball control (*soccer-specific* – 3 variables)	Partially - authors briefly consider how predictive value may differ for different age categories, but do not discuss homogeneity of their sample, due to preselection.
Van Yperen (2009)	15 years	U15 – U18	18 47	Successful Unsuccessful	*Low fidelity:* Goal commitment, coping, social support (*psychological* – 3 variables) *Other:* Assessment of initial performance by coaches (1 variable)	No, but the author did control for initial performance level.

2

Table 2.1 (continued)

Study	Prognostic period (follow up)	Age at assessment	*N*	Criterion	Predictors	Considers restriction of range
Martinez-Santos et al. (2016)	2 – 18 years	Adult	74 161	First/second division Semi professional	*Low fidelity:* Speed, strength (*physiological – 3 variables)*	No

U = Under, i.e., U18 means under the age of 18 years, *SSG* = Small Sided Game, *NA* = Not Available.

[a] The exact number of players per performance level could not be retrieved

[b] An individual soccer criterion measure, instead of performance- or skill level

Note: Electronic databases (MEDLINE, SPORTDiscus, Google Scholar) were searched between 2000 and 2018 for empirical studies on talent identification, using the following combination of terms: talent identification OR selection OR prediction and performance and soccer OR football. Additionally, snowballing was used to identify other relevant studies. Studies were included if they met the following criteria: (1) focused on soccer or Association Football (2) aimed to relate empirically multidimensional abilities and skills (e.g., physical, physiological, psychological, technical, tactical) or assessment methods to soccer performance or skill level, (3) were peer reviewed journal articles written in English. To restrict our sample, we excluded studies that focused predominantly on other types of football (e.g., futsal, American Football, Australian Rules football), and goalkeepers. Moreover, we excluded studies that mainly focused on the effects of relative age, maturity and genetic disposition. Although these topics are highly relevant for understanding talent development, we believe they warrant their own discussion and are, therefore, not within the scope of this paper. Finally, both cross-sectional and longitudinal studies were included. Although the empirical value of cross-sectional studies is limited compared to those with longitudinal designs, the methodological topics that are addressed in this paper apply to those studies as well.

Thus, it is not clear whether predictors of perceptions of successful performance are also valid predictors of individual in-game performance after selection (Baker et al., 2018).

There are only a few studies within the talent identification literature that used individual soccer performance as an outcome measure. Examples include structured ratings of in-game performance (Fenner et al., 2016; Rowat et al., 2017; Zuber & Conzelmann, 2014), and metrics based on successful and unsuccessful skill involvements during games (Pappalardo et al., 2019; Wilson et al., 2017). As we will discuss in section 3.1, we believe that the validity and reliability of such measures requires closer examination in future research. Taken together, we argue that the criterion measures that are currently used in most talent identification studies are intuitive and straightforward, but have their shortcomings and are insufficiently validated for studies that aim to identify and understand what factors predict individual soccer performance. In contrast, a reliable and objective soccer-specific criterion measure is complicated to operationalize, but allows for measurement of individual performance differences, so that the predictive value of different measures can be determined more meaningfully.

2.2.2 PREDICTORS OF SOCCER PERFORMANCE

The predictors that have been studied in soccer talent identification research are strongly influenced by the classification scheme proposed by Williams and Reilly (Reilly et al., 2000; Williams & Reilly, 2000), who classified predictors of individual soccer performance into four sport science disciplines: physical, physiological, psychological, and sociological. Examples of predictors include height, weight, and body composition (physical; e.g., Figueiredo et al., 2009; Gil et al., 2014; Vaeyens et al., 2006), speed, strength and endurance (physiological; e.g., Gonaus & Müller, 2012; Höner & Votteler, 2016; Huijgen et al., 2014; Martinez-Santos et al., 2016), self-regulation, motivation, task- and ego orientation, and cognitive functions (psychological; e.g., Baláková et al., 2015; Höner & Feichtinger, 2016; Huijgen et al., 2014; Reilly et al., 2000; Toering et al., 2009; Van Yperen, 2009; Verburgh et al., 2014; Vestberg et al., 2012, 2017; Zuber et al., 2015), and hours of practice and perceived social support (sociological; e.g., Haugaasen et al., 2014; Van Yperen, 2009). Other predictors that are derived from this classification scheme are technical skills, such as dribbling and passing technique, and self-assessed tactical skills (e.g., Coelho e Silva et al., 2010; Deprez, Fransen, et al., 2015; Höner et al., 2017; Huijgen et al., 2013; Kannekens et al., 2011; Le Moal et al., 2014; Reilly et al., 2000; see Table 2.1).

Given the multifaceted nature of soccer performance, it makes sense to investigate the extent to which these variables combined predict success and individual performance. Different studies have demonstrated that some of these skills and abilities are able to discriminate between players of varying performance levels (Johnston et al., 2018; Murr, Raabe, et al., 2018; Rees et al., 2016; Sarmento, Anguera, et al., 2018). More importantly, the major advantage of this approach in talent identification procedures is that skills and abilities, such as intermittent endurance capacity, dribbling technique, and passing ability, are relatively straightforward to measure in a standardized and reliable way (Ali, 2011; Mirkov et al., 2008; Visscher et al., 2006).

Although many studies have examined the predictive relevance of these variables in soccer, the reported effect sizes are generally small to moderate (Gonaus & Müller, 2012; Höner et al., 2017; Höner & Votteler, 2016; Murr, Raabe, et al., 2018). An explanation from selection psychology for the limited predictive validities in soccer talent identification research may be related to the 'fidelity' of the predictors, that is, the extent to which the performance task mimics the criterion behavior in content and context. On one side of the fidelity continuum are low fidelity predictors, which have relatively little overlap with the criterion in terms of the behavior the player should show and the context in which the player must perform (Callinan & Robertson, 2000; Lyons et al., 2011). These low fidelity predictors measure distinct, general performance components that are thought to be related to the criterion behavior. Such low fidelity predictors are referred to as 'signs' in the selection psychology literature (Wernimont & Campbell, 1968). Thus, most of the predictors classified by Williams and Reilly (Williams & Reilly, 2000), such as height, speed, and motivation, can be characterized as signs, because they measure distinct components and lack fidelity to the criterion of soccer performance in terms of the task and/or the context in which they are assessed (Lyons et al., 2011).

The selection psychology literature shows that the predictive validity of assessment procedures often improves when the degree of fidelity increases, that is, when the predictor becomes more similar to the criterion in terms of behavior, task, and contextual constraints (Den Hartigh, Niessen, et al., 2018; Lievens & De Soete, 2012; Vaeyens et al., 2008). The underlying rationale is the notion of behavioral consistency: 'the best predictor of future behavior is similar past or current behavior' (Meehl, 1989; Ouellette & Wood, 1998; Van der Flier, 1992; Wernimont & Campbell, 1968). Tests that assess soccer-specific technical skills, such as dribbling and passing technique, possess higher fidelity to the criterion of soccer performance than variables such as height, speed, and motivation. Accordingly, there is evidence that these predictors have better prognostic relevance (Höner et al., 2017; Huijgen et al., 2013), and discriminate more consistently between skill groups than the latter group of variables (Höner et al., 2017; Murr, Feichtinger, et al., 2018; Wilson et al., 2017). Still, these tests measure distinct skills, and do not incorporate many of the necessary contextual constraints of in-game soccer performance, such as the task of scoring goals and the presence of moving opponents. In other words, such tests may still not mimic the criterion of interest, which is in-game soccer performance, to a large enough extent (Phillips et al., 2010). For example, the Loughborough Soccer Passing Test, a test frequently used to assess the passing ability of soccer players (Ali, 2011; Huijgen et al., 2013), was recently found to be a poor predictor of in-game passing performance (Serpiello et al., 2017).

An important avenue, therefore, is to develop predictors that minimize the ‘inferential leap’ from the predictor to the criterion further, and thus possess even higher fidelity. One approach to establish such predictors in soccer is to take a ‘sample’ of the criterion performance in a highly representative context (Callinan & Robertson, 2000; Lyons et al., 2011), for example, in small-sided games (SSGs). SSGs are games played on reduced pitch areas and with fewer players (e.g., 4 vs. 4, or 7 vs. 7) than in an official game. Individual performance in SSGs can be considered a sample-based predictor, because it is obtained based on behavior, task, and contextual constraints similar to those present in the criterion performance.

An important conclusion from the selection psychology literature is that sample-based assessments can be very good predictors of future performance (Hunter & Hunter, 1984; Niessen et al., 2016; Roth et al., 2005; Schmidt & Hunter, 1998), especially in homogeneous samples and for multidimensional outcome measures (Sackett et al., 2017). Because soccer talent identification research is often based on homogenous samples (e.g., players who are already in a talent program), and soccer performance is multidimensional (Williams & Reilly, 2000), a samples approach to prediction is expected to result in greater predictive value (Den Hartigh, Niessen, et al., 2018). Accordingly, several recent studies have related performance or skill level to predictors that we would characterize as sample-based, such as attempted and completed actions (i.e., event data) within SSGs or regular games (Bennett et al., 2018; Waldron & Murphy, 2013; Waldron & Worsfold, 2010). These sample-based predictors were relatively successful in distinguishing between groups of elite and sub-elite or non-elite players, and these results demonstrate how high-fidelity methods may be useful as alternatives to isolated components in predicting soccer performance (Bennett et al., 2018; Waldron & Murphy, 2013; Waldron & Worsfold, 2010). However, similar to individual soccer performance criterion measures, the reliability of individual performance assessed through SSGs needs to be addressed in future studies (see section 3.2).

Finally, the suggestion of samples as predictors of performance is also directly in accordance with theoretical developments in the field of motor learning and talent development regarding the use of representative designs for learning and assessment purposes (Davids, Araújo, Correia, et al., 2013; Davids, Araújo, Vilar, et al., 2013; Den Hartigh, Niessen, et al., 2018; Pinder et al., 2011). Several authors have already suggested that talent identification procedures should include more representative measures (Breitbach et al., 2014; Johnston et al., 2018; Unnithan et al., 2012; Vaeyens et al., 2008). In using samples as predictors of soccer performance, the interaction between different performance components is embedded in behavior that is representative of the criterion performance, thereby closing the gap between predictor and criterion.

In conclusion, soccer talent identification research has generally focused on low- or moderate fidelity, sign-based predictors of soccer performance, which has resulted in some interesting findings, but also in an inconsistent body of evidence that does not provide clear guidelines for stakeholders in practice. The selection psychology literature suggests that high-fidelity, sample-based measures may enhance the predictive value of talent identification procedures, but such methods are not often applied in the soccer talent identification literature yet.

2.2.3 RESTRICTION OF RANGE

Talent identification studies often compare samples that are already highly restricted in terms of talent or skill, such as elite against sub-elite athletes. In such cases, empirical relationships between performance indicators used as predictors and the criterion performance often deviate from relationships in the population (Ackerman, 2014). This is a problem when, due to selection, a relatively homogenous sample that is not representative of the population of interest (containing all candidates, selected and not selected) is used to establish predictor-criterion relations (Baker et al., 2018). As a result, predictor-criterion relationships obtained from such samples are usually underestimated because of ‘restriction in range’ (Sackett & Yang, 2000).

To illustrate the effect of range restriction, we consider the study by le Gall et al. (2010). The authors examined anthropometric and physical characteristics of highly trained U14 - U16 soccer players in a national academy, who, upon leaving the academy, achieved either international or professional status, or remained amateurs. They investigated the mean differences for 17 dependent variables, ranging from height, weight, and maturity measurements, to sprint- and endurance performance and lower body explosiveness. Although statistically significant mean differences were found for some variables, there were no large differences between the groups on most performance indicators within age categories. For instance, in the U16 category, maximal anaerobic power and height distinguished between future internationals and amateurs with moderate effect sizes, but there was no strong evidence for vertical jump, 10-, 20-, 30-, and 40-meter sprint, and lower body explosiveness distinguishing between any combination of international, professional, and amateur players.

Based on these findings, the conclusion may be that these variables are not very useful for differentiating future career success in elite-level U16 players. However, it would be false to conclude that these characteristics are not important for attaining soccer-specific success in general (Ackerman, 2014). It is likely that the sample of academy players were exposed to the same training routine, had similar practice histories, and were (directly or indirectly) pre-selected on at least some of the variables in this study. This preselection in a homogenous group of athletes in terms of physical performance results in a reduction in variance in the predictors and in the criterion. If the same predictors were studied in a more heterogeneous group of soccer players, larger effect sizes would likely have been found for at least some of these predictors (e.g., Franks et al., 1999; Williams & Reilly, 2000).

Although the issue described above sounds straightforward, the effects of range restriction are often not explicitly taken into account in talent identification research. Range restriction is generally an issue when the aim of a study is to generalize results obtained from a specific selected group of elite players to a more general group, which is often the case when we study relationships between a performance criterion and predictors. Aside from general issues such as insufficient power, careful consideration of the homogeneity of the participant group, in terms of the predictors the study examines, is also required to accurately interpret why certain relationships were or were not found. This is important because the ability of predictors to differentiate between players also depends on the degree of restriction in the sample. For example, some evidence suggests that a physiological sign-based predictor such as sprinting ability is more suitable for differentiating between performance levels for relatively younger (e.g., U14 – U16) than for older (e.g., U17 – U19) skilled players (Deprez, Fransen, et al., 2015; Gil et al., 2007; Vaeyens et al., 2006), probably because the former group is more physically diverse, less exposed to systematic training, and not as strongly pre-selected on this variable. Some talent identification researchers relate their findings to the homogeneity of the sample and acknowledge that the discriminating or predictive value likely changes with the competitive level (Deprez, Fransen, et al., 2015; Gonaus & Müller, 2012; Vaeyens et al., 2006). However, findings so far have been too inconsistent across studies to accurately determine what is important for any specific age group or skill level.

Thus, restriction of range is common in talent identification research, but is rarely considered explicitly when the generalizability of predictive validities is discussed (see Table 2.1).

2.2.4 THE BASE RATE AND THE UTILITY OF TALENT IDENTIFICATION PROGRAMS

Successful talent identification procedures strive to select individuals who will attain excellent performance and reject individuals who will not (Breitbach et al., 2014). The focus of talent identification research is on the predictive value of different performance indicators. However, the practical usefulness or utility of these predictors, in terms of correctly identified players, is often not considered when evaluating the effectiveness of talent identification programs (Ackerman, 2014; Güllich & Cobley, 2017).

The utility of selection procedures is greatly affected by contextual factors, especially the base rate and the selection ratio. The base rate is the proportion of individuals in the population of interest who are able to reach satisfactory criterion performance, that is, the proportion of individuals performing successfully if there is no selection (Taylor & Russell, 1939). Thus, the base rate is the prior probability of success for any given candidate (Meehl & Rosen, 1955). Naturally, the base rate depends on the population of interest (i.e., the candidate pool) and on the criterion of interest. For example, several prospective cohort studies aimed to predict elite adult or late adolescent soccer success on the basis of performance indicators in groups of early adolescent players who were selected from large populations (Höner et al., 2017; Höner & Votteler, 2016). This context is characterized by a very low base rate, because very few young players have the ability to attain the elite adult level (Güllich, 2014). The base rate is higher when we consider, for example, strongly pre-selected older players in an elite youth academy, and when our criterion is operationalized as progressing to next year's age class in the academy (Aquino et al., 2017; Gil et al., 2007; Huijgen et al., 2014).

The selection ratio is defined as the proportion of players in the population of interest that is selected (Taylor & Russell, 1939). The selection ratio and the base rate are easily confounded in the soccer talent identification literature, because the selection decision is often used as the criterion measure in this research field, as discussed in Section 2.1. Yet, they are essentially different, and need to be defined separately in order to estimate the utility of a predictor.

The base rate, the selection ratio, and an unrestricted correlation coefficient between the predictor and the criterion can be used in utility models to estimate the gain in criterion performance as a result of using a particular predictor (Ackerman, 2014; Niessen & Meijer, 2017). There are several utility models, mostly developed in the context of personnel selection (Ashton et al., 1968; Lawshe et al., 1958; Taylor & Russell, 1939). As an example, we provide a description of the simplest model, the Taylor-Russell model (Taylor & Russell, 1939).

In the Taylor-Russell model, a continuous criterion variable is dichotomized into a 'successful' and an 'unsuccessful' group, based on a certain cutoff value used to define successful performance. Subsequently, utility is defined as the proportional increase in successful soccer players among those who are selected (the success ratio), resulting from using a specific selection procedure, compared to having no selection procedure (the base rate), or compared to the success ratio that would result from using a different selection procedure. In selection decisions, four groups can thus be distinguished: selected athletes who are successful (true positives), selected athletes who are unsuccessful (false positives), unselected athletes who would have been successful (false negatives), and unselected athletes who would not have been successful (true negatives). Accordingly, the proportion of true positives among all selected candidates corresponds to the sensitivity of a selection procedure, whereas the proportion of true negatives among all unselected candidates corresponds to the specificity. These terms are often used in medical research. Figure 2.1 visually represents these areas. In general, procedures with a high predictive validity, applied in contexts with a low selection ratio and a base rate that yields balanced groups of 'suitable' and 'unsuitable' players (around .50), yield the highest utilities. In addition, even when an assessment procedure has high predictive validity, utility will be relatively low when the selection ratio is high, and/or when the base rate is either very high or very low (Meehl & Rosen, 1955; Taylor & Russell, 1939).

Consider the following example. Assume that around 5000 U12 competence center players are selected annually from a total of 100,000 amateur club players (e.g., Höner and Votteler, 2016), resulting in a selection ratio of 5%. Furthermore, they are selected based on a procedure that shows an unrestricted correlation of $r = .4$ with elite adult soccer performance. Note that $r = .4$ suggests relatively high predictive validity, especially considering the complexity in predicting a performance outcome of young players several years in the future from the time of testing (Ackerman, 2014). In addition, only 1% of the population of U12 players (i.e., 1000 players) has the ability to obtain excellent elite adult soccer performance (the base rate).

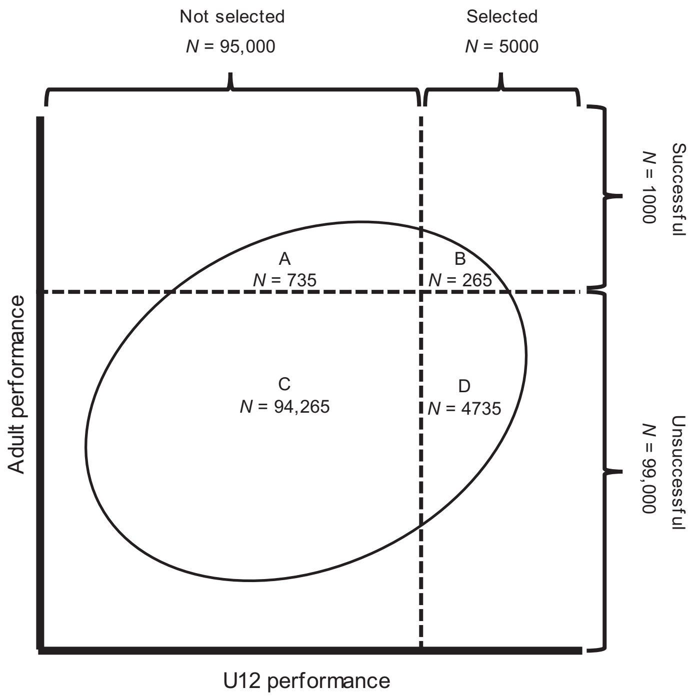


Figure 2.1 Visual representation of the example regarding the selection procedure of talented U12 players (N = 100,000). Adapted from Taylor & Russell (1939), with permission. Note: A = Wrongfully rejected (false negatives), B = Rightfully accepted C = Rightfully rejected, D = Wrongfully accepted (false positives). B / (B + D) = sensitivity, whereas C / (C + A) = specificity.

With this information, the success ratio resulting from the talent identification procedure can be computed (for example, by using an online Theoretical Expectancy Calculator; McLellan, 1996).

The results based on this example are shown in Figure 2.1. We obtain a success ratio of 5.3%, which means that only 5.3% (265 / 5000) of the selected players will be successful in achieving elite adult soccer performance. This may seem like a modest result. However, compared to the base rate of 1%, this may be a substantial increase. Moreover, 73.5% (735 / 1000) of all 'suitable' players among the population of U12 players are not selected. Conversely, of the 99,000 players who do not have the ability to be successful approximately 95% (94,265 / 99,000) are not selected.

This example demonstrates how the base rate and the selection ratio can influence expectations regarding the utility of talent identification procedures for performance predictions (Güllich & Cobley, 2017).

To date, the talent identification literature has not generally taken this into account. We were able to identify one study within the talent identification literature that considered utility (Höner & Votteler, 2016), whereas the effect of the base rate on the usefulness of the examined predictors was not discussed in the other studies in Table 2.1.

2.3 DISCUSSION AND SUGGESTIONS FOR FUTURE RESEARCH

The aim of this position paper was to evaluate the methodology in the soccer talent identification literature based on common principles from selection psychology that are relevant for talent identification research. We are aware that talent identification, in particular at younger ages, is very difficult (Güllich & Cobley, 2017; Güllich & Emrich, 2014). Yet, we also believe that selection in general can provide players with realistic opportunities for successful development, and is often necessary from a practical point of view (Larkin & Reeves, 2018). An important challenge, therefore, is to develop best-practice selection methods with clearly established predictive validity and reliability. The realization of a coherent body of knowledge regarding the prediction of soccer performance should ultimately provide guidelines for stakeholders and practitioners in talent identification. Considering the four topics discussed in this paper, we suggest that future talent identification studies in soccer consider the following points in order to help advance research practices and increase their practical and scientific impact.

2.3.1 DEVELOP CRITERION MEASURES OF INDIVIDUAL SOCCER PERFORMANCE

First, we suggest that future studies pay more attention to the criterion variables used in talent identification research and develop individual soccer performance measures. More specifically, future studies may develop criterion measures that are not essentially selection decisions, and that can describe individual differences within selected groups of players to investigate what performance indicators are related to which kind of soccer performance.

It should be emphasized that the development of such methods is a complicated task, because of the dynamic nature of soccer. Elite individual soccer performance emerges through the complex interactions between the person and environmental constraints (Davids, Araújo, Vilar, et al., 2013; Phillips et al., 2010). As of yet, there is simply no single, objective measure of soccer-performance available that can capture these complex interactions. Individual performance is dependent on the abilities of both teammates and opponents, which makes valid and reliable measurements very challenging (Ackerman & Beier, 2006). The comparison of individuals' soccer performance is complicated even further when we consider that different positions require different tasks and skills (Baker et al., 2015).

Despite the challenges, we believe that efforts to devise meaningful criterion measures are necessary to establish clear predictor – criterion relationships. The literature is limited in providing measures that can describe individual performance differences, keep the person-task-environment relation intact, and account for the complex interactions between teammates and opponents (Travassos et al., 2013). Yet, there are several ways to obtain individual soccer performance measures that may provide a useful step in the right direction. For example, notation data on the frequency and quality of in-game events (e.g., Waldron and Worsfold, 2010; van Maarseveen et al., 2017) may be weighted and combined to assess performance per position. The weights of the events that are relevant for different positions can be determined by experts, such as coaches or scouts, or through machine learning approaches when large amounts of data are available (Pappalardo et al., 2019). Furthermore, positional data (e.g., Frencken et al., 2011; Memmert et al., 2017) may be used to quantify spatial-temporal patterns of play, which may be related to individual in-game success. Both these tools can be used to construct composite measures of 'general' soccer performance (Pappalardo et al., 2019), or to measure a specific aspect of performance, such as passing (Goes et al., 2018), when the emphasis is on assessing the tasks of a specific player position (Lyons et al., 2011). Finally, simpler measures such as structured expert ratings are efficient tools for quantitatively evaluating individual performance (Musculus & Lobinger, 2018), but it should be kept in mind that these also introduce more subjectivity, which can lead to biases and low inter-rater reliability (Newman et al., 2004). Most importantly, studies are warranted that evaluate the validity and reliability of criterion measures, before they are implemented in predictive talent identification research.

2.3.2 CLOSE THE GAP BETWEEN PREDICTOR AND CRITERION VARIABLES

Secondly, we suggest that future studies explore the use of predictors that are more in line with the criterion. Specifically, talent identification research may broaden its current focus on low-fidelity signs as predictors to include high-fidelity samples as predictors of performance. With respect to the notion of behavioral consistency, several recent studies have demonstrated that prior competitive success in different sports is a relatively good predictor of short-term (i.e., 1-2 years) success (Barreiros et al., 2014; Güllich & Emrich, 2014; Kearney & Hayes, 2018; Li et al., 2018). However, studies on soccer generally based individual performance on the highest (inter)national level of competition reached, which is less relevant for talent identification procedures, and also suffers from limitations regarding the categorization of players. Therefore, it will be interesting to see whether samples of past soccer performance as predictors yield higher predictive validities of future individual soccer performance, compared to signs.

In-game event data, positional data, and structured ratings can also be used to develop predictors by quantifying performance in sample-based assessment procedures, such as SSGs or 11-a-side games. It is important to note, however, that similar to using an individual soccer criterion measure, measurements based on sample-based predictors may pose challenges related to the complex nature of soccer performance, including the dependence of individual performance on teammates and opponents, comparing different positions and competitions, and biases related to judgment. The reliability of such measurements needs to be investigated in future studies to develop optimally valid measures. Accordingly, recent efforts have been made to develop reliable structured rating forms to measure performance in SSGs (Cobb et al., 2018; Van Maarseveen et al., 2017). As mentioned by other researchers (Breitbach et al., 2014; Leyhr et al., 2018; Vaeyens et al., 2008; Williams & Reilly, 2000), performance should preferably be assessed longitudinally over a series of games, in order to obtain reliable assessments of individual soccer performance based on these samples. In addition, when a researcher aims to investigate in-game performance for a given group of players, and has control over the organization of the games, the performance level of opponents and teammates can be controlled for by reorganizing players into different teams after each (small-sided) game, as was done by Fenner et al., 2016.

2.3.3 CONSIDER RESTRICTION OF RANGE

Thirdly, future studies should take into account the potential effect of range restriction on their conclusions by carefully considering the homogeneity of their study participants in terms of physical, physiological, and other soccer-related characteristics. Subsequently, researchers should clearly state the population to which findings may be generalized. In strongly restricted samples, the absence of observed predictor-criterion relationships does not necessarily imply that a predictor is not positively related to attaining elite performance in the general population, or to the initial performance level prior to the selection decision. In addition, which predictors are useful for differentiating between players probably depends on the level of expertise, and hence, the degree of pre-selection, in the population of interest. Future research could pay close attention to which predictors work in which specific populations.

It should be noted that correcting for the effects of range restriction has been challenging in talent identification research. Range restriction is an issue that occurs in most selection contexts, including personnel- and educational selection. In a typical selection study, the entire candidate pool would be assessed on the predictor variables, but criterion performance data are only available for the candidates who were selected. The resulting underestimated predictor-criterion relationship can be corrected using several available formulas (Sackett & Yang, 2000; Schmidt et al., 2006), which yield estimates of the predictor-criterion relationship in the unrestricted population of interest (Sackett & Yang, 2000; Schmidt & Hunter, 2014). These corrections are often applied in the selection psychology literature (American Educational Research Association et al., 2014). However, they have not been used in a talent identification context, which is most likely due to the design of most talent identification studies; because performance level or a selection decision functions as the criterion, range restriction does not occur within the sample(s) under study. Accordingly, when the design of future studies includes soccer criterion measures that can differentiate between individual players' performance after selection, range restricted relationships can be accounted and corrected for using correction formulas that take the variance in the candidate pool into account (Sackett & Yang, 2000; Schmidt & Hunter, 2014).

2.3.4 IDENTIFY THE UTILITY OF PREDICTORS

Finally, we suggest that future studies discuss the potential utility of predictors more often and consider realistic estimates of contextual factors such as the base rate and the selection ratio. For instance, future studies may investigate how novel predictors compare to current selection decisions made by coaches and scouts, in terms of incremental validity and utility. We acknowledge that it is difficult to obtain estimates of the base rate based on empirical data. However, an educated guess about a range of plausible values of the base rate (Niessen & Meijer, 2016) can be obtained based on interactions with experts, such as by asking several coaches or scouts to estimate the proportion of players who they think have the potential to obtain excellence. That range of plausible values can be used in utility models. Since this base rate is generally very low in talent identification contexts (Ackerman, 2014; Höner & Votteler, 2016), and arguably often lower than the selection ratio, not all selected players can become successful, regardless of the predictor's validity. Therefore, we believe that utility estimates will help to create realistic expectations for researchers and stakeholders about talent identification procedures.

2.4 CONCLUSION

In the current position paper, we discussed several methodological issues common in the soccer talent identification literature, and provided suggestions to improve the methodological quality and robustness of research practices in future talent identification studies. We hope that the general principles discussed here will also transfer to practical selection contexts, and we believe that researchers have an important responsibility to communicate the reliability and validity of talent identification procedures to the sports field (Drenth, 2008). Thinking critically about the methodology and design of studies in sports opens the door for innovative research that advances this exciting field, and hopefully leads to a more coherent scientific and practical framework for talent identification.


00:13:12

CHAPTER 3

THE VALIDITY OF SMALL-SIDED GAMES IN PREDICTING 11-VS-11 SOCCER GAME PERFORMANCE

This chapter was published as:

Bergkamp, T. L. G., den Hartigh, R. J. R., Frencken, W. G. P., Niessen, A. S. M., & Meijer, R. R. (2020). The validity of small-sided games in predicting 11-vs-11 soccer game performance. *PLOS ONE*, *15*(9), e0239448. https://doi.org/10.1371/journal.pone.0239448

ABSTRACT

Predicting performance in soccer games has been a major focus within talent identification and development. Past research has mainly used performance levels, such as elite vs. non-elite players, as the performance to predict (i.e., the criterion). Moreover, these studies have mainly focused on isolated performance attributes as predictors of soccer performance levels. However, there has been an increasing interest in finer grained criterion measures of soccer performance, as well as representative assessments at the level of performance predictors. In this study, we first determined the degree to which 7-vs-7 small-sided games can be considered as representative of 11-vs-11 games. Second, we assessed the validity of individual players' small-sided game performance in predicting their 11-vs-11 game performance on a continuous scale. Moreover, we explored the predictive validity for 11-vs-11 game performance of several physiological and motor tests in isolation. Sixty-three elite youth players of a professional soccer academy participated in 11 to 17 small-sided games and six 11-vs-11 soccer games. In-game performance indicators were assessed through notational analysis and combined into an overall offensive and defensive performance measure, based on their relationship with game success. Physiological and motor abilities were assessed using a sprint, endurance, and agility test. Results showed that the small-sided games were faster paced, but representative of 11-vs-11 games, with the exception of aerial duels. Furthermore, individual small-sided game performance yielded moderate predictive validities with 11-vs-11 game performance. In contrast, the physiological and motor tests yielded small to trivial relations with game performance. Altogether, this study provides novel insights into the application of representative soccer assessments and the use of continuous criterion measures of soccer performance.

3.1 INTRODUCTION

Professional soccer organizations strive to identify, select, and develop players who have the potential to become elite soccer players. In order to establish evidence-based selection procedures, talent selection and identification studies often aim to determine the extent to which distinct skills and abilities are related to future performance (Larkin & Reeves, 2018; Vaeyens et al., 2008). This has led to a plethora of studies examining the predictive value of many different kinds of attributes across different performance categories, such as height and weight (i.e., anthropometric attributes), sprint speed, endurance capacity, and agility (i.e., physiological and motor skills), dribbling and passing skills (i.e., technical skills), and motivation and self-regulation (i.e., personality-related or psychological; Johnston et al., 2018; Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018; Sarmento, Anguera, et al., 2018). These attributes are typically assessed in laboratory settings or field tests, and in isolation of in-game soccer constraints (Breitbach et al., 2014). Moreover, the value of these attributes as indicators of 'talent' is assessed by examining how well they discriminate between players with different (future) performance levels (e.g., elite versus non-elite players), or between selected and deselected academy players (Bergkamp et al., 2019). As discussed below, the way the predictors and criterion-performance have been defined in previous studies has limitations. Consequently, there has been an increasing interest in finer grained criterion measures of soccer performance, and more ecologically valid assessments at the level of performance predictors (Bergkamp et al., 2018, 2019; Den Hartigh, Niessen, et al., 2018; Johnston et al., 2018; Unnithan et al., 2012; Vaeyens et al., 2008).

3

3.1.1 SOCCER PERFORMANCE CRITERION

Using performance levels as the criterion (i.e., the outcome variable and performance to predict) is understandable from a practical standpoint, but has a few disadvantages (Bergkamp et al., 2019). First, a disadvantage of this approach is that there are often inconsistencies in the definition of performance levels, which may impede comparisons across studies. For example, definitions of elite athletes have ranged from international to regional level competitors, and strongly depend on the competitiveness of the sport in the athlete's country (Swann et al., 2015). Second, since talent research ultimately aims to identify players who have the potential to excel in soccer games (Unnithan et al., 2012), it can be argued that the environments of interest are competitive 11-vs-11 games. It follows that the relevant criterion is, ideally, individual performance within these games (Bergkamp et al., 2019; Unnithan et al., 2012). However, while coaches or scouts – responsible for grouping players into performance levels – arguably decide what talented in-game performance looks like, the validity of these judgments is not well established, and is often even biased (Den Hartigh, Niessen, et al., 2018; Meylan et al., 2010; Wiseman et al., 2014). For instance, judges (e.g., coaches) are easily influenced by factors unrelated to performance, such as the athlete's appearance or reputation (Findlay & Ste-marie, 2004; Stone et al., 1997). The bias of coaches to select more mature players, or players born earlier in the calendar year, has also been well established in soccer (Helsen et al., 2012). Finally, and importantly, dichotomizing the criterion into performance levels provides no information on the differences between individuals within the same level on an in-game soccer performance outcome (Bergkamp et al., 2019; Wilson et al., 2017). Therefore, talent identification researchers are facing the question whether they can define in-game soccer performance criteria that are not based on grouping performance levels, and that are able to distinguish between individual players on a continuous scale (Bergkamp et al., 2019; Phillips et al., 2010; Piggott et al., 2019; Unnithan et al., 2012).

There are multiple ways to quantify different aspects of individual in-game soccer performance. Global and local positioning systems may be used to quantify physiological in-game performance characteristics, such as high intensity meters, total distance run, and accelerations (Vieira et al., 2019). By extracting spatio-temporal information of the players on the pitch, these systems may also be used to assess tactical performance indicators, such as the space created with a pass (Memmert et al., 2017). A more straightforward technique that does not demand advanced technologies is notational analysis. This technique lends itself particularly well to assess on-ball technical and tactical performance indicators, by manually coding observed events (Hughes & Bartlett, 2002; Van Maarseveen et al., 2017). Recent work suggests that performance indicators derived through this technique, such as passes, duels, and shots, are related to game success (i.e., winning; Pappalardo & Cintia, 2017). This opens promising opportunities for operationalizing soccer performance at the criterion level, as well as assessing performance at the predictor level.

3.1.2 ASSESSMENTS IN SOCCER

The attributes assessed in the talent identification literature resulted in various levels of success in discriminating between performance levels (Höner & Votteler, 2016; Murr, Feichtinger, et al., 2018). For example, a recent systematic review evaluated the discriminatory value of different physical and physiological attributes (Murr, Raabe, et al., 2018). The authors found median effect sizes across studies of d = .37 for sprint speed (< 20m), d = .41 for endurance capacity, and d = .42 for change of direction, which can be considered low (Cohen, 1988). In contrast, repeated sprinting ability and sprint speed (> 20m) had effect sizes of d = 1.21 and d = .57, which can be considered as strong and medium, respectively.

Nevertheless, it has recently been argued that assessments that are representative of competitive 11-vs-11 games may result in better performance predictions compared to abilities that are tested in isolation (Bergkamp et al., 2019; Breitbach et al., 2014; Burgess & Naughton, 2010; Den Hartigh, Niessen, et al., 2018; Pinder et al., 2011, 2013). Representative assessment is described as a design that maintains, or 'samples', the personal, environmental, and task constraints of the performance environment of interest (Pinder et al., 2011, 2013). When the criterion is operationalized as performance in 11-vs-11 games, a representative context incorporates environmental constraints in these games, such as the presence of moving opponents and the task to score goals. At the same time, it simulates soccer-specific motor, physiological, technical, tactical, and perceptual-cognitive in-game performance behaviors for the player (Araújo et al., 2007; Bergkamp et al., 2019; Den Hartigh, Niessen, et al., 2018). Thereby, representative assessments do justice to the idea that the mechanism underlying elite soccer performance is characterized by how the player acts upon, and interacts with environmental constraints (Den Hartigh, Niessen, et al., 2018).

3

By simulating 11-vs-11 games, a representative assessment also builds on the notion of behavioral consistency. That is, the assumption that the best predictor of future behavior is similar behavior in the past (Meehl, 1989; Wernimont & Campbell, 1968). Predictors that are similar to the criterion in content and context are said to be high in fidelity. Accordingly, research in sports has repeatedly demonstrated that predictive validity increases when the fidelity of the predictor increases (Callinan & Robertson, 2000; Lievens & De Soete, 2012; Lyons et al., 2011). Tests that measure attributes that are less similar to the criterion behavior (i.e., 11-vs-11 game performance) may be considered as lower-fidelity attributes, and are described as 'signs' (Lievens & De Soete, 2012; Pinder et al., 2011; Stoffregen et al., 2003). From this point of view, representative assessments would provide higher-fidelity predictors than sign-based tests measuring motor, physiological, technical, tactical, and perceptual-cognitive attributes in isolation.

An example of representative, 'sample-based' assessments in soccer are small-sided games (SSGs; Davids, Araújo, Correia, et al., 2013; Den Hartigh, Niessen, et al., 2018; Unnithan et al., 2012). SSGs are games played with fewer players and on a smaller pitch size compared to 11-vs-11 games. However, the degree of representativeness may be dependent on variations in the specific number of players and pitch size (Sarmento, Clemente, et al., 2018). It is, therefore, important to evaluate the degree to which SSGs are representative of 11-vs-11 game. To the best of our knowledge, one study has been conducted in this direction. Results from Olthof et al. (2019) suggest that the tactical demands of SSGs for Under-13 year old (U13), U15, U17, and U19 players reflect those of 11-vs-11 games, when teams consist of 6 or 8 players and when a match derived relative pitch area of 320 m^2 per player is used.

Interestingly, the few studies that have explored the concurrent or predictive validity of individual SSG performance mainly included smaller SSGs. Fenner et al. (Fenner et al., 2016) and Unnithan et al. (Unnithan et al., 2012) showed that 4-vs-4 SSG performance for U10 and U16 players, based on matches won and goals scored, had a moderate-to-strong relationship with technical skills, as determined by a scouting tool ($r = .76$ and $r = .39$, respectively). Moreover, Bennett et al. (Bennett et al., 2018) demonstrated that on-ball skill proficiencies, such as dribbles, passes, touches, and shots, discriminated significantly between high and low-level soccer players in 4-vs-4 SSGs. While these studies provide important first clues on how individual SSG performance may be utilized for performance assessment, an exploration of performance in larger SSGs as predictors of performance in 11-vs-11 games has not been conducted yet. Furthermore, the previous studies correlated overall SSG performance with subjective scout ratings or performance levels (Bennett et al., 2018; Fenner et al., 2016; Unnithan et al., 2012), whereas more objective in-game indicators may better serve as a criterion measure.

3.1.3 THE CURRENT STUDY

The current study expands the previous literature by quantifying in-game soccer performance on a continuous scale. By doing so, we first examined the degree to which performance indicators in large-scaled, 7-vs-7 SSGs can be considered representative of performance indicators in competitive 11-vs-11 games. The concept of representative assessment suggests that predictive validity is driven by using predictors that are highly representative for the criterion. Therefore, the representativeness of SSGs for 11-vs-11 games can be considered a prerequisite for their predictive validity. Second, we explored the value of the SSGs as a high-fidelity sample-based predictor, by assessing the validity of individual players' in-game SSG performance in predicting their 11-vs-11 game performance. In addition to our two primary aims, we explored the validity of physiological and motor attributes that are frequently used in the talent literature and by soccer teams in monitoring and predicting performance, namely sprint, agility, and endurance capacity tests (Altmann et al., 2019; Sporis et al., 2010). Because these tests may be considered as low-fidelity signs in relation to individual performance in soccer games, relatively low correlations with the criterion could be expected.

3.2 MATERIALS AND METHODS

3.2.1 PARTICIPANTS

Elite youth players from the U15, U17, U19, and U23 teams of a professional soccer academy in the Netherlands were recruited to participate in the study. Recruitment started two months before the start of the 2018-2019 competitive soccer season, and was conducted after approval from the youth players, the coaches, the academy's technical director and the club's head of performance. All players belonging to the U15 to U23 teams were eligible to participate in the study, resulting in n = 87 who participated in at least one SSG over the course of the season. However, we excluded players who did not play any minutes in the 11-vs-11 games or played in few SSGs (i.e., more than 2 standard deviations below the average number of SSGs played per team; see Table 3.1), due to injury, dropping out of the academy, or other circumstances. This resulted in a total of n = 63 players from the U15 (n = 17), U17 (n = 15), U19 (n = 16), and U23 (n = 15) teams who were included in the analyses.

3

Table 3.1 Descriptives (mean, *SD* in brackets) for the elite players (n = 63) included in the study, classified by age category (i.e., team).

Team	*n*	Age (yrs)	Height (cm)	Weight (kg)	SSGs (number)	Playing time SSG (min)	Playing time 11-v-11 (min)
U15	17	14.04 (.40)	161.29 (5.85)	47.29 (5.18)	16.00 (4.51)	96.00 (27.08)	127.00 (71.78)
U17	15	15.97 (.58)	176.60 (7.57)	64.01 (7.16)	11.47 (2.20)	68.80 (13.22)	162.80 (91.13)
U19	16	17.45 (.39)	181.94 (7.47)	70.34 (8.83)	17.75 (4.80)	106.50 (28.77)	131.25 (71.21)
U23	15	19.41 (1.05)	181.29 (5.18)	74.74 (7.38)	14.80 (3.97)	88.80 (23.81)	153.53 (70.55)

Table 3.1 presents descriptive information of the included players per team. The players of the different teams had comparable practice schedules. They had four or five technical and tactical practice sessions and one or two physical practice session per week, resulting in 7.5 to 10.5 hours of practice per week. Additionally, the teams played one competitive match each week. The U17 and U19 teams competed at the highest and second highest national level within their respective youth competition, the U15 team competed at the third highest national level. Players in the U23 team competed at the highest adult amateur level. Thus, participants in this study played at an elite level given their age, and our sample is considered to be representative of the population of elite soccer players in the U15 to U23 age categories. Written informed consent was acquired from the players (and

their parents when necessary) prior to the start of the study. The protocol of the study was approved by the Ethical Committee of Psychology, University of Groningen (Research code: 17197-O).

3.2.2. PROCEDURE AND MEASURES

Predictor: SSGs

The SSGs for this study were organized approximately once per month, over the course of 8 months, as part of the regular technical and tactical training sessions for each team. The SSGs were scheduled in consultation with the teams' physical trainers. Depending on the physical load scheduled for the teams by the physical trainers, 3 to 6 SSGs per team were organized per training session. Due to uncontrollable circumstances, such as the cancellation of training sessions due to bad weather, the absence of players due to illness or injuries, or players dropping out, players within and across teams could not participate in the exact same number of SSGs. Therefore, players in the U15, U17, U19, and U23 teams played on average in 16, 11, 17, and 14 SSGs, respectively (see Table 3.1).

The SSGs were played outdoors on the teams' usual practice grounds, with the U23 and U19 teams playing on natural turf and the U17 and U15 teams playing on artificial turf. The pitch size was constrained to 80 m x 56 m, which corresponds to the match-derived relative pitch area of 320 m^2 (Olthof et al., 2019). Each SSG lasted 6 minutes, with 2 minutes of rest in between SSGs, and included standard soccer rules, such as throw-ins, off-side, free kicks, and corner shots. The games were filmed using a Canon Legria HF R68.

Finally, to control for the strength of opposition and the quality of the team, players were reorganized into different teams after each SSG (cf. Fenner et al., 2016). This was done semi-randomly, by accounting for the position (i.e., attack-midfield-defense) of the players in order to avoid teams consisting of mainly one playing position. Thus, players played each game with a different set of teammates.

We used notational analysis to assess performance in the SSGs (Hughes & Bartlett, 2002). A coding scheme detailing offensive and defensive indicators was developed by the first author and the soccer club's head of performance and data analyst. The head of performance and the data analyst each had more than 7 years of experience managing, processing, and analyzing event data (i.e., data on in-game soccer performance indicators, regardless of outcome). The coding scheme contained in-game performance indicators that are positively correlated with game success (Pappalardo & Cintia, 2017), and were deemed to present an accurate picture of an individual's in-game on-ball performance, namely passes forward, offensive and defensive duels, assists, key passes, shots on target, applying pressure, and pass interceptions (see Table A3.1 in the appendix).

Performance indicators in the SSG videos were coded independently by one researcher and two graduate students using Noldus The Observer XT (Noldus Information Technology, Wageningen, the Netherlands). The researcher and graduate students prepared and practiced with coding for a week, in order to make slight adjustments to the definitions of performance indicators and obtain familiarity with the coding scheme. Then, three of the total $k = 82$ SSGs were coded by both the researcher and the students to assess the reliability between the raters. This yielded a Cohen's kappa of .77, which indicates acceptable reliability.

Predictor: Physiological and motor tests

Physiological and motor testing was conducted approximately two months after the beginning of the season. Players' sprinting ability was measured by a maximal 30-meter linear sprint, with a local position measurement system tracking the position and time of the players (Inmotio Object Tracking BV, Amsterdam, the Netherlands). Timing gates were placed at the 0, 10, and 30 m mark. Players positioned themselves .5 m behind the first timing gate, and were instructed to run as fast as possible. Each player performed 2 sprints. The fastest time was recorded and used for analysis (Altmann et al., 2019).

3

To assess each athlete's interval endurance capacity, players performed the Interval Shuttle Run Test (ISRT; Lemmink et al., 2004). During this test, players were required to run back and forth on a 20 m course, with pylons set 3 m before the turning lines. Sound signals on a prerecorded disc indicated the pace at which the players had to reach the 3 m turning lines. The running speed, dictated by the frequency of these signals, was increased by 1 km/hr every 90 s from a starting point of 10km/hr and by .5 km/hr every 90 s from 13 km/hr onwards. Each 90 s period was divided into two 45 s periods in which players ran for 30 s and walked for 15 s. Players were instructed to complete as many tracks as possible, and were told to stop when they could not follow the pace or felt unable to complete the run. The maximum number of completed tracks was recorded and used for analysis.

Finally, players' agility was measured using a modified version of the agility T-test (Haj-Sassi et al., 2011; Pauole et al., 2000). Four cones were arranged in a T shape, with a cone placed 5 m from the starting cone and 5 m on either side of the second cone. Players were instructed to sprint from the starting cone to the second cone, sprint to a side-cone, sprint to the opposite side-cone, sprint back to the second cone, and finally sprint back to the starting cone. This test was conducted twice, with players turning either right or left around the cones, to obtain a right and left agility estimate, respectively. Thus, in this modified version, players had to sprint around, instead of shuffle between the outer cones. Times were recorded using the local position measurement system. An average agility estimate was computed by taking the mean of the left and right estimate, which was used for further analyses.

Criterion: 11-vs-11 games

Criterion data was obtained by analyzing participants' performance in 11-vs-11 games. The 11-vs-11 games were played as part of the team's regular competitions, and were filmed by a staff member of the club. In deciding the number of 11-vs-11 games to analyze, we aimed to match approximately the number of analyzed minutes in the SSGs and 11-vs-11 per team. This would result in analyzing three full 11-vs-11 games per team. However, in order to have sufficient variability in opponent strength, as well as in the performance of the participants, we instead analyzed one half of six different 11-vs-11 games.

Games were selected based on each team's placement in their competition standings: we selected two games against higher placed opponents, two games against lower placed opponents, and two games against opponents with approximately the same placement. For each game we randomly selected either the first or second half. All selected games were played in the last four months of the same season in which the SSGs were played.

Individual soccer performance in the 11-vs-11 games was assessed using the same notational analysis procedure and coding scheme as for the SSGs. Thus, we coded the same performance indicators in the 11-vs-11 games as in the SSGs. The coding process was conducted by the same researcher and graduate students.

3.2.3 DATA PREPARATION

The performance indicators 'dribbles' and 'take-ons' were summed to create an 'offensive duel' indicator; 'tackles' and 'in-fronts' were summed to create an 'defensive duel' indicator (see Table 3.2). More than half of the players did not have any recorded events on offensive and defensive aerial duels in the SSGs. Therefore, these indicators were excluded from the individual performance analysis.

In order to compare performance between players who varied in total minutes played, the indicators that were counted 'when they occurred' (i.e., interceptions, applying pressure, chances created, shots on target) were transformed to a rate statistic, by computing the number of events per bout of six minutes (i.e., the duration of each SSG). To operationalize each player's performance on the indicators that had a successful or unsuccessful outcome (i.e., passes forward, offensive duels, and defensive duels) we applied a rigorous statistical approach. Specifically, we estimated a random intercept multilevel logistic regression model for these indicators in both SSGs and 11-vs-11 games, in which the intercepts were allowed to vary across players. The advantage of this model is that it does not require an equal number of observations for each individual (e.g., simply dividing successful passes by total number of passes may lead to over- or underestimations of a player's performance (Hox, 2010). In addition to the random intercepts, 'team' was included as a categorical covariate. This model predicts the probability of a successful outcome on the indicator (i.e., the dependent variable, for example, a successful pass) for each player simply by their intercept (i.e., the model's fixed effect intercept plus a random effect for each player) and their team effect. Thus, these 'posterior' estimates can be seen as a measure of each player's performance on the performance indicators (see Table A3.2 in the appendix for a summary of the multilevel models).

Finally, we combined the offensive and defensive performance indicators to obtain an overall measure of offensive and defensive in-game performance for each player, respectively. The weights for each indicator were derived from its team-wise correlation with a proxy for in-game offensive and defensive success, namely shots on target and shots on target conceded (i.e., a shot on target by the opposite team, both including goals; cf. Pappalardo et al., 2019). Specifically, we assessed the team's performance on the performance indicators in each SSG and 11-vs-11 game, and computed Spearman's rank correlations between the indicators their respective in-game success proxy (see Table 3.1 and Table A3.3 in the appendix). To account for differences in the number of observations and performance levels across age groups, the correlations were aggregated using a random effect meta-analysis.

The correlation coefficients for each indicator were in the expected direction, meaning that greater performance on the offensive indicators was positively associated with shots on target, while greater performance on the defensive indicators was negatively associated with shots on target conceded (see Table 3.1). Therefore, we transformed the performance indicators for the players to z-scores within each team, multiplied their score with the correlation coefficient, and summed the scores (Pappalardo et al., 2019). Additionally, we added the individual player's shots on target to the offensive performance measure, giving it a weight of 1. These overall performance measures can be seen as a player's contribution to in-game success.

Table 3.2 Definitions and weights for offensive and defensive performance indicators.

	Offense					
Indicator	Pass forward	Dribble	Take on	Chance created	Shot on target (incl. goals)	Offensive aerial duel
Definition	A pass attempt in the forward (i.e., opponent's goal) direction.	An attempt by the attacker with the ball to drive by a defender. No dribble is awarded if the attacker dribbles in 'open space' and does not attempt to drive by a defender.	An attempt by the attacker with the ball to maintain ball-control/ possession, and/or create space, when in contest with a defender.	The final pass that leads to the recipient of the ball having a shot on target (i.e., key pass) or scoring a goal (i.e., assist).	A scoring attempt that goes into the net (i.e., a goal) or an attempt that clearly would have gone into the net, but was saved by the goalkeeper or a player who is the last line of defense.	An attempt by the attacker (i.e., the player whose team was in ball possession) to maintain control/ possession of the ball, when in contest with a defender in the air.
Merged	-	Offensive duels		-	-	-
Outcome	Successful - unsuccessful	Successful - unsuccessful		Counted when occurs	Counted when occurs	Successful - unsuccessful
Weight[a]	.21	.17		.50	1	-
Formula[b]	Offensive performance = Passes forward * .21 + Offensive duels * .17 + Chances created * .50 + 1 * Shot on target					

Table 3.2: (continued)

	Defense				
Indicator	Tackle	Staying in front	Applying pressure	Interception	Defensive aerial duel
Definition	An attempt by the defender to obtain ball control/ possession of an attacking player with the ball	An attempt by the defender to stay in front of an attacking player, in order to prevent a dangerous offensive (e.g., goal scoring) opportunity.	A situation in which the defender puts pressure on an attacking player with the ball, thereby making the opposing player lose the ball (e.g., through an unsuccessful pass attempt).	A situation in which the defender 'reads' the pass of the opposing player and moves into the line of the intended the pass, thereby intercepting the pass.	An attempt by the defender (i.e., the player whose team was not in possession) to obtain ball control/ possession, when in contest with an attacker in the air.
Merged	Defensive duel		-	-	-
Outcome	Successful - unsuccessful		Counted when occurs	Counted when occurs	Successful - unsuccessful
Weight[a]	-.14		-.11	-.06	-
Formula[b]	Defensive performance = (Defensive duels * -.14 + Interceptions * -.06 + Applying pressure * -.11) *-1				

[a] *Weights* indicate the aggregated correlation of the performance indicator with shots on target (offensive) and shots on target conceded (defensive).
[b] *Formula* indicates the computation for the individual overall offensive and defensive performance. Performance indicators in the formula row indicate standardized (z) scores. The defensive score was multiplied by -1 such that a higher score indicates a better defensive performance.

3.2.4 STATISTICAL ANALYSES

To evaluate the extent to which SSGs are representative for 11-vs-11 games in terms of the assessed performance indicators (i.e., aim 1), we first computed the mean number of times an event occurred per 6 minutes of playing time, for each performance indicator, in each game format. Second, we conducted a chi-square goodness of fit test to compare the total number of observed events per performance indicator in the SSGs (i.e., the empirical distribution) against the relative frequency of the observed events on the performance indicators in the 11-vs-11 games (i.e., treating this as the theoretical distribution). We checked the observed and expected events, as well as the Pearson standardized residuals to evaluate which performance indicators differed most in incidence in the SSGs and 11-vs-11 games. Given that effect sizes for chi-square tests are often difficult to interpret (Cohen, 1988), we computed a Spearman's rank correlation (rs) between the total number of observed events in both game formats to assess the degree of association between the distributions.

To assess the predictive validity of SSG performance (i.e., aim 2), we computed Spearman's rank correlations between the performance indicators in the SSGs and 11-vs-11 games. Moreover, to assess the predictive validity of physiological and motor performance, we computed Spearman's rank correlations between the physiological and motor tests and overall offensive and defensive performance in the 11-vs-11 games. Players with partially missing data (i.e., on either the ISRT, sprint, or agility tests) were still included in analyses for which they had sufficient data. Four players did not have enough offensive duel events and 2 players did not have defensive duel events in the 11-vs-11 games. In addition, 6 players could not participate in the sprint- and agility tests due to illness or injury, including 1 that could also not participate in the ISRT. One player had missing data on both the sprint test and offensive duels. This yielded sample sizes of $55 < n < 63$ for the different analyses.

To account for possible differences between players across teams, correlations were first computed within each team. Then, in order to draw inferences on the overall strength of the predictor-criterion relationships across our sample ($55 < n < 63$), we combined the coefficients from the different teams using a random effect meta-analysis. The random effect meta-analysis accounts for the heterogeneity across coefficients, as well the sample size per team, resulting in a weighted average correlation coefficient (Borenstein et al., 2010). We refer to the weighted average coefficients as the aggregated correlation coefficient.

We computed Spearman's rank correlations instead of Pearson correlations, because we are interested in the association between the rankings on the predictors and criterion, and want to account for any potential outliers. The correlations' magnitudes were interpreted according to the thresholds suggested by Cohen (Cohen, 1988), with $r = 0 - .1$ indicating a trivial, $r_s = .1 - .3$ indicating a small, $r_s = .3 - .5$ a moderate, and $r_s > .5$ a large relationship. Finally, while we report p-values, we aim to avoid dichotomizing results as 'significant' or not, and focus on the point estimates and confidence intervals (McShane et al., 2019; Wasserstein et al., 2019).

3.3 RESULTS

3.3.1 REPRESENTATIVENESS OF SSGS

Figure 3.1 presents the mean number of events per 6 minutes for each performance indicator, per SSG and 11-vs-11 game (see Table A3.4 in the appendix for a table with this information). With the exception of aerial duels and pass interceptions, there were more events per 6 minutes for every performance indicator in an average SSG, compared to an average 11-vs-11 game.

3

Table 3.3 presents results from the chi-square goodness of fit test. The chi-square goodness of fit test indicated that the total number of observed events per indicator in the SSGs was not consistent with the distribution of events in the 11-vs-11 games, χ^2 (10, N = 6060) = 923.79, $p < .01$. By examining the expected number of events and the standardized residuals in table 3, it can be seen that this finding is mainly driven by both aerial duels, the shots on target, chances created, and staying in front. Specifically, there were substantially fewer aerial duels in the SSGs than in the 11-vs-11 games, whereas shots on target, chances created and staying in front were observed more often in the SSGs (see also Figure 3.1). However, while there were differences on these performance indicators between the observed and expected events, we found that the overall association between the distributions was strong ($r_s = .78$, 95% CI = .35; .94). The overall high degree of representativeness of the SSGs is also supported by the finding that the removal of aerial duels reduces the chi-square value by approximately a half (χ^2 (8, N = 5973) = 422.52, $p < .01$), and increases the correlation to $r_s = .98$, (95%, CI = .92; 1). Together, these results suggest that, with the exception of aerial duels, the distribution of events is similar in the SSGs compared to the 11-vs-11 games. However, the SSGs yield more opportunities for events on the performance indicators, particularly in terms of shots on target and chances created.

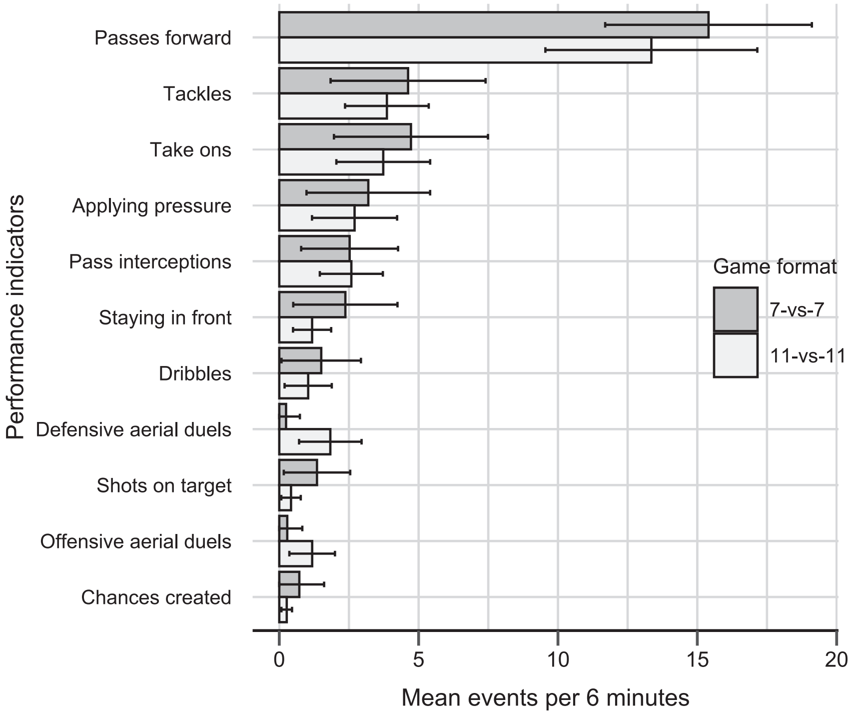


Figure 3.1 Mean events per 6 minutes for the performance indicators in 7-vs-7 SSGs and 11-vs-11 games.

3.3.2 INDIVIDUAL SSG PERFORMANCE

Table 3.4 displays the aggregated Spearman's correlations between the players' performance on the different indicators in the SSGs and the 11-vs-11 games (see Table A3.5 in the appendix for correlations per team). With respect to the aggregated coefficients, individual performance in the SSGs and 11-vs-11 games was moderately-to-largely correlated for 6 of the 9 performance indicators.

Table 3.4 displays the aggregated Spearman's correlations between the players' performance on the different indicators in the SSGs and the 11-vs-11 games (see Table A3.5 in the appendix for correlations per team). With respect to the aggregated coefficients, individual performance in the SSGs and 11-vs-11 games was moderately-to-largely correlated for 6 of the 9 performance indicators. The largest relationship was found for performance on pass interceptions (r_s = .53, 95% CI = .25; .73).

Table 3.3 Results from the chi-square goodness of fit test.

Performance indicator	Observed events 11-vs-11[a]	Prop. 11-vs-11	Observed events SSG[a]	Prop. SSG	Expected events SSG	St. residuals
Passes forward	2167	.42	2526	.42	2519.09	.18
Tackles	619	.12	758	.12	719.57	1.53
Take-ons	601	.11	775	.13	698.65	3.07
Applying pressure	439	.08	524	.08	510.33	.63
Pass interceptions	418	.08	414	.07	485.92	-3.40
Defensive aerial duel	303	.06	40	.01	352.23	-17.14
Staying in front	195	.04	389	.06	226.68	10.99
Offensive aerial duel	195	.04	47	.01	226.68	-12.16
Dribbles	165	.03	247	.04	191.81	4.05
Shots on target	68	.01	222	.04	79.05	16.18
Chances created	43	.01	118	.02	49.99	9.66

Test result: χ^2 (10, N = 6060) = 923.79, $p < .01$.
Prop = proportion; *st.* = standardized
[a]used to assess the correlation between the distribution of events in both game formats.

Individual forward passing performance (r_s = .38, 95% CI = .11; .59), offensive duel performance (r_s = .35, 95% CI = .08; .58), shots on target (r_s = .38, 95% CI = .05; .63), successfully applying pressure (r_s = .40, 95% CI = .13; .61), and overall offensive performance (r_s = .46, 95% CI = .20; .65) in the SSGs and 11-vs-11 games were moderately correlated. A small correlation was found for overall defensive performance (r_s = .28, 95% CI = 0; .52), while trivial correlations were found for defensive duel performance (r_s = .02, 95% CI = -.26; .30) and chances created (r_s < .01, 95% CI = -.27; .26). Moreover, the confidence intervals for every indicator were relatively wide, ranging from a positive small to positive large association for the indicators with a moderate-to-large point estimate. In sum, these results suggest that the predictive validity of individual SSG performance is moderate-to-large but that there is variability across performance indicators.

Table 3.4 Aggregated Spearman's correlations between the performance indicators in the SSGs and 11-vs-11 games.

Performance indicator	r_s (95 % CI)	p	n
Forward passing	.38 (.11; .59)	< .01	63
Chances created	< .01 (-.27; .26)	.98	63
Shots on target	.38 (.05; .63)	.03	63
Pass interceptions	.53 (.25; .73)	< .01	63
Applying pressure	.40 (.13; .61)	.005	63
Offensive duels	.35 (.08; .58)	.01	59
Overall offensive performance	.46 (.20; .65)	< .01	59
Defensive duels	.02 (-.26; .30)	.88	61
Overall defensive performance	.28 (.0; .52)	.05	61

r_s = aggregated spearman correlation coefficient; *CI* = Confidence Interval

3.3.3 PHYSIOLOGICAL AND MOTOR PERFORMANCE

Table 3.5 presents Spearman's correlations between the players' performance on the physiological and motor tests and the overall offensive performance (left), and the overall defensive performance (right) in the 11-vs-11 games (see Table A3.6 in the appendix for correlations per team). The aggregated coefficients were negative small or trivial for 10 m sprint and 11-vs-11 performance (r_s = -.19, 95% CI = -.47; .12; r_s = .05, 95% CI = -.24; .34), 30 m sprint and 11-vs-11 performance (r_s = -.20, 95% CI =-.54; .20; r_s = .02, 95% CI = -.26; .31), and agility and offensive performance (r_s = -.11, 95% CI = -.46; .29). A small positive aggregated correlation was found for offensive performance and ISRT (r_s = .15, 95% CI = -.22; .48). Moreover, a small negative aggregated correlation was found between ISRT and defensive performance (r_s = -.12, 95% CI = -.38; .17), and a small positive correlation for defensive performance and agility (r_s = .11, 95% CI = -.18; .39). Additionally, the confidence intervals were wide, and ranged from a (small-to-large) negative to (small-to-moderate) positive association for all physiological and motor tests. In sum, the point estimates suggest that the predictive validity of physiological and motor test performance varies between small and negative to small and positive, with respect to our operationalization of overall offensive and defensive performance in the 11-vs-11 games.

3.4 DISCUSSION

In the current study we aimed to take novel steps in quantifying in-game soccer performance, and in assessing the representativeness of SSG performance for 11-vs-11 game performance. First, we examined whether 7-vs-7 SSGs provided a representative assessment context for 11-vs-11 games, in terms of various performance indicators. Second, we determined the predictive validity of individual soccer SSG performance with respect to performance in 11-vs-11 games. Moreover, we explored the predictive validity of physiological and motor tests for performance in 11-vs-11 games.

Table 3.5 Aggregated Spearman's correlations between physiological and motor tests and overall offensive (left) and defensive performance (right) in 11-vs-11 games.

Physiological and motor performance	Overall offensive performance (11-vs-11)			Overall defensive performance (11-vs-11)		
	r_s (95 % CI)	*p*	*n*	r_s (95 % CI)	*p*	*n*
10 m sprint	-.19 (-.47; .12)	.23	55	.05 (-.24; .34)	.72	56
30 m sprint	-.20 (-.54; .20)	.32	55	.02 (-.26; .31)	.87	56
ISRT	.15 (-.22; .48)	.43	58	-.12 (-.38; .17)	.42	60
Agility	-.11 (-.46; .29)	.62	55	.11 (-.18; .39)	.45	56

r_s = aggregated spearman correlation coefficient; CI = Confidence Interval
Note: a lower time on the sprinting and agility tests indicates a better performance, hence a negative correlation indicates that faster sprinting and agility is related to better overall performance in 11-vs-11.

We found strong associations between the distribution of observed events across the performance indicators in both game formats. Additionally, we found that, on average, more events per 6 minutes occur in the SSGs than in the 11-vs-11 games. This was the case for almost all performance indicators, the main exceptions being aerial duels, which occurred considerably more often in the 11-vs-11 games. Together, these results suggest that the SSGs are representative for 11-vs-11 games in terms of assessed indicators, but that they are generally faster paced than 11-vs-11 games. While the relative pitch area was constrained to match those of official games (Olthof et al., 2019), the smaller absolute pitch size and lower number of players may still lead to a faster offensive play, as shown by the increase in shots, chances created, and staying in front of a player on the defensive end. Likewise, an explanation for the exception of aerial duels is that the smaller pitch size changes the environmental constraints of the soccer game. This may alter the affordances, for instance of aerial goal-kick possibilities, which typically result in aerial duels (Katis & Kellis, 2009; Kelly & Drust, 2009). Although unanticipated, these results can be interesting and relevant to talent identification and development in soccer. Given that high-paced handling is crucial for modern day professional soccer (Wallace & Norton, 2014), the large scaled 7-vs-7 SSGs may provide ample opportunities as a practice context. It is also plausible that such patterns are reinforced when pitch or team sizes are reduced even further. Therefore, it would be interesting to assess the extent to which small scaled 4-vs-4 SSG, as used in other studies (Bennett et al., 2018; Fenner et al., 2016), can be considered representative of 11-vs-11 games.

When looking at the predictive validity of SSG performance, performance on pass interceptions, forward passes, applying pressure, shots on target, offensive duels and overall offensive performance were positively and moderately correlated, meaning that individual performance on these indicators in the SSGs was related to performance in the 11-vs-11 games. In contrast, trivial and small correlations were found for performance on chances created, overall defensive performance, and defensive duels. These results suggest that 7-vs-7 SSGs are particularly useful for assessing and predicting offensive 11-vs-11 performance. The small correlation for overall defensive performance seems a logical result of defensive duels: This indicator received the largest weight in creating the defensive performance indicator, but defensive duels in the SSGs and 11-vs-11 games were not correlated.

More generally, the variability in correlations and relatively large confidence intervals across indicators is likely due to the natural variation around in-game technical and tactical performance (Rampinini et al., 2007). While players across age categories played in multiple SSGs and 11-vs-11 games, the sample size in terms of both minutes played and number of players was still relatively small. This could have made it difficult to obtain stable validity estimates for the performance indicators, particularly for chances created, defensive duels, and defensive performance. Still, the moderate predictive validities based on a relatively small sample size are encouraging of using 7-vs-7 SSGs as representative contexts for predicting performance in 11-vs-11 games.

These findings are in accordance with our hypothesis that a sample-based predictor that mimics the criterion behavior in content and context enhances predictive validity (i.e., behavioral consistency). This is reinforced by the finding that the physiological and motor tests yielded trivial-to-small correlations with offensive and defensive performance, as assessed through the indicators. These results, therefore, make intuitive and theoretical sense; they suggest that a predictor based on a representative assessment may be more suitable for making predictions than results of isolated physiological and motor tests, at least when soccer performance is defined in terms of the assessed performance indicators. In sports, these findings correspond to Lyons et al. (Lyons et al., 2011), who studied the predictive validity of physiological and motor performance and collegiate performance on in-game American football performance. The authors found that collegiate performance was a more valid, and more consistent predictor of American Football performance than physiological tests. Furthermore, the trivial correlations for physiological and motor performance are in accordance with Wilson et al. (Wilson et al., 2017), who showed that athletic ability had a very weak association with performance in 11-vs-11 games, as determined by similar performance indicators.

Although the predictive validity of the physiological and motor tests was small in our study, these results do not mean that physiological and motor performance is unimportant for elite soccer performance in general. For example, range restriction in the physiological and motor variables likely attenuated their relationship with 11-vs-11 performance. This means that physiological and motor performance is most likely related to soccer performance in the general population of all youth players. However, there is not enough variance in physiological and motor performance among the elite soccer players to meaningfully differentiate between them, as it is likely that the elite players have, explicitly or implicitly, been preselected on these variables (Bergkamp et al., 2019). Thus, stronger relationships may have been found if the physiological and motor variables were studied in a more heterogeneous group of players.

3.5 STRENGTHS AND LIMITATIONS

In this study, we developed a finer-grained measure of soccer performance. At the same time, our operationalization of soccer performance cannot be considered a 'complete' measure of in-game performance (Travassos et al., 2013; Vilar et al., 2012). We measured in-game performance using performance indicators that could be coded based on recordings of games. For instance, we were not able to reliably define off-the-ball movements for each player at each moment (Sarmento, Clemente, et al., 2018), or include physiological measures such as high-intensity sprints on the field, or total distance ran. Integrating such (physiological) measures into our on-ball 11-vs-11 performance metrics could have increased the predictive validities of the physiological and motor tests (Redkva et al., 2018). In addition, note that although off-ball performance actions, such as positioning, deciding, and running actions were not explicitly assessed, they are often intertwined with other indicators we assessed (e.g., forward passes). Furthermore, and more importantly, we focused on on-ball performance, because this has been shown to predict game success (i.e., game outcome) in soccer (Pappalardo & Cintia, 2017). Our study further supports these findings; we also found positive and negative correlations between the offensive and defensive performance indicators, and shots on target and shots on target conceded, respectively. In contrast, evidence for the relationship between physiological in-game performance indicators and game success has been mixed (Chmura et al., 2018; Gomez-Piqueras et al., 2019; Hoppe et al., 2015).

Other limitations pertain to the notational analysis method used to assess soccer performance. This is a relatively intensive method to assess performance and its reliability depends on a common interpretation of indicators by each coder. Although the reliability was acceptable in our study, it is almost unavoidable that particular definitions of indicators (e.g., 'applying pressure') leave room for interpretation. Additionally, using the same observers to code both the predictor and criterion data could have positively affected the correlations between the indicators. Integrating physiological or tactical information derived through local or global positioning systems into the predictor or criterion may offer more reliable information. This could improve soccer performance assessments, and future research should consider if this is feasible. Furthermore, performance in the SSGs and 11-vs-11 was assessed in a single season, which could have increased the correlations between performance in both game formats. Finally, while SSG and 11-vs-11 performance was moderately correlated overall, we did not account for positional differences. Thus, more research is needed assessing the extent to which SSG performance transfers to position-specific roles in 11-vs-11 games.

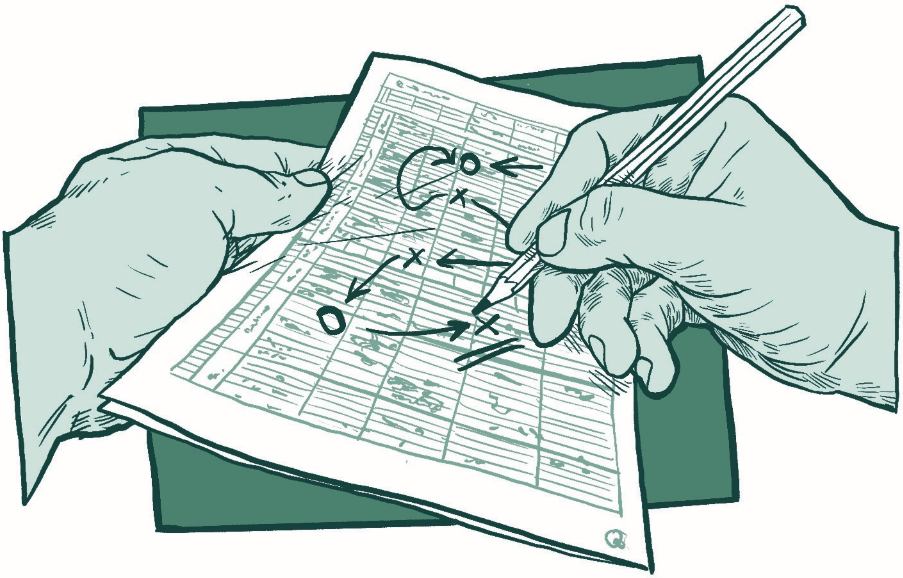
3.6 CONCLUSION

This study provides encouraging first results on the usefulness of SSG performance in predicting 11-vs-11 game performance. We demonstrated that SSGs are faster paced, but representative of 11-vs-11 soccer games in terms of the distribution of performance indicators. Moreover, we found that the in-game performance indicators are correlated with game success. Based on these correlations, we used a novel approach to quantify overall offensive and defensive in-game performance, and showed that individual SSG performance was moderately predictive of 11-vs-11 performance. Finally, in line with the notion of behavioral consistency, we found that SSG performance yielded higher predictive validities than physiological and motor tests that are often used in soccer science and practice.

The current study provides a novel step in operationalizing the criterion as in-game performance, in relation to predicting performance based on a representative assessment. However, since the predictive validities in SSGs can still not be considered as large based on our result, we would not (yet) recommend solely using scores on SSGs for talent identification and selection purposes. We encourage researchers to further examine the validity of SSGs. More importantly, future researcher should give further emphasis to quantifying in-game soccer performance at the criterion and predictor level, thereby incorporating physiological and tactical (off-the-ball) parameters. We expect that the rapid technological advancements in soccer analytics can be fruitfully used in future research on talent selection.

ACKNOWLEDGEMENTS

We would like to thank Marieke Timmerman for her helpful suggestions regarding the data analysis, Sil Piek for assisting in the conceptualization of the performance indicators, Mees van der Linde and students at FC Groningen for assisting in the data collection, and Lilli Schrijber and Sem Otten for assisting in the notational analysis.



CHAPTER 4

HOW SOCCER SCOUTS IDENTIFY TALENTED PLAYERS

This chapter was published as:

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, 22*(7), 994 - 1004. https://doi.org/10.1080/17461391.2021.1916081

ABSTRACT

Scouts of soccer clubs are often the first to identify talented players. However, there is a lack of research on how these scouts assess and predict overall soccer performance. Therefore, we conducted a large-scaled study to examine the process of talent identification among 125 soccer scouts. Through an online self-report questionnaire, scouts were asked about 1) the players' age at which they can predict players' soccer performance, 2) the attributes they consider relevant, and 3) the extent to which they predict performance in a structured manner. The most important results are as follows. First, scouts who observed 12-year-old and younger players perceived they could predict at older ages (13.6 years old, on average) whether a player has the potential to become a professional soccer player. This suggests that scouts are aware of the idea that early indicators of later performance are often lacking, yet do advise on selection of players at younger ages. Second, when identifying talented players, scouts considered more easily observable attributes, such as technical attributes. However, scouts described these often in a broad sense rather than in terms of specific predictors of future performance. Finally, scouts reported that they assess attributes of players in a structured manner. Yet, they ultimately based their prediction (i.e., final score) on an intuitive integration of different performance attributes, which is a suboptimal strategy according to existing literature. Taken together, these outcomes provide specific clues to improve the reliability and validity of the scouting process.

4.1 INTRODUCTION

Talent identification in soccer is the complex process of recognizing and selecting players that have the greatest potential to excel in the future (Johnston et al., 2018). In practice, talented players are often identified by talent scouts (Reeves et al., 2018). These scouts typically play an important role in the initial phases of a club's talent identification process. Scouts mainly observe and assess players who are not yet recruited by elite soccer academies in soccer trials or games (Reeves et al., 2019). Based on assessments of current soccer performance, they make predictions of players' future performance to advise on selection decisions (Larkin et al., 2020). In this sense, the task of a scout differs from that of a coach, who is typically (also) involved in long-term player development processes (Johansson & Fahlén, 2017).

In order to make valid and reliable performance predictions, and to ultimately decide whether a player has the potential to excel, scouts need to go through a process in which they address different issues. Specifically, they must 1) define the age cohort of talented players for which they can predict performance, 2) consider what soccer-specific attributes are relevant predictors of performance and how to assess them, and 3) form an overall performance prediction based on assessments on these predictors. However, little is known about the way in which scouts address these important issues (Larkin & Reeves, 2018; Reeves et al., 2018). Therefore, we conducted a large scaled study examining the processes of talent identification among soccer scouts.

4.1.1 PREDICTING PERFORMANCE

The scout's task to predict future performance of young players is incredibly difficult (Bergkamp et al., 2019). Across different sports, research has shown that athletes develop in different – often nonlinear – ways, and that reliable indicators of future elite performance are often not yet present or developed in young players (Baker et al., 2018; Den Hartigh et al., 2016; Güllich, 2014). Still, soccer scouts are mainly assigned by their club to identify young (e.g., 13-15-year-old) to very young (younger than 12-year-old) players (Ford et al., 2020). An interesting first question is then whether scouts' beliefs align with their scouting practices. Specifically, for which age cohort of players do scouts perceive they can make reliable predictions of future soccer performance in the first place?

A second important question, specifically focused on the act of scouting, concerns the operationalization of soccer-specific performance predictors. Although there is a large body of literature on the predictive value of various attributes (Ivarsson et al., 2020; Murr, Raabe, et al., 2018; O'Connor et al., 2016), only a few studies have examined what soccer scouts and coaches consider relevant attributes for future performance (Larkin et al., 2020). Larkin and O'Connor (2017), for instance, found that Australian scouts and coaches (n = 20) perceived technical (e.g., first touch, 1-vs-1), psychological (e.g., positive attitude, personality), and several miscellaneous (e.g., X-factor) attributes as most important when identifying under (U)-13 soccer players. In contrast, they deemed motor skills (e.g., speed), physical attributes (e.g., strength), and defensive ability less important within the talent identification process. These findings are in accordance with a recent study by Roberts, McRobert, et al., (2019), who found that scouts and coaches (n = 99) considered decision-making, positioning, and passing accuracy more important for central midfielders than physiological attributes such as stamina. Finally, Jokuschies et al. (2017) found that coaches (n = 5) most often named personality-related attributes as talent criteria, whereas few named motor skills or physical attributes. Still, findings across these studies and their included samples were relatively small and diverse. Hence, studying what a large sample of soccer scouts considers key attributes to predict performance is warranted.

A third major question is how scouts score and combine information on these predictors into an overall performance assessment. Since these assessments (and, therefore, predictions) are essential in the decision to select a player, it is important that they are valid and reliable. Although scouts and coaches account for multidimensional attributes, research suggests that they generally do not assess these attributes in a structured manner when predicting performance. Qualitative studies showed that coaches primarily assessed performance based on their overall impression, intuition, or 'coaches eye' (Roberts, Greenwood et al., 2019). In other words, coaches did not use explicit criteria and relied on holistic performance assessments (Johansson & Fahlén, 2017). Coaches reported that they were able to recognize patterns that resonated with their ideal performance image based on their impressions (Christensen, 2009), and 'knew it when they saw it' (Miller et al., 2015). Yet, they had difficulty verbalizing what these patterns of performance looked like exactly and how they weighed the performance attributes (Christensen, 2009).

It is interesting to note that the holistic approach can be sub-optimal, because it typically leads to inconsistent assessments within and between decision-makers (Dawes et al., 1989; Den Hartigh, Niessen, et al., 2018). Relatedly, there is a large body of evidence that shows that reliability and predictive validity improve when assessment processes increase in structure (Dana & Rick, 2006; Huffcutt & Arthur, 1994). Strategies such as explicitly defining criteria, systematically scoring information, and combining scores according to a decision rule are valid ways to improve assessments (Arkes et al., 2006; Meijer et al., 2020). In sports, few studies have evaluated to what extent scouts apply these strategies to reach their final performance assessment (see MacMahon et al., 2019 for an exception).

4.1.2 THE CURRENT STUDY

Based on the questions above, we aimed to explore – through a self-report measure – how soccer scouts identify talented players. In line with the difficulty of predicting future performance of young players, we first examined at what age scouts perceive they can predict a player's performance. Second, we analyzed what attributes scouts consider to be important for future performance. Finally, we examined to what extent scouts report scoring and combining this information in a structured manner. We therefore conducted a large-scaled study among soccer scouts across the Netherlands.

4.2 MATERIALS AND METHODS

4.2.1 PARTICIPANTS

Ethical approval was granted by the Ethical Committee of Psychology, University of Groningen (code PSY- 1819-S-0024). We recruited professional and part-time scouts from professional soccer clubs and scouts associated with The Royal Dutch Football Association (KNVB). First, heads of scouting of ten different clubs in the Dutch Eredivisie were approached by e-mail, of which four distributed a digital questionnaire to their organization's scouts. These scouts are responsible for identifying players for the club's developmental academy or first team. Second, four scouting coordinators of the KNVB were approached and agreed to distribute the questionnaire to their regional scouts. These regional scouts are responsible for identifying players for KNVB's 'Youth Plan Netherlands' (JPN) program. JPN is a platform which targets talented youth players from under U11 to U17 (for girls U16) who have not yet been recruited by a professional soccer club. A total of 125 scouts responded and completed the questionnaire. Almost all scouts (n = 123, 98%) indicated they were male, and most of them (n = 110, 88%) scouted male players. Scouts were on average 58.2 years old (SD = 12.3), had 11.2 years of experience (SD = 8.39). Furthermore 63 (50%) observed players in the U12 and younger age cohort, 45 (26%) in the U13-U15 cohort, 9 (7%) in the U16-U18 cohort and 7 (6%) observed adult players.

4.2.2 MEASURES

A digital questionnaire was distributed via Qualtrics (Qualtrics, Provo, Utah). Before distribution, the questions were reviewed by four JPN scouts and two scouts of a professional soccer academy – who were also included in the sample – to improve terminology, consistency, and clarity. In total, the questionnaire consisted of 8 questions (2 open-ended, 1 rank, and 5 multiple-choice questions) divided across three sections.

Table 4.1 presents the different questions and response scales per section of the questionnaire. Participants completed the questionnaire at their own discretion. The questionnaire opened on 11-03-2019 and closed on 31-05-2019. In the first section scouts were asked “at what age can you reliably predict if a player has the potential to participate in professional soccer?” The second section consisted of two questions asking scouts about the information they take into account when assessing performance. Finally, the third section contained five statements focusing on the extent to which scouts assess performance in a structured manner. Previous studies in other contexts (e.g., in job interviewing, Chapman & Zweig, 2005) found that applying structure was not a unidimensional construct, but consisted of different components. As such, we analyzed the single-item scores, instead of treating the statements as one or multiple scales (see Table 4.1).

Table 4.1 Questions in the questionnaire, per section of the questionnaire

Section	Question number[a]	Question	Scale
Scouts’ perception of predicting performance	1	“At what age can you reliably predict if a player has the potential to participate in professional soccer?”	Age in years (e.g., 14 or 17 years old)
Attributes relevant for future performance	1	“Describe a maximum of five attributes that you take into account when observing a player in your respective age cohort and that you consider to be predictive of future soccer performance”	Open
	2	“Please rank the attributes you described in the previous question from 1 = most predictive to 5 = least predictive	Rank
Scoring and combining information	1	“Before observing a player, I already know which attributes I will evaluate”	Likert (1 = never to 5 = always)
	2	“When observing a player, I evaluate each attribute I find important separately”	
	3	“I evaluate different players - of the same age and playing position - on the same attributes”	
	4	“After observing a player, I sum my scores on the independently evaluated attributes to form my final assessment”	
	5	“After observing a player, I use my overall impression of the player’s attributes to form my final assessment”	

[a] The question number per section of the questionnaire

4.2.3 STATISTICAL ANALYSIS

We computed means and standard deviations to examine the spread in age at which scouts perceived they could predict if a player has the potential to participate in professional soccer. These responses were stratified according to the age cohorts typically observed by the scouts.

In order to assess the frequency, variety, and importance of the attributes that scouts considered predictive of future soccer performance, the first two authors simultaneously categorized each attribute based on its descriptive content. Five performance categories emerged when exploring the attributes, namely 1) technical, 2) tactical and perceptual-cognitive skills, 3) personality-related and mental skills, 4) physical, physiological, and motor skills, and 5) 'miscellaneous' attributes. Similar categories are frequently identified in the soccer talent literature when discussing potential performance predictors (e.g., Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018; Williams & Reilly, 2000). Answers that varied in description, but were similar in content and context, were grouped together in a single attribute construct (e.g., 'positioning on offense' and 'moving without the ball in offense') based on previous literature (Larkin & O'Connor, 2017; Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018; Roberts, McRobert, et al., 2019). Then, we assessed the frequency of each attribute, as well as the number of times that attribute was considered to be the most important predictor of future performance by a scout (i.e., being ranked as the first attribute). Finally, to assess the level of detail in the scouts' answers, each attribute was either rated as 'general' when describing a domain (e.g., 'technical skills or abilities') or 'specific,' when describing a skill or ability (e.g., 'pass accuracy'). In order to assess the inter-rater reliability of this coding process, a random sample of $k = 90$ answers (approximately 15%) were translated, grouped together, and rated on specificity, by the first and last author, independently. This yielded a Cohen's Kappa of .94, which indicates excellent reliability. The remaining answers were coded by the first author.

Finally, to examine the extent to which scouts score and combine information in a structured manner, we first looked at the response percentages to each statement on the structure of the talent identification process. Then, we computed Spearman's correlations between the statements. These correlations provide information on whether the scouts apply the different statements uniformly and consistently. For instance, do scouts who know beforehand which attributes to assess also assess each attribute separately?

4.3 RESULTS

4.3.1 SCOUTS' PERCEPTIONS OF PREDICTING FUTURE PERFORMANCE

Figure 4.1 presents scouts' answers on the age at which they can predict if a player has the potential to participate in professional soccer (i.e., predict future performance). The findings are stratified according to the age cohort in which each scout typically observed players. The results show that the average age at which scouts perceived they could predict future performance increased depending on the age cohort they observed players in. More specifically, scouts who typically observed U12 and younger players perceived, on average, that they could reliably predict a player's future performance at 13.6 (*SD* = 2.10) years old; for scouts who observed U13-U15 players this was 14.2 (*SD* = 1.84) years old; for scouts who observed U16 – U18 year old players this was 15 (*SD* = 1.80) years old, and for scouts who observed adult players this was 16.8 (*SD* = 1.28) years old. Interestingly, most of the scouts (63 out of 125) observed players in the U12 and younger cohort. Thus, the largest group of scouts perceived they could predict future performance for players that were older (i.e., 13.6 years on average) than the players they typically observed in practice.

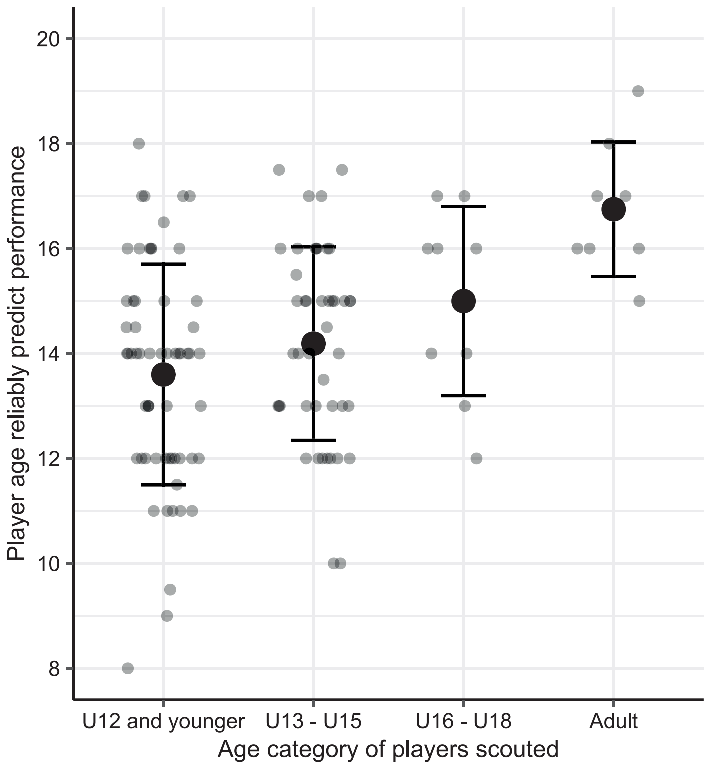


Figure 4.1 Age at which scouts perceive they can predict performance, stratified by age cohort of players scouted (error bars indicate ± SD). Each dot indicates a scout answer.

4.3.2 ATTRIBUTES RELEVANT FOR FUTURE PERFORMANCE

The attributes that scouts considered predictive of future performance were very similar across age cohorts (see Tables A4.1 to A4.4 in the appendix). Therefore, we present results for the total sample here. Table 4.2 presents the frequency (k) with which each attribute was mentioned and the number of times each attribute was considered to be most predictive (1st) of future performance, grouped by performance category. The scouts mentioned a wide variety of attributes: after grouping similar answers together, a total of 58 attributes were identified.

The nine most frequently named attributes were technical skills or technique with the ball (k = 82, 1st = 34), game sense and awareness (k = 53, 1st = 11), physiological or motor skills (k = 38, 1st = 15), sprinting speed, (k = 36, 1st = 4), winning mindset or mentality (k = 32, 1st = 6), drive or intrinsic motivation (k = 31, 1st = 9), ball control (k = 25 1st = 6), speed of handling (k = 23, 1st = 7), and physical attributes (k = 23, 1st = 2). Thus, scouts provided both general, non-specific attributes (e.g., technical skills or technique with the ball and physiological or motor skills) and more specific attributes (e.g., sprinting speed, ball control, and winning mindset or mentality).

4

Table 4.2 Attributes scouts considered predictive of future soccer performance, in terms of total frequency (*k*) and the number of times each attribute was considered most predictive (1st).

Performance category	Attribute	*k*	1st
Technical	Technical skills or technique with the ball[a]	82 (50%)	34 (74%)
	Ball control	25 (15%)	6 (13%)
	(Skills related to) transitioning[a]	11 (7%)	1 (> 2%)
	(Skills related to) defending[a]	9 (6%)	1 (> 2%)
	Pass intention or accuracy	9 (6%)	1 (> 2%)
	First touch	6 (%)	0 (0%)
	(Skills and abilities related to) attacking[a]	5 (3%)	1 (> 2%)
	Shooting or shot technique	5 (3%)	0 (0%)
	Two legged	3 (2%)	1 (> 2%)
	Dribbling	2 (1%)	0 (0%)
	Applying pressure	1 (< 1%)	0 (0%)
	Blocking	1 (< 1%)	0 (0%)
	Building up offensively	1 (< 1%)	0 (0%)
	Disrupting the offensive build up	1 (< 1%)	1 (> 2%)
	Preventing goal scoring opportunities	1 (< 1%)	0 (0%)
	Scoring goals	1 (< 1%)	0 (0%)
	Performance category total	163 (28%)	46 (37%)
Tactical and perceptual-cognitive	Game sense and awareness	53 (40%)	11 (39%)
	Speed of handling	23 (17%)	7 (25%)
	Positioning or moving without the ball	19 (14%)	2 (7%)
	Vision, perception, seeing teammates and opponents, gaze behavior	19 (14%)	2 (7%)
	Decision-making	8 (6%)	5 (18%)
	Tactical skills[a]	6 (5%)	0 (0%)
	Soccer intelligence	4 (> 3%)	1 (4%)
	Performance category total	132 (22%)	28 (22%)
Physical, physiological, and motor skills	Physiological or motor skills[a]	38 (30%)	15 (58%)
	Sprinting speed	36 (28%)	4 (15%)
	Physical attributes[a]	23 (18%)	2 (8%)
	Coordination	7 (5%)	0 (0%)
	Body composition or athletic build	6 (5%)	2 (8%)
	Agility	4 (3%)	1 (< 4%)

Table 4.2 (continued)

Performance category	**Attribute**	***k***	**1st**
Physical, physiological, and motor skills	Strength in duels	4 (3%)	0 (0%)
	Explosiveness	3 (2%)	0 (0%)
	Length	3 (2%)	1 (< 4%)
	Mobility	2 (2%)	0% (0)
	Movement rhythm	1 (1%)	1 (< 4%)
	Stability	1 (1%)	0 (0%)
	Performance category total	128 (22%)	26 (21%)
Personality-related and mental skills	Winning mindset or mentality	32 (26%)	6 (33%)
	Drive or intrinsic motivation	31 (25%)	9 (50%)
	Personality-related attributes[a]	17 (14%)	1 (< 6%)
	Perseverance, resilience, or toughness	11 (9%)	1 (< 6%)
	Behavior on and off the pitch	7 (6%)	1 (< 6%)
	Coachability, fast learner, or leadership	7 (6%)	0 (0%)
	Assertiveness or dominance	5 (4%)	0 (0%)
	Coaching other players or leadership	5 (4%)	0 (0%)
	Positive attitude	4 (3%)	0 (0%)
	Performance or goal oriented	2 (< 2%)	0 (0%)
	Focus or concentration	2 (< 2%)	0 (0%)
	Self-confidence	1 (< 1%)	0 (0%)
	Performance category total	124 (21%)	18 (14%)
Miscellaneous	Team understanding, involving teammates	12 (26%)	1 (14%)
	Communication	10 (21%)	0% (0)
	Undefined[b]	8 (17%)	3 (43%)
	X-factor	5 (11%)	0% (0)
	Innate talent (nature)	3 (6%)	2 (29%)
	Adaptability	2 (4%)	0 (0%)
	Biological age	2 (4%)	0 (0%)
	Calendar age	2 (4%)	1 (14%)
	Appearance	1 (> 2%)	0% (0)

4

Table 4.2 (continued)

Performance category	Attribute	*k*	1st
	Education level	1 (> 2%)	0 (0%)
	Lifestyle	1 (> 2%)	0 (0%)
	Performance category total	47 (8%)	7 (6%)
	Grand total	594	125

Results are presented as absolute number of answers with percentage in brackets. Percentages per attribute refer to the percentage within performance category, whereas percentage for performance category total row refer to percentage of grand total number of answers. Note: the total frequency for the attributes does not sum to k = 625 (i.e., 5 x 125), because multiple scouts listed fewer than 5 predictors.
[a] indicates an answer that can be considered a ‘general´ domain, rather than a more specific predictor
[b] answers that did not contain enough content information to be considered a predictor and could not be assigned to a performance category (e.g., “matching the playing style of club [..]”).

Concerning the general performance categories, scouts mainly considered attributes in the technical performance category as predictors of future performance: A total of 163 (28%) answers belonged to this category. This was followed by 132 (22%) answers that belonged to the tactical and perceptual-cognitive skills, 128 (22%) to physical, physiological, and motor skills, 124 (21%) to personality-related and mental skills, and 47 (8%) to the miscellaneous category. Moreover, 46 of the 125 scouts (37%) ranked an attribute in the technical category as the most important predictor, followed by a tactical and perceptual-cognitive skill (n = 28, 22%), a physical, physiological, and motor skill (n = 26, 21%), a psychological or personality-related attribute (n = 18, 14%), and a miscellaneous attribute (n = 7, 6%). Thus, a technical skill was mentioned most often as the most important predictor. Tactical and perceptual-cognitive skills, physical, physiological, and motor skills and psychological or personality-related attributes were roughly equally distributed as the most important among the remaining scouts, and a small minority mentioned a miscellaneous attribute as most predictive.

4.3.3 SCORING AND COMBINING INFORMATION

Figure 4.2 presents the response percentages to the statements on the different aspects of structure in scouts' talent identification process. Overall, the scouts indicated that they applied a very structured process when observing players. Approximately 74% of the scouts indicated that they 'always' or 'very frequently' evaluated different players – of the same age and playing position – on the same attributes, and 73% indicated that they already knew which attributes they would evaluate before they observed a player. Moreover, 69% of the scouts indicated to always or very frequently evaluate different attributes separately, when observing a player. Although the scouts seemed to apply a structured approach in defining and evaluating separate skills and abilities, they mainly used their overall impression of the player's attributes to form their final assessment, as 68% always or very frequently took this approach. Accordingly, a minority of 41% always or very frequently summed the independently evaluated attributes to form their final assessment.

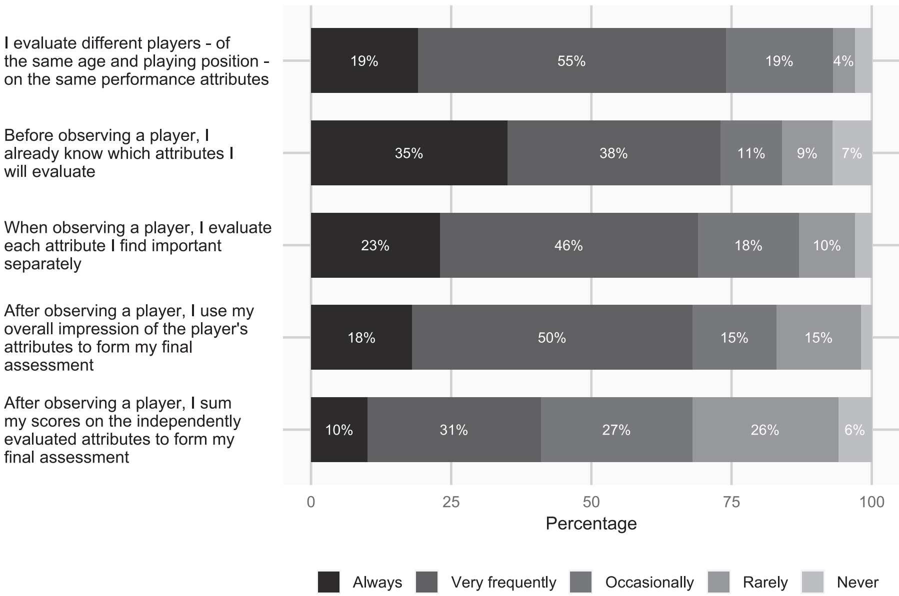


Figure 4.2 Response percentages to the statements on the different aspects of structures that scouts apply when assessing players. Note: Response percentages smaller than 4% are not displayed.

The Spearman's correlations between different aspects of structure that scouts used when identifying talent were relatively low (between .03 and .45, see Table A4.5). This suggests that applying structure cannot be seen as a single construct, and that scouts did not uniformly apply all aspects when observing players. For example, we found a relatively small correlation between the statements 'When observing a player, I evaluate each attribute I find important separately' and 'Before observing a player, I already know which attributes I will evaluate' (r_s = .23, 95% CI = .05; .39).

4.4 DISCUSSION

The current study examined three issues that are important to how soccer scouts identify talented players. Based on self-report data, we analyzed at which players' age soccer scouts perceive they can predict future performance; what attributes they consider to be important for future performance; and to what extent they score and combine assessments on these attributes in a structured manner.

Our results showed that the average age at which scouts perceive they could predict performance increased depending on the age cohort in which they scouted. The average age to predict performance fell within the age interval of these cohorts, with the exception of the largest cohort, that is, the U12 and younger cohort. Notably, there was a discrepancy between the player's age at which many scouts in this cohort perceived they could reliably predict future performance (i.e., 13.6 years old, on average), and the actual age at which they scouted players (i.e., younger than 12 years old). This finding suggests that these scouts are aware of the idea that early indicators of later performance are often lacking or hard to predict (Abbott et al., 2005; Den Hartigh et al., 2016). Yet, scouts do assess and advise on selection of players at younger ages.

One explanation for this discrepancy is that – given the difficulty of predicting future performance directly – those who scout in the younger age cohorts may be more concerned with finding the best current player, rather than finding the best player for the future (Ford et al., 2020). However, given that clubs invest substantial resources in developing these players over non-selected players, this approach seems to rely on the assumption that the best current young players are also those that have the highest potential for excellence in the future. It should also be noted that this assumption implies an inconsistency of thought: scouts are still indirectly making a prediction when assuming that the best current players are also the ones with the highest potential. Moreover, since the attributes needed for excellence are often unstable, develop non-linearly over time, and may not even be present in young players (Abbott et al., 2005; Den Hartigh et al., 2016; Simonton, 1999), selecting the best current players at a young age could harm the selection process. In sum, the finding that many scouts do not perceive they can predict performance for the players they scout, raises questions about the early (i.e., pre-pubertal) talent identification process (Güllich & Cobley, 2017).

Furthermore, although there seemed to be no apparent differences between scouts in the different age cohorts, we showed that scouts across cohorts consider a multidimensional range of soccer-related attributes when predicting performance. The five most frequently named attributes covered four major performance categories: technical skills or technique with the ball (i.e., technical) game sense and awareness (i.e., tactical and perceptual-cognitive skills), physiological or motor skills and sprinting speed (i.e., physical, physiological, and motor skills), and winning mindset or mentality (i.e., personality-related and mental skills). When examining the general performance categories, scouts mainly reported considering attributes in the technical performance category as predictors of future soccer performance. This was followed by tactical and perceptual-cognitive skills, physical, physiological, and motor skills, personality-related and mental skills, which were considered most important approximately equally often. The emphasis on technical attributes is encouraging, as these attributes have been shown to have relatively good predictive value in match play (Bergkamp et al., 2020), and in specific technical tasks where they may be less influenced by maturational timing (Murr, Feichtinger, et al., 2018; Vandendriessche et al., 2012).

On the other hand, the relative importance given to physical, physiological and motor skills differs from findings by Larkin and O'Connor (2017), Roberts et al. (2019) and Jokuschies et al. (2017). For instance, sprinting speed was a frequently named attribute in our sample (named by 36 of the 125 scouts), but was excluded from the final list (together with agility and strength) by Larkin and O'Connor (2017), because it was not considered important enough by the coaches and scouts. It can also be argued that the tendency of clubs to systematically select older or more mature players indicates that scouts (implicitly) consider physical attributes as most important in practice. The emphasis on physical and physiological attributes in this way can be particularly problematic for young players, because of the large inter-individual differences that result from maturity status and relative age, which reduce after puberty (Deprez, Buchheit, et al., 2015). Therefore, both biological and calendar age need to be taken into account when assessing the physical and physiological attributes of young players (Meylan et al., 2010).

Interestingly, scouts generally indicated that they predict performance by assessing the attributes in a structured manner. A majority of scouts indicated to a) always, or very frequently, evaluate different players – of the same age and playing position – on the same attributes, b) know which attributes they would assess before observing a player, and c) evaluate different attributes separately. These aspects are important for maintaining high levels of inter and intra-rater reliability when assessing performance, and are therefore encouraging (Huffcutt & Arthur, 1994). However, there are three remarks regarding this finding.

First, while scouts claimed to systematically assess players on different attributes, it remains an open question how well they define those attributes, and if they do this explicitly or implicitly. It appeared that scouts often placed general domains on the attribute list (e.g., technical skills or technique with the ball) while fewer provided specific examples of skills and abilities that belonged to those domains. Thus, scouts may have had difficulty verbalizing in detail what attributes they considered important predictors of future performance, which suggests that they implicitly integrate various attributes in their mind. This would be in line with the way coaches identify talent (Christensen, 2009; Johansson & Fahlén, 2017), and is an indication of the holistic approach to predicting performance (Dana & Rick, 2006). For example, it is likely that skills and abilities considered to belong to 'technique,' such as passing, dribbling, tackling, differ from scout to scout. Consequently, when assessing technique in this way, it may affect the reliability within and between scouts (cf. Chapman & Zweig, 2005).

Second, most scouts combined their assessments into an overall assessment based on their overall impression, as opposed to a sum of the independently assessed attributes. While predictions based on combining attributes according to a decision rule (e.g., summing scores on attributes) have been shown to outperform predictions based on overall impressions and intuition in holistic approaches (Arkes et al., 2006; Kuncel et al., 2013), the latter are commonplace across selection contexts (Dana et al., 2013). Therefore, it is not surprising that scouts in this study also applied this approach. Nevertheless, the predictive validity and reliability of scouts' performance assessments may improve further if they use a decision rule to combine information (i.e., mechanical or actuarial judgment, see Den Hartigh, Niessen, et al., 2018; for an explanation outside sports Meijer et al., 2020).

Finally, the low correlations between the statements suggest that scouts did not uniformly apply all aspects of structure. For example, most scouts who knew beforehand which attributes they were going to assess did not also evaluate different players – of the same age and position – on the same attributes, or evaluate each attribute separately. Thus, different scouts applied different aspects of structure, whereas literature suggests that predictions may become more consistent if scouts apply all aspects (Chapman & Zweig, 2005).

4.5 LIMITATIONS AND CONCLUSION

The main limitation of this study is that it assessed the talent identification process of scouts through self-report. This carries the risk that respondents are constrained in their self-knowledge (Paulhus & Vazire, 2007) or provide socially desirable responses. Including qualitative data could have provided additional insights into why scouts hold the perceptions that were found in this study and whether these align with what scouts do in practice (cf. Larkin et al., 2020; MacMahon et al., 2019; Roberts, Greenwood et al., 2019). For example, in-depth interviews or think-aloud protocols could reveal what type of player scouts generally are selecting for (i.e., best player available or best long-term prospect), and their perception on how these selection strategies relate to each other (cf. Reeves et al., 2019). Additionally, observing scouts in practice could show to what extent their perceptions of applying structure align with what they actually do. Finally, an interesting avenue for future research is to consider the reliability and validity of scouts' judgments. In such a design it would be necessary to collect the predictions of scouts and relate these to the future performance of players longitudinally (e.g., see whether players they picked actually reached the professional status).

A second possible limitation concerns the lack of detail in the predictors considered by scouts. This lack of detail may relate to the instruction in the questionnaire, as we did not want to steer scouts in a specific direction in section two of the questionnaire. Therefore, scouts were free to describe predictors in any way they wished, which resulted in varying levels of specificity for the attributes described. A final limitation is that we measured different aspects of structure using single item-scores, for brevity purposes. However, this meant that we were not able to compute reliability estimates over these items. Future research should consider measuring different aspects of structure with multiple items to compute reliability estimates (Chapman & Zweig, 2005).

The current study concludes the following regarding the process of talent identification in soccer scouts. First, previous literature has shown that early indicators of later performance can be unreliable (Den Hartigh et al., 2016; Simonton, 1999). In line with this literature, we showed that most scouts who observe younger players (i.e., U12 and younger) perceive they cannot reliably predict performance for the players they typically scout. Accordingly, we recommend that soccer organizations invest in the continuous (de)-selection of players across all age cohorts, and consider targeting post-pubertal players more often than is currently the norm (Güllich, 2014).

Second, considering the predictors that scouts say they find relevant, they value a multidimensional collection of attributes, but mostly account for general technical soccer attributes. Additionally, they seem to have difficulty formulating specific predictors of performance and likely integrate various attributes in their mind. Third, scouts report adopting a generally structured approach when scouting players, but do not apply the different structuring approaches uniformly, and mainly use their overall impression of the attributes to form their final predictions (i.e., holistic assessment). Given previous literature demonstrating that predictions based on overall 'intuitive' impressions are non-optimal in terms of reliability and validity, we recommend that scouts are trained in a more consistent use of the different aspects of structure when predicting performance. For instance, soccer organizations could create more opportunities for scouts to train themselves in formulating specific predictors of future performance, and to systematically score and combine these predictors according to a decision-rule (Den Hartigh, Niessen, et al., 2018). We believe these recommendations will improve the reliability and predictive validity of scouts' predictions in the future.

ACKNOWLEDGEMENTS

We would like to thank Maurice Hagebeuk and the other Talent Performance Coaches of the KNVB's Jeugdplan Nederland for their helpful suggestions regarding the conceptualization of the survey and for their help in recruiting scouts as participants.

CHAPTER 5

EXAMINING THE RELIABILITY AND PREDICTIVE VALIDITY OF PERFORMANCE ASSESSMENTS BY SOCCER COACHES AND SCOUTS:

THE INFLUENCE OF STRUCTURED COLLECTION AND MECHANICAL COMBINATION OF INFORMATION

This chapter is based on the manuscript:

Bergkamp, T. L. G., Meijer, R. R., den Hartigh, R. J. R., Frencken, W. G. P., & Niessen, A. S. M. (2022). *Examining the reliability and predictive validity of performance assessments by soccer coaches and scout: the influence of structured collection and mechanical combination of information.* Manuscript submitted for publication.

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-rate 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.

5.1 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, Niessen, et al., 2018; 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 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).

5.1.1 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 (Dawes et al., 1989; Kahneman et al., 2016; Karelaia & Hogarth, 2008). 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 (Roberts et al., 2019b). 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 (Roberts et al., 2019b). 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., 2021).

5.1.2 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., 2021).

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, Niessen, 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).

5.2.3 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 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.

5.1.4 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:

H_1: Structured-mechanical performance ratings yield the highest inter-rater reliability, followed by structured-holistic ratings, followed by unstructured-holistic ratings.

H_2: Structured-mechanical performance ratings yield the highest predictive validity, followed by structured-holistic ratings, followed by unstructured-holistic ratings.

5.2 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.

5.2.1 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.

5.2.2 MATERIALS AND MEASURES

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 were 1) all full backs 2) younger than 23 years old at the time and 3) had played at least 10 full 90-minute 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.

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). While estimated market values are influenced by a multitude of factors, we considered these estimates an adequate proxy for players' performance, as research has shown that they are strongly correlated with on-field technical soccer performance (Müller et al., 2017), expert ratings of soccer performance (Herm et al., 2014), and actual transfer fees (Torgler & Schmidt, 2007). 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.

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., 2021, Larkin & O'Connor, 2017; Roberts et al., 2019) and in collaboration with the KNVB (see Table 5.1).

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.

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 unstructured-holistic rating.

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., .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.

Table 5.1 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). 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.

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 α

= .68; structured-mechanical α = .83; Structured-holistic α = .84; No-observation α = .81).

5.2.3 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. The no-observation condition was the final condition in both versions. Participants were randomly allocated to a version (See OSF preregistration, section 2.4, 'randomization').

After they provided consent and answering 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 5.1). 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.

5.2.4 STATISTICAL ANALYSIS

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.

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 (rs) between the performance ratings from each assessment condition and players' market value in the 2018-2019 season. We assessed whether the difference between two coefficients was statistically significant using the method for dependent correlation coefficients – common index - described by Steiger (1980).

5

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.

Model of participants' structured 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, 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. 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).

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.

5.3 RESULTS

5.3.1 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 = .14, 95% CI = -.04; .39), followed by the structured-holistic rating (ICC = .07, 95% CI = -.09; .31) and the structured-mechanical rating (ICC = .04, 95% CI = -.11; .27). Because the differences were not in the expected direction, we did not test the ICC differences for statistical significance.

5.3.2 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 = .31, 95% CI = .11; .48, $p < .01$), followed by the structured-mechanical rating (r_s = .25, 95% CI = .06; .43, $p = .01$) and the structured-holistic rating (r_s = .22, 95% CI = .02; .40, $p = .03$). 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 difference = .03, $p = .38$).

5.3.3 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 = .41, 95% CI = .22; .56, $p < .01$), followed by predictions from the unstructured condition (r_s = .38, 95% CI = .19; .54, $p < .01$) and the no-observation condition (r_s = .25, 95% CI = .05; .43, $p < .01$). Differences in correlation coefficients between the no-observation condition and the two other assessment conditions were small and not statistically significant (see Table A5.1, in the appendix). Hence, we found no evidence that observing soccer players in games hurt or helped validity, but the differences point more towards 'helps' than 'hurts.'

5

5.3.4 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 = .53, R^2_{adj} = .49, $F(7, 88) = 14.26$, $p < .01$; see Table A5.2 and A5.3 in the appendix for the regression results and correlation matrix, respectively). Figure 5.1 presents the relative importance of each predictor in explaining the variance in participants' 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

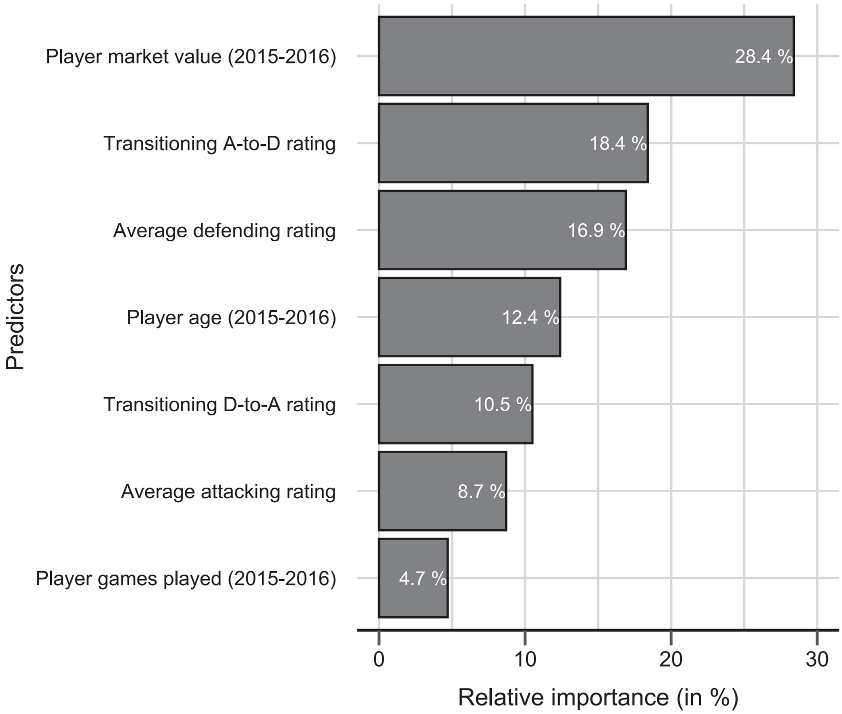


Figure 5.1 Relative importance of each predictor in explaining participants' 2018-2019 market value prediction. Note: Relative importance is scaled to sum to 100%;

rating (contribution = 18.4%) and average defending rating (contribution = 16.9%) having the largest contribution.

5.3.5 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 < .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 = .62) than the mean of the unstructured-holistic (M = 3.23, SD = .55), structured-mechanical (M = 3.16, SD = .51), and structured-holistic approach (M = 3.29, SD = .45). Comparisons among the other assessment approaches did not differ significantly (see Table A5.4 in the appendix).

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 < .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 = .83), structured-mechanical (M = 3.11, SD = .81), and structured-holistic approach (M = 3.30, SD = .68). Comparisons among the latter three assessment approaches also did not differ significantly (see Table A5.5 in the appendix).

5.4 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.

5.4.1 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 reliability and the predictive validity of the ratings overall were poor and small-to-moderate, respectively.

The absence of systematic differences in reliability and predictive validity 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 = .14 compared to ICC = .15 by Arkes), the ICC of the structured-mechanical rating was much smaller (ICC = .04 compared to ICC = .31). Furthermore, 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 of the same player. 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; Roberts 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 H1 and H2. 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.

5.4.2 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 = .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.

5.5 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 minutes), and it took participants' approximately 1.5 to 2 hours 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 sport 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.

5.6 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.

ACKNOWLEDGEMENTS

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.

CHAPTER 6

STRUCTURING PERFORMANCE ASSESSMENTS IN SOCCER:

A CASE STUDY AT A PROFESSIONAL YOUTH SOCCER ACADEMY

This chapter describes a practically oriented case study and was based on (in Dutch):

Bergkamp, T. L. G., Niessen, A. S. M., Hartigh, den, R. J. R., Meijer, R. R., & Frencken, W. G. P. (2020). (On)terecht buitenspel gezet. Sportprestaties voorspellen door systematische en gestructureerde beoordelingen. *SportGericht*, *74*(4), 36 - 40.

ABSTRACT

An important finding from selection psychology is that predictions are more reliable and accurate when based on structured assessment and combination of information through a decision-rule (i.e., mechanical combination), than when based on intuition or general impressions (i.e., holistic combination). A Dutch professional soccer club has recently applied this principle when assessing youth players. In this article, we show that assessing the performance of youth players is very difficult, as assessments by soccer coaches yielded low inter-rater reliabilities that were insufficient to accurately predict performance. However, coaches' assessments became somewhat more reliable when they used a structured assessment procedure paired with mechanical combination of information.

6.1 INTRODUCTION

Imagine the following scenario. A youth coach and scout of a professional soccer club travel to a regional club. The coach provides a training and the club aims to select grassroots players based on the coach's and scout's observations. One of the players enters the pitch 10 minutes after the beginning of training. He is late, because the bus that he normally takes was cancelled, but raced on his bike in an attempt to make it on time. The youth coach – who is in the middle of organizing the training – is irritated by this 'excuse.' He doubts whether the player has the right motivation and mindset, and advises the club not the select him. The scout overhears the conversation, but does not question the player's motivation. Moreover, he observes an excellent soccer player. He believes the player has great potential and advises the club to select him.

Although the coach and scout should ideally arrive at the same assessment based on this training, this was not the case. This inconsistency is common in human decision-making and selection processes. In this paper, we describe strategies to improve performance assessments and illustrate this with a practical implementation in the selection of youth players from a regional soccer school.

6.1.1 NOISE IN SELECTION PROCESSES

Assessing soccer performance and selecting soccer players successfully is difficult. This is because assessment and selection are inherently tied to predicting future performance (Bergkamp et al., 2019; Den Hartigh, Niessen, et al., 2018). In psychology, many studies have been conducted on selection processes and performance predictions. Specifically, various studies examined the effect of different assessment approaches on the reliability and validity of performance predictions (cf., Conway et al., 1995; Dawes et al., 1989; Grove & Meehl, 1996).

One of the approaches to assess performance is based on the general impression of the decision-maker. This approach is the most common method to select players in sports, and hence to (implicitly) predict future sports performance (Johansson & Fahlén, 2017; A. H. Roberts et al., 2019). Using a general impression implies that decision-makers weight and combine the information on which they base their prediction 'in their head' to form their final assessment: they combine the information holistically and use their intuition, experience, or gut feeling (Dawes et al., 1989). However, human decision-makers are often not good at consistently weighting and combining information holistically. Research has shown that predictions made at different time points, but made by the same decision-maker and based on the same information, tend to differ substantially (Karelaia & Hogarth, 2008). Predictions made by different decision-makers, as in the example at the beginning of this paper, often differ to an even greater extent (Kahneman et al., 2016; Viswesvaran et al., 1996). An important reason for this inconsistency is that different decision-makers tend to include different performance indicators in their general impressions. Moreover, decision-makers are often strongly influenced by information that is not, or only weakly related to future performance, such as appearance, body language, or prejudices about a player (Dana & Rick, 2006; Dawes et al., 1989; Den Hartigh, Niessen, et al., 2018). In short, predictions based on the general impression of the decision-makers are susceptible to noise, and this can have important implications for the accuracy of the predictions (Kahneman et al., 2016).

6.1.2 STRUCTURE AND DECISION-RULES

Predictions by human decision-makers are never perfectly reliable. However, optimizing reliability is an important aim for those involved in the selection process, because reliability is a prerequisite for validity. How can coaches, scouts, and staff achieve this aim?

Psychological research showed that systematically scoring information through structured assessment and combining information mechanically through a decision-rule often yields better predictions (Ægisdóttir et al., 2006; Huffcutt & Arthur, 1994; Kuncel et al., 2013). This approach does not have to be complex and can be created by coaches, scouts, and staff with a simple step-by-step plan (see Figure 6.1; Den Hartigh et al., 2018; Kahneman, 2011; Meijer et al., 2020). For example, coaches use a structured assessment approach if they define and score separate performance indicators when observing players' performance. They combine information mechanically if they subsequently take the average or sum of the scores on the indicators.

In theory, this structured assessment approach paired with mechanical combination leads to more consistency among decision-makers, and therefore higher reliability (Arkes et al., 2006; Conway et al., 1995). The player in the introductory example would likely still receive an unfavorable rating on the indicator 'motivation' by the youth coach. However, by rating multiple performance indicators separately and combining the ratings according to predefined weights, the final assessments of the coach and scout would probably be more consistent. In addition, by rating pre-defined performance indicators separately, the coach and scout will assess the same indicators, on which they agreed that they are relevant. This reduces the tendency to include irrelevant information in their assessment.

Comparing structured assessment approaches based on holistic and mechanical combination of information is a new avenue in the field of sports (Den

Hartigh, Niessen, et al., 2018). Below we describe a setting where we applied a structured assessment approach at the youth scouting of a professional soccer club, and specifically examined the reliability in the predictions of soccer coaches. We should note that the primary aim of this example is to implement the theory described above in a practical setting in which selecting players is a challenge in itself. The aim of this paper is not to offer solutions for the fundamental issues that are inherent in the selection of (very) young soccer players (cf. Abbott et al., 2005; Breitbach et al., 2014; Güllich & Cobley, 2017).

6.2 IMPLEMENTATION IN PRACTICE

The example discussed in the intro roughly corresponds with the selection process at FC Groningen. This professional soccer club selects male youth players for the youngest youth team, Under-12, from regional soccer schools. The players train at their amateur club, but have an extra training with coaches from FC Groningen on Wednesdays and Sundays. For this field study, 19 head- and assistant coaches assessed the performance of 50 players during these training sessions, based on both their general impression and on separate performance indicators.

We developed an instrument with four indicators in collaboration with staff members of the club. The performance indicators were operationalized as attacking, defending, movement, and toughness (step 1 in Figure 6.1). Next, the staff decided on the importance of each indicator, which determined their weights.

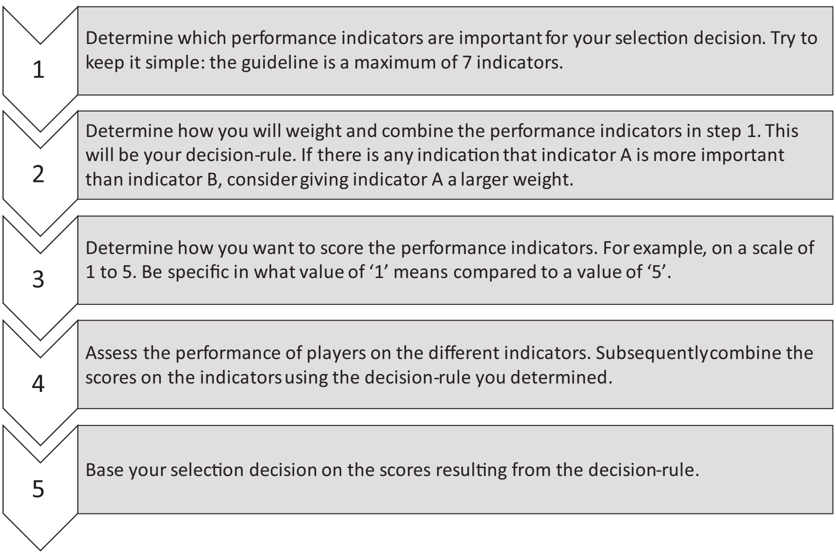


Figure 6.1 Step-by-step guide for a structured assessment approach paired with mechanical combination through a creating a decision-rule.

The first three indicators were considered equally important by the club, and were therefore given the same weight (step 2). In addition to rating these four indicators, the coaches also gave an ‘overall’ rating based on their general impression. We referred to this rating as the holistic rating. For each player, the holistic rating and separate indicator scores were provided on a 5-point scale (step 3). Specifically, for each performance indicator, the coaches were asked to ‘assess that a player’s future performance is best suited for a 1 = small amateur club, 2 = large amateur club, 3 = small professional club, 4 = medium professional club, 5 = large professional club (see Figure 6.2). ‘Toughness’ was evaluated with a pass or fail, because the club perceived this indicator to be equally important for each performance level. We took the average of the attacking, defending, and movement rating, and added .33 for a pass on toughness to arrive at our mechanical rating.

Over the course of 12 weeks, every player was assessed at 5 or 6 different moments by different coaches, which resulted in around 250 independent ratings. Coaches were not allowed to discuss players’ performance with each other, in order to collect ratings as independently as possible. To obtain an estimate of the reliability of the coaches’ predictions, we examined the inter-rater reliability of the ratings. We found a reliability estimate of .20 (95% Confidence interval, CI = .07; .37) for the holistic rating. This is a very low reliability, indicating that different predictions for the same player differed substantially. On the other hand, the reliability of the mechanical ratings was .27 (95% CI = .13; .43). This is still insufficient according to reliability guidelines, where a reliability of .8 is often considered acceptable (Koo & Li, 2016). However, it is higher than the reliability of the holistic rating based on general impression of the coaches. In other words, even in a complex practical situation, coaches’ predictions were somewhat more aligned when the ratings on attacking, defending, movement and toughness were combined via a simple decision-rule.

‘I assess that a player’s future performance is best suited for …’

	Small AC =	Large AC =	Small PC =	Medium PC =	Large PC =
Indication future club level →	**1**	**2**	**3**	**4**	**5**

Name player	‘Overall’	Attacking	Defending	Movement	Toughness
					P / F

Figure 6.2 Example of the rating instrument used by the coaches to assess performance. AC = Amateur club; PC = Professional club.

Another advantage of the approach was that it led to more transparency in the assessment process. Because a database of the coaches' ratings became available, we could analyze the relationship between the different performance indicators and the holistic ratings. A relative importance analysis showed that these ratings were most affected by the score on attacking (29%), followed by defending (22%), movement (20%), and toughness (11%). Thus, approximately 80 percent of the variance in the holistic rating could be explained on the basis of the performance indicators. This means that the instrument 'captures' a large part of what coaches observe and assess, but that a small part of their ratings included information that was not part of the model.

6.3 WHAT CAN WE LEARN FROM THIS?

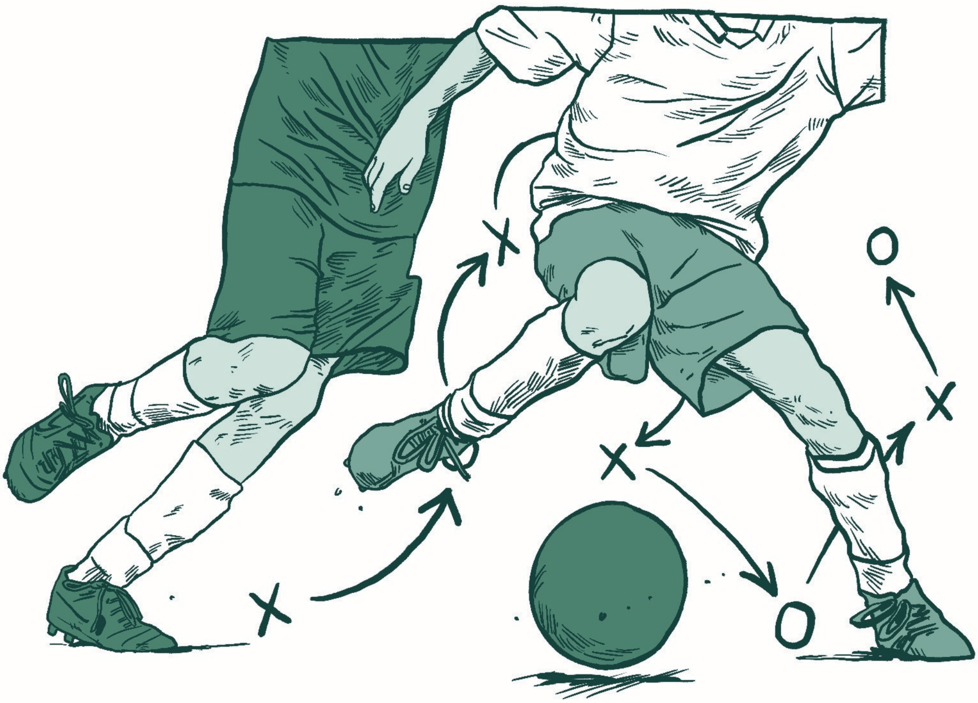
This practical implementation of structured assessment and mechanical combination has multiple implications. First, it shows the difficulty of soccer talent identification (Vaeyens et al., 2008). The overall low reliability suggests that the process of predicting future soccer performance for young players is characterized by a lot of noise (Güllich & Cobley, 2017). At the same time, the evaluation shows the importance of establishing different performance indicators for selection purposes. In this complex practical setting the predictions of coaches became somewhat more reliable by combining separate ratings on these indicators via a simple decision-rule. By defining, rating, and weighting separate performance indicators – instead of using the general impression of the coach – it became more transparent to evaluate which player fitted the club's philosophy the best. Coaches, scouts, and staff who typically have to account for a wide variety of performance indicators made this information explicit, ultimately resulting in better-informed selection decisions.

In addition, the evaluation shows decision-makers can play an important role in the development of a decision-rule. There is often a lot of resistance against the use of decision-rules or algorithms in selection contexts, because it may restrict the autonomy of the decision-maker (Nolan & Highhouse, 2014). Especially when the decision-rule is based on 'objective information' (such as in an optimal algorithm), the lack of subjective input is often considered a shortcoming by decision-makers. However, the input of the decision-maker was very important in this setting. Soccer coaches were involved in defining the performance indicators, observing the players, and providing feedback on the instrument. Thus, there is plenty of room for subjective contributions when using decision-rules, even in contexts that mainly rely on objective information, such as data scouting. The condition for less noise is that the criteria are defined beforehand and consistently weighted according to the step-by-step guide in Figure 6.1.

Despite the advantages of the structured assessment procedure, there is still much room for improvement. The inter-rater reliabilities of the holistic and mechanical ratings were insufficient to also make accurate predictions. This raises the question whether the current talent identification and development model that most clubs use is effective. In this model, many resources are invested in a small group of youth players, most of whom are unlikely to progress to the first team (Güllich, 2014). Therefore, it is interesting to think about alternative training and scouting models. For example, for several years Swedish club AIK Fotboll has been working with a model in which all amateur players younger than 14 years are welcome at the academy (De Hoog, 2020a). On the other hand, the selection and scouting of players – both young and older – remains something that most clubs must account for at some point in time. It is simply impossible to select all players for the first team. Given the advantages mentioned above, it seems better to do this based on a decision-rule, rather than the general impression of the decision-makers. That said, this method probably lends itself better to scouting or selecting older players for whom predictions of future performance are relatively easier to make (Den Hartigh, Niessen, et al., 2018; Kearney & Hayes, 2018).

A possible explanation for the low reliability is that the players were assessed at different time points by different coaches. That is, the coaches based their assessments on different observations, which likely influenced the reliability estimates. However, we deliberately opted for this design, because it accurately reflects soccer practice: predictions of players are often not based on one, but on series of observations. In addition, the performance indicators (e.g., ‘attacking’) were relatively ‘broadly’ defined, which means that there was room for interpretation among coaches. Further specifying or expanding the performance indicators (e.g., splitting up ‘attacking’ into ‘attacking with the ball’ and ‘attacking without the ball’) may possibly further reduce the noise between coaches. However, this may also negatively affect the accessibility of the instrument. FC Groningen will take these elements into account when aiming to improve their structured assessment process.

Finally, to return to the example which started this paper, being late for practice may be included in selection decisions. However, decision-makers should agree in advance which indicators matter and which do not, and how much weight each indicator should receive. It would be a shame if the club misses a future star because of the disagreement between coaches and scouts.



CHAPTER 7

DISCUSSION

7.1 DISCUSSION

The central question of this thesis was how soccer performance could be predicted reliably and validly. Specifically, this question was addressed through the lens of selection psychology. The field of selection psychology offers various principles that may improve our understanding of predicting soccer performance, enhance reliability and predictive validity, and ultimately lead to more robust research practices for studying soccer selection processes (Den Hartigh, Niessen, et al., 2018). These principles from selection psychology also fit well with recent developments in the field of sport sciences on the dynamic person-environment relations that give rise to soccer performance (Duarte et al., 2013; Vilar et al., 2013). Principles from selection psychology are, therefore, very relevant to the study of soccer selection- and talent identification, but have hardly been considered in this field.

In this thesis, I was specifically interested in principles that may optimize performance predictions in terms of 1) which predictors are used and 2) how information on predictors of future performance is collected and combined. Regarding predictors, research from selection psychology showed that high-fidelity, sample-based tests that mimic the criterion performance are often good predictors of future performance. This is particularly the case in homogenous (i.e., preselected) samples, such as elite soccer players (cf., Lievens & De Soete, 2012b; Sackett et al., 2017). In the context of soccer, a commonly used training format such as small-sided games (SSG) could serve as such a high-fidelity predictor. Therefore, the first aim of this thesis was to examine the predictive validity of small-sided game (SSG) performance.

With respect to information collection and combination, research from selection psychology repeatedly showed that structured collection of information and mechanical combination through a decision-rule outperform unstructured information collection and intuitive, holistic combination (Dawes et al., 1989; McDaniel et al., 1994). These principles apply to predictors which typically result in quantitative scores (e.g., standardized tests), but also to predictors which can be quantified by decision-makers (e.g., scouts and coaches) through assessments of observations (Arkes et al., 2006, 2010; Dana & Rick, 2006). Since soccer scouts and coaches regularly use their own assessments of observations to make selection decisions and predict players' performance (Jokuschies et al., 2017), the second aim of this thesis was to examine whether structured collection of information and mechanical combination optimized scouts' and coaches' assessments.

In the following section I will summarize the main findings of the different chapters in the thesis. Furthermore, I will reflect on these findings and the challenges of selection and prediction in a soccer context, and provide some suggestions for future research. Finally, I will describe some limitations of this thesis.

7.2 WHAT DID WE FIND?

7.2.1 PREDICTORS

Literature review

A large body of research aimed to explain and predict future soccer performance on the basis of different soccer performance indicators. Based on a thorough and systematic search of the literature, we reflected on the methodological quality of these talent identification studies, and provided several suggestions for improvement in chapter 2.

First, we observed that soccer performance was often operationalized as a categorical variable indicating performance level (e.g., elite vs non-elite). Because this hinders discrimination between players within selected and non-selected groups, we suggested to operationalize the criterion as individual, in-game soccer performance. An individual in-game performance criterion is more relevant for talent identification studies that aim to use the predictors to select the best soccer performer relative to others.

Second, talent identification research mainly focused on soccer skills and characteristics tested in isolation as predictors of performance level. Such predictors are defined as sign-based predictors (Bergkamp et al., 2019; Wernimont & Campbell, 1968). Yet the predictive value of these predictors in the context of soccer has generally been low-to-moderate and inconsistent (Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018). These isolated skills are relatively dissimilar to the criterion in terms of behavior, task, and contextual constraints (Phillips et al., 2010). Therefore, we suggested that talent identification studies examine the use of high-fidelity, sample-based predictors of performance that maintain the dynamic person-environment interaction, such as SSG performance (Bennett et al., 2018; Fenner et al., 2016; Van Maarseveen et al., 2017).

Third we observed that the talent identification literature rarely considered issues related to range restriction. Range restriction is an issue that occurs in soccer – and most other selection contexts – when the included sample is strongly preselected (i.e., homogenous) on the predictors of interest. As a result, predictor-criterion relationships obtained from such samples are usually underestimated (Sackett & Yang, 2000). We suggested to apply correctional formulas for range restriction when possible.

Fourth, high-level soccer typically deals with a (very) low base rate, as there are only a few players from the candidate pool that would be successful if no selection took place. A low base rate can significantly affect the utility of a predictor or selection procedure (Ackerman, 2014; Meehl & Rosen, 1955). We proposed to use an educated guess of (range of) base rate(s) to more explicitly examine its influence on utility in soccer.

SSG Performance

In chapter 3, we aimed to address some of the issues described in chapter 2. We examined the validity of individual SSG performance in predicting 11-vs-11 soccer performance. As suggested in chapter 2, we used a continuous measure of in-game performance by assessing different in-game performance indicators (e.g., passes forward, dribbles, interceptions) through notational analysis. We used these indicators to differentiate between individuals at the predictor and criterion level. We found a strong relationship between the SSGs and 11-vs-11 game formats in terms of the relative frequency with which different actions on the performance indicators were performed. Moreover, we found that individual performance in the SSGs yielded moderate-to-large predictive validities for individual performance in 11-vs-11 games, particularly for offensive performance. In contrast, typical physiological and motor skills tested in isolation yielded trivial-to-low predictive validities for 11-vs-11 performance. These outcomes suggest that a high-fidelity predictor that mimics the criterion performance in context and content enhances predictive validity over the physiological sign-based predictors, as suggested in chapter 2.

7.2.2 COLLECTING AND COMBINING INFORMATION

Survey on soccer scouts

While the interest in in the decision-making process of soccer scouts and coaches is rapidly gaining popularity (e.g., Lath et al., 2021; Roberts et al., 2019), little empirical research in this area existed when the studies in this thesis were drafted. Particularly, studies on how soccer scouts assess and select players were scarce (Larkin & O'Connor, 2017). Based on a large-scaled survey in chapter 4, we examined how Dutch soccer scouts identify talented players.

The survey yielded three main findings. First, soccer scouts who scout young players (i.e., U12 and younger) often reported that the player age on which they can reliably predict performance was higher (i.e., 13.6 years old) than the age category they scouted in. This suggests that scouts are aware that indicators of future performance may not be present in (very) young players (Abbott et al., 2005; Den Hartigh et al., 2016), but do still advise on selection of such players.

Second, we found that scouts considered a wide range of performance indicators as predictors of future performance, including specific and general technical (e.g., technique, passing ability), tactical (e.g., game sense and awareness, positioning) and physical and physiological (e.g., sprinting speed) indicators. Technical performance indicators were considered the most important. Notably, scouts did not often describe specific predictors of technical performance (e.g., passing ability), but simply named general indicators such as 'technique' as the most important predictor of future performance (cf. Roberts et al., 2019).

Finally, most scouts reported that they collected information at least in a somewhat structured manner, by (a) evaluating different players in the same position on the same performance indicators, (b) determining which indicators they would assess beforehand, and (c) evaluating different indicators separately. However, scouts did not apply these strategies in conjunction; different scouts applied different structuring strategies. Most scouts also indicated that they combined ratings on the different performance indicators based on their intuition or overall impression, rather than based on a decision-rule.

Although scouts' structured approach to assessment can be considered as positive, research from selection psychology describes how an increase in the degree of structure in information collection and the application of mechanical combination can improve scouts' predictions even further (Arkes et al., 2006; Dawes et al., 1989; Huffcutt & Arthur, 1994). Therefore, a logical next step in chapters 5 and 6 was to study how these principles affected the inter-rater reliability and predictive validity of performance assessments by soccer scouts and coaches.

Performance assessments of scouts and coaches

In chapter 5, soccer scouts and coaches observed soccer players' performance on video and rated their performance in an unstructured and structured manner. In the unstructured condition, participants gave a single, 'overall' performance rating (i.e., unstructured assessment with holistic combination), whereas in the structured condition participants rated eight specific performance indicators. We combined the ratings in the structured condition mechanically. Finally, participants also gave a single, overall rating based on their overall impression, after rating the distinct indicators in the structured condition This resulted in three types of ratings: unstructured-holistic, structured-mechanical, and structured-holistic rating.

Contrary to our expectations, we did not find that the structured-mechanical (ICC = .04) or structured-holistic ratings (ICC = .07) yielded larger inter-rater reliabilities and predictive validities than the unstructured-holistic rating (ICC = .14). The unstructured-holistic rating was slightly more reliable and predictively valid, but the differences were not statistically significant. Overall, the reliabilities of each type of rating were very low, meaning that participants did not agree in their assessment of the same players. Predictive validities were all small-to-moderate (.22 < r_s < 31).

The results of Chapter 5 did not align with prior research on the use of structured information collection and mechanical combination of information (Arkes et al., 2006; Wiesner & Cronshaw, 1988). The ambiguous findings are hard to interpret, which makes it difficult to formulate clear implications on how scouts and coaches can best apply structured information collection and mechanical combination of information. Although the study aimed to mimic a soccer scouting context, its ambitious experimental design may have made parts of the task overly complex for participants, which could have affected results. For instance, the experiment was relatively long, the predictors were formulated in a detailed and complex way, and scouts and coaches from diverse organizations participated in the study.

In chapter 6, we adopted a simpler approach to study structure and mechanical combination within a professional soccer club. More specifically, we asked soccer coaches of a professional academy to rate players in a structured way. In contrast to chapter 5, only five, broadly defined performance indicators were assessed: attacking, defending, movement, toughness (i.e., these four indicators were combined into a mechanical rating) and 'overall performance potential' (i.e., a holistic rating). All coaches were associated with a single club and were made familiar with the rating sheet through a presentation before each practice.

Similar to chapter 5, we found that the overall reliabilities of the ratings were poor, but the reliability of the mechanical rating (ICC = .27) was somewhat larger than the holistic rating (ICC = .20). Compared to chapter 5, the reliability of the mechanical rating in chapter 6 was also substantially larger, although this difference should be interpreted with caution given the small sample size and differences in design. Although the reliabilities of both ratings were insufficient to make valid predictions, these findings carefully suggest that a mechanical assessment procedure may improve the reliabilities of soccer coaches' assessments, under the condition that these practitioners are made familiar with the procedure beforehand and are associated with the same club or organization.

7.3 WHAT CAN WE LEARN FROM THESE FINDINGS?

Below, I will reflect on the findings in this thesis and discuss some challenges for future research with regards to 1) measuring in-game performance, 2) the use of samples-based and sign-based predictors, and 3) value of structure and mechanical combination.

7.3.1 MEASURING IN-GAME PERFORMANCE

In chapter 1, I defined soccer performance as “all observable and measurable actions, behaviors, and outcomes that soccer players engage in and which contribute to the team’s tasks within a soccer game (p.7).” Thus, in line with my suggestion regarding the operationalization of the criterion in chapter 2, this definition prioritizes individual performance within soccer games. It includes specific performance categories, such as physical, physiological, technical, tactical, and psychological skills and abilities (as reflected in the scouts’ answers in chapter 4, Williams & Reilly, 2000). At the same time, similar to task performance in jobs, I expected that these performance categories share common variance and that there is a ‘general’ soccer performance factor (Kharrat et al., 2019; Pappalardo et al., 2019; Viswesvaran et al., 2005).

I believe it is worthwhile to operationalize the criterion as individual soccer performance after a selection decision, rather than the selection decision itself (e.g., elite vs. non-elite players). If the aim of soccer research is to use the variables or procedures under study to inform selection decisions, then in-game soccer performance is, and should be, the outcome of interest (Wilson et al., 2017). More so than the selection decision, an in-game criterion caters to the complexity of soccer performance, as it maintains the ongoing interactions between performers and their environment (Travassos et al., 2013). In this sense, the field of soccer performance predictions can continue to draw from selection psychology on the operationalization of the criterion.

I proposed and used different operationalizations of individual in-game soccer performance in this thesis. These included manually notating the quality and frequency of in-game performance indicators, and combining these into an overall attacking and defending performance measure (i.e., chapter 3). Moreover, I used ratings by soccer scouts and coaches on relatively specific (i.e., chapter 5) and general (i.e., chapter 6) performance indicators. The relevance of these indicators was derived through careful and structured analyses with soccer scouts and coaches, similar to a job analysis (Hough & Oswald, 2000). Finally, I used individual market value as a proxy for soccer performance (i.e., chapter 5), as this is strongly related to the in-game performance indicators used in chapter 3 (Müller et al., 2017).

Although the performance operationalizations differed in terms of the included performance indicators, they corresponded most closely to what can be defined as technical-tactical performance. This type of performance is particularly relevant, as it is related to game success (i.e., game outcome; Pappalardo & Cintia, 2017). Furthermore, I would argue that these operationalizations included the complexity of soccer performance to a large extent, but were relatively simple in concept.

That said, defining a reliable and valid individual in-game performance criterion remains a challenge. Notational analysis is time consuming and requires a difficult decision on which performance indicators to include and which to exclude (Travassos et al., 2013). Furthermore, the (structured) performance ratings yielded low inter-rater reliabilities. The difficulty also lies in the fact that soccer is a fluid and dynamic sport that is not characterized by a series of discrete events (Travassos et al., 2013). Individual in-game performance emerges from functional interactions between players and the performance environment (Vilar et al., 2012). Moreover, game-to-game soccer performance can be highly variable (Rampinini et al., 2007), and in-game success can be achieved in different ways because of the interaction between performance dimensions (e.g., technical, tactical, physical, Den Hartigh, Hill, et al., 2018; Travassos et al., 2013).

Given the challenges described above, is it more difficult to operationalize the criterion of interest in soccer than in other performance domains? It might be more difficult compared to higher education, where GPA is generally available as a straightforward and highly reliable measure of performance (Beatty et al., 2015). In addition, personnel selection deals with similar challenges surrounding the operationalization of the criterion, such as the relatively low inter-rater reliabilities of supervisory job performance ratings (Salgado & Moscoso, 1996; Viswesvaran et al., 1996). However, the reliabilities of job performance ratings are generally still substantially higher (i.e., ICC = .52) than those for soccer performance reported in chapters 5 and 6. Finally, operationalizing performance in soccer is definitely harder compared to sports where performance is expressed in terms of racing times or distances, such as swimming or track and field (Kearney & Hayes, 2018; Mitchell et al., 2018) or team-based sports that have more discrete possessions compared to soccer, such as basketball and baseball. These team-sports allow for the computation of impact or efficiency measures for individual players (Sill, 2010; Tango et al., 2007).

In sum, the operationalizations of in-game performance in this thesis do not yet suggest a new 'norm' or 'gold standard' to measure in-game performance reliably and validly. Yet, and more importantly, they do demonstrate how a criterion that differentiates effectively between individuals after a selection decision corresponds more closely to what we aim to predict. These operationalizations then allow for statements on the prediction of individual soccer performance, thereby yielding more meaningful predictor-criterion relationships.

7.3.2 SAMPLE- AND SIGN-BASED PREDICTORS

The development of reliable in-game soccer performance measures does not only offer opportunities for operationalizing the performance criterion. Similar measures can also be employed as predictors of soccer performance in future studies. For instance, chapter 3 showed that performance in 7-vs-7 SSGs can be a good predictor of soccer performance in 11-vs-11 games.

As demonstrated in Chapter 3, we found larger predictive validities for SSG performance than for the sign-based physiological and motor indicators. An explanation for these findings from the field of ecological dynamics in the sports sciences is that SSG performance can be considered a predictor based on a representative design (Pinder et al., 2011). The ecological dynamics approach posits that soccer performance emerges through the continuous interaction between the performer (e.g., traits and abilities) and environment (e.g., the presence of moving opponents and teammates and the task to score goals; Davids et al., 2013). This interaction – and thereby the coupling between perception and action – should remain intact in the predictor context and content, resulting in a design that is representative of the criterion context and behavior (Davids, Araújo, Correia, et al., 2013; Pinder et al., 2013). Physiological and motor performance and SSG performance were used as predictors of a 11-vs-11- performance, but since the action-perception coupling only remained intact in the SSGs, this may have yielded larger predictive validities for this predictor. Another possible explanation for the findings in chapter 3 is that the sample of players in the study was relatively homogenous, meaning they were preselected on the physical and motor indicators. As described in chapter 2, this may have resulted in lower predictive validities for the physical predictors. Accordingly, chapter 3 then provides an empirical example of how sample-based predictors can be valuable for homogenous samples of players (Sackett et al., 2017).

That said, sign-based predictors can also have their value in particular situations. For instance, signs can be effective in heterogeneous samples when athletes have not been preselected on the predictors of interest. It is likely that even an isolated test to assess motor ability or dribbling performance has predictive value in a sample of post-pubertal players containing recreational-amateur to elite levels players (Murr, Feichtinger, et al., 2018). Moreover, signs can be effective for sports in which successful performance relies more on the traits and skills “inside” the athlete than on their continuous adaptation and interaction with the performance environment. For example, height and weight were found to be good predictors of rowing and swimming performance (Mitchell et al., 2018; Schranz et al., 2012).

Nevertheless, given the logical fit between the value of representative designs in sports and the use of sample-based predictors in homogeneous groups, future research should build upon chapter 3 in developing in-game performance measures at the predictor (and simultaneously at the criterion) level.

7.3.3 THE VALUE OF STRUCTURE AND MECHANICAL COMBINATION

In chapters 5 and 6 we found mixed results regarding the benefit of structured information collection and mechanical combination of information on performance assessments by scouts and coaches; while we did not find differences in the expected direction in chapter 5, the results in chapter 6 indicated higher reliability when using a structured, mechanical approach, although the difference was small and the overall reliability still very low. Given the strong evidence of the benefit of these strategies in personnel selection (Conway et al., 1995; Grove et al., 2000; Kuncel et al., 2013), these findings were unexpected. The marginal differences between approaches only allow us to speculate. Yet, possible explanations for these findings are that the coaches in chapter 6 were part of the same organization and were provided with a presentation on the use of structured assessment before each practice. In contrast, participants in chapter 5 were part of many different organizations, were not as familiar with the procedure and the indicators they were asked to assess, and participated in the experiment online. Therefore, it might be beneficial for future research to recruit participants from the same organization, provide a training beforehand, and monitor their assessments 'live'.

The findings in the chapters also suggest that structured collection and mechanical combination of information may be more difficult to implement in assessing in-game soccer performance, compared to (for example) hiring interviews used in personnel selection. As discussed in chapter 5, the dynamic nature of soccer implies that the 'tasks' that each player encounters are not standardized. Therefore, they are not consistent across games or players, which makes increasing the level of structure by assessing in-game performance on a narrower task level and developing explicit benchmarks (i.e., as can be done in hiring interviews) very difficult.

Despite these difficulties, I believe it is worthwhile for research to continue to examine how these strategies can improve the assessments of scouts and coaches. The main argument is that the task of predicting soccer performance does not satisfy the conditions for the alternative approach (i.e., unstructured-holistic assessments) to function well. Specifically, Kahneman and Klein (2009) discussed the conditions under which skilled intuitive (i.e., unstructured-holistic) judgments can arise. They stated that skilled intuitive judgment requires an environment that 1) is highly predictable, in the sense that highly valid cues are available, and 2) offers decision-makers immediate, clear and complete feedback on their judgments and predictions, which is necessary to learn what the valid cues are. However, at the moment, there are generally no highly valid cues available for soccer scouts and coaches to predict performance, and feedback on predictions is typically not obtained immediately (i.e., in the case of scouting youth players, feedback is obtained many years later). In fact, it is likely that Kahneman and Klein (2009) themselves would not consider soccer to satisfy the conditions for skilled intuitive judgments, as they listed baseball – a sport arguably less noisy than soccer – as an environment which is 'insufficiently regular' or in which practitioners have not mastered the valid cues.

While chapter 5 suggests that scouts and coaches can extract some valid information when they assess soccer performance, it is unlikely that these cues are as 'obvious' and strong for practitioners to consistently rate them and make accurate performance predictions, respectively. If that was the case, we should have observed much higher inter-rater reliabilities and predictive validities in chapters 5 and 6. Future research should therefore aim to find effective ways to use structured information collection and mechanical combination of information in soccer. Increasing the level of structure while accounting for the dynamic nature of soccer is challenging, but given the evidence base in selection psychology it is plausible that these approaches will ultimately yield superior soccer performance assessments over unstructured-holistic approaches.

7.4 IMPLICATIONS FOR PRACTICE

The main practical contribution of this thesis was that it provided a practical application of scientific assessment principles from selection psychology on predictors and the way information is collected and combined. For example, the thesis demonstrated how SSGs can be organized within a professional academy, while accounting for differences in the performance level of players by reorganizing the teams. It also showed how performance in SSGs and regular games can be measured, both through notational analysis and ratings of scouts and coaches. Paired with the finding that SSG performance can be a valid predictor of performance in 11-vs-11 games, the demonstration of these applications is valuable for clubs and coaches who aim to explore the use of SSGs as an assessment tool, instead of solely a training format (Unnithan et al., 2012).

This thesis also brought issues on the reliability and validity of current soccer selection procedures to the attention of the KNVB and various professional Dutch clubs. High-level soccer selection decisions are typically based on the assessments of scouts and coaches who use little-to-no structure and base their final assessment on their intuition or overall impression. Studies in this thesis that pointed out the limitations of this approach caught the attention of multiple soccer practitioners. As described in chapter 6, this led to efforts to integrate structured information collection and mechanical combination of information at a professional club. These examples show how practitioners can use the principles in this thesis in a practical setting.

7.5 LIMITATIONS

The studies in this thesis have several limitations. While the sample sizes are relatively large for soccer research in sports, the studies are still underpowered. For example, we recruited $n = 96$ scouts and coaches from high-level soccer organizations as participants in chapter 5. These participants are relatively difficult to recruit, as there are only a finite number of them in the Netherlands. Yet, this yielded a power of 64% for the experiment, far removed from the typical desired power level of 80% in social sciences. The small sample sizes can affect the stability and replicability of the results, although it was impossible to recruit more participants that met the inclusion criteria for the studies throughout this thesis.

There are also some limitations for each study that need elaboration. A limitation of chapter 3 on the predictive validity of SSG performance was that we did not include a technical performance indicator tested in isolation, such as dribbling or passing performance, as part of the sign-based tests (Huijgen et al., 2010). Although it was practically impossible to include such a predictor, this would have arguably resulted in a 'fairer' comparison of the sign-based and samples-based (i.e., SSG performance) predictors in terms of predictive value, as performance in the SSGs was operationalized in what is often considered 'technical' on-ball performance indicators (Klingner et al., 2021).

A limitation of the survey on soccer scouts was that we examined the tendencies of scouts to use structured assessment and mechanical combination at the item level. Expanding on these questions and examining whether they can be seen as specific dimensions would have resulted in more robust inferences. For example, Chapman and Zweig (2005) concluded that structure in hiring interviews was best described by four dimensions. It is interesting to explore whether these dimensions also translate to scouting in soccer.

In the chapters on structure and mechanical combination, chapter 6 did not include an unstructured-holistic rating, whereas chapter 5 did. Due to time constraints, coaches gave their holistic rating and structured assessment at the same time when assessing a player, and it was practically impossible for coaches to provide an additional holistic rating that was independent from the structured assessment. As a result, we do not know whether coaches provided an overall-rating, or ratings on the performance indicators first. Thus, the designs of chapters 5 and 6 differ in that the structured approaches are not entirely identical and chapter 6 did not include an unstructured-holistic rating.

7.6 CONCLUDING REMARKS

The present thesis examined soccer performance predictions through the lens of selection psychology. We studied the predictive validity of different types of predictors and investigated how information derived from observations of players can be collected and combined to make predictions. Clearly, the chapters in this thesis do not serve as an 'endpoint' on these topics. There are ample opportunities to explore how the use of a continuous in-game performance criterion, SSGs as sample-based tests, and structure and mechanical combination, can optimize soccer performance predictions. Specifically, I believe there are important avenues for combining these principles into an evidence-based 'toolbox' to be used for selection decisions in soccer.

For example, spatio-temporal data, collected by measuring players' time and position on the pitch, can be included in future operationalizations of individual in-game predictor and criterion measures (Frencken et al., 2010; Goes et al., 2021). Combined with event data from notational analysis, this allows the development of measures of performance that include the person-environment interactions to a larger extent (Travassos et al., 2013; Vilar et al., 2012). Resources to obtain spatio-temporal metrics are currently not widely available for grassroots or professional youth players, but technological advances suggest that it is only a matter of time before they become more accessible and can be used in talent identification studies more easily (Herold et al., 2019; Memmert et al., 2017). In the meantime, it is interesting to explore more easily available measures of performance, such as so called 'top-down' measures which assess players' impact on the outcomes of (small-sided) soccer games. These top-down measures could be used as a practical, but objective methods to assess individual in-game soccer performance (cf. Fenner et al., 2016; Wilson et al., 2021). Finally, future research can continue to examine the value of structured information collection and mechanical combination. Regardless of any technological advances, subjective judgments by soccer scouts and coaches will likely remain an important part of the soccer selection process in practice. Future research can find ways to combine subjective judgments with more objective data from (small-sided) games to make evidence-based selection decisions.

To conclude with a personal note, soccer is often referred to as the beautiful game. One of the things that makes it beautiful is that the performance and development of players often seems unpredictable. There is nothing more captivating than the story of an underdog – such as Virgil van Dijk – who beat all odds and became a world-class player against expectations. At the same time, such stories drive us to find aspects in performance that are predictive of the future; what if we could get a sense of what players will contribute in the future, even if it was only a glimpse? Thus, I also believe there is beauty in trying to solve the complex puzzle of making reliable and valid performance predictions, despite their inevitable imperfections. With this thesis I aimed to make a key pass, hopefully my colleagues in science and practice will contribute to scoring the goal.

SUMMARY (IN DUTCH)

REFERENCES

APPENDICES

DANKWOORD

CURRICULUM VITAE & LIST OF PUBLICATIONS

SAMENVATTING

Professionele voetbalclubs zijn continu op zoek naar de grootste voetbaltalenten. Gebaseerd op observaties van wedstrijden zoeken voetbalscouts en coaches naar spelers die de potentie hebben voor een carrière in het betaald voetbal (Jokuschies et al., 2017). Jeugdspelers die 'geschikt' worden bevonden worden vaak geselecteerd voor professionele opleidingen van de clubs (Till & Baker, 2020). Hier worden zij voorzien van uitgebreide voetbalinhoudelijke en fysieke trainingen, professionele verzorging, en high-tech materialen, met als doel om hun voetbalontwikkeling te stimuleren.

Zoals elk selectievraagstuk is de selectie van voetballers voornamelijk een voorspellingsvraagstuk. Scouts en coaches doen (impliciet) een voorspelling wanneer zij hun observaties gebruiken om in te schatten of een speler geschikt is voor het eerste team of een carrière in het betaald voetbal (Den Hartigh, Niessen, et al., 2018). Het maken van dit soort voorspellingen is echter erg moeilijk. Spelers worden regelmatig op (zeer) jonge leeftijd geselecteerd, wat betekent dat de voorspellingen een groot tijdsinterval beslaan. De wetenschappelijke literatuur zegt echter dat voorspellingen over langere tijdsintervallen steeds minder accuraat worden (Güllich, 2014; Vaeyens et al., 2008). Dit heeft er onder andere mee te maken dat de kenmerken en vaardigheden die een indicatie geven van toekomstige voetbalprestaties vaak nog niet aanwezig of ontwikkeld zijn bij jonge spelers (Baker et al., 2018; Den Hartigh et al., 2016). Modellen en methoden voor selectiebeslissingen in het voetbal krijgen dan ook veel aandacht in het maatschappelijk en wetenschappelijk debat (Abbott et al., 2005; De Hoog, 2020b).

Ondanks dat het voorspellen van voetbalprestaties moeilijk is, is het, gezien de mogelijk grote impact op spelers en clubs, belangrijk dat die voorspellingen zo betrouwbaar en accuraat mogelijk zijn. Daarnaast is de realiteit dat de meeste professionele voetbalclubs op een gegeven moment selectiebeslissingen moeten nemen; niet elke speler kan simpelweg in het eerste team spelen. Het is daarom belangrijk om te kijken hoe selectiebeslissingen geoptimaliseerd kunnen worden en hoe we 'evidence-based' methoden kunnen integreren in het selectieproces (Bergkamp et al., 2019). Met andere woorden; 'hoe kunnen voetbalprestaties betrouwbaar en valide worden voorspeld?' Dit was de centrale vraag van dit proefschrift.

Psychologisch onderzoek naar selectie (i.e., selectiepsychologie) biedt verschillende methoden en principes om deze vraag te beantwoorden (Hough & Oswald, 2000). Hoewel er in de sportliteratuur de afgelopen decennia veel aandacht is besteed aan het voorspellen van voetbalprestatie(niveau)s, ontbrak de toepassing van principes uit de selectiepsychologie. De principes zijn echter zeer relevant voor de sport. Zo hebben ze een natural fit met recente inzichten over hoe voetbalprestaties op het veld tot stand komen vanuit dynamische persoon-omgeving interacties (Pinder et al., 2011; Vilar et al., 2013). Specifiek licht ik twee gebieden uit waar principes uit de selectiepsychologie relevant zijn voor de voetbalpraktijk, namelijk 1) wanneer en welke soorten voorspellers goed werken en 2) hoe informatie het beste verzameld en gecombineerd kan worden om tot betrouwbaardere en accuratere voorspellingen te komen.

Naast de het meten van specifieke vaardigheden en eigenschappen, zegt de selectiepsychologie dat representatieve, sample-based testen die het relevante criteriumgedrag nabootsen vaak goede voorspellers zijn (Born et al., 2022; Robertson & Smith, 2001; Schmidt & Hunter, 1998). Deze sample-based benadering is gebaseerd op het principe van behavioral consistency: de beste voorspeller van toekomstig gedrag is soortgelijke gedrag in het verleden (Meehl, 1989). Daarnaast zijn sample-based testen vaak goede voorspellers in steekproeven waar de deelnemers zijn voorgeselecteerd op relevante eigenschappen (i.e., homogene steekproeven), zoals het geval is bij jeugdspelers in een professionele opleiding (Lievens & De Soete, 2012; Sackett et al., 2017). In de voetbalcontext zou een small-sided game (i.e., kleine partijvorm) mogelijk als een sample-based voorspeller gebruikt kunnen worden (Davids, Araújo, Correia, et al., 2013). Een small-sided game is een voetbalwedstrijd met minder spelers, van kortere duur, en gespeeld op een kleiner veld dan een reguliere 11-tegen-11 wedstrijd (Sarmento, Clemente, et al., 2018; Van Maarseveen et al., 2017). Onderzoek naar prestatievoorspellingen in het voetbal heeft echter nog weinig aandacht besteed aan small-sided games (Fenner et al., 2016; Unnithan et al., 2012). Het eerste doel van deze thesis was dan ook om de voorspellende waarde van voetbalprestaties in deze kleine partijvormen te onderzoeken.

Met betrekking tot het verzamelen en combineren van informatie laat de selectiepsychologie zien dat voorspellingen gebaseerd op gestructureerde informatieverzameling en ‘mechanisch’ gecombineerd (aan de hand van een beslisregel) vaak betrouwbaarder en accurater zijn dan voorspellingen gebaseerd op de algemene indruk van een beoordelaar (Conway et al., 1995; Dawes et al., 1989; Kuncel et al., 2013). Dit geldt voor voorspellingen waarin de informatie vaak kwantitatief is (i.e., test scores), maar ook waar beoordelaars zelf informatie moeten kwantificeren op basis van observaties (Ægisdóttir et al., 2006; Arkes et al., 2006; Dana & Rick, 2006). Coaches die spelers observeren maken bijvoorbeeld gebruik van gestructureerde informatieverzameling wanneer zij verschillende prestatie-indicatoren apart van elkaar beoordelen. Zij kunnen vervolgens de beoordelingen op de indicatoren mechanisch combineren middels een beslisregel, bijvoorbeeld door het gemiddelde of de som van de scores te nemen om tot een eindoordeel te komen (Den Hartigh, Niessen, et al., 2018; Meijer et al., 2020). Voorspellingen die op deze manier gemaakt worden zijn vaak betrouwbaarder en accurater dan voorspellingen gebaseerd op de algemene indrukken, omdat informatie consistenter verzameld en gewogen wordt (Dana & Dawes, 2004). Hoewel dit robuuste bevindingen zijn in de personeelsselectie en selectie voor het hoger onderwijs, zijn deze principes nog niet onderzocht in de sport. Het tweede doel van dit proefschrift was dan ook om te onderzoeken of gestructureerde informatieverzameling en mechanische combinatie de betrouwbaarheid en validiteit van voorspellingen van coaches en scouts verbeteren, ten opzichte van voorspellingen op basis van de algemene indruk.

HOOFDSTUK 2

In hoofdstuk 2 beschouwden we de talentidentificatieliteratuur door de lens van selectiepsychologie. Op basis van die beschouwing identificeerden we vier methodologische limitaties. Ten eerste observeerden we dat eerder onderzoek voornamelijk heeft gekeken naar welke factoren een onderscheid kunnen maken tussen voetballers van verschillende prestatieniveaus, zoals elite versus niet-elite spelers (Sarmento, Anguera, et al., 2018). Op deze manier kan er echter geen onderscheid gemaakt worden tussen spelers binnen hetzelfde niveau wat betreft voetbalprestaties op het veld (Wilson et al., 2017). Omdat het doel van talentidentificatie is om te voorspellen hoe goed voetballers zullen presteren ten opzichte van andere voetballers, stelden wij dat het gebruik van individuele voetbalprestaties op het veld een relevanter criterium is dan prestatieniveau.

Ten tweede vonden we dat de sportliteratuur voornamelijk heeft gekeken naar de voorspellende waarde van sign-based voorspellers. Dit zijn specifieke voetbal eigenschappen en vaardigheden gemeten in geïsoleerde testen, zoals sprintsnelheid, dribbel vaardigheden, en uithoudingsvermogen (Murr, Feichtinger, et al., 2018; Murr, Raabe, et al., 2018). De voorspellende waarde van deze voorspellers was echter laag-tot-middelgroot en inconsistent (Breitbach et al., 2014). Daarnaast lijken deze signs relatief weinig op het criterium wat betreft voetbalgedrag, taak, en context (Pinder et al., 2011; Renshaw et al., 2019). We stelden daarom voor om onderzoek te doen naar sample-based voorspellers – zoals small-sided games – die de dynamisch persoons-omgeving interacties behouden (Davids, Araújo, Correia, et al., 2013; Olthof et al., 2019).

Ten derde observeerden we dat de literatuur weinig rekening hield met range restriction. Range restriction komt voor wanneer participanten in de steekproef (sterk) zijn voorgeselecteerd op de relevante voorspellers. In dat geval worden relaties tussen de voorspellers en het criterium vaak onderschat (Sackett & Yang, 2000). Omdat de talentidentificatieliteratuur vaak steekproeven bevat van 'elite' spelers die (impliciet of expliciet) zijn voorgeselecteerd op de relevante voorspellers (e.g., sprintsnelheid, technische vaardigheden), is range restriction een veelvoorkomend probleem. We moedigden onderzoekers daarom aan om – waar mogelijk – correcties voor range restriction toe te passen.

Ten slotte hebben we bij selectie in het professioneel voetbal te maken met een (erg) lage base rate. Dit houdt in dat er slechts erg weinig spelers in poule van kandidaten (e.g., het amateurvoetbal) geschikt zijn om het niveau van betaald voetbal te halen (Ackerman, 2014). De lage base rate heeft een grote invloed op de effectiviteit van een voorspeller of selectieprocedure, wat betreft het percentage extra geïdentificeerde succesvolle spelers (Meehl & Rosen, 1955). We stelden daarom voor om weloverwegen schatting van de base rate te gebruiken om het effect van het gebruik van een selectieprocedure of voorspeller concreet te maken.

HOOFDSTUK 3

In hoofdstuk 3 hebben we een aantal suggesties uit hoofdstuk 2 geïmplementeerd in een empirische studie. Hier onderzochten we de voorspellende waarde van voetbalprestaties in 7-tegen-7 small-sided games – een sample-based voorspeller – voor prestaties in reguliere, 11-tegen-11 wedstrijden. Daarnaast onderzochten we de voorspellende waarde van een aantal geïsoleerde eigenschappen die veel in de sportliteratuur gebruikt zijn, namelijk sprintsnelheid, wendbaarheid, en uithoudingsvermogen. Voetbalprestaties in de kleine- en reguliere wedstrijden werden gemeten door het noteren van on-ball prestatie indicatoren, waaronder pass- dribbel-, en duel vaardigheden. Hierdoor konden we onderscheid maken tussen individuen op het niveau van de voorspellers en het criterium.

De relatieve frequentie waarmee de vaardigheden werden uitgevoerd in de small-sided games kwam sterk overeen met verdeling in de 11-tegen-11 wedstrijden. Dit suggereert dat 7-tegen-7 partijen representatief zijn voor 'echte wedstrijden' (cf. Olthof et al., 2019). Daarnaast vonden we middelgrote correlaties tussen de prestatie indicatoren in beide spelvormen, maar slechts zwakke correlaties tussen de fysieke eigenschappen en prestaties in de 11-tegen-11 wedstrijden. Deze resultaten suggereren dat prestaties in de kleine partijen een relatief goede voorspeller zijn voor prestaties in 11-tegen-11 wedstrijden, en dat een representatieve context de voorspellingen ten goede komt, zoals geopperd in hoofdstuk 2 (cf. Wilson et al., 2017).

HOOFDSTUK 4

In hoofdstuk 4 t/m 6 onderzochten we het besluitvormingsproces van voetbalcoaches en scouts; een onderwerp waar ten tijde van het plannen van deze studies nog erg weinig aandacht aan was besteed. In hoofdstuk 4 startten we daarom met een surveyonderzoek naar de perceptie van Nederlandse voetbalscouts op het talentidentificatieproces en prestatievoorspellingen in het voetbal.

Als eerste vonden we dat de leeftijd van spelers waarop de scouts dachten betrouwbare voorspellingen te kunnen maken en de leeftijd van spelers waarop scouts daadwerkelijk scoutten, niet overeenkwamen. Scouts in de Onder-(O)12 leeftijdscategorie geloofden pas vanaf 13.6 jaar betrouwbaar te kunnen voorspellen of een speler geschikt was oor een carrière in het betaald voetbal. Dit suggereert dat scouts zich bewust zijn dat indicatoren van toekomstige prestaties nog niet aanwezig zijn in (erg) jonge spelers, maar dat zij toch advies uitbrengen rondom de selectie van deze spelers (Abbott et al., 2005; Den Hartigh et al., 2016).

Ten tweede vonden we dat scouts een grote verscheidenheid aan prestatie indicatoren meenamen in hun beoordelingen, waaronder algemene en specifieke technische (e.g., techniek en passvaardigheden), tactische (tactiek en spelinzicht), en fysieke vaardigheden (fysieke voorwaarden en sprintsnelheid). Techniek of technische indicatoren werden daarbij het meest belangrijk gevonden. Scouts beschreven de indicatoren echter vaak in globale termen. Zo omschreven zij zelden specifiek technische voorspellers (e.g., pass intentie- of nauwkeurigheid), maar noemden simpelweg het woord ‘techniek’ als meest belangrijke voorspeller (cf. Roberts et al., 2019).

Ten slotte gaven scouts aan op een enigszins gestructureerde manier informatie te verzamelen door (a) verschillende spelers van dezelfde positie op dezelfde indicatoren te beoordelen, (b) voorafgaand te weten welke indicatoren zij gingen beoordelen (c) verschillende indicatoren apart van elkaar te beoordelen. Scouts pasten deze strategieën echter niet tegelijk toe, maar verschillende scouts gebruikten verschillende strategieën. Daarnaast gaven de meeste scouts aan tot een eindoordeel te komen op basis van hun algemene indruk van de verschillende indicatoren. Er valt dus duidelijk nog een slag te slaan door eindbeoordelingen te baseren op mechanische gecombineerde scores via een beslisregel. Een logische volgende stap was dan ook om de invloed van gestructureerde informatieverzameling en mechanisch combinatie van informatie op de daadwerkelijke speler beoordelingen van scouts en coaches te onderzoeken.

HOOFDSTUK 5

In hoofdstuk 5 bekeken voetbalscouts en coaches van de KNVB en verschillende betaald voetbalorganisaties video's van professionele voetbalspelers en beoordeelden hun prestaties. Dit deden zij op zowel een ongestructureerde en gestructureerde manier. In de ongestructureerde conditie gaven zij slechts één totaalbeoordeling op basis van hun algemene indruk (i.e., ongestructureerde beoordeling gepaard met holistisch combinatie van informatie). In de gestructureerde conditie beoordeelden zij acht verschillende prestatie indicatoren die mechanisch werden gecombineerd. Ten slotte gaven participanten in de gestructureerde conditie ook nog een totaalbeoordeling op basis van hun algemene indruk. Dit resulteerde in drie soorten beoordelingen: ongestructureerd-holistisch, gestructureerd-mechanisch, en gestructureerd-holistisch.

Tegen onze verwachtingen in resulteerde de gestructureerd-mechanisch beoordeling niet in de hoogste interbeoordelaarsbetrouwbaarheid (i.e., intraclass correlatie coëfficiënt, ICC = .04) en predictieve validiteit (i.e., r_s = .25). De ongestructureerd-holistische beoordeling had de hoogste betrouwbaarheid en predictieve validiteit (ICC = .14; r_s = .31), maar de verschillen tussen de beoordelingen waren erg klein. De betrouwbaarheid was bij alle methoden van beoordelen erg laag (ICC < .15), wat betekent dat de beoordelingen van dezelfde spelers erg van elkaar afweken. Tegelijkertijd vonden we lage-tot-middelgrote correlaties (.22 < r_s < .31) tussen de beoordelingen en marktwaarde van de spelers drie seizoenen later.

Samenvattend waren de resultaten in hoofdstuk 5 niet in overeenstemming met eerder onderzoek naar het gebruik van gestructureerde beoordelingen en mechanische combinatie van informatie (Arkes et al., 2006; Conway et al., 1995; Dana & Rick, 2006). De ambigue resultaten zijn moeilijk te interpreteren en maken het lastig om duidelijke aanbevelingen te doen naar coaches en scouts op het gebruik van deze principes. Hoewel we hebben geprobeerd om een scoutcontext na te bootsen, kan het zijn dat de beoordelingsopdracht te complex was voor de deelnemers. Mogelijke verklaringen voor de resultaten zijn dan ook het experiment relatief lang duurde, de voorspellers gedetailleerd, maar complex geformuleerd waren, en dat scouts en coaches afkomstig waren van veel verschillende organisaties.

HOOFDSTUK 6

In hoofdstuk 6 gebruikten we een simpelere opzet om de waarde van gestructureerde beoordelingen en mechanische combinatie te onderzoeken. We vroegen coaches van een professionele opleiding om de prestatie van jeugdspelers op een gestructureerde manier te beoordelen. In tegenstelling tot hoofdstuk 5 werden er slechts vier 'brede' prestatie-indicatoren gedefinieerd, namelijk aanvallen, verdedigen, bewegen, en strijdvaardigheid (deze werden gecombineerd tot een mechanische beoordeling). Daarnaast werden spelers op hun 'globale potentie' beoordeeld (i.e., holistische beoordeling). De coaches waren onderdeel van dezelfde organisatie en kregen een kregen voorafgaand aan elke training informatie rondom het invullen van het beoordelingsformulier.

Net als in hoofdstuk 5 vonden we ook een lage interbeoordelaarsbetrouwbaarheid voor de beide manieren van beoordelen. Echter vonden we in deze studie wel dat de betrouwbaarheid van de mechanische beoordeling wat hoger was (ICC = .27) dan die van de holistische (ICC = .20). Vergeleken met hoofdstuk 5 was de mechanische beoordeling ook hoger, maar deze observatie moet voorzichtig worden geïnterpreteerd gegeven de relatief kleine steekproefgrootte en verschillen in designs. Deze resultaten suggereren voorzichtig dat een mechanische beoordeling de betrouwbaarheden van coaches kan verhogen, wanneer zij bij dezelfde organisatie werkzaam zijn en bekend zijn met de gestructureerde beoordelingsstrategie. De betrouwbaarheid was echter ook bij gestructureerd beoordelen nog onvoldoende om ook valide voorspellingen op te leveren.

Dit proefschrift biedt nieuwe inzichten over het betrouwbaar en valide voorspellen van voetbalprestaties middels principes uit de selectiepsychologie. Met betrekking tot het eerste doel laten de studies zien dat een sample-based voorspeller, zoals prestaties in small-sided games, een valide voorspeller van prestaties in 11-tegen-11 wedstrijden kan zijn. Met betrekking tot het tweede doel laten de studies zien dat het implementeren van structuur en mechanische combinatie in het voetbal moeilijk is, gezien de lage betrouwbaarheid van de beoordelingen van voetbalcoaches en scouts. Hoewel er nog veel werk aan de winkel is om deze strategieën te implementeren in de praktijk, ben ik overtuigd van hun waarde voor selectie in het voetbal (Kahneman & Klein, 2009). Ten slotte heeft de huidige thesis ook bijgedragen aan bewustwording onder de KNVB en clubs over het belang van (evidence-based) methoden van assessment en selectie in het voetbal; iets waar ik trots op ben.

REFERENCES

Abbott, A., Button, C., Pepping, G., & Collins, D. (2005). Unnatural selection: Talent identification and development in sport. *Nonlinear Dynamics, Psychology, and Life Sciences*, *9*(1), 61–88.

Ackerman, P. L. (2014). Nonsense, common sense, and science of expert performance: Talent and individual differences. *Intelligence*, *45*(1), 6–17. https://doi.org/10.1016/j.intell.2013.04.009

Ackerman, P. L., & Beier, M. E. (2006). Methods for studying the structure of expertise: Psychometric approaches. In A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (1st ed., pp. 213–232). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796

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

Ali, A. (2011). Measuring soccer skill performance: A review. *Scandinavian Journal of Medicine and Science in Sports*, *21*(2), 170–183. https://doi.org/10.1111/j.1600-0838.2010.01256.x

Altmann, S., Ringhof, S., Neumann, R., Woll, A., & Rumpf, M. C. (2019). Validity and reliability of speed tests used in soccer: A systematic review. *PLOS ONE*, *14*(8), e0220982. https://doi.org/10.1371/journal.pone.0220982

American Educational Research Association, American Psychological Association, & National Council on Measurment in Education. (2014). *Standards For Educational and Psychological Testing*. American Educational Research Association.

Aquino, R., Alves, I. S., Padilha, M. B., Casanova, F., Puggina, E. F., & Maia, J. (2017). Multivariate profiles of selected versus non-selected elite youth brazilian soccer players. *Journal of Human Kinetics*, *60*(1), 113–121. https://doi.org/10.1515/hukin-2017-0094

Araújo, D., Davids, K., & Hristovski, R. (2006). The ecological dynamics of decision making in sport. *Psychology of Sport and Exercise*, *7*(6), 653–676. https://doi.org/10.1016/j.psychsport.2006.07.002

Araújo, D., Davids, K., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology, 19*(1), 69–78. https://doi.org/10.1080/10407410709336951

Arkes, H. R., González-Vallejo, C., Bonham, A. J., Kung, Y.-H., & Bailey, N. (2010). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making, 23,* 250–270. https://doi.org/10.1002/bdm

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

Ashton, G. C., Cronbach, L. J., & Gleser, C. (1968). Psychological tests and personnel decisions. *Biometrics, 24*(2), 442. https://doi.org/10.2307/2528052

Baker, J., Schorer, J., & Wattie, N. (2018). Compromising talent: Issues in identifying and selecting talent in sport. *Quest, 70*(1), 48–63. https://doi.org/10.1080/00336297.2017.1333438

Baker, J., Wattie, N., & Schorer, J. (2015). Defining expertise: A taxonomy for researchers in skill acquisition and expertise. In Baker, Joe & D. Farrow (Eds.), *Routledge handbook of sports expertise* (1st ed., pp. 183–195). Routledge.

Baláková, V., Boschek, P., & Skalíková, L. (2015). Selected cognitive abilities in elite youth soccer players. *Journal of Human Kinetics, 49*(1), 267–276. https://doi.org/10.1515/hukin-2015-0129

Barreiros, A., Côté, J., & Fonseca, A. M. (2014). From early to adult sport success: Analysing athletes' progression in national squads. *European Journal of Sport Science, 14*(Suppl. 1), S178-82. https://doi.org/10.1080/17461391.2012.671368

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice, 34*(4), 31–40. https://doi.org/10.1111/emip.12096

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology, 54*(4), 343–349.

Bennett, K. J. M., Novak, A. R., Pluss, M. A., Stevens, C. J., Coutts, A. J., & Fransen, J. (2018). The use of small-sided games to assess skill proficiency in youth soccer players: a talent identification tool. *Science and Medicine in Football, 2*(3), 231–236. https://doi.org/10.1080/24733938.2017.1413246

Bergkamp, T. L. G., den Hartigh, R. J. R., Frencken, W. G. P., Niessen, A. S. M., & Meijer, R. R. (2020). The validity of small-sided games in predicting 11-vs-11 soccer game performance. *PLOS ONE*, *15*(9), e0239448. https://doi.org/10.1371/journal.pone.0239448

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*. *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. (2018). Comment on: "Talent identification in sport: A systematic review." *Sports Medicine*, *48*(6), 1517–1519. https://doi.org/10.1007/s40279-018-0868-6

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Born, M. P., Stegers-Jager, K. M., & van Andel, C. E. E. (2022). Inferring signs from purposeful samples: The role of context in competency assessment. *Medical Education*, *56*(1), 117–126. https://doi.org/10.1111/medu.14669

Breitbach, S., Tug, S., & Simon, P. (2014). Conventional and genetic talent identification in sports: Will recent developments trace talent? *Sports Medicine*, *44*(11), 1489–1503. https://doi.org/10.1007/s40279-014-0221-7

Brueckl, M., & Heuer, F. (2021). *irrNA: Coefficients of Interrater Reliability - Generalized for Randomly Incomplete Datasets* (0.2.2). R package. https://cran.r-project.org/package=irrNA

Burgess, D. J., & Naughton, G. A. (2010). Talent development in adolescent team sports: A review. *International Journal of Sports Physiology and Performance*, *5*(1), 103–116. https://doi.org/10.1123/ijspp.5.1.103

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*(4), 248–260. https://doi.org/10.1111/1468-2389.00154

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

Chmura, P., Konefał, M., Chmura, J., Kowalczuk, E., Zajac, T., Rokita, A., & Andrzejewski, M. (2018). Match outcome and running performance in different intensity ranges among elite soccer players. *Biology of Sport, 35*(2), 197–203. https://doi.org/10.5114/biolsport.2018.74196

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

Cobb, N. M., Unnithan, V., & McRobert, A. P. (2018). The validity, objectivity, and reliability of a soccer-specific behaviour measurement tool. *Science and Medicine in Football, 2*(3), 196–202. https://doi.org/10.1080/24733938.2017.1423176

Coelho e Silva, M. J., Figueiredo, A. J., Simoes, F., Seabra, A., Natal, A., Vaeyens, R., Philippaerts, R., Cumming, S. P., & Malina, R. M. (2010). Discrimination of U-14 soccer players by level and position. *International Journal of Sports Medicine, 31*(11), 790–796. https://doi.org/https://dx.doi.org/10.1055/s-0030-1263139

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. In *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). 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. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics, 29*(3), 317–331.

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.

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

Davids, K., Araújo, D., Correia, V., & Vilar, L. (2013). How small-sided and conditioned games enhance acquisition of movement and decision-making skills. *Exercise and Sport Sciences Reviews, 41*(3), 154–161. https://doi.org/10.1097/JES.0b013e318292f3ec

Davids, K., Araújo, D., Vilar, L., Renshaw, I., & Pinder, R. (2013). An ecological dynamics approach to skill acquisition: Implications for development of talent in sport. *Talent Development and Excellence, 5*(1), 21–34.

Davids, K., Button, C., & Bennett, S. (2008). *Dyanamics of skill acquisition: A constraints-led approach*. Human Kinetics.

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

De Hoog, M. (2020a). Deze Zweedse topclub wil talent vinden door het níét te zoeken. *De Correspondent*.

De Hoog, M. (2020b). Nee, je kunt helemaal niet zien of je achtjarige de nieuwe Messi is. *De Correspondent*.

Den Hartigh, R. J. R., Hill, Y., & Van Geert, P. L. C. (2018). The development of talent in sports: A dynamic network approach. *Complexity*. https://doi.org/https ://doi.org/10.1155/2018/92801 54

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

Den Hartigh, R. J. R., Van Der Steen, S., Hakvoort, B., Frencken, W. G. P., & Lemmink, K. A. P. M. (2017). Differences in game reading between selected and non-selected youth soccer players. *Journal of Sports Sciences, 36*(4), 422–428. https://doi.org/https://dx.doi.org/10.1080/02640414.2017.1313442

Den Hartigh, R. J. R., Van Dijk, M. W. G., Steenbeek, H. W., & Van Geert, P. L. C. (2016). A dynamic network model to explain the development of excellent human performance. *Frontiers in Psychology, 7,* 532. https://doi.org/10.3389/fpsyg.2016.00532

Deprez, D. N., Buchheit, M., Fransen, J., Pion, J., Lenoir, M., Philippaerts, R. M., & Vaeyens, R. (2015). A longitudinal study investigating the stability of anthropometry and soccer-specific endurance in pubertal high-level youth soccer players. *Journal of Sports Science and Medicine*, *14*(2), 418–426. https://doi.org/10.1080/02640414.2011.652654

Deprez, D. N., Fransen, J., Lenoir, M., Philippaerts, R. M., & Vaeyens, R. (2015). A retrospective study on anthropometrical, physical fitness, and motor coordination characteristics that influence dropout, contract status, and first-team playing time in high-level soccer players aged eight to eighteen years. *Journal of Strength and Conditioning Research*, *29*(6), 1692–1704. https://doi.org/10.1519/JSC.0000000000000806

Drenth, P. J. D. (2008). Psychology: Is it applied enough? *Applied Psychology*, *57*(3), 524–540. https://doi.org/10.1111/j.1464-0597.2008.00337.x

Duarte, R., Araújo, D., Folgado, H., Esteves, P., Marques, P., & Davids, K. (2013). Capturing complex, non-linear team behaviours during competitive football performance. *Journal of Systems Science and Complexity*, *26*(1), 62–72. https://doi.org/10.1007/s11424-013-2290-3

Fenner, J. S. J., Iga, J., & Unnithan, V. (2016). The evaluation of small-sided games as a talent identification tool in highly trained prepubertal soccer players. *Journal of Sports Sciences*, *34*(20), 1983–1990. https://doi.org/10.1080/02640414.2016.1149602

Figueiredo, A. J., Gonçalves, C. E., Coelho e Silva, M. J., & Malina, R. M. (2009). Characteristics of youth soccer players who drop out, persist or move up. *Journal of Sports Sciences*, *27*(9), 883–891. https://doi.org/10.1080/02640410902946469

Findlay, L. C., & Ste-marie, D. M. (2004). A reputation bias in figure skating judging. *Journal of Sport & Exercise Psychology*, *26*(1), 154–166. https://doi.org/10.1123/jsep.26.1.154

Ford, P. R., Bordonau, J. L. D., Bonanno, D., Tavares, J., Groenendijk, C., Fink, C., Gualtieri, D., Gregson, W., Varley, M. C., Weston, M., Lolli, L., Platt, D., & Di Salvo, V. (2020). A survey of talent identification and development processes in the youth academies of professional soccer clubs from around the world. *Journal of Sports Sciences*, *38*(11–12), 1269–1278. https://doi.org/10.1080/02640414.2020.1752440

Franks, A., Williams, A. M., Reilly, T., & Nevill, A. (1999). Talent identification in elite youth soccer players: Physical and physiological characteristics. *Journal of Sports Sciences*, *17*(10), 812.

Frencken, W. G. P., Lemmink, K. A. P. M., & Delleman, N. J. (2010). Soccer-specific accuracy and validity of the local position measurement (LPM) system. *Journal of Science and Medicine in Sport, 13*, 641–645. https://doi.org/10.1016/j.jsams.2010.04.003

Frencken, W. G. P., Lemmink, K., Delleman, N., & Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. *European Journal of Sport Science, 11*(4), 215–223. https://doi.org/10.1080/17461391.2010.499967

Gil, S. M., Ruiz, F., Irazusta, A., Gil, J., & Irazusta, J. (2007). Selection of young soccer player in terms of antropometric and physiological factor. *Journal of Sports Medicine and Physical Fitness, 47*(1), 25–32.

Gil, S. M., Zabala-Lili, J., Bidaurrazaga-Letona, I., Aduna, B., Lekue, J. A., Santos-Concejero, J., & Granados, C. (2014). Talent identification and selection process of outfield players and goalkeepers in a professional soccer club. *Journal of Sports Sciences, 32*(20), 1931–1939. https://doi.org/10.1080/02640414.2014.964290

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. P. M. (2018). Not every pass can be an assist: A data-driven model to measure pass effectiveness in professional soccer matches. *Big Data, 6*(4). https://doi.org/10.1089/big.2018.0067

Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., & Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science, 21*(4), 481–496. https://doi.org/10.1080/17461391.2020.1747552

Gomez-Piqueras, P., Gonzalez-Villora, S., Castellano, J., & Teoldo, I. (2019). Relation between the physical demands and success in professional soccer players. *Journal of Human Sport and Exercise, 14*(1), 1–11. https://doi.org/10.14198/jhse.2019.141.01

Gonaus, C., & Müller, E. (2012). Using physiological data to predict future career progression in 14- to 17-year-old Austrian soccer academy players. *Journal of Sports Sciences, 30*(15), 1673–1682. https://doi.org/10.1080/02640414.2012.713980

Goto, H., Morris, J. G., & Nevill, M. E. (2015). Match analysis of U9 and U10 English premier league academy soccer players using a global positioning system. *Journal of Strength and Conditioning Research*, *29*(4), 954–963. https://doi.org/10.1519/JSC.0b013e3182a0d751

Gravina, L., Gil, S. M., Ruiz, F., Zubero, J., Gil, J., & Irazusta, J. (2008). Anthropometric and physiological differences between first team and reserve soccer players aged 10-14 years at the beginning and end of the season. *Journal of Strength and Conditioning Research*, *22*(4), 1308–1314. https://doi.org/https://dx.doi.org/10.1519/JSC.0b013e31816a5c8e

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

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gulbin, J., Weissensteiner, J., Oldenziel, K., & Gagné, F. (2013). Patterns of performance development in elite athletes. *European Journal of Sport Science*, *13*(6), 605–614. https://doi.org/10.1080/17461391.2012.756542

Güllich, A. (2014). Selection, de-selection and progression in German football talent promotion. *European Journal of Sport Science*, *14*(6), 530–537. https://doi.org/10.1080/17461391.2013.858371

Güllich, A., & Cobley, S. (2017). On the efficacy of talent identification and talent development programmes. In J. Baker, S. Cobley, J. Schorer, & N. Wattie (Eds.), *Routledge Handbook of Talent Identification and Development in Sport* (pp. 80–98). Routledge. https://doi.org/10.4324/9781315668017

Güllich, A., & Emrich, E. (2014). Considering long-term sustainability in the development of world class success. *European Journal of Sport Science*, *14*(Suppl.1), 383–397. https://doi.org/10.1080/17461391.2012.706320

Haj-Sassi, R., Dardouri, W., Gharbi, Z., Chaouachi, A., Mansour, H., Rabhi, A., & Mahfoudhi, M.-E. (2011). Reliability and validity of a new repeated agility test as a measure of anearobic and explosive power. *Journal of Strength and Conditioning Research*, *25*(2), 472–480.

Haugaasen, M., Toering, T., & Jordet, G. (2014). From childhood to senior professional football: A multi-level approach to elite youth football players' engagement in football-specific activities. *Psychology of Sport and Exercise, 15*(4), 336–344. https://doi.org/10.1016/j.psychsport.2014.02.007

Helsen, W. F., Baker, J., Michiels, S., Schorer, J., van Winckel, J., & Williams, A. M. (2012). The relative age effect in European professional soccer: Did ten years of research make any difference? *Journal of Sports Sciences, 30*(15), 1665–1671. https://doi.org/10.1080/02640414.2012.721929

Henderson, D., & Denison, R. (1989). Stepwise regression in social and psychological research. *Psychological Reports, 64*(1981), 251–257. https://doi.org/10.2466/pr0.1989.64.1.251

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

Herold, M., Goes, F. R., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science and Coaching, 14*(6), 798–817. https://doi.org/10.1177/1747954119879350

Höner, O., & Feichtinger, P. (2016). Psychological talent predictors in early adolescence and their empirical relationship with current and future performance in soccer. *Psychology of Sport and Exercise, 25*, 17–26. https://doi.org/10.1016/j.psychsport.2016.03.004

Höner, O., Leyhr, D., & Kelava, A. (2017). The influence of speed abilities and technical skills in early adolescence on adult success in soccer: A long-term prospective analysis using ANOVA and SEM approaches. *PLOS ONE, 12*(8), e0182211. https://doi.org/10.1371/journal.pone.0182211

Höner, O., & Votteler, A. (2016). Prognostic relevance of motor talent predictors in early adolescence: A group- and individual-based evaluation considering different levels of achievement in youth football. *Journal of Sports Sciences, 34*(24), 2269–2278. https://doi.org/10.1080/02640414.2016.1177658

Hoppe, M. W., Slomka, M., Baumgart, C., Weber, H., & Freiwald, J. (2015). Match running performance and success across a season in German Bundesliga soccer teams. *International Journal of Sports Medicine, 36*(7), 563–566. https://doi.org/10.1055/s-0034-1398578

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future--remembering the past. *Annual Review of Psychology, 51*(1), 631–664. https://doi.org/10.1146/annurev.psych.51.1.631

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). Routledge.

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

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences, 20*(10), 739–754. https://doi.org/10.1080/026404102320675602

Huijgen, B. C. H., Elferink-Gemser, M. T., Ali, A., & Visscher, C. (2013). Soccer skill development in talented players. *International Journal of Sports Medicine, 34*(8), 720–726. https://doi.org/10.1055/s-0032-1323781

Huijgen, B. C. H., Elferink-Gemser, M. T., Lemmink, K. A. P. M., & Visscher, C. (2014). Multidimensional performance characteristics in selected and deselected talented soccer players. *European Journal of Sport Science, 14*(1), 2–10. https://doi.org/10.1080/17461391.2012.725102

Huijgen, B. C. H., Elferink-Gemser, M. T., Post, W., & Visscher, C. (2010). Development of dribbling in talented youth soccer players aged 12–19 years: A longitudinal study. *Journal of Sports Sciences, 28*(7), 689–698. https://doi.org/10.1080/02640411003645679

Huijgen, B. C. H., Leemhuis, S., Kok, N. M., Verburgh, L., Oosterlaan, J., Elferink-Gemser, M. T., & Visscher, C. (2015). Cognitive functions in elite and sub-elite youth soccer players aged 13 to 17 years. *PLOS ONE, 10*(12), e0144580. https://doi.org/10.1371/journal.pone.0144580

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72–98. https://doi.org/10.1037/0033-2909.96.1.72

Ivarsson, A., Kilhage-Persson, A., Martindale, R., Priestley, D., Huijgen, B. C. H., Ardern, C., & McCall, A. (2020). Psychological factors and future performance of football players: A systematic review with meta-analysis. *Journal of Science and Medicine in Sport, 23*(4), 415–420. https://doi.org/10.1016/j.jsams.2019.10.021

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 and Coaching, 12*(4), 470–480. https://doi.org/10.1177/1747954117718020

Johnston, K., Wattie, N., Schorer, J., & Baker, J. (2018). Talent identification in sport: A systematic review. *Sports Medicine, 48*(1), 97–109. https://doi.org/10.1007/s40279-017-0803-2

Jokuschies, N., Gut, V., & Conzelmann, A. (2017). Systematizing coaches' 'eye for talent': Player assessments based on expert coaches' subjective talent criteria in top-level youth soccer. *International Journal of Sports Science and Coaching, 12*(5), 565–576. https://doi.org/10.1177/1747954117727646

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). Macmillan. https://doi.org/10.18177/sym.2015.55.ca.10990

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise. *Harvard Business Review, 60*(8), 40–46.

Kannekens, R., Elferink-Gemser, M. T., & Visscher, C. (2011). Positioning and deciding: Key factors for talent development in soccer. *Scandinavian Journal of Medicine and Science in Sports, 21*(6), 846–852. https://doi.org/10.1111/j.1600-0838.2010.01104.x

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

Katis, A., & Kellis, E. (2009). Effects of small-sided games on physical conditioning and performance in young soccer players. *Journal of Sport Science and Medicine, May*, 374–380.

Kavussanu, M., White, S. A., Jowett, S., & England, S. (2011). Elite and non-elite male footballers differ in goal orientation and perceptions of parental climate. *International Journal of Sport and Exercise Psychology, 9*(3), 284–290. https://doi.org/10.1080/1612197X.2011.614854

Kearney, P. E., & Hayes, P. R. (2018). Excelling at youth level in competitive track and field athletics is not a prerequisite for later success. *Journal of Sports Sciences*, *36*(21), 2502–2509. https://doi.org/10.1080/02640414.2018.1465724

Kelly, D. M., & Drust, B. (2009). The effect of pitch dimensions on heart rate responses and technical demands of small-sided soccer games in elite players. *Journal of Science and Medicine in Sport*, *12*, 475–479. https://doi.org/10.1016/j.jsams.2008.01.010

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kharrat, T., McHale, I. G., & Peña, J. L. (2019). Plus–minus player ratings for soccer. *European Journal of Operational Research*, *283*(2), 726–736. https://doi.org/10.1016/j.ejor.2019.11.026

Klingner, F. C., Huijgen, B. C. H., Den Hartigh, R. J. R., & Kempe, M. (2021). Technical–tactical skill assessments in small-sided soccer games: A scoping review. *International Journal of Sports Science and Coaching*. https://doi.org/10.1177/17479541211049532

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

Lago-Penas, C., Rey, E., Casais, L., & Gomez-Lopez, M. (2014). Relationship between performance characteristics and the selection process in youth soccer players. *Journal of Human Kinetics*, *40*, 189–199. https://doi.org/https://dx.doi.org/10.2478/hukin-2014-0021

Larkin, P., Marchant, D., Syder, A., & Farrow, D. (2020). An eye for talent: The recruiters' role in the Australian Football talent pathway. *PLOS ONE*, *15*(11), e0241307. https://doi.org/10.1371/journal.pone.0241307

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), e0175716. https://doi.org/10.1371/journal.pone.0175716

Larkin, P., & Reeves, M. J. (2018). Junior-elite football: time to re-position talent identification? *Soccer and Society, 19*(8), 1183–1192. https://doi.org/10.1080/14660970.2018.1432389

Lath, F., Koopmann, T., Faber, I., Baker, J., & Schorer, J. (2021). Focusing on the coach's eye; towards a working model of coach decision-making in talent selection. *Psychology of Sport and Exercise*, *56*, 1–13. https://doi.org/10.1016/j.psychsport.2021.102011

Lawshe, C. H., Bolda, R. A., Brune, R. L., & Auclair, G. (1958). Expectancy charts II. Their theoretical development. *Personnel Psychology, 11*(4), 545–559. https://doi.org/10.1111/j.1744-6570.1958.tb00040.x

le Gall, F., Carling, C., Williams, A. M., & Reilly, T. (2010). Anthropometric and fitness characteristics of international, professional and amateur male graduate soccer players from an elite youth academy. *Journal of Science and Medicine in Sport, 13*(1), 90–95. https://doi.org/10.1016/j.jsams.2008.07.004

Le Moal, E., Rué, O., Ajmol, A., Abderrahman, A. B., Hammami, M. A., Ounis, O. B., Kebsi, W., & Zouhal, H. (2014). Validation of the Loughborough soccer passing test in young soccer players. *Journal of Strength and Conditioning Research, 28*(5), 1418–1426. https://doi.org/10.1519/JSC.0000000000000296

Lemmink, K. A. P. M., Visscher, C., Lambert, M., & Lamberts, R. P. (2004). The interval shuttle run test for intermittent sport players: Evaluation of reliability. *Journal of Strength and Conditioning Research, 71*(4), 737–767. https://doi.org/10.1002/fut

Leyhr, D., Kelava, A., Raabe, J., & Höner, O. (2018). Longitudinal motor performance development in early adolescence and its relationship to adult success: An 8-year prospective study of highly talented soccer players. *PLOS ONE, 13*(5), e0196324. https://doi.org/https://dx.doi.org/10.1371/journal.pone.0196324

Li, P., De Bosscher, V., Pion, J., Weissensteiner, J. R., & Vertonghen, J. (2018). Is international junior success a reliable predictor for international senior success in elite combat sports? *European Journal of Sport Science, 18*(4), 550–559. https://doi.org/10.1080/17461391.2018.1439104

Lidor, R., Côté, J., & Hackfort, D. (2009). ISSP position stand: To test or not to test? The use of physical skill tests in talent detection and in early phases of sport development. *International Journal of Sport and Exercise Psychology*, *7*(2), 131–146. https://doi.org/10.1080/1612197X.2009.9671896

Lievens, F., & De Soete, B. (2012). Simulations. In N Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection* (pp. 383–410). Oxford University Press.

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

Lyons, B. D., Hoffman, B. J., Michel, J. W., & Williams, K. J. (2011). On the predictive efficiency of past performance and physical ability: The case of the national football league. *Human Performance*, *24*(2), 158–172. https://doi.org/10.1080/08959285.2011.555218

MacMahon, C., Bailey, A., Croser, M., & Weissensteiner, J. (2019). Exploring the skill of recruiting in the Australian Football League. *International Journal of Sports Science and Coaching*, *14*(1), 72–81. https://doi.org/10.1177/1747954118809775

Martindale, R. J., Collins, D., & Daubney, J. (2005). Talent development: A guide for practice and research within sport. *Quest*, *57*(4), 353–375. https://doi.org/10.1080/00336297.2005.10491862

Martinez-Santos, R., Castillo, D., & Los Arcos, A. (2016). Sprint and jump performances do not determine the promotion to professional elite soccer in Spain, 1994–2012. *Journal of Sports Sciences*, *34*(24), 2279–2285. https://doi.org/10.1080/02640414.2016.1190460

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.

McLellan, R. A. (1996). *Theoretical expectancy calculator*. http://www.hrsoftware.net/cgi/TheoreticalExpectancy.cgi

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician*, *73*(Suppl.1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

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

Meehl, P. E. (1989). Law and the fireside inductions (with postscript): Some reflections of a clinical psychologist. *Behavioral Sciences & the Law*, *7*(4), 521–550. https://doi.org/10.1002/bsl.2370070408

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*(3), 194–216. https://doi.org/10.1037/h0048070

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*, 3002. https://doi.org/10.3389/fpsyg.2019.03002

Memmert, D., Lemmink, K. A. P. M., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, *47*(1), 1–10. https://doi.org/10.1007/s40279-016-0562-5

Meylan, C., Cronin, J., Oliver, J., & Hughes, M. (2010). Talent identification in soccer: The role of maturity status on physical, physiological and technical characteristics. *International Journal of Sports Science and Coaching*, *5*(4), 571–592. https://doi.org/10.1260/1747-9541.5.4.571

Miller, P. K., Cronin, C., & Baker, G. (2015). Nurture, nature and some very dubious social skills: An interpretative phenomenological analysis of talent identification practices in elite English youth soccer. *Qualitative Research in Sport, Exercise and Health*, *7*(5), 642–662. https://doi.org/10.1080/2159676X.2015.1012544

Mirkov, D., Nedeljkovic, A., Kukolj, M., Ugarkovic, D., & Jaric, S. (2008). Evaluation of the reliability of soccer-specific field tests. *Journal of Strength and Conditioning Research*, *22*(4), 1046–1050. https://doi.org/https://dx.doi.org/10.1519/JSC.0b013e31816eb4af

Mitchell, L. J. G., Rattray, B., Saunders, P. U., & Pyne, D. B. (2018). The relationship between talent identification testing parameters and performance in elite junior swimmers. *Journal of Science and Medicine in Sport*. https://doi.org/10.1016/j.jsams.2018.05.006

Mol, S. T., Born, M. P., Willemsen, M. E., & Van Der Molen, H. T. (2005). Predicting expatriate job performance for selection purposes: A quantitative review. *Journal of Cross-Cultural Psychology*, *36*(5), 590–620. https://doi.org/10.1177/0022022105278544

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

Murr, D., Feichtinger, P., Larkin, P., O'Connor, D., & Höner, O. (2018). Psychological talent predictors in youth soccer: A systematic review of the prognostic relevance of psychomotor, perceptual-cognitive and personality-related factors. *PLOS ONE, 13*(10), e0205337. https://doi.org/10.1371/journal.pone.0205337

Murr, D., Raabe, J., & Höner, O. (2018). The prognostic value of physiological and physical characteristics in youth soccer: A systematic review. *European Journal of Sport Science, 18*(1), 62–74. https://doi.org/10.1080/17461391.2017.1386719

Musch, J., & Hay, R. (1999). The relative age effect in soccer: Cross-cultural evidence for a systematic discrimination against children born late in the competition year. *Sociology of Sport Journal, 16*(1), 54–64. https://doi.org/10.1123/ssj.16.1.54

Musculus, L., & Lobinger, B. H. (2018). Psychological characteristics in talented soccer players - Recommendations on how to improve coaches' assessment. *Frontiers in Psychology, 9*, 41. https://doi.org/10.3389/fpsyg.2018.00041

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

Newman, D. A., Kinney, T., & Farr, J. L. (2004). Job performance ratings. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment* (Vol. 4, pp. 373–389). John Wiley & Sons, Inc.

Niessen, A. S. M., & Meijer, R. R. (2016). Selection of medical students on the basis of nonacademic skills: Is it worth the trouble? *Clinical Medicine, Journal of the Royal College of Physicians of London, 16*(4), 339–342. https://doi.org/10.7861/clinmedicine.16-4-339

Niessen, A. S. M., & Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. *Perspectives on Psychological Science, 12*(3), 436–448. https://doi.org/10.1177/1745691616683050

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting performance in higher education using proximal predictors. *PLOS ONE, 11*(4), e0153663. https://doi.org/10.1371/journal.pone.0153663

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLOS ONE*, *13*(6), e0198746. https://doi.org/10.1371/journal.pone.0198746

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

O'Connor, D., Larkin, P., & Mark Williams, A. (2016). Talent identification and selection in elite youth football: An Australian context. *European Journal of Sport Science*, *16*(7), 837–844. https://doi.org/10.1080/17461391.2016.1151945

Olthof, S. B. H., Frencken, W. G. P., & Lemmink, K. A. P. M. (2019). A match-derived relative pitch area facilitates the tactical representativeness of small-sided games for the official soccer match. *Journal of Strength and Conditioning Research*, *33*(2), 523–530. https://doi.org/10.1519/JSC.0000000000002978

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life. *Psychological Bulletin*, *124*(1), 54–74.

Pankhurst, A., & Collins, D. (2013). Talent identification and development: The need for coherence between research, system, and process. *Quest*, *65*(1), 83–97. https://doi.org/10.1080/00336297.2012.727374

Pappalardo, L., & Cintia, P. (2017). Quantifying the relation between performance and success in soccer. *Advances in Complex Systems*, *20*(4), 1–30. https://doi.org/10.1142/S021952591750014X

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 1–27. https://doi.org/10.1145/3343172

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). The Guilford Press.

Pauole, K., Madole, K., Garhammer, J., Lacourse, M., & Rozenek, R. (2000). Reliability and validity of the t-test as a measure of agility, leg power, and leg speed in college-aged men and women. *Journal of Strength and Conditioning Research*, *14*(4), 443–450. https://doi.org/10.1519/00124278-200011000-00012

Pearson, D. T., Naughton, G. A., & Torode, M. (2006). Predictability of physiological testing and the role of maturation in talent identification for adolescent team sports. *Journal of Science and Medicine in Sport, 9*(4), 277–287. https://doi.org/10.1016/j.jsams.2006.05.020

Phillips, E., Davids, K., Renshaw, I., & Portus, M. (2010). Expert performance in sport and the dynamics of talent development. *Sports Medicine, 40*(4), 271–283. https://doi.org/10.2165/11593020-000000000-00000

Piggott, B., Müller, S., Chivers, P., Papaluca, C., & Hoyne, G. (2019). Is sports science answering the call for interdisciplinary research? A systematic review. *European Journal of Sport Science, 19*(3), 267–286. https://doi.org/10.1080/17461391.2018.1508506

Pinder, R. A., Davids, K., Renshaw, I., & Araújo, D. (2011). Representative learning design and functionality of research and practice in sport. *Journal of Sport and Exercise Psychology, 33*(1), 146–155. https://doi.org/10.1123/jsep.33.1.146

Pinder, R. A., Renshaw, I., & Davids, K. (2013). The role of representative design in talent development: A comment on “Talent identification and promotion programmes of Olympic athletes.” *Journal of Sports Sciences, 31*(8), 803–806. https://doi.org/10.1080/02640414.2012.718090

Rampinini, E., Coutts, A. J., Castagna, C., Sassi, R., & Impellizzeri, F. M. (2007). Variation in top level soccer match performance. *International Journal of Sports Medicine, 28*(12), 1018–1024. https://doi.org/10.1055/s-2007-965158

Redkva, P. E., Paes, M. R., Fernandez, R., & Da-Silva, S. G. (2018). Correlation between match performance and field tests in professional soccer players. *Journal of Human Kinetics, 62*(1), 213–219. https://doi.org/10.1515/hukin-2017-0171

Rees, T., Hardy, L., Güllich, A., Abernethy, B., Côté, J., Woodman, T., Montgomery, H., Laing, S., & Warr, C. (2016). The great British medalists project: A review of current knowledge on the development of the world’s best sporting talent. *Sports Medicine, 46*(8), 1041–1058. https://doi.org/10.1007/s40279-016-0476-2

Reeves, M. J., Littlewood, M. A., McRobert, A. P., & Roberts, S. J. (2018). The nature and function of talent identification in junior-elite football in English category one academies. *Soccer and Society, 19*(8), 1122–1134. https://doi.org/10.1080/14660970.2018.1432385

Reeves, M. J., McRobert, A. P., Lewis, C. J., & Roberts, S. J. (2019). A case study of the use of verbal reports for talent identification purposes in soccer: A Messi affair! *PLOS ONE*, *14*(11), e0225033. https://doi.org/10.1371/journal.pone.0225033

Reilly, T., Williams, A. M., Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences*, *18*(9), 695–702. https://doi.org/https://dx.doi.org/10.1080/02640410050120078

Renshaw, I., Davids, K., Araújo, D., Lucas, A., Roberts, W. M., Newcombe, D. J., & Franks, B. (2019). Evaluating weaknesses of “perceptual-cognitive training” and “brain training” methods in sport: An ecological dynamics critique. *Frontiers in Psychology*, *9*, 2468. https://doi.org/10.3389/fpsyg.2018.02468

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

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Annual Review of Psychology*, *48*, 229–337. https://doi.org/10.1002/0470048204.ch17

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

Roth, P. L., Bobko, P., & Mcfarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*(4), 1009–1037. https://doi.org/10.1111/j.1744-6570.2005.00714.x

Rowat, O., Fenner, J., & Unnithan, V. (2017). Technical and physical determinants of soccer match-play performance in elite youth soccer players. *The Journal of Sports Medicine and Physical Fitness*, *57*(4), 369–379. https://doi.org/10.23736/S0022-4707.16.06093-X

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435–1447. https://doi.org/10.1037/apl0000236

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills, 83*, 1195–1201. https://doi.org/10.1146/annurev.ps.43.020192.003211

Sarmento, H., Anguera, M. T., Pereira, A., & Araújo, D. (2018). Talent identification and development in male football: A systematic review. *Sports Medicine, 48*(4), 907–931. https://doi.org/10.1007/s40279-017-0851-7

Sarmento, H., Clemente, F. M., Harper, L. D., Costa, I. T. da, Owen, A., & Figueiredo, A. J. (2018). Small sided games in soccer–a systematic review. *International Journal of Performance Analysis in Sport, 18*(5), 693–749. https://doi.org/10.1080/24748668.2018.1517288

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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage publications.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*(2), 281–305. https://doi.org/10.1111/j.1744-6570.2006.00037.x

Schranz, N., Tomkinson, G., Olds, T., Petkov, J., & Hahn, A. G. (2012). Is three-dimensional anthropometric analysis as good as traditional anthropometric analysis in predicting junior rowing performance? *Journal of Sports Sciences, 30*(12), 1241–1248. https://doi.org/10.1080/02640414.2012.696204

Serpiello, F. R., Cox, A., Oppici, L., Hopkins, W. G., & Varley, M. C. (2017). The Loughborough soccer passing test has impractical criterion validity in elite youth football. *Science and Medicine in Football, 1*(1), 60–64. https://doi.org/10.1080/02640414.2016.1254810

Sill, J. (2010). Improved NBA adjusted + / - using regularization and out-of-sample testing. *MIT Sloan Sports Analytics Conference*, 1–7.

Simonton, D. K. (1999). Talent and its development: An emergenic and epigenetic model. *Psychological Review, 106*(3), 435–457.

Sporis, G., Jukic, I., Milanovic, L., & Vucetic, V. (2010). Reliability and factorial validity of agility tests for soccer players. *Journal of Strength and Conditioning Research, 24*(3), 679–686. https://doi.org/10.1519/JSC.0b013e3181c4d324

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

Stevens, T. G. A., De Ruiter, C. J., Beek, P. J., & Savelsbergh, G. J. P. (2016). Validity and reliability of 6-a-side small-sided game locomotor performance in assessing physical fitness in football players. *Journal of Sports Sciences, 34*(6), 527–534. https://doi.org/https://dx.doi.org/10.1080/02640414.2015.1116709

Stoffregen, T. A., Bardy, B. G., Smart, L. J., & Pagulayan, R. J. (2003). On the nature and evaluation of fidelity in virtual environments. In L. J. Hettinger & M. W. Haas (Eds.), *Virtual and adaptive environments: Applications, implications, and human performance issues* (pp. 111–128). Lawrence Erlbaum Associates Publishers.

Stone, J., Perry, Z. W., & Darley, J. M. (1997). "'White men can't jump'": Evidence for the perceptual confirmation of racial stereotypes following a basketball game. *Basic and Applied Social Psychology, 19*(3), 291–306. https://doi.org/10.1207/15324839751036977

Swann, C., Moran, A., & Piggott, D. (2015). Defining elite athletes: Issues in the study of expert performance in sport psychology. *Psychology of Sport and Exercise, 16*(P1), 3–14. https://doi.org/10.1016/j.psychsport.2014.07.004

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2007). *The Book: Playing the percentages in baseball*. Potomac Books. http://www.eskom.co.za/CustomerCare/TariffsAndCharges/Documents/RSA Distribution Tariff Code Vers 6.pdf%0Ahttp://www.nersa.org.za/

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*(5), 565–578. https://doi.org/10.1037/h0057079

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525–534. https://doi.org/10.1177/0013164495055004001

Till, K., & Baker, J. (2020). Challenges and [possible] solutions to optimizing talent identification and development in sport. *Frontiers in Psychology, 11,* 664. https://doi.org/10.3389/fpsyg.2020.00664

Toering, T. T., Elferink-Gemser, M. T., Jordet, G., & Visscher, C. (2009). Self-regulation and performance level of elite and non-elite youth soccer players. *Journal of Sports Sciences, 27*(14), 1509–1517. https://doi.org/10.1080/02640410903369919

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

Travassos, B., Davids, K., Araújo, D., & Esteves, P. T. (2013). Performance analysis in team sports: Advances from an Ecological Dynamics approach. *International Journal of Performance Analysis in Sport, 13*(1), 83–95. https://doi.org/10.1080/24748668.2013.11868633

Tukey, J. W. (1980). We need both explanatory and confirmatory. *The American Statistician, 34*(1), 23–25.

Unnithan, V., White, J., Georgiou, A., Iga, J., & Drust, B. (2012). Talent identification in youth soccer. *Journal of Sports Sciences, 30*(15), 1719–1726. https://doi.org/10.1080/02640414.2012.731515

Vaeyens, R., Lenoir, M., Williams, A. M., & Philippaerts, R. M. (2008). Talent identification and development programmes in sport: Current models and future directions. *Sports Medicine, 38*(9), 703–714. https://doi.org/10.2165/00007256-200838090-00001

Vaeyens, R., Malina, R. M., Janssens, M., Van Renterghem, B., Bourgois, J., Vrijens, J., & Philippaerts, R. M. (2006). A multidisciplinary selection model for youth soccer: The Ghent Youth Soccer Project. *British Journal of Sports Medicine, 40*(11), 928–934. https://doi.org/10.1136/bjsm.2006.029652

Van der Flier, H. (1992). *Hebben wij eigenschappen nodig? "Signs" en "samples" in het psychologisch selectie-onderzoek.* Vrije Universiteit.

Van Maarseveen, M. J. J., Oudejans, R. R. D., & Savelsbergh, G. J. P. (2017). System for notational analysis in small-sided soccer games. *International Journal of Sports Science and Coaching, 12*(2), 194–206. https://doi.org/10.1177/1747954117694922

Van Yperen, N. W. (2009). Why some make it and others do not: Identifying psychological factors that predict career success in professional adult soccer. *Sport Psychologist, 23*(3), 317–329. https://doi.org/10.1123/tsp.23.3.317

Vandendriessche, J. B., Vaeyens, R., Vandorpe, B., Lenoir, M., Lefevre, J., & Philippaerts, R. M. (2012). Biological maturation, morphology, fitness, and motor coordination as part of a selection strategy in the search for international youth soccer players (age 15-16 years). *Journal of Sports Sciences, 30*(15), 1695–1703. https://doi.org/10.1080/02640414.2011.652654

Verburgh, L., Scherder, E. J. A., Van Lange, P. A. M., & Oosterlaan, J. (2014). Executive functioning in highly talented soccer players. *PLOS ONE, 9*(3), e91254. https://doi.org/10.1371/journal.pone.0091254

Vestberg, T., Gustafson, R., Maurex, L., Ingvar, M., & Petrovic, P. (2012). Executive functions predict the success of top soccer players. *PLOS ONE*, *7*(4), e34731. https://doi.org/https://doi.org/10.1371/journal.pone.0034731

Vestberg, T., Reinebo, G., Maurex, L., Ingvar, M., & Petrovic, P. (2017). Core executive functions are associated with success in young elite soccer players. *PLOS ONE, 12*(2), e0170845. https://doi.org/10.1371/journal.pone.0170845

Vieira, L. H. P., Carling, C., Barbieri, F. A., Aquino, R., & Santiago, P. R. P. (2019). Match running performance in young soccer players: A systematic review. *Sports Medicine*, *49*(2), 289–318. https://doi.org/10.1007/s40279-018-01048-8

Vilar, L., Araújo, D., Davids, K., & Bar-Yam, Y. (2013). Science of winning soccer: Emergent pattern-forming dynamics in association football. *Journal of Systems Science and Complexity, 26*(1), 73–84. https://doi.org/10.1007/s11424-013-2286-z

Vilar, L., Araujo, D., Davids, K., & Button, C. (2012). The role of ecological dynamics in analysing performance in team sports. *Sports Medicine, 40*(12), 1019–1035. https://doi.org/10.2165/11536850-000000000-00000

Vinchur, A. J., & Bryan, L. L. K. (2012). A History of Personnel Selection and Assessment. In Neil Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection* (pp. 9–30). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0002

Visscher, C., Elferink-Gemser, M. T., & Lemmink, K. A. P. M. (2006). Interval endurance capacity of talented youth soccer players. *Perceptual and Motor Skills, 102*(1), 81–86. https://doi.org/10.1080/03014460802570584

Visser, W. (2018). Hoe Virgil van Dijk alsnog de wereld veroverde. *De Volkskrant*.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108–131. https://doi.org/10.1037/0021-9010.90.1.108

Waldron, M., & Murphy, A. (2013). A comparison of physical abilities and match performance characteristics among elite and subelite under-14 soccer players. *Pediatric Exercise Science, 25*(3), 423–434. https://doi.org/10.1123/pes.25.3.423

Waldron, M., & Worsfold, P. (2010). Differences in the game specific skills of elite and sub-elite youth football players: Implications for talent identification. *International Journal of Performance Analysis in Sport, 10*(1), 9–24. https://doi.org/10.1080/24748668.2010.11868497

Wallace, J. L., & Norton, K. I. (2014). Evolution of World Cup soccer final games 1966-2010: Game structure, speed and play patterns. *Journal of Science and Medicine in Sport, 17*(2), 223–228. https://doi.org/10.1016/j.jsams.2013.03.016

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond " p < 0.05." *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*(5), 372–376. https://doi.org/10.1037/h0026244

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*(4), 275–290. https://doi.org/10.1111/j.2044-8325.1988.tb00467.x

Williams, A. M., Ford, P. R., & Drust, B. (2020). Talent identification and development in soccer since the millennium. *Journal of Sports Sciences, 38*(11–12), 1199–1210. https://doi.org/10.1080/02640414.2020.1766647

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sports Sciences, 18*(9), 657–667. https://doi.org/10.1080/02640410050120041

Wilson, R. S., David, G. K., Murphy, S. C., Angilletta, M. J., Niehaus, A. C., Hunter, A. H., & Smith, M. D. (2017). Skill not athleticism predicts individual variation in match performance of soccer players. *Proceedings of the Royal Society B: Biological Sciences, 284*(1868). https://doi.org/10.1098/rspb.2017.0953

Wilson, R. S., Hunter, A. H., Camata, T. V., Foster, C. S. P., Sparkes, G. R., Moura, F. A., Santiago, P. R. P., & Smith, N. M. A. (2021). Simple and reliable protocol for identifying talented junior players in team sports using small-sided games. *Scandinavian Journal of Medicine and Science in Sports, 31*(8), 1647–1656. https://doi.org/10.1111/sms.13969

Wilson, R. S., James, R. S., David, G., Hermann, E., Morgan, O. J., Niehaus, A. C., Hunter, A., Thake, D., & Smith, M. D. (2016). Multivariate analyses of individual variation in soccer skill as a tool for talent identification and development: utilising evolutionary theory in sports science. *Journal of Sports Sciences, 3421*(21), 2074–2086. https://doi.org/10.1080/02640414.2016.1151544

Wiseman, A. C., Bracken, N., Horton, S., & Weir, P. L. (2014). The difficulty of talent identification: Inconsistency among coaches through skill-based assessment of youth hockey players. *International Journal of Sports Science and Coaching, 9*(3), 447–455. https://doi.org/10.1260/1747-9541.9.3.447

Zibung, M., Zuber, C., & Conzelmann, A. (2016). The motor subsystem as a predictor of success in young football talents: A person-oriented study. *PLOS ONE, 11*(8), e0161049. https://doi.org/10.1371/journal.pone.0161049

Zuber, C., & Conzelmann, A. (2014). The impact of the achievement motive on athletic performance in adolescent football players. *European Journal of Sport Science, 14*(5), 475–483. https://doi.org/https://dx.doi.org/10.1080/17461391.2013.837513

Zuber, C., Zibung, M., & Conzelmann, A. (2015). Motivational patterns as an instrument for predicting success in promising young football players. *Journal of Sports Sciences, 33*(2), 160–168. https://doi.org/10.1080/02640414.2014.928827

Zuber, C., Zibung, M., & Conzelmann, A. (2016). Holistic patterns as an instrument for predicting the performance of promising young soccer players - A 3-years longitudinal study. *Frontiers in Psychology, 7*, 1088. https://doi.org/10.3389/fpsyg.2016.01088

APPENDICES

Table A3.1 Detailed coding scheme and event definitions of performance indicators.

Indicator	Outcome	Definition
Pass forward	Successful – Unsuccessful	A situation in which the attacker attempts to play the ball to a teammate in the forward (i.e., opponent's goal) direction, by means of his foot/leg/head/torso/sliding. A pass is deemed successful if it reaches the intended teammate and is not touched by a defender. If the ball is touched by an opponent, the passing player will only be awarded a successful pass when it is clear that the pass would have also reached the intended teammate without the deflection. It is deemed unsuccessful if it does not reach the intended teammate, or does reach an unintended teammate but was touched/changed direction by an opponent.
Dribble	Successful – Unsuccessful	A contest between two or more players in which the attacker attempts to drive by a defender. It is deemed successful if the attacker drives by the defender and maintains possession of the ball. It is deemed unsuccessful if the attacker loses possession of the ball (e.g., often through a successful tackle by the defender). No dribble is awarded if the attacker dribbles in 'open space' and does not attempt to drive by a defender.
Take on	Successful – Unsuccessful	A contest between two or more players in which the attacker is challenged by the defender, often through physical contact, and aims to maintain control/possession of the ball and/or create space by actions that are not dribbles (e.g., a feint or 'trick'). It is deemed successful when the attacker maintains control of the ball or creates space to successfully pass to a teammate. It is deemed unsuccessful when the attacker loses possession of the ball.
Offensive aerial duel	Successful – Unsuccessful	A contest in the air between two players or more where the attacker (i.e., the player whose team was in possession) attempts to maintain control of the ball, either through passing to a teammate (e.g., by means of a header) or a successful touch. The attempt is deemed successful when the attacker or his teammate maintain possession. It is deemed unsuccessful when he loses possession.
Key pass	Counted when it occurs	The final pass that leads to the recipient of the ball having a successful shot attempt without scoring (i.e., a shot on target).
Assist	Counted when it occurs	The final pass that leads to the recipient of the ball scoring a goal.
Shot on target	Counted when it occurs	A scoring attempt that goes into the net (i.e., a goal) or an attempt that clearly would have gone into the net, but was saved by the goalkeeper or a player who is the last line of defense.

Table A3.1 (continued)

Indicator	Outcome	Definition
Tackle	Successful – Unsuccessful	A contest between two or more players in which the defender attempts to gain ball possession of an opposing player who is in possession, often through physical contact (e.g., a sliding). The tackle is deemed successful when he successfully takes the ball away from the opposing player, when his teammate gains possession, or when the ball goes out of play and is 'safe'. It is deemed unsuccessful when he does not gain possession or makes a foul.
Defensive aerial duel	Successful – Unsuccessful	A contest in the air between two players or more where the defender (i.e., the player whose team was not in possession) attempts to gain control of the ball, either through passing to a teammate (e.g., by means of a header) or a successful touch. The attempt is deemed successful when the defender or his teammate gain possession, or when the ball goes out of play and is 'safe'. It is deemed unsuccessful when he does not gain possession or makes a foul.
Interception	Counted when it occurs	A situation in which the defender 'reads' the pass of the opposing player and moves into the line of the intended the pass, thereby intercepting the pass. It is deemed successful when the defender gains possession, or when the ball goes out of play and is 'safe'. No interception is awarded if the defender accidently receives the ball from the opposing player (e.g., when the defender did not read the pass line, such as picking up a clearance).
Applying pressure	Counted when it occurs	A situation in which the defender puts pressure on an opposing player who has ball possession. It is successful when the player in possession loses the ball, often through an unsuccessful pass attempt. A successful pressure attempt can be followed by a tackle, when the defender attempts to conquer the ball through physical contact.

Table A3.2 Multilevel logistic regression analyses for the performance indicators with a successful – unsuccessful outcome in 7-vs-7 and 11-vs-11 gam

Performance indicators		**Forward passes (7-vs-7)**		**Forward passes (11-vs-11)**		**Offensive duels (7-vs-7)**		**Offensive duels (11-vs-11)**		**Defensive duels (7-vs-7)**		**Defensive duels (11-vs-11)**	
		Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Fixed Effects	Intercept	0.78	0.10	0.54	0.12	0.66	0.15	0.61	0.15	-0.21	0.14	0.05	0.18
	Team (U17)	0.57	0.17	-0.42	0.20	0.29	0.25	-0.3	0.23	-0.29	0.24	0.08	0.26
	Team (U19)	0.54	0.15	0.14	0.20	0.61	0.24	-0.16	0.24	0.07	0.21	0.65	0.29
	Team (U23)	0.99	0.16	0.20	0.20	0.27	0.24	-0.23	0.21	0.20	0.22	0.53	0.27
Random Effect (SD)	Intercept	0.22		0.40		0.35		0.14		0.37		0.41	

Coeff. = Estimated Regression Coefficient; *SD* = Standard Deviation; *SE* = Estimated Standard Error; The reference group for the factor 'Team' is the Under 15 (U15) age category.

Table A3.3 Spearman's correlations (95% CI in brackets) between the offensive performance indicators and shots on target (top), and defensive performance indicators and shots on target conceded (bottom), per age category and game format

Team	Game format	Passes forward	Offensive duels	Chances created	Data points
U15	SSG	0.04 (-0.26; 0.34)	0.26 (-0.05; 0.52)	0.57 (0.33; 0.75)	42
	11-vs-11	-0.26 (-0.89; 0.70)	0.32 (-0.66; 0.90)	0.40 (-0.61; 0.92)	6
U17	SSG	0.41 (0.03; 0.69)	0.06 (-0.34; 0.43)	0.39 (0.01; 0.68)	26
	11-vs-11	0.49 (-0.53; 0.93)	-0.12 (-0.85; 0.76)	0.28 (-0.69; 0.89)	6
U19	SSG	0.23 (-0.06; 0.47)	0.15 (0.13; 0.42)	0.39 (0.13; 0.60)	50
	11-vs-11	-0.06 (-0.83; 0.79)	-0.26 (-0.89; 0.70)	0.78 (-0.09; 0.97)	6
U23	SSG	0.26 (-0.04; 0.51)	0.20 (-0.10; 0.47)	0.57 (0.33; 0.74)	46
	11-vs-11	< .01 (-0.81; 0.81)	0.24 (-0.71; 0.88)	0.61 (-0.40; 0.95)	6

Table A3.3 (continued)

Team	Game format	Defensive duels	Pass interceptions	Applying pressure	Data points
U15	SSG	-0.15 (-0.43; 0.16)	-0.01 (-0.31; 0.30)	0.04 (-0.26; 0.34)	42
	11-vs-11	0.14 (-0.76; 0.85)	0.08 (-0.78; 0.84)	0.49 (-0.53; 0.93)	6
U17	SSG	-0.28 (-0.60; 0.13)	-0.02 (-0.40; 0.37)	-0.16 (-0.52; 0.24)	26
	11-vs-11	-0.6 (-0.95; 0.42)	0.67 (-0.32; 0.96)	-0.48 (-0.93; 0.54)	6
U19	SSG	-0.06 (-0.33; 0.22)	-0.07 (-0.35; 0.21)	-0.14 (-0.40; 0.14)	50
	11-vs-11	-0.13 (-0.85; 0.76)	-0.13 (-0.85; 0.76)	-0.39 (-0.91; 0.61)	6
U23	SSG	-0.18 (-0.44; 0.12)	-0.21 (-0.47; 0.09)	-0.17 (-0.44; 0.13)	46
	11-vs-11	0.31 (-0.67; 0.90)	0.12 (-0.76; 0.85)	-0.19 (-0.87; 0.74)	6

Table A3.4. Mean (and SD) events per 6 minutes on the performance indicators across all age categories (top) and per age category (bottom).

Performance indicator	SSG		11-vs-11	
	Mean (SD)	95% CI	Mean (SD)	95% CI
Passes forward	15.40 (3.71)	14.83; 15.97	13.35 (3.80)	11.83; 14.87
Tackles	4.62 (2.78)	4.20; 5.05	3.86 (1.50)	3.26; 4.46
Take ons	4.73 (2.76)	4.30; 5.15	3.73 (1.68)	3.06; 4.40
Applying pressure	3.20 (2.22)	2.86; 3.53	2.70 (1.53)	2.09; 3.31
Pass interceptions	2.52 (1.74)	2.26; 2.79	2.59 (1.13)	2.13; 3.04
Defensive aerial duels	0.24 (0.50)	0.17; 0.32	1.82 (1.12)	1.31; 2.28
Staying in front	2.37 (1.87)	2.09; 2.66	1.18 (0.68)	0.90; 1.45
Offensive aerial duels	0.29 (0.54)	0.20; 0.37	1.18 (0.82)	0.85; 1.51
Dribbles	1.43 (4.73)	0.70; 2.15	0.85 (3.73)	0*; 2.34
Shots on target	1.35 (1.19)	1.17; 1.54	0.42 (0.35)	0.28; 0.56
Shots	2.20 (1.48)	1.97; 2.43	0.80 (0.48)	0.60; 0.99
Chances created	0.72 (0.89)	0.58; 0.86	0.27 (0.19)	0.19; 0.34

Table A3.4- (continued)

Team	Game format	Passes forward	Tackles	Dribbles	Take ons	Staying in front	Offensive aerial duels	Defensive aerial duels	Applying pressure	Pass intercepti ons	Shots on target	Chances created
15	SSG	14.33 (3.84)	6.19 (3.00)	1.88 (1.67)	5.76 (3.16)	2.07 (1.94)	0.21 (0.47)	0.19 (0.40)	2.95 (2.27)	3.02 (1.99)	1.17 (1.31)	0.79 (1.00)
	11-vs-11	18.34 (1.97)	5.34 (0.75)	1.60 (1.20)	5.69 (1.48)	1.00 (0.40)	1.26 (0.76)	1.80 (1.29)	3.51 (1.77)	3.31 (0.97)	0.66 (0.49)	0.37 (0.20)
17	SSG	15.65 (4.19)	4.54 (1.96)	1.50 (1.39)	4.96 (2.34)	2.39 (1.58)	0.54 (0.76)	0.35 (0.56)	3.69 (2.60)	2.39 (1.55)	1.42 (1.30)	0.77 (0.86)
	11-vs-11	10.55 (3.13)	4.55 (1.10)	1.15 (0.84)	2.75 (1.32)	1.43 (1.21)	1.00 (0.49)	1.75 (0.71)	2.45 (1.97)	2.88 (0.62)	0.28 (0.28)	0.30 (0.25)
19	SSG	15.76 (3.49)	4.72 (2.78)	1.60 (1.47)	4.60 (2.67)	2.58 (2.19)	0.30 (0.51)	0.20 (0.50)	2.84 (2.02)	2.82 (1.84)	1.42 (1.21)	0.54 (0.73)
	11-vs-11	13.42 (0.99)	2.47 (1.17)	0.78 (0.54)	3.69 (1.39)	0.91 (0.50)	1.29 (1.09)	2.22 (1.47)	2.91 (1.25)	2.04 (1.41)	0.42 (0.26)	0.22 (0.14)
23	SSG	15.85 (3.46)	3.13 (2.17)	1.07 (1.02)	3.78 (2.40)	2.41 (1.59)	0.20 (0.45)	0.28 (0.54)	3.52 (2.11)	1.83 (1.22)	1.41 (1.00)	0.83 (0.95)
	11-vs-11	11.09 (2.53)	3.09 (0.99)	0.62 (0.44)	2.80 (0.61)	1.38 (0.18)	1.18 (1.00)	1.56 (1.08)	1.93 (0.71)	2.11 (1.10)	0.33 (0.28)	0.18 (0.14)

Table A3.5 Spearman's correlations (95% CI in brackets) between the performance indicators in the SSGs and 11-vs-11 games, per age category (i.e., team)

Team	Passes forward	Chances created	Shots on target	Pass Interceptions	Pressure	Offensive duels	Overall offensive performance	Defensive duels	Overall defensive performance
U15	0.08 (-0.42; 0.54)	-0.1 (-0.56; 0.40)	0.69 (0.32; 0.88)	0.71 (0.34; 0.89)	0.18 (-0.33; 0.61)	0.49 (-0.01; 0.79)	0.53 (0.04; 0.81)	0.08 (-0.42; 0.54)	0.20 (-0.31; 0.62)
U17	0.33 (-0.22; 0.72)	0.18 (-0.36 - 0.64)	0.13 (-0.41; 0.60)	0.32 (-0.23; 0.71)	0.47 (-0.06; 0.79)	-0.08 (-0.59; 0.47)	0.49 (-0.05; 0.81)	0.13 (-0.41; 0.60)	0.19 (-0.36; 0.64)
U19	0.40 (-0.12; 0.75)	-0.11 (-0.58 - 0.41)	0.43 (-0.08 - 0.76)	0.69 (0.30; 0.89)	0.22 (-0.31; 0.65)	0.38 (-0.19; 0.76)	0.53 (0; 0.83)	-0.02 (-0.53; 0.49)	0.27 (-0.28; 0.69)
U23	0.65 (0.21; 0.87)	0.04 (-0.48; 0.54)	0.09 (-0.44; 0.58)	0.24 (-0.31; 0.67)	0.68 (0.25; 0.88)	0.51 (0; 0.81)	0.24 (-0.31; 0.67)	-0.10 (-0.60; 0.46)	0.45 (-0.11; 0.79)

Table A3.6 Spearman's correlations (95% CI in brackets) between physiological and motor tests and overall offensive (top) and defensive performance (bottom) in 11-vs-11 games, per age category (i.e., team)

Team	10 m sprint	30 m sprint	ISRT	Agility
U15	-0.07 (-0.62; 0.52)	0.02 (-0.56; 0.59)	-0.26 (-0.68; 0.29)	0.08 (-0.51; 0.63)
U17	-0.58 (-0.85; -0.06)	-0.66 (-0.88; -0.20)	0.53 (-0.01; 0.83)	-0.55 (-0.84; -0.03)
U19	0.04 (-0.50; 0.56)	0.09 (-0.46; 0.59	0.33 (-0.24; 0.73)	0.31 (-0.27; 0.72)
U23	-0.07 (-0.56; 0.46)	-0.1 (-0.58; 0.43)	-0.03 (-0.53; 0.49)	-0.17 (-0.63; 0.37)
U15	0.04 (-0.55; 0.60)	-0.11 (-0.64; 0.50)	-0.07 (-0.55; 0.44)	0.20 (-0.42; 0.70)
U17	0.10 (-0.43; 0.58)	0.10 (-0.43; 0.58)	-0.30 (-0.70; 0.25)	0.22 (-0.33; 0.65)
U19	0.02 (-0.49; 0.53)	-0.11 (-0.59; 0.43)	0.01 (-0.51; 0.52)	-0.24 (-0.67; 0.31)
U23	0.05 (-0.49; 0.57)	0.19 (-0.38; 0.66)	-0.09 (-0.59; 0.46)	0.31 (-0.26; 0.72)

Table 4.1 Attributes scouts in the U12 age category considered predictive of future soccer performance.

Performance category	**Attribute**	***k***	**1st**
Technical	Technical attributes or technique with the ball[a]	37 (45%)	17 (68%)
	Ball control	16 (19%)	5 (20%)
	Pass intention or accuracy	7 (8%)	1 (4%)
	(Skills and abilities related to) transitioning[a]	4 (5%)	0 (0%)
	First touch	4 (5%)	0 (0%)
	(Skills and abilities related to) defending[a]	3 (4%)	0 (0%)
	(Skills and abilities related to) attacking[a]	2 (2%)	1 (4%)
	Shooting or shot technique	2 (2%)	0 (0%)
	Two legged	2 (2%)	0 (0%)
	Blocking	1 (> 1%)	0 (0%)
	Building up offensively	1 (> 1%)	0 (0%)
	Disrupting the offensive build up	1 (> 1%)	1 (4%)
	Dribbling	1 (> 1%)	0 (0%)
	Preventing goal scoring opportunities	1 (> 1%)	0 (0%)
	Scoring goals	1 (> 1%)	0 (0%)
	Performance category total	83 (28%)	25 (40%)
Tactical and perceptual-cognitive	Game sense and awareness	28 (42%)	7 (64%)
	Vision, perception, or seeing teammates and opponents, gaze behavior	12 (18%)	1 (9%)
	Positioning or moving without the ball	10 (15%)	2 (18%)
	Speed of handling	9 (14%)	1 (9%)
	Tactical skills	4 (6%)	0 (0%)
	Soccer intelligence	2 (3%)	0 (0%)
	Decision-making	1 (2%)	0 (0%)
	Performance category total	66 (22%)	11 (17%)
Physical, physiological, and motor skills	Physiological or motor skills[a]	20 (32%)	10 (59%)
	Running speed	18 (29%)	4 (24%)
	Physical attributes[a]	6 (10%)	0 (0%)
	Coordination	5 (8%)	0 (0%)
	Agility	4 (6%)	1 (< 6%)
	Body composition or athletic build	3 (5%)	1 (< 6%)
	Explosiveness	3 (5%)	0 (0%)
	Length	1 (< 2%)	0 (0%)
	Mobility	1 (< 2%)	0 (0%)
	Movement rhythm	1 (< 2%)	1 (< 6%)

Table 4.1 (continued)

Performance category	Attribute	*k*	1st
Personality-related and mental skills	Drive or intrinsic motivation	22 (34%)	5 (71%)
	Winning mindset or winning mentality	13 (20%)	2 (29%)
	Perseverance, resilience, or toughness	8 (13%)	0 (0%)
	Personality-related attributes[a]	7 (11%)	0 (0%)
	Behavior, on and off the pitch	5 (8%)	0 (0%)
	Assertiveness or dominance	4 (6%)	0 (0%)
	Positive attitude	3 (5%)	0 (0%)
	Coaching other players or leadership	1 (< 2%)	0 (0%)
	Performance or goal oriented	1 (< 2%)	0 (0%)
	Performance category total	64 (21%)	7 (11%)
Miscellaneous	Communication	6 (27%)	0 (0%)
	Team understanding, involving teammates	5 (23%)	0 (0%)
	Innate talent (nature)	3 (14%)	2 (67%)
	Undefined[b]	3 (14%)	1 (33%)
	Adaptability	2 (9%)	0 (0%)
	Coachability, fast learner, or growth mindset	2 (9%)	0 (0%)
	Education level	1 (< 5%)	0 (0%)
	Performance category total	22 (> 7%)	3 (5%)
	Grand total	298	63

Results are presented as absolute number of answers with percentage in brackets. Percentages per attribute refer to the percentage within performance category, whereas percentage for performance category total row refer to percentage of grand total number of answers.

[a] indicates an answer that can be considered a 'general´ domain, rather than a more specific predictor.

[b] answers that did not contain enough content information to be considered a predictor and could not be assigned to a performance category (e.g., "matching the playing style of club [..]").

Table 4.2 Attributes scouts in the U13 – U15 age category considered predictive of future soccer performance.

Performance category	Attribute	*k*	1st
Technical	Technical attributes or technique with the ball[a]	33 (52%)	14 (82%)
	(Skills and abilities related to) transitioning[a]	7 (11%)	1 (6%)
	(Skills and abilities related to) defending[a]	6 (9%)	1 (6%)
	Ball control	6 (9%)	1 (6%)
	(Skills and abilities related to) attacking[a]	3 (5%)	0 (0%)
	Shooting or shot technique	3 (5%)	0 (0%)
	First touch	2 (3%)	0 (0%)
	Pass intention or accuracy	2 (3%)	0 (0%)
	Applying pressure	1 (< 2%)	0 (0%)
	Dribbling	1 (< 2%)	0 (0%)
	Performance category total	64 (29%)	17 (38%)
Tactical and perceptual-cognitive	Game sense and awareness	17 (36%)	3 (25%)
	Speed of handling	9 (19%)	4 (33%)
	Positioning or moving without the ball	7 (15%)	0 (0%)
	Decision-making	6 (13%)	4 (33%)
	Vision, perception, or seeing teammates and opponents, gaze behavior	6 (13%)	1 (> 8%)
	Soccer intelligence	1 (2%)	0 (0%)
	Tactical skills[a]	1 (2%)	0 (0%)
	Performance category total	47 (22%)	12 (27%)
Physical, physiological, and motor skills	Physiological or motor skills[a]	14 (31%)	4 (67%)
	Running speed	12 (27%)	0 (0%)
	Physical attributes[a]	10 (22%)	1 (< 17%)
	Body composition or athletic build	3 (7%)	1 (< 17%)
	Strength in duels	2 (4%)	0 (0%)
	Coordination	1 (> 2%)	0 (0%)
	Length	1 (> 2%)	0 (0%)
	Mobility	1 (> 2%)	0 (0%)
	Stability	1 (> 2%)	0 (0%)
	Performance category total	45 (21%)	6 (13%)
Personality-related and mental skills	Winning mindset or winning mentality	12 (30%)	2 (29%)
	Personality-related attributes[a]	8 (20%)	1 (14%)
	Drive or intrinsic motivation	6 (15%)	3 (43%)

Table 4.1 (continued)

Performance category	**Attribute**	***k***	**1st**
Personality-related and mental skills	Coachability, fast learner, or growth mindset	4 (10%)	0 (0%)
	Coaching other players or leadership	3 (8%)	0 (0%)
	Behavior, on and off the pitch	2 (5%)	1 (14%)
	Focus or concentration	1 (< 3%)	0 (0%)
	Performance or goal oriented	1 (< 3%)	0 (0%)
	Perseverance, resilience, or toughness	1 (< 3%)	0 (0%)
	Positive attitude	1 (< 3%)	0 (0%)
	Self-confidence	1 (< 3%)	0 (0%)
	Performance category total	40 (18%)	7 (16%)
Miscellaneous	Team understanding, involving teammates	5 (24%)	1 (> 33%)
	Communication	4 (19%)	0 (0%)
	X-factor	4 (19%)	0 (0%)
	Undefined[b]	3 (14%)	1 (> 33%)
	Biological age	2 (10%)	0 (0%)
	Calendar age	2 (10%)	1 (> 33%)
	Lifestyle	1 (< 5%)	0 (0%)
	Performance category total	21 (10%)	3 (< 7%)
	Grand total	217	45

Results are presented as absolute number of answers with percentage in brackets. Percentages per attribute refer to the percentage within performance category, whereas percentage for performance category total row refer to percentage of grand total number of answers.
[a] indicates an answer that can be considered a 'general´ domain, rather than a more specific predictor.
[b] answers that did not contain enough content information to be considered a predictor and could not be assigned to a performance category (e.g., "matching the playing style of club [..]").

Table 4.3 Attributes scouts in the U16 – U18 age category considered predictive of future soccer performance.

Performance category	**Attribute**	***k***	**1st**
Technical	Technical attributes or technique with the ball[a]	6 (75%)	2 (67%)
	Ball control	1 (< 13%)	0 (0%)
	Two legged	1 (< 13%)	1 (33%)
	Performance category total	8 (19%)	3 (33%)
Tactical and perceptual-cognitive	Game sense and awareness	3 (38%)	0 (0)
	Decision-making	1 (< 13%)	1 (50%)
	Positioning or moving without the ball	1 (< 13%)	0 (0%)
	Speed of handling	1 (< 13%)	1 (50%)
	Tactical skills[a]	1 (< 13%)	0 (0%)
	Vision, perception, or seeing teammates and opponents, gaze behaior	1 (< 13%)	0 (0%)
	Performance category total	8 (19%)	2 (22%)
Physical, physiological, and motor ability	Physical attributes	5 (45%)	0 (0%)
	Running speed	4 (36%)	0 (0%)
	Coordination	1 (> 9%)	0 (0%)
	Length	1 (> 9%)	1 (100%)
	Performance category total	11 (26%)	1 (11%)
Personality	Winning mindset or winning mentality	4 (40%)	1 (50%)
	Personality-related attributes	2 (20%)	0 (0%)
	Coaching other players or leadership	1 (10%)	0 (0%)
	Drive or intrinsic motivation	1 (10%)	0 (0%)
	Focus or concentration	1 (10%)	0 (0%)
	Perseverance, resilience, or toughness	1 (10%)	1 (50%)
	Performance category total	10 (24%)	2 (22%)
Miscellaneous	Appearance	1 (20%)	0 (0%)
	Coachability, fast learner, or growth mindset	1 (20%)	0 (0%)
	Team understanding, involving teammates	1 (20%)	0 (0%)
	Undefined[b]	1 (20%)	1 (11%)
Miscellaneous	X-factor	1 (20%)	0 (0%)
	Performance category total	5 (12%)	1 (11%)
	Grand total	42	9

Results are presented as absolute number of answers with percentage in brackets. Percentages per attribute refer to the percentage within performance category, whereas percentage for performance category total row refer to percentage of grand total number of answers.

Table 4.4 Attributes scouts in the adult age category considered predictive of future soccer performance.

Performance category	Attribute	*k*	1st
Technical	Technical attributes or technique with the ball[a]	6 (75%)	1 (1%)
	Ball control	2 (25%)	0 (0%)
	Performance category total	8 (22%)	1 (13%)
Tactical and perceptual-cognitive	Game sense and awareness	5 (45%)	1 (33%)
	Speed of handling	4 (36%)	1 (33%)
	Soccer intelligence	1 (> 9%)	1 (33%)
	Positioning or moving without the ball	1 (> 9%)	0 (0%)
	Performance category total	11 (30%)	3 (38%)
Physical, physiological, and motor ability	Physiological or motor skills[a]	4 (44%)	1 (50%)
	Physical attributes[a]	2 (22%)	1 (50%)
	Running speed	2 (22%)	0 (0%)
	Strength in duels	1 (> 11%)	0 (0%)
	Performance category total	9 (24%)	2 (25%)
Personality	Winning mindset or winning mentality	3 (43%)	1 (50%)
	Drive or intrinsic motivation	2 (29%)	1 (50%)
	Assertiveness or dominance	1 (14%)	0 (0%)
	Perseverance, resilience, or toughness	1 (14%)	0 (0%)
	Performance category total	7 (19%)	2 (25%)
Miscellaneous	Team understanding, involving teammates	1 (50%)	0 (0%)
	Undefined	1 (50%)	0 (0%)
	Performance category total	2 (5%)	0 (0%)
	Grand total	37	8

Results are presented as absolute number of answers with percentage in brackets. Percentages per attribute refer to the percentage within performance category, whereas percentage for performance category total row refer to percentage of grand total number of answers.
[a] indicates an answer that can be considered a ‘general´ domain, rather than a more specific predictor
[b] answers that did not contain enough content information to be considered a predictor and could not be assigned to a performance category (e.g., “matching the playing style of club [..]”).

Table A4.5 Spearman's correlations between the statement scores on the different aspects of structure.

	Q1	Q2	Q3	Q4	Q5
Q1 - Before observing a player, I already know which attributes I will evaluate	-				
Q2 - When observing a player, I evaluate each attribute I find important separately	0.23 (0.05; 0.39)	-			
Q3 - I evaluate different players - of the same age and playing position - on the same attributes	0.37 (0.21; 0.52)	0.34 (0.18; 0.49)	-		
Q4 - After observing a player, I use my overall impression of the player's attributes to form my final prediction	0.11 (-0.07; 0.28)	0.26 (0.09; 0.42)	0.22 (0.04; 0.38)	-	
Q5 - After observing a player, I sum my scores on the independently evaluated attributes to form my final prediction	0.03 (-0.14; 0.21)	0.45 (0.29; 0.58)	0.27 (0.10; 0.42)	0.12 (-0.05; 0.42)	-

Note: 95% CI in brackets.

Table A5.1: Difference in correlations of participants' market value predictions, between conditions 1

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 A5.2 Results from regression model predicting the logarithm of participants prediction 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

$R^2 = 0.53$, $R^2_{adj} = 0.49$, $F(7, 88) = 14.26$, $p < 0.01$
β = beta coefficient, *SE* = standard error,
[a] natural logarithm of player market value; [b] in the 2015-2016 soccer season; [c] Average of three attacking and defending ratings, respectively.
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%; predictors ordered by relative importance.

Table A5.3 Pearson's correlations between different predictors in regression model predicting participants' prediction of players' market value.

	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	0.50	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	0.40	0.32	0.40	0.84	0.51	0.40	0.63	1	-	-	-	-	-
Avg def. rating	0.53	0.87	0.82	0.81	0.57	0.80	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	0.20	0.07	0.06	0.14	0.19	0.13	1	-	-
Games played (2015-2016)[a]	0.24	-0.06	0.20	-0.04	0.13	0.14	0.10	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	0.20	0.14	1

Note: *att.* = attacking, *avg.* = average, *trans.* = transitioning, *def.* = defending,

[a] denotes background information of the player in the 2015 – 2016 soccer season. Order of att. and def. rating refers to order in Table 5.1

APPENDICES | 193

Table A5.4 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 A5.5 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.

DANKWOORD

Dankwoorden staan doorgaans bol van de clichés die – hoe kant het ook anders - ook de sportwereld niet vreemd zijn: "Dit was een echte team effort, ik had dit nooit kunnen bereiken zonder jullie hulp, bedankt dat jullie er voor me waren in stressvolle tijden," enzovoorts. Clichés zijn echter clichés omdat ze een kern van waarheid bevatten (wat op zichzelf ook weer een cliché is). Ik wil deze uitspraken daarom hier ongegeneerd herhalen en iedereen bedanken die ervoor gezorgd heeft dat ik de afgelopen vijf jaar met zoveel plezier aan dit proefschrift heb kunnen werken.

Rob, Susan, Ruud, en Wouter: ik had me geen beter team van (co-)promotoren kunnen wensen. Jullie zijn simpelweg fijne mensen en ik heb genoten van onze samenwerking. Bij jullie stond de deur altijd open om te sparren over nieuwe ideeën of, onder het genot van een kop koffie, het weekend te bespreken. Bedankt voor jullie geduld wanneer ik weer eens mijn PhD-frustraties moest uiten en voor jullie begrip wanneer ik weer eens een deadline had gemist.

Rob, bedankt voor de creatieve vrijheid die je deze 'kluns' gaf. Je was altijd enthousiast en supportive over mijn onderzoeksvoorstellen. Het uitvoeren verliep soms wat hobbelig, maar hierdoor heb ik geleerd om beter op mijn eigen benen te staan. Ik ga je tirades over 'godslasterende onderzoekers die de leer van de Heilige Paul Meehl verwerpen' missen. Susan, je bent een van de slimste mensen die ik ken. Je bewaarde altijd het overzicht, bracht je nuance, en wist vaak belangrijke details te benoemen die ik over het hoofd had gezien. Daarmee bewaakte je de (methodologische) kwaliteit van de hoofdstukken. Dank daarvoor! Ruud, bedankt voor alle falafel lunches, fijne gesprekken, en dagelijkse supervisie. Toegegeven, wanneer ik als feedback op een artikel "dat kan nog wel een stukkie strakker" kreeg, wilde ik soms mijn laptop uit het raam gooien. Maar wanneer ik er dan nog een keer goed voor ging zitten kwamen je opmerkingen de papers altijd enorm ten goede. Wouter, je talent om een brug te slaan tussen de wetenschap en de voetbalpraktijk is voor mij een inspiratie. Wanneer ik tijdens een onderzoek het overzicht kwijt was wist je altijd een stap terug te nemen, en in simpele taal tot de kern van het probleem te komen. Ik verliet onze gesprekken dan weer met nieuwe energie om aan de slag te gaan. Allen, ik hoop van harte dat we elkaar in de toekomst nog regelmatig zien en spreken.

Beste leescommissie, prof. dr. Born, prof. dr. Savelsbergh en prof. dr. Van Yperen, bedankt voor de tijd die jullie hebben besteed aan het lezen en beoordelen van mijn proefschrift.

Roomies Marvin en Jasmine, bedankt dat jullie kamer 0212 met mij wilden delen. Marvin, de house-en-techno dj-sessie doen we de volgende keer in Grunn! Ook bedankt aan alle collega's van de afdeling, in het bijzonder Henk, Casper, en Marieke: alle hulp, tips, en feedback op de hoofdstukken werd erg gewaardeerd!

Ook dank ik alle collega's van de Talent Development and Creativity-groep, in het bijzonder Nico, Barbara, Yannick, Joske, en Rick. Bedankt voor jullie gezelligheid; ik liep altijd graag een verdieping naar boven om bij jullie te buurten.

Dank aan alle staf, scouts, en (talent performance) coaches van de KNVB die hebben bijgedragen- of deelgenomen aan het onderzoek. In het bijzonder natuurlijk Jan en Frederike: jullie hulp in het faciliteren van de verschillende studies was van onschatbare waarde. Ik kijk er naar uit om straks met jullie samen te werken en het voetbalonderzoek bij de KNVB verder op de kaart te zetten.

Dank aan alle spelers, trainers, scouts, en staf van FC Groningen, in het bijzonder Mees, Sil, Luca, Byron, en Arno. Jullie hulp tijdens het uitvoeren van de onderzoeken was fantastisch. Misschien nog wel belangrijker; van jullie heb ik ontzettend veel geleerd over de voetbalpraktijk. Onze gesprekken maakten onderzoek doen bij een professionele club voor mij tot een nog rijkere ervaring.

Beste Hajo, *gooozer*, ik had niet verwacht dat een willekeurig ochtend ongeconcentreerd samenwerken (met name te danken aan mijn brakheid) zou resulteren in zo'n fijne vriendschap. Bedankt voor alle afleiding (zie, Tour de France kijken tijdens werkuren). Op nog duizend wilde plannen, waarvan we er waarschijnlijk geen één uitvoeren. Lisanne, we gaan way back, en ik ben ontzettend blij dat we nog steeds lekker gaan. Dank voor alle steun, niet alleen als paranimf, maar vooral als *mien moat*.

Basketball boys van de LAN-party: bedankt dat jullie me weer met open armen in Utrecht hebben ontvangen. Speciale shoutout naar Jitse en Marloes; bedankt dat ik in jullie woonkamer aan de laatste loodjes van dit proefschrift mocht werken.

Eus, Henry, Dide en Dico, bedankt voor jullie support. Jullie maken het stereotiepe beeld van schoonfamilie niet waar (en dat bedoel ik in positieve zin).

Mam, pap en Marlies, bedankt voor jullie onvoorwaardelijke liefde en steun. Dankzij jullie kon ik mijn hart volgen. Werken gaat makkelijker wanneer je weet dat je altijd familie hebt waarop je kan terugvallen in onrustige tijden.

Marije, *You the real MVP*. Zonder goed te overleggen solliciteerde ik voor deze functie en voor de liefde verhuisde je mee naar Groningen. Er zijn niet genoeg clichés om te zeggen hoe veel ik van je hou. Ten slotte Art: ondanks dat ik nu in een staat van constante vermoeidheid leef, ben jij onze *Work of Art*.

CURRICULUM VITAE

Tom Bergkamp was born in Almelo in 1992. He successfully followed the Bachelor program 'Interdisciplinary Social Sciences (graduated in 2015) and Research Master program 'Methodologies and Statistics for the Behavioural, Biomedical, and Social Sciences (graduated in 2017) at Utrecht University. During his studies, he worked as a student teacher to assist in various methodological courses and summer schools. In 2017, he moved to Groningen to start his PhD research program 'Performance Prediction in Team Sports;' a collaboration between the University of Groningen, the KNVB, and FC Groningen. The PhD project was supervised by prof.dr. Rob Meijer, dr. Ruud den Hartigh, dr. Susan Niessen, and dr. Wouter Frencken, and focused on implementing concepts and insights from the field of selection psychology to the field of talent selection in soccer. The research conducted in this project led to various publications in high-impact scientific and professional journals. During this project, Tom presented at different international conferences, such as the World Congress on Science and Football and the FEPSAC (European Federation of Sports Psychology), as well as (non-)academic institutions, such as soccer clubs and high schools. In 2020, he was selected to be a 'Face of Science;' a project by the KNAW and De Jonge Akademie. Faces of Sciences offers a platform for young researchers, who present their research to high school students through blogs and vlogs. Tom also appeared in (soccer-related) popular media, such as VI and NPO radio 1. He will continue to be involved in soccer related research a researcher at the KNVB.

LIST OF PUBLICATIONS

Peer-reviewed publications

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, 22*(7), 994 - 1004. https://doi.org/10.1080/17461391.2021.1916081

Bergkamp, T. L. G., den Hartigh, R. J. R., Frencken, W. G. P., Niessen, A. S. M., & Meijer, R. R. (2020). The validity of small-sided games in predicting 11-vs-11 soccer game performance. *PLOS ONE, 15*(9), e0239448. https://doi.org/10.1371/journal.pone.0239448

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

Bergkamp, T. L. G., Niessen, A. S. M., den Hartigh, R. J. R., Frencken, W. G. P., & Meijer, R. R. (2018). Comment on: "Talent identification in sport: A systematic review." *Sports Medicine, 48*(6), 1517–1519. https://doi.org/10.1007/s40279-018-0868-6

Submitted

Bergkamp, T. L. G., Meijer, R. R., den Hartigh, R. J. R., Frencken, W. G. P., & Niessen, A. S. M. (2022). *Examining the reliability and predictive validity of performance assessments by soccer coaches and scout: the influence of structured collection and mechanical combination of information*.

Van Kesteren, E. J. & Bergkamp, T. L. G. (2022). *Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage*.

Professional publications

Bergkamp, T. L. G., Niessen, A. S. M., Hartigh, den, R. J. R., Meijer, R. R., & Frencken, W. G. P. (2020). (On)terecht buitenspel gezet. Sportprestaties voorspellen door systematische en gestructureerde beoordelingen. *SportGericht*, *74*(4), 36 - 40.

Media

Faces of Science

Voetbal International – Tasten in het duister: onwetendheid en Babylonische spraakverwarring bij voetbalscouts

NPO Radio 1 FOCUS – "heeft het zin om topsporters op (zeer) jonge leeftijd te scouten?"