Sports Innovation Journal, 2024, 122–142
https://doi.org/10.18060/27667

Application of an Advanced Risk Analysis Evaluation Tool of Sports Integrity in an International Sports Competition: A Quantitative Case Study

Richard McLaren, Diana Tesic, and Martin Dubbey

The estimated annual cost of corruption in the global sports industry is anywhere from $1.7 to $2.6 trillion (Carrillo, 2023; United Nations, 2023). Among the most common types of corruption is the involvement of sports officials in bribery, cheating, embezzlement, match manipulation, and final-result fixing. The authors describe an advanced risk assessment voice analytics tool and its use in identifying and countering the threat of corruption amid sport officials. The tool was implemented during an actual, multi-phase inquiry of amateur boxing officials at an international tournament. Of the 93 completed (pre- and post-competition) automated interviews, 77.4% resulted in risk-positive outcomes for knowledge of and/or involvement in corruption. Ground truth confirmed 80% of the risk-positive flags. In addition, statistical analyses verified the decision support tool's unbiased model across a group of men and women from 49 different countries. Further, an association of risk-positive flags with countries' human developmental indices—previously shown to be inversely related to corruption—provides convergent validity for the automated tool. This study's findings contribute to the international sports integrity investigations and management literature.

Keywords: risk assessment, automated technology, voice, fraud, sport integrity

---

Richard McLaren, OC, CArb, is is CEO of McLaren Global Sports Solutions, Inc., and an internationally recognized expert in sports law and arbitration. He is also a professor of law at Western University in London, Ontario; Chairman Emeritus Advisory Board, National Sports Law Institute at Marquette University; Special Advisor to Strauss Institute for Dispute Resolution, Pepperdine University; Counsel to McKenzie Lake Lawyers, LLP; and a high-ranking member of CAS, WADA, and other sports-integrity supervisory boards and organizations. Having published in a wide range of areas, including sports-related dispute resolution, he has led investigations and been involved in the adjudication of worldwide sports integrity-related disputes at both the amateur and professional levels. Email: richard.mclaren@mckenzielake.com

Diana Tesic, MA, JD, is an international sports lawyer with more than 15 years of experience in high-profile and complex sports investigations, arbitration, and legal counsel for various sports organizations. She serves as legal counsel at McLaren Global Sports Solutions, Inc., President

Introduction

In today's world of sports, corruption (e.g., match fixing, doping, bribery) is pervasive and problematic enough that it is widely covered in international news outlets ranging from *The New York Times* to the *South China Morning Post* (Panja & Draper, 2020; Zhou, 2023). Regardless of the technological implementations and efforts to address sports corruption (International Olympic Committee, 2021; Schneider, 2014), gaps in its detection still remain, reflecting the complex and multi-dimensional nature of this global phenomenon (Kihl et al., 2016; Shilbury & Ferkins, 2019).

Recently, a global sporting organization reviewed the governance, ethics, financial management, and officiating (i.e., refereeing and judging) standards of boxing during the 2016 Summer Olympics in Rio. Based on their results, the International Boxing Association (IBA) was suspended from organizing and participating in the 2020 Summer Olympics in Tokyo and the 2024 Summer Olympics in Paris. For the first time in modern Olympic history, the International Olympic Committee (IOC) assumed a leadership role in qualification events. Until IBA corrects the issues identified in the IOC reports, the suspension will continue, and boxing athletes will continue to remain under the purview of the IOC (McLaren, 2021).

In the spring of 2021, executives of the IBA approved an independent, multi-step investigation of corruption issues by McLaren Global Sports Solutions (MGSS). The initial stage of the investigation focused on the potential manipulation of sporting results at the 2016 Rio Olympics by specific IBA senior staff and officials (i.e., referees and judges). It was discovered that the bout match fixing was fine-tuned prior to the games. The illicit, usurping methodology relied upon the Draw Commission, whereby key personnel assumed controls beyond the scope of their appointed positions, with their subordinates having been complicit to the bouts that were manipulated for money, the perceived benefit

---

of the FIBA Appeals Panel, and Anti-Corruption Hearing Officer for International Tennis. She previously acted as counsel to the President of the Basketball Arbitration Tribunal. She has played key roles in major sports inquiries, including WADA's Independent Commission Investigation into doping and corruption at the IAAF, the Investigation into the International Weightlifting Federation (IWF), and international boxing. Additionally, she co-authored Canada's most comprehensive study of the national safe sport system, providing strategic recommendations for implementing an effective safeguarding mechanism in the Canadian market. Email: dianatesic@gmail.com

Martin Dubbey, is managing director at Harod Associates, where he brings a unique blend of more than 30 years of law enforcement and a decade of sports investigation experience, combined with a solid understanding of the latest technologies to support integrity, and help counter corruption in sport. He has comprehensive expertise in private sector commercial fraud and asset-tracing work, integrating his know-how of government, global connections, and the use of sophisticated technology. Email: martin.dubbey@harodassociates.com

of IBA, and political backing. In addition to evidence provided by confidential witnesses, nearly two million documents, emails, video, and audio recordings were reviewed as part of the Stage One investigative process (McLaren, 2021). At that time, there was no objective technology available to fast-track sport official screening "clears" while reliably and precisely producing "flags" to detect risk knowledge and involvement areas. In other words, corruption detection inefficiencies and blind spots have real-time, global implications in amateur boxing, which require the evaluation and implementation of additional innovative tools.

By the autumn of 2021, as part of subsequent investigative steps of candidates applying to officiate at the IBA-sponsored 2021 World Boxing Championship, leaders of MGSS and Harod Associates piloted a novel artificial intelligence (AI) assisted risk-detection technology (i.e., designated codename: Challenger) to assess its validity, utility, and ability to enhance their risk-detection protocols and strategies. That particular use case was chosen for retrospective analysis in this article as (1) the sample size allowed for adequate statistical power, (2) multiple variables afforded a rich analysis, (3) empirical verification was available for flagged results, and (4) the real-life environment reflected the usage conditions for which the tool described was designed.

The Present Case Study

The primary goal of this retrospective study was to evaluate a new AI-assisted automated voice analytics technology as a fast, objective, reliable, and precise flagging tool of sports corruption knowledge or involvement when implemented in a screening organization's established risk assessment protocol with sports officiating candidates. In particular, the decision support tool described identifies vocal signals to simple yes/no responses to high-stakes questions about knowledge or involvement in corruption. These signals are analyzed to produce a risk assessment measure, thereby 'flagging' interviews recommended for follow-up. The secondary purpose was to demonstrate the non-biased generalizability of the technology across various demographic variables (i.e., age, gender, language, position) and to provide convergent validity for the tool by correlating a country-level indicator of corruption (i.e., country developmental status) with individual-level risk flags.

In support of the study's primary purpose, the authors tested the following hypotheses:

- H1a: The AI-assisted tool would produce risk assessment output variations.
- H1b: The flagging precision of the automated technology would be significantly higher than chance.

- H1c: No particular question would outperform the others (in risk-output effectiveness).
- H1d: The perception of stakes would negatively correlate with the number of affirmative responses made.

In support of the secondary study purpose, the authors additionally tested the following hypotheses:

- H2a: Age, gender, language, or officiating position would not predict overall risk outcomes for interviews.
- H2b: Age, gender, language, or officiating position would not predict the type of risk detected (i.e., none, knowledge, involvement, or knowledge and involvement).
- H2c: The representative country developmental status (a well-documented, inverse indicator of corruption; Akcay, 2006) would be negatively associated with risk-positive results.

Methodology

Participants

The sample consisted of $N = 93$ complete interviews conducted on-site with IBA officials of the 2021 World Boxing Championships during six days in the late autumn of 2021. The sample included $n = 70$ pre-competition interviews[1] ($n = 40$ on October 22, $n = 27$ on October 23, $n = 2$ on October 24, and $n = 1$ on October 25) focused on the effects of corruption on tournament outcomes (e.g., knowledge or involvement in cheating or bribery) within the past five years and $n = 23$ post-competition interviews ($n = 5$ on November 3, $n = 18$ on November 4) focused on involvement in cheating or bribery specific to the 2021 World Boxing Championships.

A total of $n = 72$ consenting sports officials (i.e., referees, judges, and international technical officers) participated in the automated interviews. Of these officials, $n = 49$ completed a pre-competition interview only, $n = 2$ completed a post-competition interview only, and $n = 21$ officials completed both. Five participants originally scheduled to take an automated interview withdrew (i.e., resulting from flight difficulties or COVID-19); a different set of $n = 5$ participants replaced them. Participants were on average 50.26 years old ± 11.73 S.D.

[1] Five pre-competition interviews were excluded from the analysis because technical issues prevented them from being completed.

(male average: 51.97 years ± 11.54 S.D.; female average: 41.27 years ± 8.42 S.D.; range 19 to 76, $n = 3$ unknown).

Of the $n = 49$ countries of citizenship represented among the $n = 72$ officiates who completed Challenger interviews, from highest to lowest, the frequencies (and respective n) distributed as: 4.2% ($n = 3$) each from Canada, Hungary, India, Russia, South Korea, Ukraine, and Uzbekistan, 2.8% ($n = 2$) each from Algeria, Argentina, Brazil, Croatia, Guatemala, Ireland, Italy, Thailand, and Wales, and 1.4% ($n = 1$) each from Austria, Azerbaijan, Bulgaria, Cuba, Denmark, Egypt, England, Estonia, Finland, France, Germany, Guinea, Kazakhstan, Kyrgyzstan, Lesotho, Moldova, Mongolia, Morocco, Netherlands, Poland, Puerto Rico, Republic of China, Serbia, Slovakia, Spain, Sri Lanka, Sweden, Tajikistan, Trinidad and Tobago, Tunisia, Turkey, United States of America, and Zimbabwe.

Remaining demographic characteristics (i.e., gender, officiating position, certification level, and language used) are listed in Table 1.

Ethical Standards and Informed Consent

All procedures followed the ethical standards of the responsible committee concerned with human experimentation and the Helsinki Declaration of 1975, as revised in 2000 (Snežana, 2001). After applicants were fully informed of all steps of the two-phase screening process aimed at clearing them for officiating, and that their privacy would be protected, under witness, they verbally consented to voluntarily participate. Not originally collected for research purposes, the secondary data from the evaluation represented real-world research. Minimal data collection methods were employed, and consenting candidates' privacy was respected and protected according to internationally established ethical guidelines (Robson & McCartan, 2016), including strict adherence to standards of the Belmont Report (National Commission, 1979), with personally identifiable information either anonymized, kept strictly confidential (via legally binding agreements), or not collected at all.

Importantly, the anonymously provided voice samples were solely used for the purpose of conducting risk assessments, and not to collect personally identifying information or biometrically identifiable data (i.e., unique physical characteristics such as voiceprints to recognize, verify, and/or authenticate the identities of interviewees). The voice utterances received for analysis (very short yes and no responses) were not technically sufficient for identifying specific persons. Further, research protocols dictate the isolation and protection of all voice data, which is never shared with third parties. Secure methods were implemented with respect to privatizing, securing, and storing the audio-recorded interviews of study participants that included encryption of the sound files containing the responses. Only yes or no responses were recorded by the system, which were

Table 1. Demographic Characteristics of Participants

	n	%
Gender		
Female	11	15.3
Male	61	84.7
Officiating Positions		
Referees, Judges (R&J)	44	61.1
International Technical Officials (ITO)	28	38.9
Certification level		
3-star	48	66.7
2-star	1	1.4
1-star	0	0
Unrecorded/unknown	23	31.9
Interview Language		
English	34	48.6
Russian	16	22.9
Spanish	10	14.3
French	6	8.6
Korean	3	4.3
Chinese	1	1.4

too short to be used for biometric purposes and were encrypted during transmission and at rest.

Materials

Decision Support Investigative Tool

The Challenger is an innovative, commercially available system stemming from a private risk assessment company located in North America, with global security interests. The Challenger, henceforth referred to as "the automated technology" is an enterprise-level, automated analytics tool that swiftly assesses

an individual's risk level concerning explicit issues via an automated telephonic interview. The system detects and quantifies the presence or absence of voice-based risk reactions to pre-developed and defined questions by evaluating specific voice and speech macro-characteristic outputs. The technology makes use of foundational NeuroIS principles voice analytics and technical processes to evaluate responses to specific questions asked during the interview (Riedl & Leger, 2016; Singh, 2019).

Unique Aspects of the Technology

The automated technology identifies specific risk alerts based on an individual's vocal responses in any language or dialect. The technology uses issue-specific questions posed during an automated telephonic interview to evaluate the presence or absence of vocal risk signatures. The literature substantiates that the voice is a conduit for perceptions and cognitions (Simon-Thomas et al., 2009). Further, there is a growing body of evidence that these reactions are detectable within half-to-one second voice and speech responses (Boril et al., 2010; Quatieri et al., 2015; Singh, 2019; Yu et al., 2014).

In the automated process for the technology described, the vocal signals evaluated are downstream effects of neurocognitive reactions to specific screening queries linked to distinct neural pathways, including those implicated in experiential memories and associations (Farrow et al., 2013), not just mendacity. Objective macro-level features of the voice (i.e., an expanded view of prosody) are combined in an automated way to come to a risk decision. Other decision support systems in the literature also use physiological input, but evaluation of the data is often subjective and can be biased (Elaad et al., 1994).

Technical Process

The automated telephone-based process uses Session Initiation Protocol that safely and securely executes many concurrent telephonic interviews from anywhere in the world. The chief requirement to use the system is a regular landline or cellular/mobile telephone connection. After the completion of each interview, an encryption system packages the vocal outputs and securely transfers them to a neural-network (NN) based AI-assisted system, which was trained via supervised learning using labeled data. Foundationally, the non-generative, discriminative (i.e., classification) AI model takes advantage of recent advances in NN technologies by utilizing a large, pretrained audio/speech model. The latter is continually fine-tuned, alongside some aspects of proprietary heuristic models with datasets.

The aforementioned process also implements a series of quality control measures to improve response performance metrics. Following this, the system generates an evaluation report. The model encrypts all data points at rest or in transit.

Continuum of Individual Responses and Overall Results

The automated tool's risk framework boundaries remain constant and involve assessment output results categorized into one of four risk levels along a spectrum: low risk (LR), which equates to negligible risk; average risk (AR), which indicates minimal risk; potential risk (PR), meaning a mid-level of risk; and high risk (HR). In this evaluation, three pertinent questions (PQs) were asked in each interview (pre- and post-competition). Accordingly, each interview produced three risk-reaction results, with one of four risk scores generated per question.

Interview Outcome Categories

The highest risk level across all PQs in an interview among all responses determines the overall risk assessment. As such, the outcome of each interview is categorized along a spectrum from LR to HR. Additional outcome categories include affirmation (AF) and not completed (NC). The latter is due to quality issues impacting evaluation or scoring. An affirmative response (AF) was the result of a "yes" to any PQs asked. An NC interview was usually the result of a technical issue (e.g., static or subpar telephone connection) that transpired during the interview. Risk-negative interviews were those in the LR and AR ranges. Interviews with PR, HR, or AF outcomes are typically recommended for follow-up. In this particular project, the HR and AF outcomes were given the highest priority for follow-up.

Procedure

Design

All candidates who applied to officiate at the 2021 World Boxing Championships were informed of the requirement to participate in the two-phase MGSS screening process, including at least one Phase 2 (pre- or post-competition) Challenger automated interview.

Having taken place several weeks before the boxing event, the initial (Phase 1) screening consisted of a deep dive into digital evidence about each candidate, weighed against officiating standards. For "ethics and rules" training, selected officiating candidates then traveled to a designated hotel proximal to the tournament event.

The sole selection criterion for the Phase 2 interview was any official cleared by Phase 1 and selected to officiate. Each consenting participant completed the Phase 2 (pre- and post-competition) automated interviews during the training stage. Each Challenger automated interview consisted of three PQs posed in one of eight languages (to meet participant fluencies). Each interview took 10 minutes or less to complete and was associated with a risk-evaluation score turnaround of fewer than 24 hours.

Candidates who completed both pre- and post-competition interviews arrived at the pre-tournament training session on time and progressed enough (in their respective bouts) that they were available for post-competition interviews, too. However, some non-advancing bout officials returned to their native countries before the tournament ended and were therefore not present for post-competition interviews. Further, some officials only completed post-competition interviews due to late arrivals to the tournament.

A subset of officials who flagged on Phase 2 pre- or post-competition automated interviews also completed follow-up, in-person interviews by MGSS experts. These focused, tête-à-tête interviews provided flagged candidates the opportunity to expand on details relevant to their affirmative responses and/or mitigate reasons for flagged scores. The investigative team used empirical evidence collected in Phase 1 to confirm or negate flags and additional details provided during follow-up interviews to calculate the positive predictive performance of the automated technology.

The process steps of both phases are highlighted in Figures 1 and 2.

Follow-up Interview Approach

Each follow-up interview was based on an established methodological approach (Hughes, 2017). During the follow-up process, the investigative experts were unblinded to Phase 1 and Phase 2 results, available via spreadsheet and a web application. In each ~30-minute follow-up interview, one or two investigators of the MGSS team (consisting of one attorney and three former investigative

Cyber-Screening (several weeks prior to boxing event)

- **Investigative experts comprehensively scrutinized officiating candidates' digital backgrounds for impactful issues**
- **The following types of digital evidence were mined:**
  - **Adverse Media** – Open-source media outlets across the world
  - **Global Sanction Lists** – Recognized sanctions on current and prior awards
  - **Social Media Profiles** – Social media footprints across various platforms, vital to behaviors (e.g., indicators of racism, extremism, gambling, nudity) scored against industry standards, IBA policies and honor codes (e.g., racism, extremism, gambling, other misbehavior)
  - **Political Exposure (PEP) Lists** – Indicators of associations with money laundering that could bring unwanted reputational damage to IBA
  - **Business Interests** – Identification of potential areas of conflict with officiating duties
  - **Red Flag Search** – Use of the MGSS in-house proprietary software program, Seeker, to expose internet background and social media indicators of criminal concern (e.g., involvement in police investigations, signs of illicit activities and associates).

Figure 1. Phase 1 of screening methodology.

Step 1: Initial Announcement (prior to boxing event)

- Officiating applicants agreed to undertake background check as part of screening process of serving as a referee, judge, or technical officer.
- Candidates informed that incriminating evidence discovered that violated tournament standards could negate being allowed to officiate
- Those who passed initial process were aware of the additional requirement to take "ethics and rules" training a few days before the tournament.
- Selected officials who arrived for training were apprised of the required step of taking automated interviews as part of a process to test a new screening tool.
- All (100%) of the selected officials consented to take the automated interviews.
- Consenting participants informed of the scheduled times and locations of interviews conducted in one of two suitable interview rooms at the same location.

Step 2: Pretest Introduction (*interview day, pre and/or post-competition*)

- At allotted time, each officiating candidate entered designated room, and requested to sit at a clean and neutral space with a desk and telephone.
- The Challenger screening specialist provided general instructions, including the need to answer all questions accurately.
- Each candidate read a list of interview questions that would be asked and underscored importance of only answering "yes" or "no" to each.
- The specialist informed candidate that upon call initiation, they would hear instructions and two iterations of the same question set were typical.
- The next step commenced after voluntary consent to take/record interview obtained.
- After introducing a unique code that deidentified the interviewee, the specialist handed the candidate the phone and stepped out of the room.

Step 3: Automated Interview (*interview day, pre and/or post-competition*)

- Upon initiation of the automated interview, candidate informed they would be asked several direct questions requiring accurate responses.
- After completing the automated interview in less than ten minutes (on average), the interviewee hung up the phone and left the room.

Step 4: Follow-up Interviews (*within several hours after automated interviews*)

- Completed by MGSS experts after completion of pre- and post-competition interviews.
- In most cases, candidates who generated the risk evaluations of interest (i.e., HR or AF)) were highest priorities for follow-up interviews (NOTE: some interviewees were not available for follow-up interviews, since they left tournament after bouts ended.)
- 1 PR interviewee subjected to follow-up result of additional data that surfaced during MGSS's screening process.

Figure 2. Phase 2 of screening methodology.

police analysts) thoroughly explored and documented explanations for flags and affirmative responses on applicable knowledge and involvement-based questions.

Establishing Ground Truth

Details collected during digital screening (e.g., documented behaviors, political associations, business interests, criminal affiliations, and background profile details derived from international open-source media outlets, sanctions, and political exposure lists) were the foundational measures of objective "ground truth" used to later confirm the veracity of flagged interviewee results and details provided during the follow-up interviews. Further, for any affirmation (i.e., Yes reply) that was made (during the automated or follow-up interview), factual/empirical verification of details relayed was a requirement of validation.

Automated Interview Foci

Only the most pressing high-stakes issues were questioned, reflected in the limited number (i.e., three) pre-competition or post-competition interview PQs developed and approved by the IBA, MGSS, and Challenger experts. These questions were based on a collective approach of a priori knowledge (i.e., of question themes and issues) and information gleaned from in-depth communications regarding the history, pervasiveness, and impact of the most critical thematic areas. In general, the goal was to determine the presence or absence of risk reaction(s) to knowledge and personal involvement questions related to sports integrity. Specifically, the foci of the automated pre-competition interviews were on knowledge and involvement questions related to cheating or taking a bribe linked to a boxing competition that transpired within the prior five years. The foci of the automated post-competition interviews were on involvement questions related to outcome

Table 2. Pertinent Questions used in Sports Integrity Pilot of Risk Analysis Tool

A. Pre-Competition Interview PQs

1. **Knowledge:** Do you know of any IBA official who has cheated in a competition in the last five years?
2. **Involvement:** In the last five years, have you taken any type of bribe to alter the outcome of a boxing match?
3. **Involvement:** In the last five years, have you cheated as a boxing official in any way?

B. Post-Competition Interview PQs

1. **Involvement:** During this competition did you try to influence the scoring decisions of any other official in any way?
2. **Involvement:** Did you purposely alter the outcome of any bout during the 2021 World Boxing Championship tournament?
3. **Involvement:** Did you accept any type of bribe to alter the outcome of any bout during the 2021 World Boxing Championship tournament?

manipulation and bribery specific to the current tournament the candidates came to officiate (see Table 2).

Results

Descriptive Statistics

Individual Response Risk Results

The n = 70 pre-competition automated interviews resulted in a total of n = 210 distinct responses, including: 17.6% (n = 37) LR; 37.1% (n = 78) AR; 24.8% (n = 52) PR; 14.8% (n = 31) HR; and 5.7% (n = 12) AF.

The n = 23 post-competition automated interviews resulted in a total of n = 69 distinct responses, including: 5.8% (n = 4) LR; 47.8% (n = 33) AR; 24.6% (n = 17) PR; 20.3% (n = 14) HR; and 1.4% (n = 1) AF.

Interview Outcome Results

In the pre-competition automated interviews, the overall interview assessment outcomes (based on highest rating across the three PQs) distributed as follows: 5.7% (n = 4) LR; 17.1% (n = 12) AR; 44.3% (n = 31) PR; 17.1% (n = 12) HR; and 15.7% (n = 11) AF.

In the post-competition automated interviews, the overall interview assessment outcomes included: 4.3% (n = 1) LR; 17.4% (n = 4) AR; 43.5% (n = 10) PR; 30.4% (n = 7) HR; and 4.3% (n = 1) AF.

Pertinent Question Risk Positive (PR, HR) Results

In the pre-competition automated interviews, involvement-based PQ2 (i.e., "In the last five years, have you taken any type of bribe to alter the outcome of a boxing match?") elicited the highest number of HR and PR responses (i.e., 42.0% and 40.4%, respectively).

In the post-competition automated interviews, involvement-based PQ2 (i.e., "Did you purposely alter the outcome of any bout during the 2021 World Boxing Championship tournament?") elicited the highest number of HR responses (50.0%), and involvement-based PQ1 (i.e., "During this competition did you try to influence the scoring decisions of any other official in any way?") elicited the highest number of PR responses (47.1%).

Inferential Statistics

Observed vs. Expected Interview Response Frequencies

To test H1a, we completed a chi-square goodness of fit test for each automated interview stage. In the pre-competition stage of Phase 2, of the n = 70 fully

completed interviews, $n = 59$[2] participants completed interviews consisting of $n = 177$ "No" responses that were assessed as LR, AR, PR, or HR. Results revealed the assessments distributed unequally across the different risk levels, χ^2 $(3, N = 177) = 25.28$, $p < 0.00001$.

In the post-competition stage of Phase 2, of the $n = 23$ fully completed interviews, $n = 22$[3] participants completed interviews consisting of $n = 66$ responses that were assessed as LR, AR, PR, or HR. Results revealed the assessments distributed unequally across the different risk levels, χ^2 $(3, N = 66) = 26.00$, $p < 0.00001$.

Consistent with H1a, the automated technology produced significant variation in risk assessment outputs. Specifically, risk negative (particularly AR) response evaluations were most common for both phases.

Flagging Precision

To test H1b, we completed a positive predictive value analysis, where the following definitions apply: "not confirmed" denotes no justification for the risk-reaction (e.g., risk-positive score) was found, and "confirmed" denotes justification for the risk-reaction (e.g., risk-positive score) was found. The latter can be further parsed to: "confirmed-validated," which denotes verification of a risk-reaction (i.e., risk-positive score) or an affirmative reply was due to factually confirmed knowledge or involvement, and "confirmed-mitigated" denotes verification of the risk-reaction (e.g., risk-positive score) resulted from more benign reasons (e.g., associations, memory triggers, delayed recall).

Of the $n = 59$ total officials whose automated interview outcomes flagged (i.e., $n = 29$ PR, $n = 18$ HR, and $n = 12$ AF), 33.8% ($n = 20$) underwent in-person follow-up interviews by experts. Although a total of $n = 12$ sports officials provided n = 13 affirmative replies ($n = 12$ in pre-competition and $n = 1$ in post-competition interviews), only $n = 4$ affirmers were subjected to the follow-up process during which time disclosure details were queried and provided.

Due to evidence discerned during the Phase 1 digital background review and confirmed testimonial details uncovered during the follow-up process, of the $n = 20$ flagged interviewees who underwent follow-up, 20% ($n = 4$) were not confirmed (i.e., potential Type I errors). Of the $n = 20$ who confirmed, 45% ($n = 9$) were confirmed-mitigated and 35% ($n = 7$) were confirmed-validated. Officiates whose flagged interview results were confirmed-validated prior to the start of the tournament were not allowed to proceed with their officiating duties.

---

[2] $n = 11$ participants were excluded from this analysis because they provided Affirmative (AF) responses during their interviews.

[3] $n = 1$ participant was excluded from this analysis because they provided an Affirmative (AF) response during their interview.

Consistent with H1b, the overall automated output and results process showcased a modest precision rate (i.e., positive predictive value or PPV) of 80%, which is higher than the rate predicted by chance (i.e., 50%).

Differences in Pertinent Questions

In pre-competition interviews, for $n = 59$ automated interviews assessed as LR, AR, PR, or HR, execution of the Friedman test showed no evidence of stochastic dominance between the PQs for score outputs, $\chi^2(2) = 1.639$, $p = .441$.

In post-competition interviews, for $n = 22$ automated interviews assessed as LR, AR, PR, or HR, the application of the Friedman test showed no evidence of stochastic dominance between the PQs for score outputs, $\chi^2(2) = 0.533$, $p = .766$.

Consistent with H1c, then, there was no evidence that any particular question outperformed the others.

Affirmative Response Frequencies and Correlation with Risk Outputs

In the pre-competition interview, 15.7% ($n = 11$) of interviews resulted in affirmative (Yes) responses to at least one PQ. With $n = 12$ affirmative replies among $n = 11$ interviewees, the average rate was 1.09 affirmative responses per affirming interviewee. The first PQ ("Do you know of any IBA official who has cheated in a competition in the last five years?") was associated with the majority (83.3%) of $n = 10$ affirmative replies. The rank order and frequency of $n = 12$ affirmative responses made by PQ were: PQ1 (83.3%, $n = 10$) > PQ2 = PQ3 (8.3%, $n = 1$, each). Most (72.7% $n = 8$) of the $n = 11$ affirmation interviews also showcased at least one risk-positive (PR or HR) non-affirmative (No) response in a different PQ. However, most (58.3%, $n = 7$) affirmative responses corresponded with risk-negative (LR, AR) outputs.

In the post-competition automated interview, 4.3% ($n = 1$) of $n = 23$ completed automated interviews resulted in affirmations during the automated interview phase. With $n = 1$ affirmative reply among $n = 1$ interviewee, the average rate was one affirmative response per affirming interviewee. PQ2 ("Did you purposely alter the outcome of any bout during the 2021 World Boxing Championship tournament?") was associated with the only affirmative reply made. Most (62%, $n = 8$) of all affirmative replies were classified (by the automated technology) as risk-negative outputs.

Relationship Between Pertinent Question Stakes and Affirmative Responses

For both pre- and post-competition interviews, the questions were designed to be of increasingly high stakes for the official to admit. That is, PQ3 > PQ2 > PQ1 with respect to stakes. Of all $n = 12$ affirmative replies made, 83.3% were associated with lower consequence (e.g., knowledge-themed) questions, while 16.7% were associated with higher consequence, involvement-themed questions. Considering the $n = 12$ AF risk responses from $n = 11$ interviews, a Spearman's

rank correlation test revealed a scatterplot, reflective of a monotonic relationship, $r_s(2) = -0.866$, $p = 0.333$.

In support of H1d, a negative correlation existed between perceived question stakes and affirmations made. However, this relationship was not statistically significant.

Feature Relationship Analyses

Effects of Age, Gender, Language, and Position on Interview Risk Outcomes

To test H2a, a cumulative odds ordinal logistic regression with proportional odds was run to determine the influence (if any) of independent factors like age, gender (i.e., two categories, coded in a single dummy variable), language (i.e., seven categories, coded in six dummy variables), and officiate position (i.e., two categories, coded in a single dummy variable) on the dependent variable of highest risk outcome achieved in pre +/- post-competition interviews.

The analysis of observations based on 72 cases demonstrated that the assumption of proportional odds was met, as evaluated by a full likelihood ratio test comparing the proportional odds location model to one with varying location parameters, $\chi^2(18) = 17.99$, $p = .456$. The deviance goodness-of-fit test demonstrated the model was a good fit to the observed data, $\chi^2(144) = 103.77$, $p = .995$. However, the final model did not statistically predict the dependent variables significantly over and above the intercept-only model, $\chi^2(9) = 16.21$, $p = .063$. Participant age was of no consequence to interview risk outcome (as determined solely by the technology), $\chi^2(1) = 0.053$, $p = 0.818$. Also, participant gender was of no consequence to interview risk outcome, $\chi^2(1) = 0.031$, $p = 0.861$. Further, the interviewee language used also had no bearing on interview risk outcomes, $\chi^2(6) = 10.873$, $p = 0.092$. Finally, officiating position had no statistically significant effect on interview risk outcome, $\chi^2(1) = 0.006$, $p = 0.939$.

In other words, age, gender, language, and officiate position had no statistically significant effect on the highest risk outcome determined by the technology.

Effects of Age, Gender, Language, and Position on Type of Risk Flagged

To test H2b, a cumulative odds ordinal logistic regression with proportional odds was executed to determine the influence (if any) of independent factors like age, gender (i.e., two categories, with one dummy variable), language (i.e., seven categories, with six dummy variables), and officiate position (i.e., two categories, with one dummy variable) on the dependent variable of risk type flagged by the technology achieved in pre +/- post-competition interviews. For this analysis, we considered the following rank order of risk type: no flags < flags on knowledge PQs < flags on involvement PQs < flags on knowledge and involvement PQs.

The analysis of observations based on 72 cases indicated that the assumption of proportional odds was met, as evaluated by a full likelihood ratio test comparing the fitted model to a model with varying location parameters, $\chi^2(18) = 25.58$, $p = .110$. The deviance goodness-of-fit test indicated the model was a good fit to the observed data, $\chi^2(159) = 129.1$, $p = .961$. However, the final model did not statistically predict the dependent variables significantly over and above the intercept-only model, $\chi^2(9) = 14.54$, $p = .105$. Participant age was of no consequence to risk type, $\chi^2(1) = 1.643$, $p = 0.200$. Participant gender was also of no consequence to risk type, $\chi^2(1) = 0.232$, $p = 0.630$. Further, the interviewee language used also had no bearing on risk type, $\chi^2(6) = 10.758$, $p = 0.096$. Finally, officiating position had no statistically significant effect on interview risk outcome, $\chi^2(1) = 1.371$, $p = 0.242$.

In other words, age, gender, language, and officiate position had no statistically significant effect on the type of sports corruption risk associated, as detected in automated interviews.

Predictive Indicators of Corruption Based on Country Represented

To test H2c, a multiple regression analysis was executed to predict risk-positive response numbers in $n = 21$ repeat interviewees (i.e., interviewees who took both pre- and post-competition interviews) based on factors like language (i.e., four categories of English, Russian, Spanish, and Chinese, with three dummy variables), officiate position (i.e., two categories of R&J and ITO, with one dummy variable), and representative country's development status. Partial regression plots and studentized residuals against predicted values indicated linearity. Further, a Durbin-Watson statistic of 0.959 indicated the independence of residuals. There was homoscedasticity, inferred by visual inspection of a plot of studentized residuals versus unstandardized predicted values. No multicollinearity was evident, as assessed by tolerance values greater than 0.1. No studentized deleted residuals were greater than ±3 standard deviations, and no values for Cook's distance were above 1. A Q-Q plot indicated having met the assumption of normality. The multiple regression model statistically predicted risk-positive response number at a significant level, $F(3, 14) = 6.410$, $p = 0.006$, adj. $R^2 = 0.49$—a medium effect size, according to Cohen's (1988) classification. Human development index (HDI) levels of countries (World Population Review, 2022) represented by participating officials was the only variable that inversely associated with the quantity of risk-positive results (see Table 3).

In order to double-check these results, a Spearman's test (which measures the strength and direction of association between two ranked variables) was additionally executed on the same ranked variables (i.e., $n = 5$ groups for quantity of

Table 3. Multiple Regression Results for Number of Risk Positive Responses

# risk+ responses	*B*	95% CI for B		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Model						0.58	0.49
Constant	3.802	2.594	5.009	0.563			
Position	0.738	-0.584	2.061	0.617	0.219		
Language	0.315	-0.114	0.745	0.2	0.279		
HDI	-0.065**	-0.104	-0.027	0.018	-0.66		

Note. $n = 21$. Model = "Enter" method in SPSS Statistics; *B* = unstandardized regression coefficient; CI = confidence level; *LL* = lower limit; *UL* = upper limit; *SE B* = standard error of the coefficient; β = standardized coefficient; R^2 = coefficient of determination; ΔR^2 = adjusted R^2. *$p \leq 0.05$, **$p < 0.01$, ***$p < 0.001$.

risk responses vs. $n = 3$ groups for HDI level). The additional results confirmed a moderate, indirect relationship between flagged response quantity and HDI, which was significant and monotonic, $r_s (21) = -0.439$, $p = 0.047$.

In other words, while language and officiate position had no statistically significant effect on the number of risk flags detected, representative country developmental status did inversely correlate at a significant level.

Discussion

In this retrospective case study executed in an international boxing environment, a cutting-edge decision support tool effectively alerted to risk in screened officials knowing about and being involved in sports corruption. This study involved a unique, real-world data set from a high-stakes sports integrity investigation, and findings suggest that AI-assisted tools can be used to support the detection of corruption in this setting with considerable precision.

Notable Findings

Most notably, 80% of risk-positive flags were confirmed on the basis of background checks and follow-up interviews. These PPV findings are consistent with research suggesting that high-stakes environments translate to vocal signals that are more intense, reliable, and easier to detect than those derived in the lab (e.g., Mendoza & Carballo 1998; Scherer, 2003; Van Puyvelde et al., 2018). This rate of precision was achieved in a short, three-question interview in which all questions performed equally well.

The high percentage of risk-positive interview outcomes (i.e., 77.1% for pre-competition and 78.2% for post-competition) suggest a high rate of risk for corruption in our sample of boxing officials and is consistent with concerns of rampant fraud within the sport (McLaren, 2008). The presence of affirmative replies was consistent with previous investigations and were more likely to occur in response to lower stakes (e.g., knowledge-based) relative to higher stakes (e.g., involvement-based) questions. Although this trend was not statistically significant (possibly due to low statistical power at $n = 13$ affirmative responses), this information might be instructional for investigative experts.

As predicted, age, gender, language used, and officiate position were neither associated with overall interview outcomes, nor the type of risk flagged. This provides evidence against modeling bias by the technology's algorithm. From ethical, legal, and financial points of view, organizations cannot afford to use decision support tools that are systemically biased toward or against particular demographics of people (Feast, 2019).

Finally, while not correlated with officiate position or language used, in repeat interview participants, two different statistical analyses provided evidence that the total number of flagged responses was indirectly and significantly correlated to officials' representative country's developmental standing. The literature has already established the use of the HDI as an inverse empirical indicator of corruption, in society and sports (Akcay, 2006). This study's significant, inverse correlation of flagged responses with HDI provides convergent validity for the automated technology described. The preliminary implication is that the country an officiate represents might predict quantifiable risk indicators of corruption.

Limitations and Future Research

It is important to address a few limitations of this case study evaluation, despite the trends and results ascertained. For instance, even though no bias was found based on gender, age, language, or officiating position on overall risk type or risk outcomes, it is possible that small or unbalanced sample size in some categories (e.g., gender) may have affected these results. It is known that low statistical power can reduce the chances of detecting a true effect in research (Button et al., 2013). Future studies with larger samples are necessary to further establish the unbiased nature of this technology.

This study also has an important blind spot that should be examined in future research. Since only risk-positive flags were followed up on, the screening metrics of precision (PPV) was evaluated as 80%. However, because cleared (i.e., risk-negative) interviews were not followed up on (typical of real-world interviews, when external validity is heightened, at the cost of internal validity; Simkus, 2023), the number of true and false negatives were not estimated.

Therefore, in this case study, the negative predictive value (NPV) and accuracy of the technology in the field were not determined.

Because this retrospective study's interviewees represented many countries and regions, empirical (factual) background data may not have been equally accessible (i.e., access to objective verification data may have varied according to geographic region of origin) for all candidates, which could have contributed to an underestimation of the technology's precision. The literature substantiates this possibility; for example, the transparent exchange of background information between law enforcement agencies of different countries is often hindered (Maennig, 2005). Indeed, the estimated precision of a technology is only as reliable as the verification procedure used. Of this study's $n = 4$ false positives, only $n = 2$ were absolutely confirmed as false, due to previously held factual intelligence. The remaining $n = 2$ estimated false positives may reflect the flawed nature of the verification process (i.e., open-source databases used to verify flags are not 100% conclusive).

On the other end of the spectrum, there is also the possibility of overestimation of verified flags, due to confirmation bias (i.e., the intentional or unintentional tendency to seek evidence that validates findings; Weisberg, 2010).

Although the use of AI-assisted systems can improve screening speeds, efficiency, and reliability measures in sports organizations seeking better risk identification strategies, there are ethical concerns to consider. Future research should consider potential ramifications of (1) human jobs replaced by machines, (2) lack of transparency of technological complexity and problems like concept drift, and (3) challenges integrating technologies with established risk management protocols and tools (Chui & Manyika, 2018; NI Business Info, 2020).

Conclusions

AI-assisted automated voice analytics technologies can serve as powerful investigative additions to the screening tool arsenal already in place for sports management establishments. In this study, the investigative team of MGSS discerned that officials who confirmed as associated with sports corruption were more prevalent than previously estimated. Therefore, the Challenger may hold promise in helping sports tournament managers gain realistic insights about the pervasiveness of high-stakes issues (including corruption) among their officials to suitably allocate resources.

Despite this, however, it is imperative to recognize the technology was not designed to hastily equate risk-positive results with "deception" and risk-negative results with "no deception." While the crème de la crème of AI-assisted neurophysiology assessment tools can identify, translate, and categorize reactions, they cannot provide the rationale or interpret why those reactions occurred. In other words, interpretative decisions should rest on human shoulders.

Disclosure of Interest

The authors declare that they have no conflicts of interest to report. However, the anonymized company from which the Challenger technology originates has strong commercial interests, in the realm of global security operations.

References

Akcay, S. (2006). Corruption and human development. *The Cato Journal, 26*(1), 29-48. https://www.researchgate.net/publication/289392968_Corruption_and_human_development

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carrillo, L. (2023, December 20). *UN: Corruption in sport worth $1.7 trillion annually.* OCCRP. https://www.occrp.org/en/daily/18321-un-corruption-in-sport-worth-1-7-trillion-annually

Chui, M., & Manyika J. (2018, April 25). *The real-world potential and limitations of artificial intelligence*. Podcast: McKinsey & Company. https://www.mckinsey.com/featured-insights/artificial-intelligence/the-real-world-potential-and-limitations-of-artificial-intelligence

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1994). The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. *Journal of Behavioral Decision Making, 7*(4), 279–292. https://doi.org/10.1002/bdm.3960070405

Farrow, T., Johnson, N., Hunter, M., Barker, A., Wilkinson, I., & Woodruff, P. (2013). Neural correlates of the behavioral-autonomic interaction response to potentially threatening stimuli. *Frontiers in Human Neuroscience, 6*(34), 1–17. https://doi.org/10.3389/fnhum.2012.00349

Feast, J. (2019, November 20). 4 ways to address gender bias in AI. *Harvard Business Review.* https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Hughes, C. (2017). *The ellipsis manual: Analysis and engineering of human behavior.* Evergreen Press.

International Olympic Committee. (2021, October 6). *Factsheet – The Integrity Betting Intelligence System (IBIS).* stillmed.olympics.com. https://stillmed.olympics.com/media/Documents/Beyond-the-Games/Factsheets/Integrity-Betting-Intelligence-System.pdf

Kihl, L. A., Skinner, J., & Engelberg, T. (2016). Corruption in sport: Understanding the complexity of corruption. *European Sport Management Quarterly, 17*(1), 1–5. https://doi.org/10.1080/16184742.2016.1257553

Maennig, W. (2005). Corruption in international sports and sport management: Forms, tendencies, extent and countermeasures. *European Sport Management Quarterly, 5*(2), 187–225. https://doi.org/10.1080/16184740500188821

McLaren, R. (2008). Corruption: Its impact on fair play. *Marquette Sports Law Review, 19*(1), 15–38. http://scholarship.law.marquette.edu/sportslaw/vol19/iss1/3

McLaren Global Sports Solutions. (2021, October). *Independent investigation of the AIBA boxing competitions prior to and during the Rio Olympic Games 2016* (pp. 1–145). Toronto, Canada.

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stress by means of cognitive workload tasks. *Journal of Voice, 12*(3), 263–273. https://doi.org/10.1016/s0892-1997(98)80017-9

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of*

human subjects of research. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

NI Business Info. (2020). *Risks and limitations of artificial intelligence in business*. https://www.nibusinessinfo.co.uk/content/risks-and-limitations-artificial-intelligence-business

Panja, T., & Draper, K. (2020, April 6). U.S. says FIFA officials were bribed to award World Cups to Russia and Qatar. *The New York Times*. https://www.nytimes.com/2020/04/06/sports/soccer/qatar-and-russia-bribery-world-cup-fifa.html

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Patel, T., Perricone, J., Mehta, D. D., Helfer, B. S., Ciccarelli, G., Ricke, D., Malyska, N., Palmer, J., Heaton, K., Eddy, M., & Moran, J. (2015). Vocal biomarkers to discriminate cognitive load in a working memory task. *Proc Interspeech 2015*. https://doi.org/10.21437/interspeech.2015-566

Riedl, R., & Léger, P.-M. (2016). *Fundamentals of NeuroIS: Information systems and the brain*. Springer.

Robson, C., & McCartan, K. (2016). *Real world research*. Wiley.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1-2), 227–256. https://doi.org/10.1016/s0167-6393(02)00084-5

Schneider, S. (2014). Integrity Betting Intelligence System. *Gaming Law Review and Economics, 18*(10), 961–962. https://doi.org/10.1089/glre.2014.18103

Shilbury, D., & Ferkins, L. (2019). *Routledge handbook of sport governance*. Taylor and Francis.

Simkus, J. (2023, December 13). Internal vs external validity in psychology. *Simply Psychology*. https://www.simplypsychology.org/internal-vs-external-validity.html

Simon-Thomas, E., Keltner, D., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion, 9*(6), 838–846. https://doi.org/10.1037/a0017810

Singh, R. (2019). *Profiling humans from their voice*. Springer.

Snežana, B. (2001). The declaration of Helsinki: The cornerstone of research ethics. *Archives of Oncology, 9*(3), 179–184. http://www.onk.ns.ac.rs/Archive/Vol9/PDFVol9/V9n3p179.pdf

United Nations. (2023, December 16). Illegal betting is the number one factor fueling corruption in sports. United Nations. https://news.un.org/en/story/2023/12/1144857

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01994

Weisberg, H. I. (2010). *Bias and causation: Models and judgment for valid comparisons*. Wiley.

World Population Review. (2022). *Human development index (HDI) by country 2022*. https://worldpopulationreview.com/country-rankings/hdi-by-country

Yu, B., Quatieri, T.F., Williamson, J. R., Mundt, J. C. (2014) Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers. *Proc. Interspeech,* 1038–1042. http://doi.org/10.21437/Interspeech.2014-270

Zhou, L. (2023, July 22). China's football corruption crackdown targets 2 more senior officials. *South China Morning Post*. https://www.scmp.com/news/china/article/3228573/chinas-football-corruption-crackdown-targets-2-more-senior-officials