

Bonetti, Leonardo et al.

Working Paper

A refined analysis of the datasets used in "Decoding the elite soccer player's psychological profile"

I4R Discussion Paper Series, No. 312

Provided in Cooperation with:

The Institute for Replication (I4R)

Suggested Citation: Bonetti, Leonardo et al. (2026) : A refined analysis of the datasets used in "Decoding the elite soccer player's psychological profile", I4R Discussion Paper Series, No. 312, Institute for Replication (I4R), s.l.

This Version is available at:

<https://hdl.handle.net/10419/342982>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

No. 312

DISCUSSION PAPER SERIES

A refined analysis of the datasets used in Decoding the elite soccer player's psychological profile

Leonardo Bonetti

Reza Jafari

Martin Ingvar

Alberto Filgueiras

Torbjörn Vestberg

Debora Seghezzi

Morten L. Kringelbach

Predrag Petrovic

This paper responds to:

Moore, K.. 2026. Concerns regarding "Decoding the elite soccer player's psychological profile". *I4R Discussion Paper Series* No. 311. Institute for Replication.

August 2026

I4R DISCUSSION PAPER SERIES

I4R DP No. 312

A refined analysis of the datasets used in Decoding the elite soccer player's psychological profile

Leonardo Bonetti^{1,2,3}, Torbjörn Vestberg⁴, Reza Jafari⁴, Debora Seghezzi³, Martin Ingvar⁴, Morten L. Kringelbach^{1,2}, Alberto Filgueiras⁵, Predrag Petrovic⁴

¹*Aarhus University, Aalborg/Denmark*

²*University of Oxford/United Kingdom*

³*University of Bologna/Italy*

⁴*Karolinska Institutet, Stockholm/Sweden*

⁵*University of Gloucestershire, Cheltenham/United Kingdom*

AUGUST 2026

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur

University of Ottawa

Jörg Ankel-Peters

RWI – Leibniz Institute for Economic Research

E-Mail: joerg.peters@rwi-essen.de

RWI – Leibniz Institute for Economic Research

Hohenzollernstraße 1-3

45128 Essen/Germany

www.i4replication.org

ISSN: 2752-1931

1 First and foremost, we would like to make the following points clear.

2 The author of the critique of the dataset reported in our original paper made a valuable
3 contribution by identifying several issues that had not been recognized during the original
4 review and publication process. These observations highlighted the need for additional
5 clarification, further analyses, and more extensive validation to determine which conclusions
6 of the original study remain well supported, which require more cautious interpretation, and
7 which would benefit from independent replication.

8 The author submitted their critique to both *Proceedings of the National Academy of Sciences*
9 (PNAS) and Institute For Replication (I4R). Before PNAS had reached a decision regarding
10 the correspondence process, I4R elected to publish the critique. As a result, we were required
11 to publish only a partial version of our response. Some of the supporting material could not be
12 disclosed at this stage, either because of ethical considerations or because it was necessary to
13 first discuss it with PNAS and coordinate the editorial process.

14 Consequently, what is fundamentally a single scientific discussion requiring careful
15 clarification has become fragmented across two publication venues, with neither venue
16 presenting the complete exchange. This is obviously non-ideal, but we do not have any control
17 on those decisions. The present document therefore attempts to provide a comprehensive
18 response to the issues raised in the critique. It discusses both the strengths and the limitations
19 of the original work in light of the concerns identified, presents the additional analyses we have
20 conducted, and clarifies which conclusions remain supported by the available evidence and
21 which should be interpreted more cautiously.

22 Attached, an additional letter of response to the points raised by the author of the critique letter
23 is provided, alongside a statement by Dr. Alberto Filgueiras, who conducted the data collection
24 for the original work.

A refined analysis of the datasets used in

Decoding the elite soccer player's psychological profile

Bonetti, Leonardo ^{1,2,3,5*}, Vestberg, Torbjörn ^{4*}, Jafari, Reza ⁴, Seghezzi, Debora ⁵, Ingvar, Martin ⁴, Kringelbach, Morten L. ^{1,2,3}, Filgueiras, Alberto ^{6*}, Petrovic, Predrag ^{4*}

¹ *Center for Music in the Brain, Department of Clinical Medicine, Aarhus University & The Royal Academy of Music, Aarhus, Aalborg, Denmark*

² *Centre for Eudaimonia and Human Flourishing, Linacre College, University of Oxford, Oxford, United Kingdom*

³ *Department of Psychiatry, University of Oxford, Oxford, United Kingdom*

⁴ *Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.*

⁵ *Department of Psychology, University of Bologna*

⁶ *Psychological Sciences, University of Gloucestershire, School of Natural, Social and Sport Sciences, QU214, Francis Close Hall Campus, Swindon Road, Gloucestershire, Cheltenham GL53 7JX, UK*

**equally contributing authors*

Introduction and overview of response

We would like to sincerely thank the editors and the independent evaluators for their careful and thorough assessment of our work. We fully recognize the seriousness of the concerns raised regarding the dataset and the conclusions derived from it, and we appreciate the opportunity to address these issues in detail.

The concerns raised in the editorial letter prompted us to immediately initiate an internal audit aimed at reviewing the data collection procedures, reconstructing the provenance of the dataset as thoroughly as possible, and reassessing the analyses and interpretations presented in the original manuscript. As part of this process, we worked directly with Alberto Filgueiras, who conducted the Brazilian data collection, in order to recover original records and reconstruct the workflow through which the published dataset was assembled.

A central objective of this audit was to verify the traceability of the collected data and to better understand how the discrepancies and inconsistencies highlighted by the evaluators may have arisen. This proved challenging because, after the data collection period, Alberto Filgueiras relocated first to the United Kingdom and later to Australia, which made it more difficult to recover and compare records across sites. In addition, since the data collection was completed several years ago, and given the relevant university policies, the original records were not routinely retained in full and were, to our knowledge, largely deleted. Fortunately, a small portion of the original material had been retained separately for demonstrational purposes, and this allowed us to gain additional insight into the data, the procedures, and the possible sources of inconsistency. This material was enclosed in the original response to PNAS but could not be enclosed here in the published response for I4R due to confidentiality and ethical reasons. We are very sorry that this response had to be scattered among two venues, a solution which we consider far from ideal.

The recovered records were overall consistent with the dataset used in the published analyses and support the conclusion that the data collection itself genuinely took place. Within this subset, we did not identify evidence suggestive of fabricated data. In addition, we collected letters from colleagues, students, and assistants stating their involvement in the data collection. These letters, together with the recoverable original data, are enclosed with our response. At the same time, the audit did identify a number of issues requiring clarification or correction. Specifically, we identified:

- a systematic scoring-related issue affecting one personality-related variable, leading to revisions relative to the original reported results;

- several isolated inconsistencies and apparent transcription or processing errors affecting some Tower of Hanoi measurements, which contribute to some of the irregularities highlighted by the evaluators;
- inaccuracies in parts of the procedural descriptions originally provided regarding aspects of the Brazilian data collection and test administration.

Importantly, these issues were not uniform across all measures included in the study. Our reassessment indicated that some analyses are more affected by these concerns than others, whereas several findings continue to show interpretable and structured patterns across independent analyses, normative comparisons, and partially recoverable source material. We therefore considered it important to reassess each major component of the study separately rather than treating the dataset as either entirely reliable or entirely invalid.

In this response, we first provide additional clarification regarding the data collection procedures and reconstruction process. We then address directly the specific concerns raised by the editors and independent evaluators. Finally, we provide a detailed reassessment of each major test and analytical component included in the original study, specifying which findings appear more robust, which require greater caution in interpretation, and which are more substantially affected by the concerns raised.

Additional methodological clarifications on data collection

To further clarify aspects of the dataset that are relevant to the interpretation of the statistical patterns discussed below, we provide a concise overview of the data collection procedures for the Brazilian cohort.

Data collection was conducted across five professional soccer clubs in Brazil (please, note that their names were shared confidentially with PNAS but could not be reported here for ethical reasons) through established collaborations with sports psychologists, physiologists, and conditioning staff affiliated with the clubs. While the overall protocol and order of tasks were standardized, data acquisition took place in field settings within club facilities rather than in controlled laboratory environments. All participants were assessed using the same core set of instruments and procedures within each cohort.

The measures included in the present study were collected over two sessions separated by approximately one week. In the first session, participants completed demographic data (e.g., age and years of education), self-reported personality measures, Five-Point Test, Simple Manual Reaction Time, Digit Span, and Tower of Hanoi. In the second session, participants completed three scales of the D-KEFS battery: Trail Making Test (number sequencing, letter sequencing, and number-letter switching), Design Fluency, and Color-Word Interference Test. This separation of sessions was implemented to reduce fatigue and ensure consistent testing conditions across participants. This procedural detail is important for interpreting the inconsistencies identified by the evaluators, which we discuss later in this letter.

Cognitive measures were collected using a combination of paper-based neuropsychological tests and tablet-based or web-based applications. Data for control participants were collected separately in a university setting using the same instruments and overall protocol, including the two-session structure and the order of tests and questionnaires administered. Testing was conducted in small groups under the supervision of experimenters, with standardized instructions applied consistently across participants.

As part of the audit process described above, we were also able to recover the names of players associated with one club. This information (which was shared with PNAS but not within the published response for I4R for ethical reasons) further supports the traceability of the participant recruitment process and confirms that the cohort included players affiliated with the professional club during the relevant seasons, including first-team players, youth players, and temporarily integrated players participating within the professional training environment.

Response to the main concerns raised by the editors

The concerns raised are grouped into four main points, according to the editor requests. We address each of them directly here and provide detailed analyses in the additional sections below.

1. Hypothesis that control data may be independently randomly generated

We appreciate the seriousness of this concern and carefully re-examined the dataset with this specific question in mind.

It is important to note that the Brazilian cohort was assessed across two separate testing sessions. The first session included personality measures, Five-Point Test, Simple Manual Reaction Time, Digit Span, and Tower of Hanoi. The second session included the D-KEFS measures (Trail Making Test, Design Fluency, and Color-Word Interference Test).

A key observation emerging from our reassessment is that the statistical irregularities highlighted by the evaluators were not uniformly distributed across the dataset. Rather, the primary concerns were concentrated in measures collected during the first testing session within the control sample. By contrast, the player dataset showed substantially clearer latent structure and inter-variable organization across measures, and the D-KEFS variables collected during the second session displayed more coherent psychometric relationships in both players and controls.

Our audit process does not provide evidence supporting the hypothesis that the dataset was fabricated or arbitrarily generated. Although the recoverable source material and participant records were limited and restricted to a subset of players from one club, the recoverable cases were consistent with actual data collection procedures and with the published dataset. While this material is insufficient to fully reconstruct or independently validate the entire dataset, it supports the conclusion that the data collection itself genuinely took place.

At the same time, we agree that parts of the control dataset, specifically those originating from the first testing session, display statistical properties that are atypical for psychometric data. In several instances, the observed relationships between variables are substantially weaker than expected from prior literature and standard psychometric behaviour. One possible interpretation is that additional measurement noise was introduced during portions of the first-session control data collection, thereby reducing the internal structure of those measures. However, this interpretation remains speculative, and we do not believe that the currently available evidence allows definitive conclusions regarding the precise origin of these statistical properties.

Precisely because of these unresolved concerns, we considered it important not to rely exclusively on the original control comparisons when reassessing the study findings. We therefore re-examined the football player data using additional approaches whenever possible, including comparisons against normative references and independent datasets. This was feasible for personality measures and working memory variables (counting for seven out of the ten variables included in the control dataset flagged by the evaluators), where the re-analyses yielded patterns broadly compatible with the original findings.

By contrast, comparable reassessment strategies were not available for the Five-Point Test, Simple Manual Reaction Time, and Tower of Hanoi (three of the ten variables). For these variables, the currently available evidence neither demonstrates that the original findings were incorrect nor provides the same degree of confidence as for the other measures. We therefore believe it is appropriate to substantially reduce the interpretative emphasis placed on those specific results.

In conclusion, the statistical irregularities highlighted by the evaluators appear primarily concentrated within the control sample for a subset of measures collected during the first testing session. By contrast, the player data and the D-KEFS measures collected during the second session retain substantially clearer structure and more coherent psychometric organization. Importantly, for seven of the ten variables mostly flagged by the evaluators, reassessment using normative references and/or independent datasets yielded patterns broadly compatible with the original findings. The remaining three variables should instead be interpreted with substantially greater caution. While the available evidence does not indicate that the original findings were necessarily incorrect, the observed relationships may have been influenced by the atypical statistical properties identified in this portion of the control dataset, thereby reducing confidence in the robustness of these specific results.

2. Distributional and correlational patterns inconsistent with prior literature

As discussed above, we acknowledge that some variables display distributional or correlational patterns that deviate from typical expectations reported in the literature. Importantly, however, these irregularities were not uniformly distributed across the dataset. Rather, they were primarily concentrated within the control sample for the measures collected during the first testing session (personality measures, Five-Point Test, Simple Manual Reaction Time, Digit Span, and Tower of Hanoi). By contrast, the player data generally retained clearer latent structure and more coherent inter-variable relationships, and the D-KEFS measures collected during the second testing session showed substantially more typical psychometric organization

in both players and controls. This observation prompted a detailed re-examination of both the data and the underlying procedures. Through this process, we identified several methodological and measurement-related factors that help clarify some of the patterns highlighted by the evaluators.

Personality measures

We thank the evaluators for raising important questions regarding the personality assessment, as these prompted a detailed re-examination of this portion of the dataset. Through this process, and in consultation with co-author Alberto Filgueiras, who conducted the Brazilian data collection, we identified a reporting error in the manuscript: the personality questionnaire used was the NEO-FFI-30 rather than the Big Five Inventory (BFI). This clarification is relevant for interpreting several distributional characteristics highlighted by the evaluators. The NEO-FFI-30 consists of 30 items, with 6 items per personality dimension rated on a 1–5 Likert scale, resulting in expected score ranges between 6 and 30. In the present dataset, at least one participant did not complete a single item, which accounts for the observed minimum value of 5 in one dimension. This resolves the apparent inconsistency in score ranges noted by the evaluators. Importantly, although the specific questionnaire was misreported, this does not alter the psychological constructs being assessed, as both the NEO-FFI-30 and the BFI measure the same five-factor personality dimensions (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness).

At the same time, we acknowledge that the personality measures in the control sample display flatter and less internally structured distributions than typically expected in psychometric datasets. By contrast, the player sample showed more coherent inter-variable relationships and distributions broadly compatible with the expected organization of personality dimensions. In addition, because standardized normative references are available for these measures, we were able to reassess the player data independently of the original control comparisons. These additional analyses yielded patterns broadly compatible with the original findings, although requiring more cautious interpretation and reassessment of some specific effects, as discussed in the corresponding sections below.

Working memory (digit span)

The evaluators correctly noted that the digit span variables displayed weaker-than-expected correlations and unusually compressed distributions, particularly within the control sample. During the audit process, we identified that the task implementation differed from the standard

WAIS administration originally described in the manuscript. Rather than employing the standard WAIS task involving multiple trials per level and partial scoring across digit span sequences, the implemented task used a single sequence for each span length, progressing from shorter to longer sequences (2 to 8 digits). Performance was recorded as the maximum span successfully completed. This implementation substantially compresses variability and reduces the number of possible values, particularly in relatively homogeneous samples such as elite football players. As a consequence, distributions become narrower and correlations between forward and backward span are expected to be attenuated relative to standard WAIS-based studies (an example of that is specifically reported below in the section focused entirely on Digit Span).

To further examine this issue, we reanalysed additional independent datasets collected in previous studies at Aarhus University, Denmark, by author Leonardo Bonetti. When these datasets were rescored using the same simplified scoring approach employed in the Brazilian cohort, they similarly showed substantially reduced variance and more polarized distributions compared to standard WAIS scoring. Thus, while the original manuscript inaccurately described the implementation of the task, the observed distributional properties become considerably more understandable under the actual administration procedure.

Importantly, despite the atypical distributional characteristics observed in the controls, the player working memory data retained coherent structure and correlations, and reassessment against external normative references and independent datasets yielded patterns broadly compatible with the original findings. To the best of our current knowledge, the evidence therefore continues to support the conclusion that the football players exhibited elevated working memory performance relative to normative expectations.

Simple Manual Reaction Time

Reaction times were collected in field conditions using tablet-based hardware and a web-based task. Under these conditions, timing precision is constrained by display refresh rates, touchscreen input latency, and software-level timing variability. These factors reduce temporal precision, compress the observable range, and reduce the occurrence of extreme values, resulting in distributions that appear flatter, less skewed, and partially discretized relative to laboratory-based reaction time data. While these characteristics plausibly account for some of the unusual distributional properties observed, we acknowledge that this measure remains difficult to validate against external normative standards under the present implementation. We

therefore believe the corresponding findings should be interpreted with substantially greater caution.

Tower of Hanoi

The Tower of Hanoi results represent the most problematic portion of the dataset identified during the audit process. During the reassessment, we identified that the completion-time and number of moves variables were derived using a non-standard scoring approach in which the practice phase (3-disk condition) and the main task (5-disk condition) were combined into single measures. This procedure differs from the more typical approach in the literature, where the primary task is analysed separately, and may introduce additional non-task-specific variance into the measure. As a consequence, the relationship between completion time and number of moves may become substantially weakened and less comparable to previously published findings. In addition, comparison of the recoverable source material with the published dataset revealed some isolated transcription or processing inconsistencies affecting this measure. Importantly, these discrepancies appeared random rather than systematic and represent the clearest example identified during the audit process of random inconsistencies between recoverable records and the final dataset. Although still limited in number, such inconsistencies further reduce confidence in the robustness of the Tower of Hanoi variables and plausibly contribute to the atypical correlational structure highlighted by the evaluators.

Taken together, these observations suggest that the distributional and correlational irregularities highlighted by the evaluators do not reflect a single uniform issue affecting the dataset as a whole. Rather, they appear to be primarily concentrated within the control sample for the measures collected during the first testing session. By contrast, the player data generally retained clearer latent structure and more coherent psychometric organization across measures. Importantly, the audit process also clarified that several of the apparent inconsistencies originated from incomplete, inaccurate, or insufficiently specific descriptions of the implemented procedures in the original manuscript, rather than from problems inherently affecting the underlying psychological constructs or the data themselves. In several cases, once the actual implementations and scoring procedures were reconstructed and reassessed, including through comparisons with normative references and additional datasets, the observed patterns became substantially more understandable and remained broadly compatible with the original findings. At the same time, we acknowledge that some measures, particularly Tower of Hanoi and Simple Manual Reaction Time, remain substantially more difficult to interpret

robustly and should therefore receive considerably reduced emphasis and more cautious interpretation than in the original manuscript.

3. Similar ranges across subgroups

We acknowledge the observation that several variables display unexpectedly similar ranges across subgroups. As with the correlational irregularities discussed above, however, these patterns were not uniformly distributed across the dataset and became substantially more understandable following the methodological reassessment conducted during the audit process.

In particular, several of the affected measures were implemented using procedures that inherently compress variability and restrict the range of observable values. This was especially the case for the working memory (digit span) variables, where performance was recorded as the maximum successfully completed span using a simplified administration procedure with only one sequence per level. Such an implementation naturally produces a limited number of discrete outcomes and substantially reduces variability, particularly within relatively homogeneous samples such as elite football players. Reanalysis of independent datasets rescored using the same procedure confirmed that this approach markedly narrows the observed range and increases clustering of scores relative to standard WAIS-style scoring.

Similarly, the Simple Manual Reaction Time task was collected under field conditions using tablet-based and web-based implementations. Device-related timing constraints, including touchscreen latency and refresh-rate discretization, reduce temporal precision and compress the observable range of reaction times, thereby limiting the occurrence of extreme values and increasing similarity in empirical boundaries across groups.

For the personality measures, the apparent range constraints also became more understandable following clarification of the actual questionnaire used (NEO-FFI-30 rather than BFI), as the scoring structure of the instrument imposes well-defined minimum and maximum values for each dimension.

At the same time, the relatively homogeneous nature of the Brazilian elite-player cohort likely further contributed to reduced variability for some measures. Importantly, however, these observations do not fully explain all of the statistical properties identified within the control sample for the first testing session, which remain more atypical than expected and are discussed separately above.

Taken together, our reassessment suggests that many of the apparent similarities in ranges across subgroups, particularly within the player sample, can be plausibly explained by the actual implementation and scoring characteristics of the measures used, combined with the

relatively homogeneous nature of the elite football cohort. In several cases, these patterns became substantially more understandable once the originally incomplete or inaccurate methodological descriptions were clarified through the audit process. By contrast, the control data from the first testing session remain more difficult to reconcile with typical psychometric expectations, consistent with the broader concerns discussed above. Precisely for this reason, we considered it important to reassess the player findings using normative references and independent datasets whenever possible, rather than relying exclusively on the original control comparisons.

4. Large effect sizes between players and controls

We acknowledge that some variables show relatively large effect sizes between elite players and controls. However, these effects are not isolated to a single measure but are observed across multiple variables and analytical approaches, including both univariate analyses and multivariate classification models.

At the same time, the magnitude of some of these effects must be interpreted in light of the statistical and methodological properties discussed above. Cohen's *D* depends not only on differences between group means but also on the variability within each group. In the present dataset, several variables, particularly working memory measures, were implemented using coarse-grained scoring procedures with a limited number of possible values. Combined with the relatively homogeneous nature of the elite football sample, this substantially reduced within-group variance and consequently amplified standardized effect sizes.

For example, the observed working memory differences correspond to relatively modest absolute shifts in performance, but because the player distributions were highly compressed, the resulting pooled standard deviation became small, thereby inflating the corresponding effect sizes. Similar, although less pronounced, effects may also contribute to the elevated effect sizes observed in some personality and reaction time measures.

Importantly, the atypical statistical properties identified in parts of the control dataset from the first testing session may also have contributed to amplifying some of these group contrasts. Precisely for this reason, we considered it important to reassess the player findings against normative references and independent datasets whenever possible, rather than relying exclusively on the original control comparisons. This new analysis showed for instance clearly significant differences between players and norms in personality which however exhibited reduced effect size compared to the original results. On the contrary, the effect size of the

working memory remained rather high even when comparing the players with normative values or additional external datasets.

The multivariate analyses further indicate that the observed effects are not driven solely by any single variable or by the measures most affected by the concerns discussed above. When questioned variables were examined in isolation, classification performance was substantially reduced. Conversely, excluding these variables did not eliminate the broader multivariate separation between players and controls, suggesting that the predictive signal was distributed across multiple measures rather than dependent on one isolated anomaly.

Taken together, these observations suggest that the large effect sizes reported in the original manuscript likely reflect a combination of genuine group differences, compressed variance associated with the implemented measurement procedures, and atypical statistical properties affecting portions of the control dataset. Consequently, while some of the originally reported effect sizes should likely be interpreted more cautiously than initially presented, the overall pattern of differences between elite players and comparison samples does not appear to depend solely on isolated artefactual effects. The new analyses using norms further reinforces the conclusion that elite players are characterised by different psychological features compared to the general population.

Additional clarification regarding recoverable source material and data reconstruction

Before addressing the individual tests and analyses in detail, we provide additional clarification regarding the recoverable source material obtained during the audit process. As described above, only a limited portion of the original Brazilian records could ultimately be recovered. Of this, informed consent and the data collected during the first session (i.e. every variable reported in the original paper¹ but D-KEFS) corresponds to 9 players from Botafogo (team 2 in the published dataset). The D-KEFS scores cannot instead be directly traced back to specific individuals. They include 11 DF Fille Dots, 15 DF Switching, 11 TMT Numbers, 8 TMT Letters, 18 TMT Switching. Please, note that correspondent numbers in the D-KEFS do not necessarily indicate a match between individuals (i.e. DFF_1 does not necessarily correspond to TMTL_1).

Importantly, the recovered material did not contain a consistent or complete identification system linking participants directly to the final dataset. In several cases, names or identifying information had already been removed as part of the anonymization process. Consequently, the recovered records had to be manually traced back to the 9 players in the published dataset through comparison of score patterns and available metadata. This reconstruction process was highly informative for understanding both the strengths and the limitations of the dataset. Importantly, most numbers found in the raw sheets matched the ones in the dataset used for the analyses, with a few discrepancies identified during the process almost entirely localized to the Tower of Hanoi variables, particularly the number-of-moves measure, where some isolated random-looking transcription inconsistencies were observed.

In addition, the reconstruction process allowed us to identify one systematic scoring-related error affecting the personality variables Agreeableness and, to a much smaller extent, Openness to Experience. Importantly, these issues appeared systematic rather than random, suggesting a scoring or coding problem rather than irregularities in the underlying data collection itself. Fortunately, because the scoring logic could be reconstructed clearly, these issues could also be corrected systematically. Specifically, the Openness to Experience issue involved one out of six items being reverse-scored incorrectly, resulting in only negligible changes to the corresponding values. By contrast, the Agreeableness variable involved four out of six items being scored in the reversed direction, leading to a substantial and systematic alteration of that measure. Because the nature of the error could be identified precisely, we were able to reconstruct and correct the corresponding values across the dataset. Additional details regarding these corrections are provided in the dedicated personality section below.

Finally, it is important to note that some information could no longer be recovered because it had already been removed during the anonymization process. In particular, the original sheets linking participant identities to anonymized records were no longer available. Similarly, for some D-KEFS measures, only partial scoring information remained recoverable. The original scoring forms included several intermediate task variables which were taken or processed for completeness but not used in the analysis (e.g. completion times, number of errors, penalty-adjusted times, number of correct designs, repeated designs, etc.) and which are no longer available. By contrast, the recoverable records contain only some final values for the CWI measures, where completion times included the addition of 3-second penalties for each error committed. Such adjusted values were computed in the context of these neuropsychological tests but, according to standard D-KEFS procedures, were not used for deriving the standardized scores, which were instead based on the original completion times. We report these details in the interest of full transparency regarding the reconstruction process and the currently available source material.

Executive functions (D-KEFS)

To evaluate the internal consistency and structure of executive function measures, we focused on the D-KEFS battery, which includes Trail Making Test (TMT), Design Fluency (DF), and Color-Word Interference (CWI) subtests. These measures are standardized and widely used, with scaled scores centred on a normative mean of 10, derived from large representative samples and commonly employed in both research and clinical settings ²⁻⁴.

In line with this, and as outlined in our previous letter ⁵, the mean (and standard deviation) performance of the control participants in the present dataset closely approximates this normative value across subtests. This indicates that the control group is representative of the general population with respect to executive functioning, thereby supporting the validity of the comparison and the overall quality of the control sample with regards to the second session of data collection. Importantly, we reiterate that the use of 10 as a reference value in our analyses is not arbitrary, as suggested by one of the evaluators, but reflects the standardized mean of D-KEFS scaled scores, a convention widely used internationally and already detailed in both the original paper and our previous response ²⁻⁵.

In the original article ¹, we did not report a direct comparison between D-KEFS measures for players and controls, as we chose to rely on standardized normative values. This decision was motivated by the absence of a control group for the Swedish cohort, making normative scores the most appropriate and consistent reference across both Swedish and Brazilian player samples.

However, in response to the concerns raised by the evaluators regarding the use of normative comparisons, we conducted additional analyses to directly compare players and controls (**Table 1** and **Figure 1**) and further reinforce our original results. These provide indeed a complementary perspective to the original approach and further support our findings: control participants closely align with the normative reference value of 10, while elite players consistently show higher performance across subtests. Importantly, we reiterate that the D-KEFS was computed on the second session of data collection, not the one where controls exhibited the most evident issues in data structure.

Test	Df	t-value	p-value	Mean players (CI)	Mean controls (CI)
<i>TMT-1</i> (<i>TMT2</i> in <i>python code</i>)	245	7.81	$< 1.7 \cdot 10^{-13}$	12.74 (12.30-13.19)	10.65 (10.35-10.94)
<i>TMT-2</i> (<i>TMT3</i> in <i>python code</i>)	245	6.02	$< 6.2 \cdot 10^{-9}$	12.79 (12.34-13.24)	11.02 (10.64-11.39)
<i>TMT-3</i> (<i>TMT4</i> in <i>python code</i>)	245	5.43	$< 1.4 \cdot 10^{-7}$	12.59 (12.17-13.00)	10.88 (10.42-11.34)
<i>DF-1</i>	245	12.53	$< 3.6 \cdot 10^{-28}$	16.20 (15.62-16.79)	10.77 (10.14-11.40)
<i>DF-2</i>	245	13.82	$< 1.6 \cdot 10^{-32}$	16.42 (15.94-16.91)	10.77 (10.13-11.42)
<i>DF-3</i>	245	11.89	$< 5.0 \cdot 10^{-26}$	15.91 (15.33-16.47)	10.69 (10.04-11.35)
<i>CWI -1</i>	245	4.72	$4.2 \cdot 10^{-6}$	11.87 (11.37-12.37)	10.03 (9.45-10.61)
<i>CWI -2</i>	245	3.76	0.0002	11.51 (11.11-11.89)	10.05 (9.39-10.71)
<i>CWI -3</i>	245	4.51	$< 9.9 \cdot 10^{-6}$	12.13 (11.61-12.65)	10.19 (9.52-10.87)
<i>CWI -4</i>	245	5.10	$< 6.7 \cdot 10^{-7}$	12.29 (11.87-12.72)	10.05(9.29-10.80)

Table 1. Enhanced executive functions in players

Results for the ten independent two-sample *t*-tests used to assess whether the performance of players in Trail Making Test, Design Fluency Test and Color Word Interference Test was different in comparison to the control group. The table shows the tests (and the relative subscales), degrees of freedom, *t*-value, *p*-value, mean with 95% low and high confidence intervals (CI) in parenthesis.

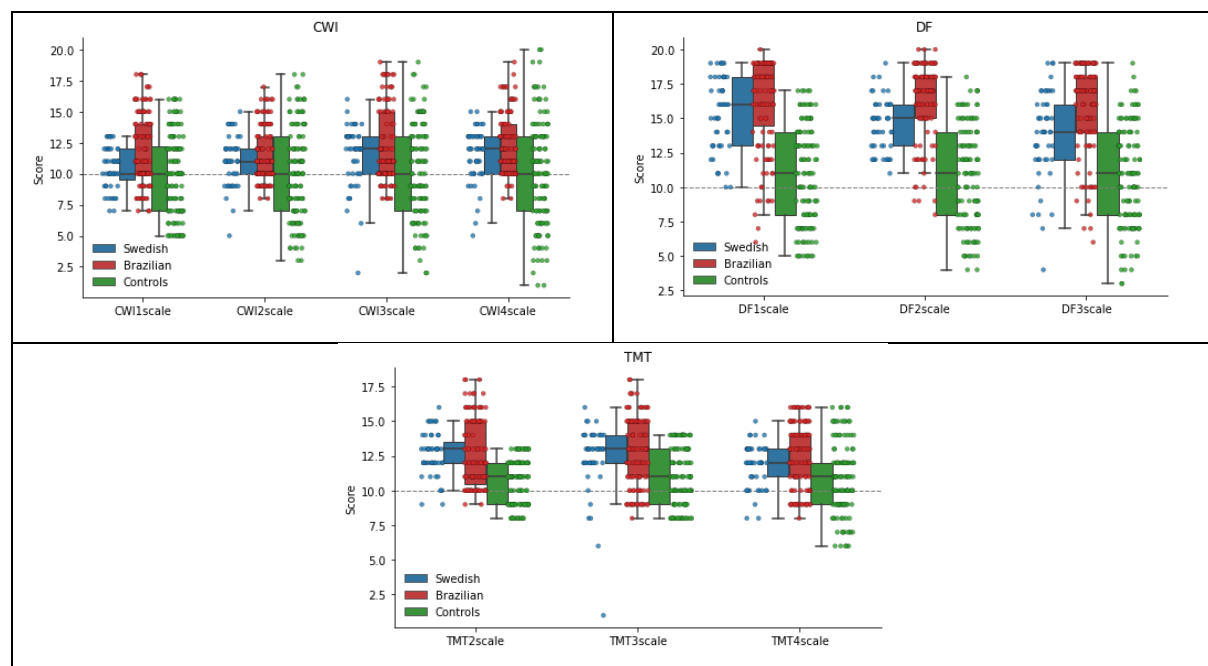


Figure 1. Distribution of D-KEFS scores across groups.

The figure shows the distribution of scores for the three D-KEFS scales, Trail Making Test (TMT), Design Fluency (DF), and Color-Word Interference (CWI), for Swedish players, Brazilian players, and control participants. For each variable, data are represented using boxplots and individual data points. The dashed horizontal line indicates the normative reference value of 10. The distributions illustrate the variability of the data within each group and confirm that scores span a broad range without evidence of artificial clustering or constraint.

After this initial analysis, we also assessed the latent structure of the D-KEFS measures using principal component analysis (PCA). This was conducted separately for the Swedish players, Brazilian players, and control samples. PCA is a well-established method for evaluating whether a set of variables reflects coherent underlying dimensions, as it captures the extent to which variance is shared across measures and identifies dominant components summarizing this structure⁶. As a matter of fact, PCA was also rightfully used by the evaluators in their analysis.

Across all three datasets, PCA revealed clear and interpretable structure (**Figure 2**). The variance explained by successive components showed a consistent decay pattern, with the first component accounting for approximately 38% of the variance in the Brazilian players, 34% in the Swedish players, and a comparable proportion in the control group. The cumulative variance explained by the first three components was similarly high across datasets, indicating robust latent structure.

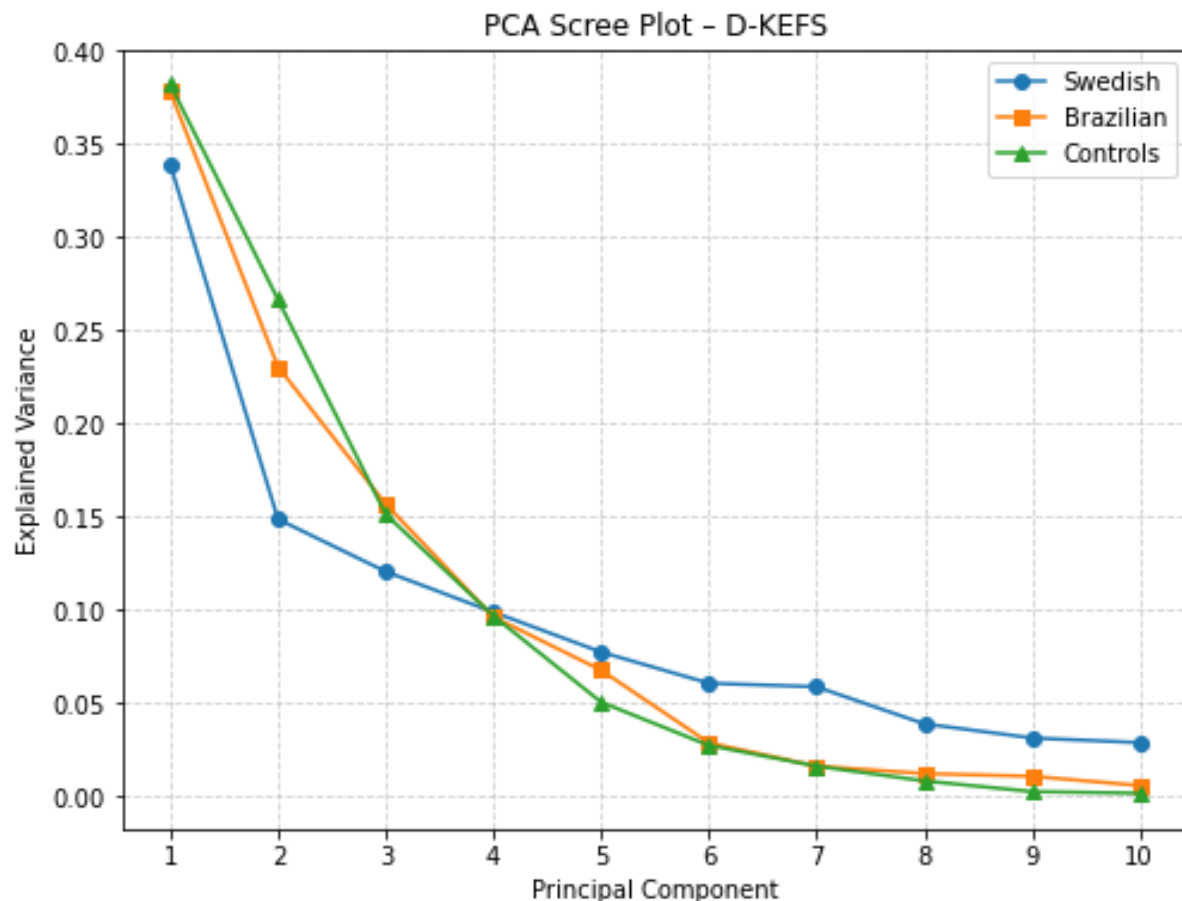


Figure 2. PCA scree plot of D-KEFS measures across groups.

The figure shows the proportion of variance explained by successive principal components (PC1–PC10) for the D-KEFS variables in Swedish players, Brazilian players, and control participants. Across all three datasets, the variance exhibits a consistent decay pattern, with the first principal component accounting for the largest proportion of variance (approximately 34% in Swedish players, 38% in Brazilian players, and a comparable value in controls). The subsequent components contribute progressively less variance, as expected in real-world multivariate data. The similarity of these patterns across independent groups indicates the presence of coherent latent structure intrinsic to the D-KEFS measures, rather than patterns driven by a specific subset of participants or random variation.

Importantly, the first principal component consistently captures the dominant structure of the data, while subsequent components account for progressively smaller proportions of variance, as expected in real-world multivariate datasets. The similarity of these patterns across independent groups, including the control sample, is particularly notable and suggests that the

observed structure is intrinsic to the measures rather than driven by a specific subset of participants.

Examination of the component weights further supports this interpretation (**Figure 3**). The first component is primarily driven by Color-Word Interference (CWI) measures, indicating a consistent and dominant latent structure. The second and third components show more distributed patterns, which still broadly reflect contributions from Design Fluency (DF) and Trail Making Test (TMT) measures, respectively. While these higher-order components are less sharply defined, as expected in real-world data, the overall organization of variables is consistent across groups and closely matches the known structure of the D-KEFS battery and indicates that the variables cluster meaningfully according to their cognitive domains. It should be noted that the sign of the weights is arbitrary in PCA; therefore, interpretation should focus on the relative grouping of variables, which remains consistent across datasets despite some variability, as expected in empirical data.

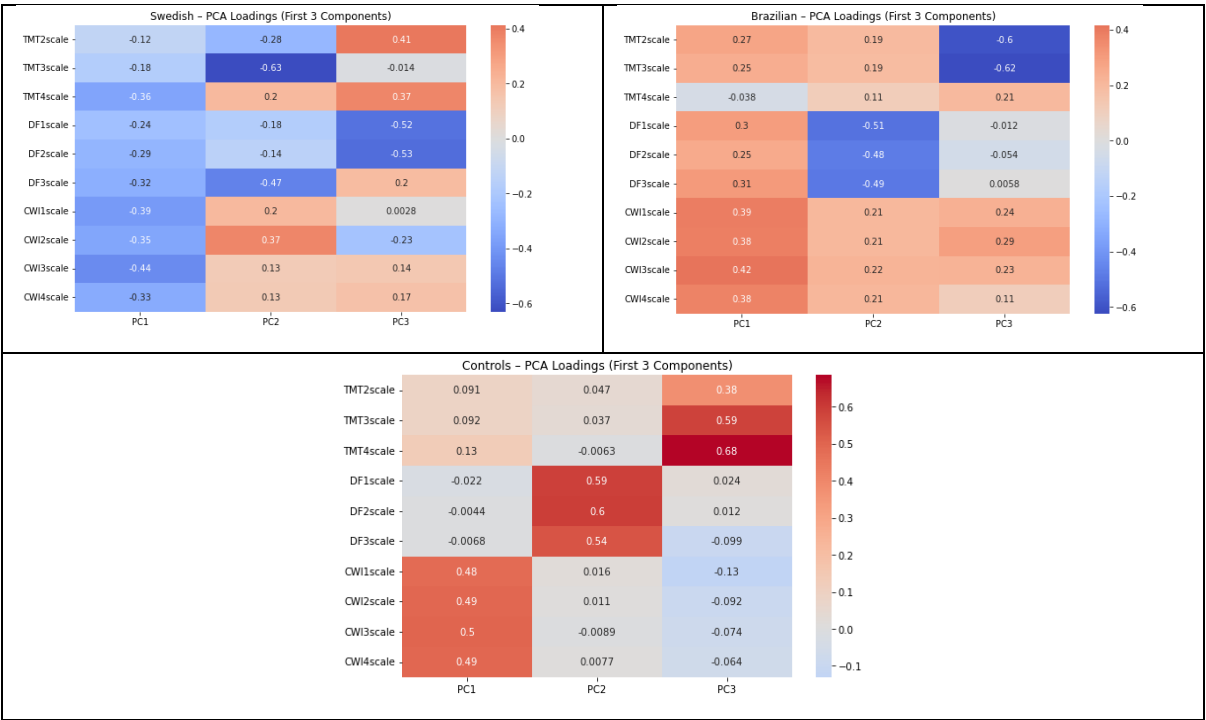


Figure 3. PCA weights of D-KEFS variables across groups.

The figure shows the weights of the first three principal components (PC1–PC3) for the D-KEFS variables in the Swedish players, Brazilian players and controls. It should be noted that the sign of PCA weights is arbitrary; therefore, interpretation should focus on the relative grouping of variables rather than the direction of the coefficients. Across all datasets, the first component, which explains the largest proportion of variance, is primarily driven by Color-Word Interference (CWI) measures, indicating a consistent and dominant latent

structure. The second and third components show more distributed patterns, still broadly reflecting contributions from Design Fluency (DF) and Trail Making Test (TMT) measures, respectively. While these higher-order components are less sharply defined, as expected in real-world data, the overall organization of variables is consistent across groups and aligns with the known structure of the D-KEFS battery.

Crucially, this pattern is not easily reconciled with what would typically be expected under independently generated or purely random data. Random data would not produce coherent component structure aligned with established cognitive domains, nor would it yield consistent patterns across multiple independent datasets. Instead, the observed organization reflects well-established relationships among executive function measures.

As pointed out by the evaluators, the strength of pairwise correlations between D-KEFS variables differed between the Brazilian and Swedish samples, with generally higher correlations observed in the Brazilian cohorts. While this pattern warrants careful consideration, several factors may plausibly contribute to these differences. The D-KEFS measures examined (Design Fluency, Trail Making Test, and Color-Word Interference) all rely to varying degrees on overlapping cognitive processes, including processing speed, attentional control, visual scanning, cognitive flexibility, and rapid response generation. In some populations, these shared demands may generate stronger covariance across tasks, resulting in elevated intercorrelations. One relevant difference between the cohorts is that the Swedish sample consisted exclusively of elite football players and included substantial linguistic and cultural heterogeneity, with at least 16 of the 52 players originating from different national backgrounds and several likely completing testing outside their native language context. Combined with the relatively small sample size and restricted range associated with elite-only sampling, these factors may plausibly reduce intercorrelations by increasing task-specific variance and measurement noise. By contrast, the Brazilian cohort was comparatively more homogeneous with respect to language and testing environment, which may have increased shared variance across executive-function measures. Importantly, however, the normative D-KEFS reference values derived from large, standardized samples lie broadly between the correlation magnitudes observed in the Brazilian and Swedish datasets (see images attached to this response and please note that the TMT scales that we used were: Number Sequencing, Letter Sequencing and Number-Letter Switching; Visual Scanning was not employed), suggesting that the observed differences represent variation in the strength of covariance rather than fundamentally incompatible psychometric structures. At the same time, we acknowledge

that some correlations in the Brazilian sample remain higher than typically reported in the literature and therefore warrant cautious interpretation. Nonetheless, the overall multivariate organization of the D-KEFS measures remained coherent and remarkably consistent across datasets, as reflected by the PCA analyses, supporting the conclusion that these variables retain meaningful latent executive-function structure rather than patterns expected under random or independently generated data.

Personality measures (Big Five)

We thank the evaluators for raising important questions regarding the personality assessment, as these prompted a detailed re-examination of this portion of the dataset. Through this process, and in consultation with co-author Alberto Filgueiras, who conducted the Brazilian data collection, we identified a reporting error in the manuscript: the personality questionnaire used was the NEO-FFI-30 ⁷ rather than the Big Five Inventory (BFI). Importantly, although the specific questionnaire was misreported, this does not alter the psychological constructs being assessed, as both the NEO-FFI-30 and the BFI measure the same five-factor personality dimensions (Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness). The clarification is nevertheless relevant for interpreting several distributional characteristics highlighted by the evaluators. The NEO-FFI-30 consists of 30 items, with 6 items per personality dimension rated on a 1–5 Likert scale, resulting in expected score ranges between 6 and 30. In the present dataset, at least one participant did not complete a single item, which accounts for the observed minimum value of 5 in one dimension and resolves the apparent inconsistency in score ranges noted by the evaluators.

During the audit process, a recoverable subset of the original questionnaires was rescored manually in order to assess the accuracy of the database transcription and scoring procedures. This re-evaluation indicated that the questionnaire responses themselves had been entered consistently into the dataset. However, a systematic scoring-related issue was identified for the Openness to Experience and Agreeableness dimensions. Specifically, the issue concerned the reverse-scoring procedure applied to negatively keyed items. For Openness to Experience, one reverse-scored item had not been inverted correctly, whereas for Agreeableness the same issue affected four out of six reverse-scored items. Because the NEO-FFI-30 dimensions are computed as sums of six items, this produced a systematic distortion of the resulting scores, particularly for Agreeableness. Importantly, the systematic nature of these discrepancies suggests a scoring or coding problem rather than inconsistencies in the underlying questionnaire responses themselves. Because the rescored subset contained both the originally stored values and the manually corrected values, it was possible to estimate the relationship between incorrect and corrected scoring. Separate calibration models were therefore derived for Openness to Experience and Agreeableness and subsequently applied to the remaining cases for which the original questionnaires could no longer be retrieved. The discrepancy observed for Openness to Experience was minimal, consistent with the fact that only one item was affected. By contrast, the Agreeableness scores were more substantially altered because

multiple reverse-scored items had been coded incorrectly. Consequently, the recalibrated Agreeableness values differed meaningfully from the original results.

At the same time, we understand the concerns raised by the evaluators regarding the relatively flat and uniform-looking distributions observed in parts of the control sample. While such patterns are not, in themselves, definitive evidence of invalid data, they are atypical for psychometric datasets and warrant caution in interpretation, particularly given the broader concerns discussed elsewhere regarding portions of the first testing session. Precisely for this reason, and because standardized normative references are available for the NEO-FFI-30, we considered it important to reassess the personality findings independently of the original control comparisons. We therefore repeated the principal analyses focusing only on the football player sample and comparing the resulting scores against published normative values.

The normative reference values were obtained from Körner and colleagues⁸, who reported population-based norms for the NEO-FFI-30 in a large German sample using the same 30-item version of the instrument. Although the use of German normative data instead of Brazilian population norms is not ideal and should be considered a limitation, these norms were derived from a large and psychometrically validated sample and therefore provide a useful independent reference framework for reassessing the player data. Trait scores for the player group were compared against the published normative means (after adjustment of the scores from 0-4 scale to 1-5 scale) for males aged 18–33 years using one-sample t-tests. Because the Openness to Experience and Agreeableness dimensions had been affected by the systematic scoring issue described above, the recalibrated values derived from the reconstruction procedure were used for these analyses. Confidence intervals and effect sizes (Cohen's d) were also computed in order to evaluate the magnitude and direction of the observed deviations from normative expectations.

Results from this analysis are reported in **Table 2**.

Trait	N	M(sample)	SD(sample)	M(norm)	SD(norm)	t	p	d	CI95(low)	CI95(high)
N	153	12.12	4.74	14.39	4.49	-5.93	1.9e-08	-.48	11.36	12.87
E	153	22.50	4.38	21.01	3.41	4.18	4.8e-05	.34	21.79	23.19
O	153	18.03	3.82	18.29	4.23	-0.85	.40	-.07	17.42	18.64
A	153	24.92	3.52	21.31	3.93	12.70	1.2e-25	1.03	24.37	25.49
C	153	23.49	3.70	22.63	3.71	2.88	.005	.23	22.90	24.08

Table 2. Personality trait scores of elite players compared against NEO-FFI-30 normative values.

Results of one-sample *t*-tests comparing the personality scores of elite football players ($N = 153$) against published normative means from Körner et al. (2015) for males aged 18–33 years using the NEO-FFI-30. The table reports sample means (M), sample standard deviations (SD), normative means and standard deviations, *t*-values, *p*-values, Cohen's *d* effect sizes, and 95% confidence intervals (CI) for the player sample means. Analyses for Openness to Experience (O) and Agreeableness (A) were conducted using recalibrated values derived from the scoring-correction procedure described in the text. Trait abbreviations: N = Neuroticism, E = Extraversion, O = Openness to Experience, A = Agreeableness, C = Conscientiousness.

Taken together, these analyses indicate that elite football players in the present sample exhibit a personality profile that differs meaningfully from normative expectations across several dimensions. Importantly, the reassessment against independent normative references broadly confirmed the original pattern of findings for Neuroticism, Extraversion, and Conscientiousness. By contrast, Openness to Experience, which already showed relatively modest effects in the original analyses, no longer appeared meaningfully different from normative expectations following recalibration. Most importantly, correction of the systematic scoring issue substantially altered the Agreeableness results, which remained strongly significant but in the opposite direction relative to the originally reported findings. These results therefore support the presence of distinct personality characteristics in the player sample, while also highlighting the importance of the methodological corrections identified during the audit process.

Five-Point Test, Simple Manual Reaction Time, and Tower of Hanoi

We group these three measures together because, unlike the D-KEFS, personality, and working memory variables, robust external normative reassessment strategies were either unavailable or substantially limited for these tasks. Consequently, although the results obtained for these measures cannot be considered necessarily not reliable, they currently provide substantially weaker support for the study conclusions and require considerably more cautious interpretation.

For the Five-Point Test, one methodological aspect related to the implementation procedure itself. The task was administered using a 2-minute version rather than the 3-minute version, for which some normative references might be available. While this does not invalidate the measure, it limits comparability with published normative datasets and therefore reduces our ability to independently reassess the findings against external reference values.

For the Simple Manual Reaction Time variable, the data were collected in field conditions using tablet-based hardware (iPad 2) and a web-based task (<https://humanbenchmark.com/tests/reactiontime>), which explicitly acknowledges that device and display characteristics can affect timing accuracy. In practical terms, such setups can introduce both fixed and variable delays⁹: for example, touchscreen input and display latency can add on the order of tens to over one hundred milliseconds, and responses are effectively registered in discrete time steps determined by the screen refresh cycle (e.g. $\approx 16\text{--}17$ ms at 60 Hz). Consequently, a true underlying reaction time (e.g. ≈ 240 ms) may be recorded as ≈ 256 ms or ≈ 272 ms depending on when it is sampled within a refresh interval, while additional device-related delays (e.g. up to ≈ 100 ms) can compress the observable range and reduce the presence of very fast or very slow responses. When combined with a relatively homogeneous adult sample and a simple detection task, these factors can plausibly yield distributions that appear flatter, less skewed and even slightly more bimodal than expected and exhibit similar empirical bounds across groups tested under identical conditions, as the evaluators pointed out. Importantly, all participants were assessed using the same devices, software, and procedures, such that any measurement constraints would have applied uniformly across groups and are therefore unlikely to bias between-group comparisons. We are also grateful for the opportunity to clarify that the Simple Manual Reaction Time variable was reported alongside the Five-Point Test primarily because the two tasks were administered sequentially during the same testing session, rather than because they were intended to measure the same psychological construct. We acknowledge that the wording in the original manuscript may have created ambiguity on this point.

The Tower of Hanoi task represents the most problematic measure identified during the audit process. During the reassessment, we identified that both completion time and number-of-moves variables were derived using a non-standard scoring approach in which the practice phase (3-disk condition) and the main task (5-disk condition) were combined into single measures. This differs from the more typical approach in the literature, where the primary task is analysed separately, and may introduce additional non-task-specific variance into the data. In addition, comparison of the recoverable source material with the published dataset revealed isolated transcription or processing inconsistencies affecting this measure. Importantly, these discrepancies appeared random rather than systematic and represent the clearest example identified during the audit process of inconsistencies between recoverable records and the final dataset. Although still limited in number, such inconsistencies further reduce confidence in the robustness of the Tower of Hanoi variables and plausibly contribute to the atypical correlational structure highlighted by the evaluators.

Importantly, to further evaluate the contribution of these measures to the overall findings, we conducted additional machine learning analyses examining the role of reaction time, Tower of Hanoi, and related variables highlighted by the evaluators. Models based primarily on these variables yielded substantially reduced, although still above-chance, classification accuracy, whereas models excluding them continued to show robust performance. This pattern suggests that, while these measures contribute to the broader multivariate structure of the dataset, the principal findings are not driven solely by them.

Taken together, these three measures differ from the other components of the dataset in that robust external normative reassessment was either unavailable or substantially constrained by the specific implementation procedures used. Consequently, although the currently available evidence does not indicate that the original findings are necessarily invalid, these measures should be interpreted cautiously and with considerably reduced emphasis relative to the D-KEFS, personality, and working memory analyses.

Digit span

We thank the evaluators and editors for raising important questions regarding the working memory measures, as these prompted a detailed re-examination of the implementation and scoring procedures used for the digit span task. During the audit process, and in consultation with co-author Alberto Filgueiras, who conducted the Brazilian data collection, we identified that the task implementation differed from the standard WAIS administration originally described in the manuscript. Rather than employing the conventional WAIS procedure involving multiple trials per span level and partial scoring across correctly recalled sequences, the implemented task used only one sequence for each span length, progressing from shorter to longer sequences (2 to 8 digits). Performance was recorded simply as the maximum span successfully completed. This implementation substantially compresses variability and reduces the number of possible values, particularly in relatively homogeneous samples such as elite football players. As a consequence, distributions become narrower and correlations between forward and backward span are expected to be attenuated relative to standard WAIS-based scoring.

At the same time, we acknowledge that some aspects of the original control distributions remain atypical and are not fully explained by the implementation procedure alone. Importantly, however, these irregularities were primarily localized within the control sample, whereas the player data retained clearer structure and correlation between forward and backward digit span. While we cannot conclude that the control data are invalid, the statistical properties of portions of the control sample do resemble unusually low-structure or partially random-looking psychometric data, consistent with the broader concerns discussed elsewhere regarding parts of the first testing session. Precisely because of these unresolved concerns, we considered it important not to rely exclusively on the original control comparisons when reassessing the working memory findings. We therefore conducted two additional analyses focused specifically on the player sample.

First, player scores for forward and backward digit span were compared against published normative values from Gregoire and Van der Linden¹⁰ using one-sample t-tests (**Table 3**).

Trait	N	M(sample)	SD(sample)	M(norm)	t	p	d	CI95(low)	CI95(high)
FDS	153	6.82	0.51	6.27	13.38	1.8e-27	1.08	6.74	6.90
BDS	153	6.54	0.73	4.80	29.35	1.7e-64	2.37	6.43	6.66

Table 3. Differences between elite players and digit span normative values.

Results of one-sample *t*-tests comparing the digit span scores of elite football players ($N = 153$) against published normative means from Gregoire and Van der Linden (1997) for individuals aged 18–34 years. The table reports sample means (*M*), sample standard deviations (*SD*), normative means, *t*-values, *p*-values, Cohen's *d* effect sizes, and 95% confidence intervals (*CI*) for the player sample means. Digit span abbreviations: FDS = Forward Digit Span, BDS = Backward Digit Span.

Second, the player sample was compared against two additional independent datasets collected in previous studies by author Leonardo Bonetti at Aarhus University, Denmark, and subsequently combined into a single comparison sample. These samples consisted primarily of 109 individuals ($n_1 = 59$, $n_2 = 50$) from Western countries (predominantly Danish), were generally highly educated (either holding or pursuing a university degree at the time of the data collection), and had a mean age of 24.92 ± 5.79 and 25.34 ± 4.52 years, respectively. The biological sex distribution was: 22 males and 14 males, respectively. Here, independent Welch two-sample *t*-tests were performed in order to account for possible differences in sample size and variance between datasets (**Table 4**).

Trait	N(B)	M(B)	SD(B)	N(A)	M(A)	SD(A)	t	p	d
FDS	153	6.82	0.51	109	5.28	1.14	13.25	2.1e-26	1.86
BDS	153	6.54	0.73	109	4.17	1.00	21.02	4.2e-51	2.77

Table 4. Differences between elite players and digit span normative values.

Results of two-sample *t*-tests comparing the digit span scores of elite football players ($N = 153$) against published normative means from Gregoire and Van der Linden (1997) for individuals aged 18–34 years. The table reports sample means (*M*), sample standard deviations (*SD*), normative means, *t*-values, *p*-values, Cohen's *d* effect sizes, and 95% confidence intervals (*CI*) for the player sample means. Digit span abbreviations: FDS = Forward Digit Span, BDS = Backward Digit Span.

In addition, because the Aarhus participants were originally administered a standard WAIS working memory task, it was possible to rescore the same datasets both according to

conventional WAIS procedures and according to the simplified scoring approach used in the Brazilian sample by Alberto Filgueiras. This directly demonstrated how the simplified implementation substantially reduced variance, compressed the range of observed values, and attenuated the correlation between forward and backward span (from .56 to .36), resulting in a similar pattern compared to the player sample from the present study. At the same time, this does not fully explain the atypical properties observed in the original control sample, consistent with the concerns already acknowledged above.

In conclusion, both reassessment strategies yielded patterns broadly compatible with the original findings, with the elite players consistently showing elevated working memory performance relative to both normative expectations and the additional external datasets. Importantly, this pattern remained evident even when comparing the players against the Aarhus samples, despite those participants having substantially higher average years of education (12.38 ± 2.15 and 13.14 ± 1.89 for the Aarhus samples, compared to 8.74 ± 1.83 in the Brazilian player sample). The groups were instead very comparable in age (26.14 ± 5.63 years of the Brazilians vs 24.92 ± 5.79 and 25.34 ± 4.52 years for the Aarhus samples). Although the Aarhus datasets included both males and females, whereas the Brazilian player sample consisted exclusively of males, it is unlikely that this difference alone could plausibly account for the observed working memory differences.

Taken together, these observations suggest that several of the unusual statistical properties highlighted by the evaluators are substantially influenced by the simplified and insufficiently described implementation procedure used for the digit span task, particularly with respect to variance compression and attenuation of correlations. While some aspects of the original control distributions remain difficult to fully explain and therefore warrant caution, the player data did not exhibit the same atypical patterns. In particular, the relatively restricted ranges observed in the player sample appear broadly consistent with the simplified scoring procedure combined with the relatively homogeneous nature of the elite football cohort. Importantly, the additional normative and external-dataset analyses further reinforced the original findings, supporting the conclusion that to the best of our knowledge the player sample exhibited elevated working memory performance relative to independent reference populations.

PCA and latent structure

To further investigate the overall internal organization of the dataset, we conducted principal component analyses (PCA) across different partitions of the data, including the full Brazilian dataset as well as the player and control groups analysed separately.

Importantly, because the dataset combines variables spanning several relatively distinct psychological domains, including personality, executive functions, working memory, reaction time, and planning, extremely strong latent structure would not necessarily be expected when all variables are analysed jointly. Rather, variance would be expected to distribute across multiple components reflecting partially independent constructs.

Consistent with this, the PCA analyses revealed only moderate overall latent structure when considering all variables together. The first principal component accounted for approximately 19% of the variance in the player sample and approximately 15% in the control sample, indicating that the dataset is not dominated by a single strong latent factor (**Figure 4**). Importantly, however, the player data consistently showed greater structure than the control data.

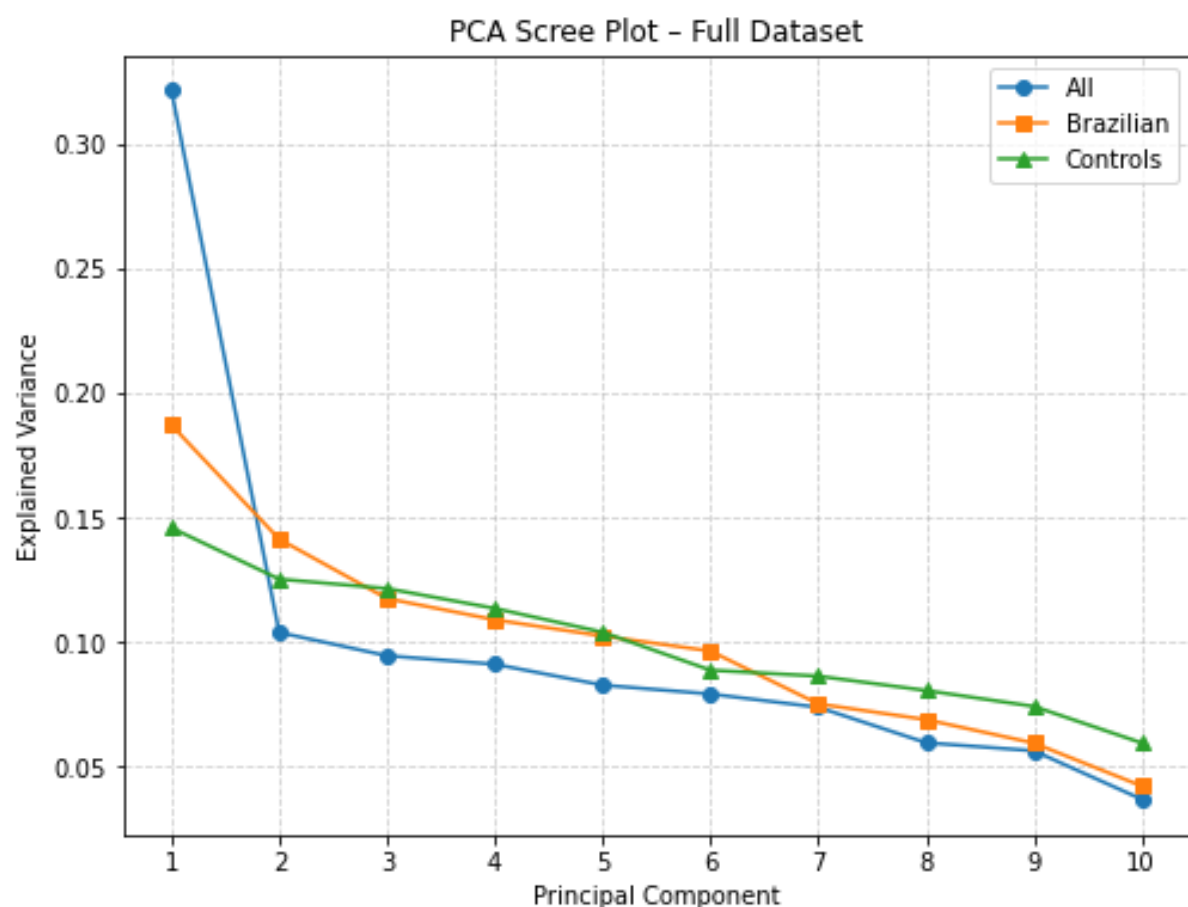


Figure 4. PCA scree plot of the variables collected in the first session.

The figure shows the proportion of variance explained by successive principal components (PC1–PC10) for the full dataset, as well as for the Brazilian players and the control group analysed separately. When all variables are considered jointly, the full dataset exhibits a more prominent first component ($PC1 \approx 0.32$), reflecting increased variance driven by between-group differences. In contrast, within-group analyses (players and controls) show a more even distribution of variance across components ($PC1 \approx 0.15$ – 0.19), consistent with the heterogeneous nature of the variables included. Importantly, the player sample retained consistently greater structure than the control sample, whereas the control data showed weaker organization and variance patterns closer to those expected under low-structure or partially random psychometric data. These findings are consistent with the broader pattern observed throughout the reassessment, namely that atypical statistical properties were primarily concentrated within portions of the control data from the first testing session rather than uniformly affecting the entire dataset.

To assess whether these levels of structure differed meaningfully from what would be expected under random or weakly structured data, we performed additional null-model analyses. Specifically, we compared the variance explained by the first principal component in the real data against two null distributions generated through: (i) permutation procedures destroying correlations between variables while preserving univariate distributions, and (ii) independently generated Gaussian random datasets matched for mean and variance.

These analyses showed that the player data retained significantly greater latent structure than both null models ($p < .001$), indicating that the observed organization is unlikely to arise from independently random or correlation-free data alone. By contrast, the control data showed substantially weaker structure and were considerably closer to the null distributions (**Figure 5**), consistent with the broader concerns discussed elsewhere regarding portions of the first testing session. At the same time, the control sample was not entirely indistinguishable from the null models, suggesting the presence of at least some residual organization rather than purely arbitrary values.

Importantly, the relatively modest structure observed across the full-variable PCA became more understandable once considering the properties of the dataset and the measures included. Several variables, particularly working memory and reaction time measures, were implemented using coarse-grained or lower-resolution procedures that reduce variance and attenuate correlations. In addition, the inclusion of heterogeneous psychological domains naturally distributes variance across multiple partially independent components rather than concentrating it into a single dominant factor.

Taken together, these findings support a heterogeneous interpretation of the dataset. The player data retain statistically meaningful latent structure beyond what would be expected under random or permuted data, whereas portions of the control data exhibit substantially weaker organization and remain more difficult to fully reconcile with standard psychometric expectations. Precisely because of these limitations, we considered it important throughout the reassessment process to complement the original control comparisons with normative references and additional independent datasets whenever possible.

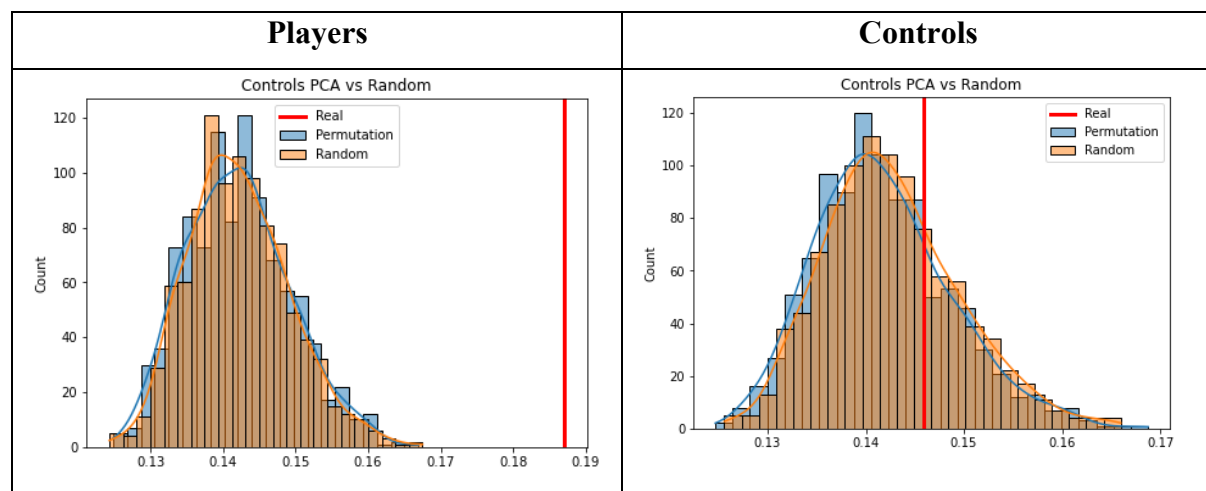


Figure 5. Comparison of PCA latent structure against null models in players and controls.

The figure compares the variance explained by the first principal component (PC1) in the real data (red vertical line) against null distributions generated from permutation procedures and independently generated Gaussian random datasets. The left panel shows the player sample, whereas the right panel shows the control sample. In the permutation analyses, correlations between variables were destroyed while preserving the original univariate distributions. In the random analyses, independent Gaussian variables matched for mean and variance were generated for each measure. The player data (left panel) exhibit substantially greater latent structure than expected under both null models, indicating that the observed organization is unlikely to arise from independently random or correlation-free data alone. By contrast, the control data (right panel) show considerably weaker structure and distributions much closer to the null models, consistent with the broader observation that atypical statistical properties were primarily concentrated within portions of the control sample from the first testing session. At the same time, the control data are not completely indistinguishable from the null models, suggesting the presence of at least partial residual organization rather than purely arbitrary values.

Robustness of multivariate classification analyses (machine learning)

To further evaluate the robustness of the principal multivariate findings, we repeated the machine learning analyses using artificial neural networks (ANNs) under multiple conditions, systematically modifying the set of input variables while maintaining the original procedure involving cross-validation with 1000 permutations to avoid overfitting. Importantly, all analyses reported below were recomputed using the recalibrated personality variables described in the previous sections (this is why analysis with the full dataset returned results that are slightly different from the results reported in the original paper). Specifically, we compared classification performance using (i) the full feature set (Brazilian samples), (ii) a reduced model excluding working memory variables, (iii) a further reduced model excluding both working memory and Tower of Hanoi measures, (iv) an even more reduced model excluding working memory, Tower of Hanoi and Simple Manual Reaction Time measures, (v) a model including only executive function measures derived from the D-KEFS battery (Brazilian and Swedish samples), and (vi) models including only the variables most questioned by the evaluators (working memory alone, Tower of Hanoi alone, and Simple Manual Reaction Time alone; Brazilian samples).

Across all configurations except condition (vi), classification performance remained substantially high and well-above chance (**Figure 6**). In the full model, the neural network achieved an average classification accuracy of approximately 88%, together with similarly elevated precision, recall, and F1 scores. Importantly, excluding working memory variables did not substantially alter performance (accuracy $\approx 82\%$), and further excluding Tower of Hanoi measures yielded comparable results (accuracy $\approx 80\%$). Removing also the Simple Manual Reaction Time variable produced a more evident reduction in performance, although classification accuracy remained clearly above chance levels (accuracy $\approx 70\%$).

Likewise, models based exclusively on executive-function measures from the D-KEFS battery continued to retain meaningful classification ability (accuracy $\approx 75\%$), indicating that multiple subsets of variables contain informative signal capable of differentiating elite players from controls, even when considered independently.

By contrast, models based exclusively on the variables most debated by the evaluators (working memory alone, Tower of Hanoi alone, or Simple Manual Reaction Time alone) showed substantially reduced performance (accuracy $\approx 64\%$, 56% , and 62% , respectively), although still above chance. This reduction was particularly evident in the corresponding confusion matrices (**Figure 7**), where control participants were substantially more likely to be misclassified as players.

956 This pattern is informative because it indicates that the debated variables contribute
957 meaningfully to the overall multivariate structure but are not sufficient on their own to account
958 for the broader classification performance. Conversely, combining multiple psychological
959 domains substantially improves classification, particularly with respect to correctly identifying
960 controls, while maintaining high accuracy for elite players.

961 Importantly, these analyses are also consistent with the broader pattern observed throughout
962 the reassessment process. The variables showing the greatest methodological limitations or
963 atypical statistical properties, particularly those associated with portions of the first testing
964 session, contribute to classification performance but do not appear to drive the principal
965 findings in isolation. Instead, the predictive signal appears distributed across several partially
966 independent domains, including executive functions, personality measures, and working
967 memory.

968 At the same time, these analyses should be interpreted cautiously, as all models ultimately
969 derive from the same underlying dataset and therefore cannot fully resolve the broader concerns
970 regarding portions of the control sample. Nonetheless, the persistence of above-chance and
971 often substantial classification performance after removal of the most debated variables
972 provides additional evidence that the principal findings are not reducible to one isolated
973 artefactual component or single problematic measure.

974 Taken together, these analyses suggest that the multivariate classification results retain
975 meaningful structure even after substantial modification of the feature space. While the overall
976 strength of the findings should be interpreted more cautiously than in the original manuscript,
977 particularly in light of the limitations discussed in the previous sections, the observed
978 multivariate signal does not appear to depend exclusively on any single variable or subset of
979 variables affected by the concerns raised by the evaluators.

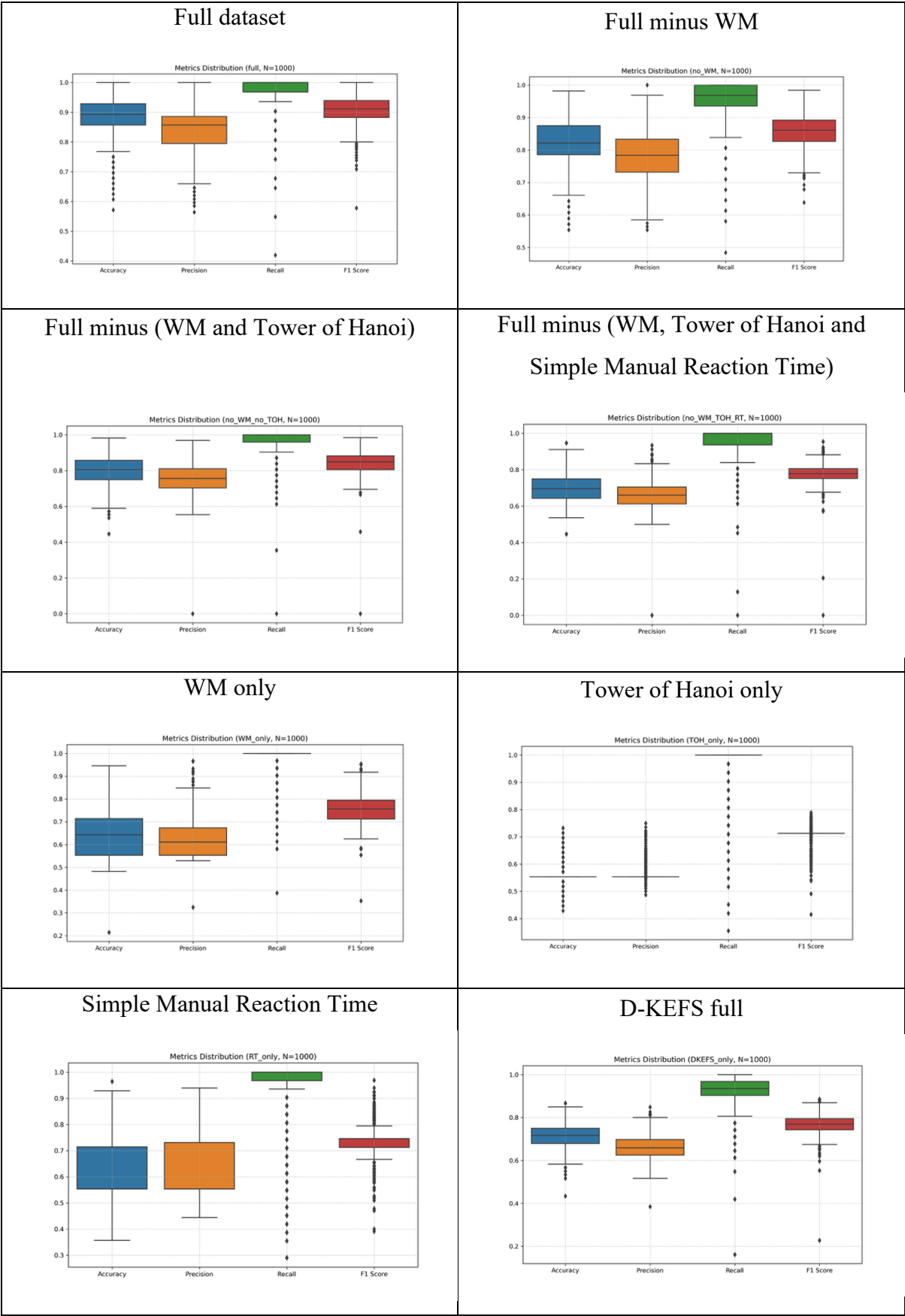
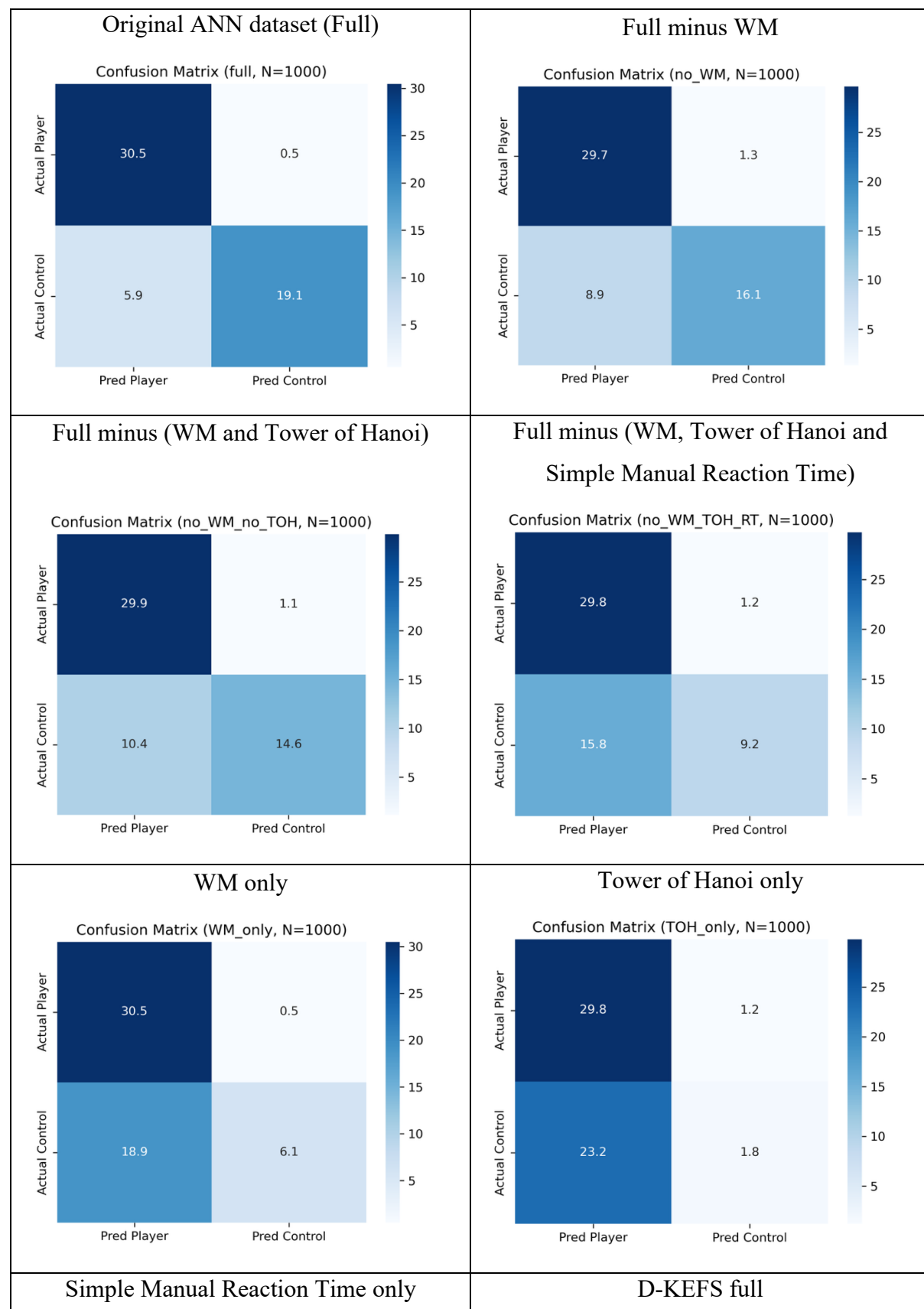


Figure 6. Robustness of machine learning classification across feature subsets.

The figure shows the distribution of classification performance metrics (accuracy, precision, recall, and F1 score) across 1000 permutations of artificial neural network models under different feature configurations: full dataset, excluding working memory (WM), excluding both WM and Tower of Hanoi (ToH), excluding WM, Tower of Hanoi (ToH) and Simple Manual Reaction Time (RT), D-KEFS only, ToH only, WM only and RT only. Across models including multiple psychological domains, classification performance remains consistently high, with strong accuracy and balanced metrics. In contrast, models based on restricted subsets, particularly WM, ToH and RT alone, show reduced performance. This pattern indicates that while individual measures capture relevant aspects of the cognitive profile, optimal performance is achieved when multiple domains are considered jointly.



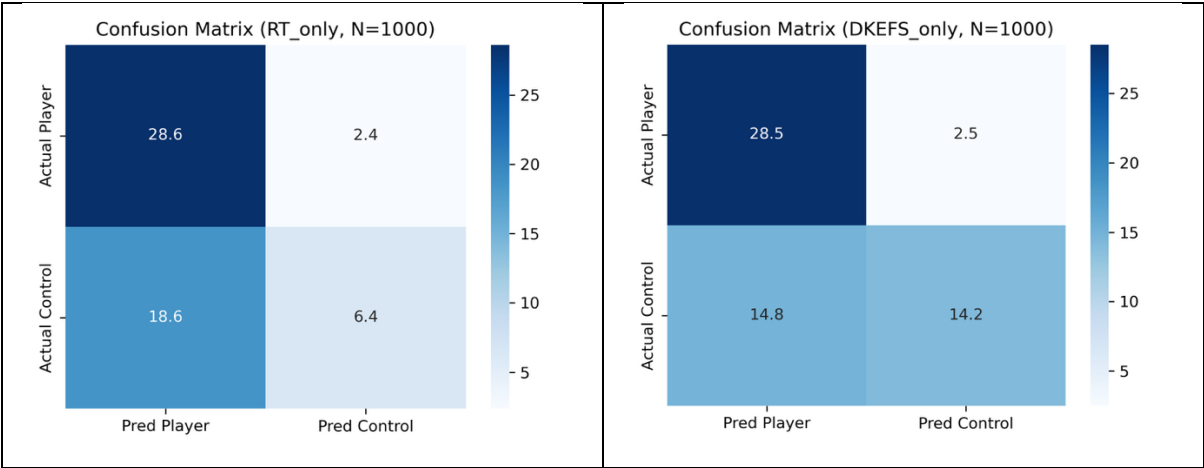


Figure 7. Confusion matrices.

The figure shows the average confusion matrices obtained across 1000 permutations of artificial neural network models under different feature configurations: full dataset, excluding working memory (WM), excluding both WM and Tower of Hanoi (ToH), excluding WM, Tower of Hanoi (ToH) and Simple Manual Reaction Time (RT), D-KEFS only, ToH only, WM only and RT only. Each matrix reports the average number of correct and incorrect classifications for elite players and control participants. Across models including multiple psychological domains (full model and reduced models), classification performance remains highly accurate, with very low misclassification rates and a balanced identification of both groups. In contrast, models based on restricted subsets of variables, particularly WM only, ToH only and RT only, show reduced performance, characterized by a marked asymmetry: elite players are still classified with high accuracy, whereas control participants are more frequently misclassified as players. This pattern indicates that individual measures capture relevant aspects of the cognitive profile associated with elite players but are not sufficient on their own to fully distinguish between groups. Optimal classification performance is achieved when multiple psychological domains are considered jointly, supporting the interpretation that the predictive signal is distributed across variables rather than driven by any single measure.

Additional remarks*1. Clarification regarding the critique by Fitzgerald et al.*

We note that one of the concerns raised by the evaluators refers to a critique previously presented in a separate letter by Fitzgerald and colleagues¹¹, to which we have already provided a detailed response⁵. As clarified in that response, the analysis proposed in that letter addresses a fundamentally different question from the one investigated in our study. Specifically, Fitzgerald et al. compare subgroups within elite players (i.e. national vs non-national team players), rather than distinguishing elite players from matched controls or the general population, which was the primary objective of our work. As such, this comparison is not directly relevant to our research question and does not constitute a valid replication of our machine learning analysis. Distinguishing between highly similar elite subgroups is inherently more challenging and typically requires substantially larger sample sizes and was not a goal of our study. Instead, our hypotheses and models were designed to capture broader differences between elite players and nonathlete individuals. We therefore believe that the analyses presented by Fitzgerald et al. address a substantially different inferential question and should not be interpreted as a direct replication or direct contradiction of the machine learning framework used in our study. While their findings are informative regarding differences within elite-player subgroups, our original analyses were designed to distinguish elite players from non-athlete comparison populations.

2. Presence of two additional variables in the shared dataset

We also note that the evaluators refer to additional variables (e.g. TMTA, TMTB, TMT23) identified in the shared dataset (Brazilian samples). These variables, generated during earlier stages of data processing, were not used in any of the analyses reported in the manuscript and do not appear in the results. As can be verified from the analysis scripts in the public repository, only the variables explicitly described in the manuscript were used. We acknowledge that these additional variables should have been removed prior to data sharing and apologize for any confusion this may have caused.

However, it is also important to emphasize that variables not included in the original analyses should not be used to draw conclusions about the validity of the reported results. The relationships highlighted by the evaluators involving these variables are therefore better understood as a consequence of exploratory analyses on unused variables, rather than as evidence of irregularities in the data.

3. Interpretation of correlations at the team level

Some of the correlations discussed at the level of individual teams should be interpreted with caution, as these subgroups consist of relatively small sample sizes (approximately 27–34 participants per team). Correlation estimates in samples of this size are known to be highly variable and sensitive to sampling fluctuations, particularly when a large number of pairwise comparisons are examined simultaneously¹². Under such conditions, some correlations may appear spuriously inflated, attenuated, or even reversed purely due to sampling variability.

To further assess the extent to which subgroup-level correlations may reflect this type of statistical instability, we conducted a simulation-based analysis in which the control sample was repeatedly partitioned into synthetic “teams” matched in size to the player subgroups. For each random partition, we computed both the mean absolute pairwise correlations between variables and the variance explained by the first principal component (PC1). Across simulations, these estimates exhibited substantial variability, spanning a broad range of values (**Figure 8**).

Importantly, the values observed across the actual player teams partly overlap with the range generated through these random partitions, indicating that a substantial degree of subgroup-level variability is expected under samples of this size. At the same time, some player teams were located toward the upper tail of the simulated distributions, reflecting comparatively stronger structure than would typically emerge from random partitioning alone.

These findings suggest that the variability of correlations at the team level is influenced in part by the degree of instability expected in relatively small subgroups and by the large number of correlations examined. At the same time, we acknowledge that some subgroup-specific values remain comparatively elevated, at least in the first-session control data, and are not fully accounted for by sampling variability alone.

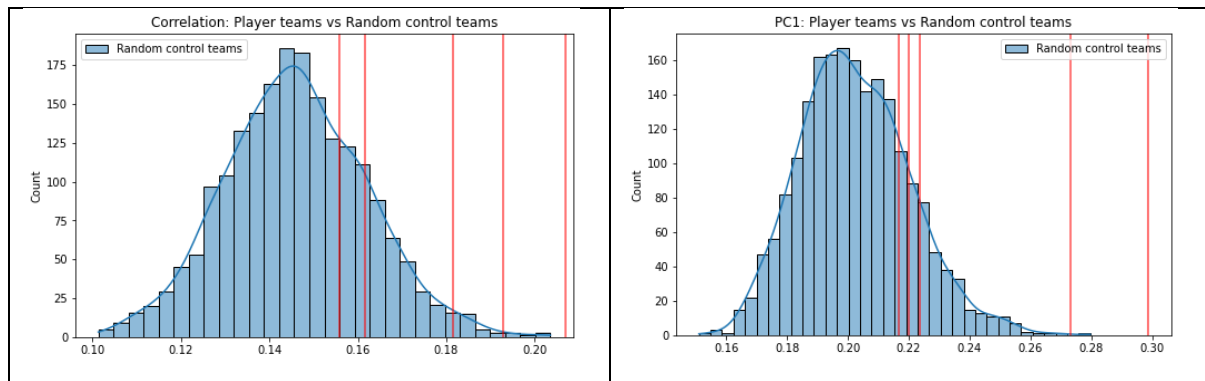


Figure 8. Subgroup variability of correlation and PCA structure compared to random control partitions.

The figure shows the distribution of (left) mean absolute correlations between variables and (right) variance explained by the first principal component (PC1) obtained from repeated random partitioning of the control sample into synthetic “teams” matched in size to the player subgroups ($n \approx 27\text{--}34$). Red vertical lines indicate the corresponding values observed for the actual player teams. The simulated distributions demonstrate substantial variability in both correlation structure and PCA estimates under random partitioning. The values observed for player teams partly overlap with this range, although some teams lie toward the upper tail of the distributions, indicating relatively stronger structure. These results suggest that subgroup-level variability is influenced by sampling effects, while also highlighting that some team-specific values remain comparatively elevated.

Concluding remarks

Taken together, the audit and reanalysis suggest that the data collection itself genuinely took place and that the study procedures were, overall, conducted appropriately, albeit with important methodological and reporting limitations identified during the reassessment process. In particular, the recoverable source material, participant records, and supporting documentation were broadly consistent with the published dataset and with the described data collection procedures. At the same time, the audit also identified several issues requiring clarification or correction, including inaccuracies in methodological descriptions, a systematic scoring issue affecting one personality dimension, and isolated inconsistencies affecting portions of the Tower of Hanoi data.

Importantly, these limitations were not uniformly distributed across the dataset. Rather, the principal concerns were concentrated primarily within the control sample for the first testing session. By contrast, the player data and the D-KEFS measures retained substantially clearer latent structure and more coherent psychometric organization. Likewise, the working-memory findings were reinforced through additional analyses against both normative references and independent external datasets, suggesting that the observed player effects are not solely dependent on the original control comparison. The personality analyses also continued to support the presence of a distinct personality profile in elite players, although the systematic scoring issue identified during the audit substantially altered the Agreeableness results and required reinterpretation of that specific dimension.

The reassessment further clarified that several apparent inconsistencies originated from incomplete or inaccurate descriptions of the implemented procedures rather than from the underlying data themselves. In particular, reconstruction of the actual scoring procedures for the personality and working-memory tasks substantially improved the interpretability of several of the statistical patterns highlighted by the evaluators. By contrast, a few measures, especially Tower of Hanoi and, to a lesser extent, Simple Manual Reaction Time and the Five-Point Test, remain substantially more difficult to interpret robustly and should therefore receive considerably reduced emphasis relative to the original manuscript.

At the same time, we believe it is important to emphasize that this dataset remains unusual and scientifically valuable. To our knowledge, it represents one of the very few relatively large datasets collected directly from truly elite football players sampled from top-level professional clubs. Importantly, the main inconsistencies and atypical statistical properties identified during the reassessment were concentrated primarily within portions of the control data from the first testing session, whereas the player data generally retained clearer structure and fewer

detectable irregularities across measures. Although these observations do not remove the limitations identified throughout this response, they support the conclusion that the dataset continues to provide meaningful information when interpreted carefully, transparently, and with appropriate caution.

Overall, we believe that the principal findings and claims of the original paper remain broadly supported, albeit with substantially greater caution, important qualifications, and reduced emphasis on some specific measures. We therefore believe that the most scientifically appropriate interpretation is not that the dataset is uniformly reliable or uniformly invalid, but rather that it contains a mixture of stronger and weaker components that should be interpreted accordingly. At the same time, we would like to sincerely thank the editors and independent evaluators again for their careful and detailed assessment of the work. This process identified important oversights, methodological inaccuracies, and limitations that required clarification and correction, and we regret that these aspects were not identified and described more precisely in the original manuscript. Importantly, however, the reassessment also reinforced that substantial portions of the dataset, particularly the player data and the executive-function measures, retain meaningful and coherent structure. We therefore hope that the present response provides a clearer, more transparent, and more balanced account of both the strengths and the limitations of the dataset and the conclusions derived from it.

References

1. Bonetti, L., *et al.* Decoding the elite soccer player's psychological profile. *Proc Natl Acad Sci U S A* **122**, e2415126122 (2025).
2. Delis, D.C. Delis-Kaplan Executive Function System Advanced. (2025).
3. Delis, D.C., Kaplan, E. & Kramer, J.H. *Executive Function System (D-KEFS) examiner's manual* (The Psychological Corporation, San Antonio, Texas, USA, 2001).
4. Delis, D.C., Kramer, J.H., Kaplan, E. & Holdnack, J. Reliability and validity of the Delis-Kaplan Executive Function System: an update. *J Int Neuropsychol Soc* **10**, 301-303 (2004).
5. Bonetti, L., *et al.* Reply to Fitzgerald et al.: Mischaracterizing our findings; the cognitive profile of elite footballers remains robust. *Proceedings of the National Academy of Sciences* **123**, e2521662123 (2026).
6. Abdi, H. & Williams, L.J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**, 433-459 (2010).
7. Körner, A., *et al.* Personality assessment with the NEO-five-factor inventory: the 30-item-short-version (NEO-FFI-30). *Psychotherapie, Psychosomatik, Medizinische Psychologie* **58**, 238-245 (2008).
8. Korner, A., *et al.* Efficient and valid assessment of personality traits: Population norms of a brief version of the NEO Five-Factor Inventory (NEO-FFI). *Archives of Psychiatry and Psychotherapy* **17**, 21-32 (2015).
9. Woods, D.L., Wyma, J.M., Yund, E.W., Herron, T.J. & Reed, B. Factors influencing the latency of simple reaction time. *Frontiers in human neuroscience* **9**, 131 (2015).
10. Grégoire, J. & Van Der Linden, M. Effect of age on forward and backward digit spans. *Aging, neuropsychology, and cognition* **4**, 140-149 (1997).
11. Fitzgerald, J., Jakobsson, N. & Brodeur, A. Revisiting the cognitive advantages of professional soccer players. *Proceedings of the National Academy of Sciences* **123**, e2515523123 (2026).
12. Schönbrodt, F.D. & Perugini, M. At what sample size do correlations stabilize? *Journal of Research in Personality* **47**, 609-612 (2013).

Before addressing the individual points raised by the evaluator, we would like to emphasize that the purpose of the present document is to provide concise, high-level responses summarizing the main conclusions of our reassessment. A much more detailed account of the audit process, methodological clarifications, statistical analyses, numerical results, and reinterpretation of the findings is provided in our accompanying general response letter and supplementary materials.

In brief, following receipt of the editorial concerns, we conducted an extensive internal reassessment of the dataset and analyses. This process identified several important methodological and reporting issues requiring clarification or correction, but it also indicated that the identified irregularities were not uniformly distributed across the dataset. Rather, the principal concerns appeared concentrated primarily within portions of the control data from the first testing session, whereas several other components of the dataset, particularly the executive-function measures and substantial portions of the player data, retained clearer latent structure and more coherent psychometric organization.

Below, we respond directly to the specific concerns raised by the evaluator. For detailed analyses, numerical results, and methodological discussion, we respectfully refer the editors and the evaluators to our full response letter.

Summary

- Inexplicably strong or weak correlations between variables
- Data for controls consistent with being entirely randomly generated
- Unexpected data distributions with highly unusual values and surprisingly similar ranges across subgroups
- Irregularities often driven by extreme data in specific subgroups
- Extreme effect sizes between controls and athlete groups
- Clear irregularities only observed for Brazilian sample

We appreciate the evaluators' detailed assessment and agree that several aspects of the Brazilian dataset required careful re-examination. As described in detail in our general response letter, these concerns prompted an extensive internal reassessment of the dataset, the underlying procedures, and the analytical workflow.

Importantly, the reassessment suggested a more heterogeneous picture than a single uniform issue affecting the entire dataset. The principal irregularities appeared concentrated primarily within portions of the control data collected during the first testing session, whereas several other components of the dataset, particularly the executive-function (D-KEFS) measures and substantial portions of the player data, retained clearer latent structure and more coherent psychometric organization.

The reassessment also clarified that several apparent anomalies became more understandable once the actual implementation and scoring procedures were reconstructed in detail. In particular, some tasks differed from the standard procedures originally described in the manuscript, including simplified scoring approaches that naturally compress variance, attenuate correlations, and reduce the range of observable values. In addition, some measures were collected under field conditions using tablet-based or web-based implementations, which likely reduced measurement precision relative to laboratory settings.

The audit process further identified specific issues requiring correction or reinterpretation, including a systematic scoring issue affecting portions of the personality measures and isolated inconsistencies affecting some Tower of Hanoi variables. We fully acknowledge these limitations and agree that they warrant substantially greater caution in the interpretation of some findings, particularly for Tower of Hanoi, Simple Manual Reaction Time, and portions of the first-session control data.

At the same time, several findings remained broadly compatible with the original conclusions when reassessed using normative references, external datasets, latent-structure analyses, and robustness checks. Importantly, the multivariate analyses remained above chance even after exclusion of the variables most questioned by the evaluators, suggesting that the broader signal was not dependent on a single isolated measure.

File 1: BrazilianSample.xlsx

Description of file

This spreadsheet contains metadata and study measures for study participants recruited in Brazil.

These participants comprise 124 controls and 153 soccer players assigned to five different teams, with each team represented by 27 to 34 players. There are 18 variables for which all individuals have data: age, years of schooling, and 16 variables derived from psychometric test results. Two of these variables are excluded from the below analysis: WM.comb.unstandardized, which is the sum of two other variables in the spreadsheet (forward digit span and backward digit span), and WM.comb.standardized, which is a normed version of WM.comb.unstandardized (Pearson r between the variables = 0.99). The below analyses also exclude 5 variables of soccer player statistics for which the controls have no data.

Observations

Unexpectedly low correlations between strongly related variables

Plotting the Pearson correlations between all pairs of unique measures split by group – five teams plus one group of controls – reveals many unusual features (Figs. 1, 2). Firstly, correlations are unexpectedly weak for pairs of variables that should be correlated both in principle and according to existing literature (Fig. 2a). It will be noted that it is extremely unlikely that these unusual correlations are caused by sampling variance or low sample size, as many deviate from expected values at extremely high significance (up to $p = 1.9 \times 10^{-30}$). Several examples are given below.

We agree that several correlations within the Brazilian dataset, particularly within portions of the control sample, are weaker than would typically be expected based on prior psychometric literature. This was one of the main issues examined during the reassessment process.

Importantly, the weaker correlations were not uniformly distributed across the dataset. They were concentrated primarily within measures collected during the first testing session in the control sample, including digit span, reaction time, Five-Point Test, and Tower of Hanoi variables. By contrast, the executive-function measures collected during the second session (D-KEFS battery) retained substantially clearer latent structure and more coherent multivariate organization across both players and controls.

The reassessment also clarified that some tasks were implemented differently from the standard procedures originally described in the manuscript. For example, the digit-span task used a simplified maximum-span procedure with only one sequence per level rather than the standard WAIS administration involving multiple trials and partial scoring. When additional independent datasets were rescored using the same simplified procedure, this markedly reduced variance and attenuated the correlation between forward and backward digit span relative to standard scoring.

Similarly, the reaction-time measures were collected under field conditions using tablet-based and web-based implementations, which likely introduced additional measurement noise and reduced temporal precision compared to laboratory-based testing. For the Tower of Hanoi variables, the reassessment identified both a non-standard scoring approach and some isolated transcription inconsistencies, which likely contributed to the atypical correlations observed for this measure.

At the same time, we acknowledge that some of the observed correlations, particularly within the original control sample, remain weaker than would normally be expected even after accounting for these methodological factors. For this reason, we reassessed the player findings using additional approaches whenever possible, including normative comparisons, external datasets, PCA analyses, and robustness analyses. These additional analyses suggested that the weaker correlations identified in portions of the control data do not generalize uniformly across the dataset as a whole.

The Tower of Hanoi (ToH) is a puzzle in which the aim is to move disks from one pole to another as efficiently as possible while following certain rules. It is administered in cognitive science research as a measure of problem-solving ability and is one of the tasks in the Bonetti et al. study. File 1 records the number of moves made by subjects to complete the puzzle as well as the time taken. It is clear that “moves made” and “time taken” should exhibit a strong positive correlation: for instance, a previously published dataset of 712 measurements (89 participants on 8 ToH tasks) (1) shows a Pearson r of 0.74 between these variables. In File 1, the correlation observed for Team 1 ($r = 0.67$) does not reject this value ($p = 0.47$ for $r = 0.74$) (Fig. 2b). Unexpectedly, however, for all other groups the correlations do not reject zero (all p -values > 0.31 for $r = 0$), but clearly reject the strong positive correlation one would expect (p -values between 6.5×10^{-6} and 1.9×10^{-30} for $r = 0.74$). The point estimate for the correlation is in fact negative for some groups, including the controls ($r = -0.09$).

We agree that the correlation pattern observed between Tower of Hanoi completion time and number of moves is unusual and warranted careful re-examination. In fact, during the reassessment process, the Tower of Hanoi variables emerged as the most problematic portion of the dataset and represent the clearest example where methodological and processing-related issues likely contributed to the atypical statistical structure highlighted by the evaluators.

Importantly, the reassessment clarified that the task was not scored using the standard approach typically employed in the literature. Rather than analysing only the principal 5-disk task, the implemented procedure combined performance from both the practice phase (3-disk condition) and the main task into single summary variables for both completion time and number of moves. This likely introduced substantial additional variance unrelated to the core problem-solving component of the task and reduced comparability with previously published studies, including the reference dataset cited by the evaluators.

In addition, comparison of the recoverable source material with the published dataset identified several isolated transcription or processing inconsistencies affecting the Tower of Hanoi variables. Importantly, these discrepancies appeared random rather than systematic, but they likely contributed additional noise to the relationship between completion time and number of moves.

We also note that the correlations discussed by the evaluators were computed within relatively small team-level subgroups (approximately 27–34 participants per team). Under these conditions, correlation estimates can become relatively unstable and sensitive to sampling variability. During the reassessment, simulation-based analyses suggested that a non-negligible degree of subgroup-level variability in correlations can emerge even under comparable data conditions. While this does not fully explain the observed pattern, it suggests caution in overinterpreting individual subgroup-level correlations.

At the same time, we acknowledge that some of the observed correlations, including the near-zero or slightly negative values in portions of the control sample, remain more atypical than would normally be expected. For this reason, and because robust external normative reassessment was not feasible for the specific implementation used, we concluded that the Tower of Hanoi findings should now be interpreted with substantially greater caution and receive considerably reduced emphasis relative to the executive-function, working-memory, and personality analyses.

Digit span measures the number of digits that a person can commit to short-term memory. Subjects can be asked to recall these digits in either forward or backward order, providing measurements of “forward digit span” (FDS) or “backward digit span” (BDS). Results on these very similar tests typically correlate at around $r = 0.5$ (2–4) (Table S1). In File 1, the observed correlation between FDS and BDS for Teams 1, 2, 3, and 5 do not reject this value (all p values > 0.48 for $r = 0.5$) (Fig. 2b). Team 4 shows a stronger correlation of 0.78 and rejects $r = 0.5$ at $p = 0.0079$, though this p -value would not pass Bonferroni correction with $\alpha = 0.05$ given the number of comparisons mentioned in this report. More surprisingly, however, the correlation observed in the controls is very slightly below zero ($r = -0.003$), rejecting $r = 0.5$ with extremely high confidence ($p = 1.2 \times 10^{-9}$).

We agree that the near-zero correlation observed between Forward Digit Span (FDS) and Backward Digit Span (BDS) within the control sample is unusual and required careful reassessment.

During the audit process, we identified that the digit-span task had not been implemented using the standard WAIS procedure originally described in the manuscript. Rather than using multiple trials per span length with partial scoring, the implemented task used a single sequence for each level, progressing from shorter to longer sequences, and performance was recorded as the maximum span successfully completed. This implementation substantially reduced the number of possible values and compressed the variability of the measures.

To better evaluate the effect of this procedure, we rescored additional independent datasets using both standard WAIS scoring and the simplified scoring procedure employed in the Brazilian sample. Under the simplified implementation, the variance of the measures was markedly reduced and the correlation between FDS and BDS became substantially weaker than under standard scoring.

At the same time, we acknowledge that the near-zero correlation observed in the original control sample remains lower than would typically be expected even under the simplified implementation procedure. Importantly, however, the player groups retained positive correlations between FDS and BDS, generally within the range expected from the simplified scoring approach.

Because of these observations, and because the atypical statistical structure appeared concentrated primarily within portions of the first-session control data, we considered it important not to rely exclusively on the original player-vs-control comparison when reassessing the working-memory findings. We therefore compared the player data against published normative references and additional independent datasets. These analyses continued to support elevated working-memory performance in the player sample relative to the comparison groups.

The Trail Making Test (TMT) is a task in which subjects must draw a path between scattered points in order. It is used to measure several domains of cognitive ability, including visual attention, processing speed, and executive function. The TMT has two variants, where subjects either follow number labels (TMT-A) or alternate between number and letter labels (TMT-B). In previous studies, results on TMT-A and TMT-B typically correlate at around $r = 0.6$ (4–6)

(Table S1). In File 1, however, much lower correlations are observed (Fig. 2b). While Team 2's correlation of $r = 0.36$ is possibly consistent with prior research ($p = 0.11$ for $r = 0.6$), correlations for all other groups do not reject zero but do reject the reference value of $r = 0.6$ with high confidence (p values between 0.0017 and 6.6×10^{-7}).

We would like to clarify that the variables referred to as TMTA, TMTB, and TMT23 were not used in any of the analyses reported in the manuscript. As noted in our response letter, these variables were generated during earlier stages of data processing and were present in the shared dataset, but they do not appear in the results and were not part of the analyses actually reported.

The Trail Making Test variables used in the paper were the D-KEFS subtests described in the manuscript, available in the dataset as: D-KEFS TMT-1 (Numbers), D-KEFS TMT-2 (Letters), D-KEFS TMT-3 (Switching). Visual Scanning was not employed.

We fully acknowledge that these unused variables should ideally have been removed from the shared dataset prior to publication in order to avoid confusion, and we regret that this was not done. At the same time, correlations involving TMTA/TMTB/TMT23 should not be interpreted as evidence regarding the validity of the reported analyses, as they concern variables that were ultimately not used in the study results.

Test scores for control samples are consistent with being independently randomly generated

The above section describes an unusually low correlation in the control group between two related variables, FDS and BDS. This is just one instance of a systematic trend: all correlations between variables in the control group are very close to zero (Fig. 1). This is especially surprising given that several pairs of variables describe very similar quantities or variations on the same test (see above for examples). Further, many of the tests administered in Bonetti et al are generally thought to load on similar factors of cognition, such as executive function or g, and thus should correlate at least modestly with each other (7, 8). Indeed, it is a fundamental assumption of Bonetti et al that each underlying "psychological feature" in question is being measured by several tests, and this is stated repeatedly (example quotations: "three tests indexing executive functions: Trail Making Test, Design Fluency, and Color Word Interference Test" [Fig. 2 caption, page 3]; "measures for... executive functions (Trail Making Test, Design Fluency, Color-Word Interference Test, 5-Point Test [Results, page 2])". If all tests in, say, the executive function battery were completely uncorrelated with each other in the control sample, this would refute the existence of a measurable "executive function" factor in that group of people to begin with. Strikingly, there does not appear to be any latent structure in the control group test scores at all. Only 5/120 (4.2%) unique variable pairs among controls are correlated at $p < 0.05$, consistent with the behaviour for 16 uncorrelated (i.e. independent) randomly generated distributions. Principal components analysis of the control group test scores reveals no obvious factors and produces eigenvalues consistent with a random matrix ($p = 0.48$ by a permutation test comparing the maximum eigenvalue observed in File 1 against the maximum eigenvalue observed for 10,000 replicates of permuted data).

We agree that the weak latent structure observed in portions of the control dataset, particularly for measures collected during the first testing session, represents one of the most concerning aspects identified during the reassessment process. Importantly, however, this pattern was not observed uniformly across the entire dataset. In particular, the executive-function measures collected during the second session (D-KEFS battery) retained substantially clearer and more interpretable latent structure across both players and controls. In our reassessment, PCA analyses of the D-KEFS variables showed coherent multivariate organization and consistent grouping of variables according to their expected cognitive domains. By contrast, the weaker structure identified by the evaluators was concentrated primarily within the first-session control data, including digit span, reaction time, Five-Point Test, personality, and Tower of Hanoi

variables. We agree that several of these measures exhibited substantially lower intercorrelations than would typically be expected in psychometric data.

During the reassessment, we identified a number of methodological factors likely contributing to this pattern. Some tasks were implemented differently from the standard procedures described in the manuscript, including simplified scoring approaches that substantially reduced variance and attenuated correlations. In addition, several measures were collected under field conditions using tablet-based or web-based implementations, which likely introduced additional measurement noise and reduced psychometric precision. At the same time, we acknowledge that these factors do not fully explain the very weak latent structure observed in portions of the original control sample. In our own reassessment, PCA analyses confirmed that the control data from the first testing session exhibited substantially weaker organization than the player data and were considerably closer to null or permuted distributions than would normally be expected in standard psychometric datasets.

For this reason, we considered it important not to rely exclusively on the original control comparisons when reassessing the principal findings. Instead, whenever possible, we re-evaluated the player data using additional approaches, including normative references, external datasets, and analyses restricted to the more robust portions of the dataset such as the D-KEFS measures. We therefore agree that the weak latent structure observed in portions of the control data represents an important limitation of the study and warrants substantial caution in interpretation. At the same time, the reassessment suggested a more heterogeneous picture than a complete absence of meaningful structure across the dataset as a whole.

Extreme correlations in specific groups

In contrast to the surprisingly low correlations mentioned above, there are unexpectedly extreme correlations between some pairs of variables for Teams 4 and 5 (Fig. 2c,d). For instance, TMT-B exhibits remarkably strong negative correlations with openness to experience, forward digit span, and backward digit span for Team 4 only (r values between -0.77 and -0.91, p -values between 2.8×10^{-7} and 6.8×10^{-13} for $r = 0$). This observation is especially surprising for several reasons. Firstly, scores on the highly related TMT-A test show no such correlation with those variables. Secondly, much weaker correlations between these variables are seen in the literature ($r \sim -0.3$ for TMT-B and digit spans, Table S1). Finally, these strong negative correlations are not observed for any other team; on the contrary, TMT-B and openness show a very strong positive correlation for Team 5 ($r = 0.75$, $p = 7.8 \times 10^{-6}$ for $r = 0$). Team 5 also shows an incredibly strong positive correlation between openness and five-point test (novel) ($r = 0.92$, $p = 7.9 \times 10^{-12}$ for $r = 0$) which is not seen for any other group. The above is not an exhaustive list of the anomalies observed in the correlation structure of the data.

We agree that some subgroup-level correlations, particularly within Teams 4 and 5, appear unusually strong and are difficult to interpret confidently. These observations were examined carefully during the reassessment process. Importantly, these correlations were observed within relatively small team-level subgroups (approximately 27–34 participants), where correlation estimates can become unstable and sensitive to individual observations. During the reassessment, simulation-based analyses suggested that substantial variability in subgroup-level correlations can emerge under samples of this size, particularly when a large number of pairwise relationships are examined simultaneously. However, we acknowledge that sampling variability alone is unlikely to fully account for some of the strongest correlations highlighted by the evaluators. At the same time, several of the variables involved in these correlations were among those collected during the first testing session, which also contained the measures showing the weakest overall latent structure in the control sample. As discussed elsewhere, the reassessment identified a number of factors likely affecting the statistical properties of these variables, including simplified scoring procedures, reduced measurement resolution, and

isolated inconsistencies affecting some measures such as Tower of Hanoi and portions of the personality variables. Importantly, these extreme subgroup-level correlations were not observed consistently across the dataset and did not generalize across teams or cohorts. In particular, the D-KEFS measures considered at the full-sample level retained substantially more coherent multivariate structure than suggested by some of the isolated team-level correlations. For these reasons, we agree that these subgroup-specific correlations should be interpreted cautiously and do not provide robust evidence regarding the broader conclusions of the study. During the reassessment, we therefore placed substantially greater emphasis on analyses conducted at the full-sample level and on multivariate approaches, which provided more stable and interpretable estimates than isolated subgroup-level correlations.

Unusual digit span distributions

More unusual features are seen in the distributions of the variables and in the differences between groups. The distribution of forward and backward digit spans will be taken as the first example. Note that while FDS and BDS scores are often presented after norming (e.g. in WAIS, to mean 10 and standard deviation 3), Bonetti et al. provide no indication in the methods or metadata that norming was performed, the data deviate strongly from the normal-like distribution one would expect after norming, and the total range of values (3 to 8) is broadly similar to the range seen for raw digit spans. As such, it appears that these values represent raw measurements and are analysed and plotted accordingly (Fig. 3). Firstly, instead of the expected normal- or Poisson-like distributions for digit spans, FDS and BDS in the controls appear to be uniformly distributed with very sharp boundaries (Figs. 3a-c). This is especially clear for the backward spans, while the forward spans may display some right skew. When plotted against each other, FDS and BDS appear very similar to a pair of independently sampled random integers from a uniform distribution (Fig. 3d), consistent with the near-zero correlation mentioned previously. The observation of uniform-like distributions is supported by goodness-of-fit tests. For instance, the distribution of FDS in File 1 fails to reject a uniform distribution (chi-square $p = 0.21$) and gives an Akaike information criterion (AIC) of 399, compared to a strong rejection (chi-square $p = 1.7 \times 10^{-4}$) and an AIC of 413 against a normal distribution.

These results can be compared to reference data from Woods et al. (3), which provides un-normed FDS measurements ($n = 763$) using two scoring schemes, maximum length (ML) and two-error maximum length (TEML). Given the uncertainty regarding exactly how FDS was scored in Bonetti et al., both scoring schemes will be considered to demonstrate that the findings are somewhat robust to methodological differences. Unlike the File 1 FDS data, the Woods et al. ML and TEML FDS scores both fail to reject a normal distribution ($p = 0.74$ and $AIC = 264$ for ML, $p = 0.37$ and $AIC = 289$ for TEML) and strongly reject a uniform distribution ($p = 3.2 \times 10^{-8}$ and $AIC = 306$ for ML, $p = 2.0 \times 10^{-10}$ and $AIC = 344$ for TEML). It could be argued that the lower sample size in the File 1 controls makes it possible to observe a uniform-like distribution by chance. However, when sub-sampling 5,000 datasets of $n = 124$ from the distributions of the Woods et al. data and testing each for goodness-of-fit to a uniform distribution, no chi-square p -value exceeds 0.05 (max $p = 4.4 \times 10^{-4}$ for ML, max $p = 0.0011$ for TEML). Similar results are obtained for BDS.

Secondly, many individuals in File 1 have measurements that are extremely rare in other datasets (Fig. 3e). The most striking anomaly is that 28/124 controls have an FDS of 3. Given that FDS scores typically show a mean of 6 to 8 and a standard deviation of around 1 (3, 9), a value of 3 represents a deviation of three or more standard deviations and (assuming a normal distribution) should be observed less than 0.3% of the time. Admittedly, digit span scoring methodology can vary between studies (e.g. taking a maximum or mean score from several attempts, or employing a final adaptive trial phase), and Bonetti et al. do not provide details regarding how digit spans were scored, so it could be argued that this discrepancy could be due

to methodological differences. However, even under very conservative assumptions, observing such a large count of 3's is overwhelmingly unlikely.

If we assume a normal distribution for FDS – a conservative assumption given the right skew empirically observed in other data – with a low mean of 6 and a generous standard deviation of 1.5, the probability of observing a single 3 (i.e. a value between 2.5 to 3.5) is 0.038. The probability of observing 28/124 values of 3 as seen in the control data can therefore be estimated at 2.4×10^{-14} (28 successes in 124 binomial trials with success probability 0.038). Less conservative assumptions result in such low probabilities that they cannot be calculated using typical methods. Citing literature and estimating these probabilities is perhaps unnecessary; any person with normal knowledge of the world should be astonished to see a claim that one in four healthy adults cannot commit four digits to short-term memory.

Relatedly, the athlete groups show ranges of values that are unusually narrow. Their FDS scores all fall within 6 and 8 (excluding one value of 5 in Team 1), and the BDS scores all fall within 5 and 7 (excluding three values of 8, also in Team 1). 33/34 of people in Team 3 have the same forward digit span of 7. Such tight distributions do not match the digit spans observed in the previously cited studies.

We agree that the digit-span distributions observed in portions of the dataset, particularly within the control sample, are unusual and required careful reassessment.

During the audit process, we clarified that the digit-span task was not administered using the standard WAIS procedure originally described in the manuscript. Rather than employing multiple trials per span length with partial scoring, the implemented task used a simplified maximum-span procedure with a single sequence for each level. Performance was recorded simply as the maximum successfully completed span. This implementation substantially reduced the number of possible values and compressed the variability of the measures.

Importantly, when additional independent datasets were rescored using the same simplified procedure, the resulting distributions became substantially narrower and less normally distributed than under standard WAIS scoring. Correlations between Forward Digit Span and Backward Digit Span also became substantially weaker. Thus, several of the unusual statistical properties highlighted by the evaluators became more understandable once the actual implementation procedure was reconstructed. At the same time, we acknowledge that the distributions observed in the original control sample remain more atypical than would generally be expected even under the simplified implementation. In particular, the near-uniform appearance of some distributions and the high frequency of very low scores are difficult to reconcile fully with standard psychometric expectations. For this reason, we agree that portions of the first-session control data should be interpreted cautiously and may be at least partially unreliable. By contrast, the player sample appeared considerably more homogeneous and showed strongly compressed ranges, particularly for Forward and Backward Digit Span. This likely reflects a combination of the simplified scoring procedure, the restricted number of possible values, and the relative homogeneity of the elite football cohort. In practice, once most participants perform near the upper end of the limited scoring range, variability becomes strongly reduced and partial ceiling effects can emerge, producing highly clustered distributions even in the absence of any artificial truncation.

Because of these concerns regarding the original control sample, we considered it important not to rely exclusively on the original player-vs-control comparison when reassessing the working-memory findings. We therefore compared the player data against published normative references and additional independent external datasets. Although these analyses should also be interpreted cautiously, they remained broadly compatible with the original observation that the player sample exhibited elevated working-memory performance relative to the comparison groups.

Extreme effect sizes between control and athlete samples

Another unusual feature of the digit span data is the extreme difference in scores between controls and soccer players. For BDS, 67/124 controls have lower values than all 153 soccer players, and 99/153 soccer players have higher values than all 124 controls. Accordingly, the Cohen's D between controls and soccer players is 2.2 – an enormous effect size to observe in psychology. A similarly large effect size of $D = 2.1$ is also seen for FDS. These values imply extraordinarily strong selection on working memory for professional athletes, which is inconsistent with prior research (10–12). While the digit span variables provide the most striking examples, there are several other instances of extreme effect sizes between athletes and controls, such as $D = 1.2$ for agreeableness.

We agree that some of the effect sizes reported in the original manuscript, particularly for Forward Digit Span (FDS) and Backward Digit Span (BDS), are unusually large and required careful reassessment.

Importantly, however, these effect sizes must be interpreted in light of the specific implementation and statistical properties of the measures used. The digit-span task was implemented using a simplified maximum-span procedure with a restricted range of possible values. Within the elite player sample, this produced highly compressed distributions and very low within-group variance. In combination with the relative homogeneity of the player cohort, this likely amplified the resulting standardized effect sizes and produced partial ceiling-like effects. At the same time, the original control sample also displayed atypical statistical properties, including unusually weak latent structure and distributions that were difficult to reconcile fully with standard psychometric expectations. It is therefore likely that the combination of compressed player variance and atypical control distributions contributed to inflating some of the observed between-group effect sizes.

For this reason, we reassessed the findings using additional approaches beyond the original control comparisons. Importantly, when the personality variables were re-analysed against external normative references, the resulting effect sizes became substantially smaller and more consistent with what would typically be expected in psychological research. In addition, the reassessment identified a systematic scoring issue affecting Agreeableness, and correction of this issue substantially altered both the direction and magnitude of the original result. By contrast, the working-memory findings remained comparatively strong even when the player sample was compared against published normative references and additional independent datasets. Although these effect sizes should still be interpreted cautiously given the restricted variance and methodological characteristics of the task implementation, the elevated working-memory performance in the player sample did not disappear when alternative comparison approaches were used. Overall, we agree that some of the originally reported effect sizes were likely inflated by methodological and statistical factors and should now be interpreted substantially more cautiously. At the same time, the reassessment suggested that the working-memory findings remained more robust than several of the other measures examined in the study.

Unusual patterns in variable ranges

For many variables in File 1, the distributions show ranges that are surprisingly different or surprisingly similar between groups. Several examples are given below (Fig. 4). Plotting scores for Big Five conscientiousness by participant group shows several unusual features (Fig. 4a). Several of these echo observations made above, namely surprisingly uniform distributions in the controls and a very large effect size between controls and teams ($D = 1.1$). Another surprising feature is that the ranges of values are remarkably similar across athlete groups: all have the same observed minimum of 17 and maximum of 29. The only exception is Team 1, which, as with BDS and FDS, is distinguished by a few slightly more extreme values (2

observations of 30). Another variable with an unusual pattern in ranges is simple reaction time (Fig. 4b). This variable also displays many of the unusual features that have been repeatedly described in this report. Firstly, an extreme effect size between athletes and controls ($D = 1.5$, though some difference in average reaction time between professional athletes and the general population is perhaps expected). Secondly, unexpectedly uniform distributions ($p = 0.73$ for chi-square fit to a uniform distribution for control data). Thirdly, unusually similar ranges for athlete groups (all minima 239-244, all maxima 297-299). Fourthly, a handful of slightly more extreme values outside this range for Team 1 specifically (one observation of 305 and another of 311). It will also be remarked here that reaction times are typically modelled as skewed distributions such as exponentially modified or inverse Gaussian, with a long tail of values above 400 and sporadic measurements below 240 ms in most samples (13–15). These features are notably lacking in the File 1 data, including the controls. Reference data from TestMyBrain (“meanRT”) (15) for 18-35 year olds (chosen to match range of ages for File 1 controls) shows extremely poor fit to a uniform distribution ($p = 1.5 \times 10^{-11}$ and $AIC = 2942$ for chi-square fit to a uniform distribution after rounding to integer values) and can be fit substantially better to an inverse Gaussian distribution ($AIC = 2767$). The final example that will be considered are the variables relating to the Tower of Hanoi test (Fig. 4c). For these, we do not see the same clear distinction between controls on the one hand and all athlete groups on the other. However, two pairs of groups show remarkably similar ranges for both variables: Team 5 and the control group, and Team 3 and Team 4. Notably, these ranges are distinctly narrower than the full range of data observed across all groups. It is difficult to explain how these features could arise in data collected and processed using typical methods. It could be argued that the notably sharp cut-offs seen at the ends of the distributions could be due to thresholding of values, but for the examples above this would require that different thresholds were used in different subsets of data, which would be extremely poor practice.

We noted that several variables in the Brazilian dataset displayed unusually restricted or similar ranges across subgroups and that these patterns required careful reassessment. However, the reassessment clarified that several of the affected measures were implemented using procedures that naturally compress variability and reduce the number of observable values.

This was particularly evident for the digit-span variables, where the simplified maximum-span scoring procedure substantially restricted the possible score range and produced highly clustered values within the relatively homogeneous elite player sample.

For the personality variables, the reassessment also identified an important methodological clarification. Specifically, the questionnaire used during data collection was the NEO-FFI-30 rather than the Big Five Inventory (BFI) originally reported in the manuscript. Although both instruments measure the same five-factor personality constructs, the NEO-FFI-30 has a different scoring structure, with six items per dimension rated on a 1–5 scale, naturally producing score ranges between 6 and 30. This clarification substantially explains several of the range characteristics highlighted by the evaluators. In addition, the reassessment identified a reverse-scoring issue affecting portions of the personality variables, particularly Agreeableness, and correction of this issue substantially altered the original effect sizes. When the recalibrated player data were compared against external normative references, the resulting personality effects became substantially smaller and more consistent with what would generally be expected in psychological research.

For the Simple Manual Reaction Time measure, the task was administered under field conditions using tablet-based hardware and a web-based implementation. Under these conditions, temporal precision is reduced by touchscreen latency, display refresh constraints, and discretization of response registration. This naturally compresses the observable range of reaction times and reduces the occurrence of extremely fast or slow responses relative to

laboratory-based measurements. In combination with the relatively homogeneous player cohort, this likely contributed to the narrow ranges observed across teams.

We also acknowledge that the original control distributions for several first-session variables remained more unusual than would normally be expected even after accounting for these methodological factors. In particular, the relatively flat distributions and weak latent structure observed in portions of the control sample are difficult to reconcile fully with standard psychometric expectations and therefore warrant substantial caution in interpretation.

For the Tower of Hanoi variables, the reassessment identified both a non-standard scoring approach and isolated inconsistencies affecting portions of the data. In addition, the subgroup analyses discussed by the evaluators were based on relatively small team-level samples, under which empirical ranges and correlations can become sensitive to individual observations. While these factors do not fully explain all of the observed patterns, they likely contributed to some of the unusually similar subgroup-level ranges highlighted in the report.

Overall, we agree that several of the range patterns identified by the evaluators represent important limitations of the dataset. At the same time, the reassessment suggested that these patterns arose from a combination of methodological characteristics, compressed variance within the elite player cohort, and atypical statistical properties concentrated primarily within portions of the first-session control data, rather than from a single uniform issue affecting the entire dataset.

Atypical Big Five scores

Another unusual observation regarding ranges is seen for Big Five extroversion. For context, Big Five factor scores are typically sums of Likert values with range 1-5 over some number of question items, with the precise count of items assigned to each factor varying across questionnaires. Accordingly, a minimum observed score of 3 means that there are at most 3 test items assigned to that factor – in other words, the explanation for that score involving the greatest number of questions is 3 questions each assigned a score of 1. For extroversion, the minimum value seen in File 1 is 5. Given the 1-5 Likert scale typically used, this implies at most 5 questions assigned to extroversion and thus a maximum possible score of 25. However, the maximum observed extroversion score in File 1 is 30, seen for two individuals in Team 1. While it is possible that this is due to use of an atypical scoring scheme, the fact that none of the other four factors display this kind of contradiction makes this explanation less likely. Furthermore, the citation provided in Bonetti et al. suggests that the questionnaire used to measure Big Five factors was the BFI-44 (16). The number of test items in BFI-44 assigned to each factor is as follows: 10 for openness, 9 for conscientiousness, 8 for extroversion, 9 for agreeableness, and 8 for neuroticism. Following the logic described above, these should also represent the minimum possible scores that can be observed for each factor. However, lower values than these are observed in File 1 for all five factors, with the minima as follows: 8 for openness, 6 for conscientiousness, 5 for extroversion, 6 for agreeableness, and 6 for neuroticism.

We appreciate the evaluators for identifying these inconsistencies, as they ultimately helped us detect an important methodological error in the original manuscript.

During the reassessment process, and in consultation with the researcher who conducted the Brazilian data collection, we identified that the personality questionnaire used was not the BFI-44 originally reported in the manuscript but rather the NEO-FFI-30. We sincerely regret this reporting error and acknowledge that the incorrect description of the questionnaire naturally created confusion regarding the observed score ranges.

Importantly, although the specific questionnaire was misreported, both the BFI and the NEO-FFI-30 assess the same five-factor personality constructs (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). However, the scoring structures of the two

instruments differ substantially. The NEO-FFI-30 consists of six items per dimension rated on a 1–5 scale, producing expected score ranges between 6 and 30. This directly explains the apparent inconsistencies highlighted by the evaluators regarding the observed minima and maxima of the personality variables.

In addition, the reassessment identified that at least one participant did not complete a single item, which explains why one variable displayed a minimum value of 5 rather than 6.

The audit process also identified a reverse-scoring issue affecting portions of the personality variables, particularly Agreeableness. Once corrected, the resulting personality analyses became more coherent and the effect sizes relative to external normative references became substantially smaller and more consistent with what would generally be expected in psychological research.

We fully acknowledge that the incorrect reporting of the questionnaire in the original manuscript represented an important oversight on our part, and we regret that this was not identified prior to publication.

File 2: SwedishBrazilianSamples.xlsx

Description of file

The other supplementary spreadsheet provides the results of executive function tests for soccer players in Sweden ($n = 51$) and Brazil ($n = 123$), labelled as dataset “1” and “2” respectively.

Observations

Missing data for Brazil samples

One unusual feature of this spreadsheet is that while the Sweden samples have values for every column, there are many columns for which every Brazil sample has an empty cell (e.g. DFlostscale, DFrepeatscale, DFTotalattempted). Most of these appear to be secondary variables which provide details on participant performance, such as the number of tasks attempted, the number of answers repeated, and accuracy percentage.

We would like to clarify that these variables were not used in any of the analyses reported in the manuscript. As can be verified from the original analysis scripts provided in the public repository, the analyses included only the executive-function variables explicitly described in the paper. We acknowledge that these unused variables should ideally have been removed from the shared dataset prior to publication in order to avoid confusion, and we regret that this was not done. However, the missing values in these columns for the Brazilian cohort do not affect the analyses reported in the manuscript.

Divergent correlation structure between Brazilian and Swedish sample sets

Additionally, viewing the correlations between variables split by country reveals a surprising feature: there are variable pairs that display extremely strong correlations ($r > 0.8$ or 0.9) for the Brazilian samples but much weaker correlations for the Swedish samples (Figs. 5a,b). This cannot be explained by lower sample size in the Swedish samples, as the confidence intervals for the correlation coefficients clearly exclude the estimates obtained from the Brazilian samples (Fig. 5b). More generally, it is very surprising to see such correlations of magnitude > 0.8 between different sub-scales – a correlation of this size would typically represent good test-retest reliability for the same measurement.

Close examination reveals a surprising feature regarding the variable TMT23scale. Its name suggests that it should represent the sum of the TMT2scale and TMT3scale columns. This appears to be the case for the Swedish samples, in which TMT23scale shows a correlation of $r = 0.99$ with the sum of TMT2scale and TMT3scale (Fig. 5c). The minor discrepancies observed are perhaps due to rounding or the inclusion of another minor variable. However, among the Brazilian samples, the TMT23scale values are consistent with being completely independent

of the sum of TMT2scale and TMT3scale ($r = -0.0047$, $p\text{-value} = 0.96$ for $r = 0$). It should be noted that high correlations in the Sweden set are generally explained by variables which represent sums of other variables (TMT23scale, DFTotalCorrect).

We would like to clarify that several of the variables discussed in this section, including TMT23scale and related derived variables, were not used in the analyses reported in the manuscript. As can be verified from the original analysis scripts made publicly available, the paper relied only on the executive-function variables explicitly described in the Methods and Results sections. In particular, variables such as TMT23scale, DFTotalCorrect, and similar derived or auxiliary measures were generated during earlier stages of data processing and were retained in the shared spreadsheet even though they were not part of the final analyses. We acknowledge that these additional variables should ideally have been removed from the shared dataset prior to publication in order to avoid confusion, and we regret that this was not done. However, we believe that the evaluation of the study should focus primarily on the variables that were actually used in the reported analyses and conclusions, namely the D-KEFS measures explicitly included in the original scripts and subsequently reassessed during the revision process.

Concluding remarks

The datasets underlying Bonetti et al. exhibit a variety of anomalies that are difficult to reconcile with authentic data collection and proper data handling. These anomalies are systematic, spanning multiple variables and multiple types of irregularity, and recur consistently across the Brazilian samples. Taken individually, these anomalies might be attributable to atypical methodology, coding bugs, or transcription errors. However, several are very difficult to explain in this way, such as the complete lack of latent structure in the control sample data. It is also hard to imagine how such a wide variety of irregularities could be generated so systematically through error alone.

It should also be noted that the net impact of these irregularities is to induce a clearly detectable difference between elite soccer players and controls. The identification of elite sportspeople, especially early in their careers before they reach elite status, is of considerable commercial interest and a major research aim in sports psychology. Resolving the concerns raised here would require more complete records of the data collection process, and additional details on methodology that are currently lacking in the Bonetti et al. manuscript.

We fully acknowledge that the reassessment identified important irregularities and limitations within portions of the dataset, particularly within parts of the Brazilian control data collected during the first testing session. We agree that several of these patterns are difficult to reconcile fully with standard psychometric expectations and warrant substantial caution in interpretation.

At the same time, the reassessment suggested a more heterogeneous picture than a single uniform issue affecting the entire dataset. In particular, several of the apparent anomalies became more understandable once the actual implementation and scoring procedures were reconstructed in detail, including the identification of simplified task implementations, lower-resolution scoring procedures, a reporting error regarding the personality questionnaire used, and isolated inconsistencies affecting some variables.

Importantly, the reassessment also indicated that the irregularities were not uniformly distributed across the dataset. The executive-function measures (D-KEFS) and substantial portions of the player data retained clearer latent structure and more coherent multivariate organization than the first-session control data. Likewise, several findings remained broadly compatible with the original interpretation when reassessed using external normative references, independent datasets, and robustness analyses.

We nevertheless agree that the original manuscript did not provide sufficient methodological detail regarding several aspects of the data collection and scoring procedures,

645 and we regret that these limitations and inaccuracies were not identified more clearly prior to
646 publication. The concerns raised by the evaluators ultimately prompted a much deeper
647 reconstruction and reassessment of the dataset than had originally been performed.

648 Overall, we believe the reassessment supports a cautious and qualified interpretation of the
649 study rather than a binary conclusion that the dataset is either entirely reliable or entirely
650 invalid. Some components of the dataset now appear substantially more robust and
651 interpretable than others, whereas some measures should receive considerably reduced
652 emphasis and be interpreted with substantial caution.

Alberto J Filgueiras
Senior Lecturer in Psychology
CQUniversity, Australia
a.filgueirasgoncalves@cqu.edu.au

Dear Editorial Board of PNAS,

I am writing regarding the concerns raised about the article “Decoding the elite soccer player’s psychological profile” published in PNAS.

First and foremost, I would like to state clearly that I take primary responsibility for the collection, organization, scoring, management, and methodological reporting of the Brazilian dataset, including the information provided to my co-authors regarding test administration procedures, scoring methods, data-processing procedures, and related methodological details. My co-authors received the dataset and methodological descriptions from me in good faith and conducted their contributions under the reasonable assumption that the material provided was accurate and appropriately documented.

The concerns raised by the editors and reviewers prompted me to carefully revisit the data collection procedures, recover as much original material as possible, and re-examine the methodological and scoring procedures used during the study. Through this process, I identified several important issues that should have been detected and clarified prior to publication. These include inaccuracies in parts of the methodological descriptions, incomplete descriptions of the implemented versions of some tests, a systematic scoring error affecting one personality variable, and several isolated transcription inconsistencies affecting portions of the Tower of Hanoi variables.

The Brazilian data collection was conducted between 2018 and 2021 while I was affiliated with Rio de Janeiro State University in Brazil. The study involved field-based data collection across multiple professional football environments and control settings, with the assistance of students and research assistants participating in portions of the testing and organizational procedures under my supervision. Although collaborators contributed to portions of the operational work, I retained primary responsibility for supervision, compilation, scoring procedures, and preparation of the dataset ultimately shared with the co-authors.

Importantly, several years have passed since the original data collection period, and during this time my own research practices, use of psychometric tools, and data-management procedures evolved substantially. Following completion of the data collection period, I relocated first to the United Kingdom and later to Australia, which further complicated the recovery and reconstruction of older locally stored materials and communications. In revisiting the study, I realized that some methodological details were incompletely documented at the time and later imperfectly recalled during manuscript preparation. This contributed, for example, to confusion regarding the precise personality questionnaire version used, the implementation and scoring procedures of the working-memory tasks, and aspects of the test administration procedures.

At the same time, I want to emphasize that the study itself was conducted in good faith and with the genuine intention of carrying out valid research under challenging field conditions involving elite football players. During the reconstruction process, I was able to provide the co-authors confidentially with supporting materials including names of clubs involved in the data collection, recoverable participant-identification material from one club cohort, names of players associated with the recoverable records, partial original raw sheets, and several

independent letters from collaborators, students, coaches, and staff members who assisted with, participated in, or witnessed portions of the data collection procedures.

Unfortunately, only a limited subset of the original raw material could ultimately be recovered. A substantial portion of the original records had already been deleted following the relevant institutional practices and the passage of several years since data collection. The recoverable material survived primarily because some copies had been retained separately as demonstrational material related to the testing procedures. While I fully recognize that this incomplete recovery cannot independently validate the entirety of the published dataset, I believe it nevertheless provides concrete evidence that substantial portions of the recruitment and data-collection activities genuinely occurred.

I also recognize that portions of the dataset, particularly parts of the first-session control data, display statistical properties that are unusual and difficult to explain fully. While I do not have evidence that intentional misconduct occurred during the collection process, I acknowledge that these statistical irregularities warrant substantial caution regarding portions of the dataset and some related interpretations.

At the same time, I believe it is important to emphasize that the concerns identified were not uniformly distributed across the entirety of the dataset. In particular, the player data retained substantially more coherent statistical structure and appeared broadly compatible with the recoverable source material, the independently recoverable player-identification records, and external normative comparisons. While limitations remain, especially regarding the first-session data for the controls, my current assessment is that the dataset should not be interpreted as entirely invalid or devoid of scientific value. Rather, different components of the study appear to differ substantially in their robustness and interpretability, and I believe that some findings still retain meaningful value when interpreted with appropriate caution and acknowledgment of the identified limitations.

I sincerely regret the methodological, documentation, scoring, and reporting shortcomings identified through this process, several of which should have been clarified or addressed more rigorously prior to publication. I fully understand the seriousness of the concerns raised by the editors and reviewers, and I remain entirely available to cooperate with any further editorial evaluation, confidential review, correction process, or additional investigation considered appropriate by the journal.

Finally, because some of the recoverable materials include potentially identifiable participant and institutional information, I respectfully ask that any broader dissemination or publication of these materials be carefully coordinated in accordance with applicable ethical requirements, confidentiality obligations, and institutional and data-protection regulations.

Yours sincerely,



Alberto J Filgueiras
Senior Lecturer in Psychology
CQUniversity, Australia
a.filgueirasgoncalves@cqu.edu.au