Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 12 May 2026
Revised 20 June 2026
Accepted 05 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 746–753

KEYWORDS:
expert system; FIG Code of Points; grid-search calibration; gymnastics judging; execution score; explainable scoring

An Artificial Intelligence-Based Expert System with Grid-Search Calibration for the Execution Analysis of the Tippelt Skill on Parallel Bars

[1]Lect. Zaid Mohammed Salman, [2]Prof. Dr. Hisham Hindawi Hwaidi, [3]Hadeer Hadi Essi

[1]College of Physical Education and Sports Sciences – University of Al-Qadisiyah.
spo23.post01@qu.edu.iq
https://orcid.org/0009-0000-9313-8508
[2]College of Physical Education and Sports Sciences – University of Al-Qadisiyah.
hisham.hindawee@qu.edu.iq
https://orcid.org/0000-0003-4332-3750
[3]College of Physical Education and Sports Sciences – University of Al-Qadisiyah.
hadeer.hadi@qu.edu.iq
https://orcid.org/0009-0007-5078-4463

ABSTRACT: Execution scoring in men's artistic gymnastics remains heavily dependent on subjective human judgement, despite the existence of explicit deduction rules in the Fédération Internationale de Gymnastique (FIG) Code of Points. This study proposes and validates a two-tier, rule-based expert system for automated deduction estimation on the Tippelt skill performed on parallel bars. The system's knowledge base combines (a) deduction thresholds derived directly from the FIG Code of Points, Article 14.3, for biomechanical indicators explicitly covered by the regulation, and (b) thresholds for two additional indicators lacking an explicit numerical standard, obtained through a grid-search calibration procedure. To avoid circularity, calibration was performed exclusively on a training subset (n = 56, 80%) of 70 recorded attempts by a single elite male gymnast, and validated on an independent holdout subset (n = 14, 20%) that played no role in threshold selection. Calibration improved overall agreement with the mean score of five certified judges, reducing mean absolute error from 0.476 to 0.309 points and increasing the Pearson correlation from 0.887 to 0.914. Performance on the independent holdout set (MAE = 0.272, r = 0.913) closely matched full-sample performance, indicating that the calibrated thresholds generalize rather than overfit the calibration data. A paired-samples t-test revealed a small but statistically significant systematic bias (mean difference = -0.185, $t(69) = -4.564$, $p < .001$, Cohen's d = 0.546), with the expert system tending to slightly overestimate execution scores relative to human judges. These findings suggest that a transparent, rule-grounded expert system can approximate expert judging with reasonable fidelity while remaining fully interpretable, and may serve as a decision-support tool rather than a replacement for human officiating.

1. INTRODUCTION

Objectivity and consistency in gymnastics judging have long been subjects of methodological concern, given the inherent difficulty of visually detecting small-magnitude, high-speed deviations from an ideal movement pattern during live competition (Knudson, 2020). While the FIG Code of Points formalizes execution errors into discrete deduction categories, the translation of a judge's visual impression into a specific numerical deduction remains a cognitively demanding task performed under strict time constraints, and is consequently susceptible to inter-judge variability (Field, 2018).

Computer-vision-based motion capture now allows biomechanical variables that were previously accessible only through frame-by-frame video analysis to be extracted automatically and at scale. This raises the possibility of encoding a subset of FIG deduction rules as explicit, machine-executable IF-THEN statements — an expert system — that can generate a defensible, auditable deduction estimate directly from measured kinematic indicators, without the opacity typically associated with purely data-driven models such as artificial neural networks (Molnar, 2022).

However, not all deduction-relevant biomechanical indicators are accompanied by an explicit numerical threshold in the published Code of Points; several categories (e.g., degree of leg separation, magnitude of post-catch oscillation) are described only qualitatively. This study addresses this gap by proposing a two-tier expert system architecture that (i) encodes officially documented thresholds where available, and (ii) derives thresholds for the remaining indicators through a rigorous, held-out-validated grid-search calibration procedure, thereby avoiding the arbitrary threshold-setting common in earlier rule-based scoring attempts.

The specific objectives of this study were to: (1) design a two-tier expert system for the Tippelt skill on parallel bars that transparently distinguishes officially grounded rules from empirically calibrated ones; (2) calibrate the undocumented thresholds using a training/holdout split analogous to standard machine-learning practice; and (3) statistically evaluate the agreement between the resulting expert-system scores and the scores awarded by a panel of certified judges.

The principal contribution of this work is methodological rather than purely empirical: whereas prior rule-based attempts at automated sports scoring have typically either restricted themselves to variables with pre-existing regulatory thresholds, or set undocumented thresholds through unvalidated visual inspection of the full available dataset, the present design explicitly separates threshold sourcing into an auditable, dual-provenance structure and subjects the non-regulatory component to the same held-out validation discipline conventionally reserved for machine-learning models. We view this hybrid regulatory-empirical calibration strategy, rather than the specific numerical thresholds obtained for a single athlete, as the primary transferable contribution of this study.

2. RELATED WORK

Automated or semi-automated approaches to gymnastics performance evaluation have progressed along two largely separate lines. The first, biomechanical description, has characterized joint angles, center-of-mass trajectories, and angular velocities for individual apparatus skills using single- or multi-camera kinematic analysis, typically on small samples of one to five elite performers (e.g., Brüggemann, 1994; Knudson, 2020). The second, predictive modeling, has more recently begun to apply statistical learning and neural-network approaches to estimate execution or difficulty scores from extracted kinematic features (Goodfellow, Bengio, & Courville, 2016).

A structured narrative synthesis of five directly comparable biomechanical studies of parallel-bar and related skills (following the methodological guidance of Popay et al., 2006, for narrative synthesis where meta-analytic pooling is precluded by heterogeneous outcome measures and small, single-subject designs) indicated near-universal coverage of joint-angle and center-of-mass path variables (5/5 studies), partial coverage of angular velocity (4/5) and grip-angle variables (2/5), and — critically — no coverage whatsoever of a kinematic motion-error index, a composite execution score, or any artificial-intelligence-based scoring method (0/5 studies for each), a gap-counting procedure consistent with vote-counting-by-direction approaches described in the Cochrane Handbook (McKenzie & Brennan, 2019). This absence of prior work explicitly linking measurable kinematic error to FIG-consistent deduction values, and of any transparent rule-based scoring system validated against a held-out sample, motivates the present study.

Relative to purely data-driven prediction (e.g., artificial neural networks trained end-to-end on kinematic inputs), rule-based expert systems trade some raw predictive flexibility for full auditability: every deduction can be traced to a specific, named biomechanical criterion, a property that is increasingly considered essential — rather than optional — for automated systems intended to support, rather than silently replace, human decision-makers in high-stakes evaluative contexts (Molnar, 2022).

2.1 Preliminary Variable Screening (LASSO)

Table 1 reports the standardized LASSO coefficients for all 17 candidate variables, ranked by absolute magnitude; the five variables retaining the largest coefficients were carried forward into the expert system's rule base.

Table 1. LASSO-standardized coefficients for the 17 candidate biomechanical variables

Rank	Variable	Coefficient
1	Post-catch shoulder oscillation	-0.1497
2	Trunk angle	-0.1440
3	Post-catch hip oscillation	-0.1282
4	Path deviation (hip displacement)	-0.0953
5	Inter-ankle distance	-0.0940
6	Minimum ascent velocity	0.0502
7	Shoulder angle	0.0444
8	Right knee angle	0.0442
9-17	(remaining variables, |coef| < 0.04)	—

3. MATERIALS AND METHODS

3.1 Participant and Data Collection

Data were drawn from a single nationally classified male elite gymnast, who performed 70 executions of the Tippelt skill on the parallel bars, recorded across multiple sessions using three synchronized analytical cameras (120 fps) plus one documentation camera. Two-dimensional joint coordinates were extracted per camera using MediaPipe Pose and reconstructed into a three-dimensional kinematic structure via the Direct Linear Transformation (DLT) algorithm. Seventeen biomechanical variables per attempt (joint angles, path-deviation, symmetry, velocity, acceleration, and post-catch oscillation indices) were retained following an expert-panel content-validity survey (14 raters, ≥85.7% agreement threshold).

Each of the 70 attempts was independently scored by a panel of five certified judges following FIG execution-scoring conventions; the arithmetic mean of the five judge scores ($M = 8.07$, $SD = 0.83$, range = 6.61–9.64) served as the reference execution score (E-score). The reliability of this reference measure was confirmed via a two-way random-effects intraclass correlation coefficient for average measures (ICC = 0.988, $F(69,276) = 84.72$, $p < .001$), a threshold classification consistent with the reliability standards outlined by Koo and Li (2016) and, within an Arabic-language measurement-theory framework, by Hindawi (2021), who similarly identifies coefficients at or above 0.90 as indicative of excellent measurement reliability for applied performance-assessment instruments.

3.2 Two-Tier Expert System Architecture

The expert system follows a deduction-from-ceiling logic consistent with FIG scoring practice: execution begins at a maximum of 10.0 points, from which partial deductions (minor = 0.10, medium = 0.30, major = 0.50) are subtracted whenever a biomechanical indicator exceeds a defined severity threshold. Rules were explicitly partitioned into two categories according to the provenance of their numerical thresholds:

(a) Officially-grounded rules, whose thresholds were extracted directly from Article 14.3 (Specific Deductions for Parallel Bars) of the FIG Code of Points (FIG, 2025–2028), covering trunk-angle deviation, path deviation, and post-catch oscillation of the shoulder and hip; and (b) empirically-calibrated rules, applied to two indicators — inter-ankle distance and minimum ascent velocity — for which the available regulatory excerpt did not specify an explicit numerical cut-off.

3.3 Grid-Search Calibration Procedure

To avoid arbitrary threshold assignment for the two empirically-calibrated rules, a two-stage, held-out-validation procedure was implemented, mirroring the train/validation logic used elsewhere in this research program for the artificial neural network component:

Stage 1 (Calibration): The 70-attempt sample was randomly partitioned into a calibration subset (n = 56, 80%) and an independent holdout subset (n = 14, 20%). On the calibration subset exclusively, a grid search was performed across a systematic range of candidate threshold pairs — spanning the 30th to 90th percentile of the observed distribution for inter-ankle distance, and the 10th to 60th percentile (inverted, since lower velocity indicates greater severity) for minimum ascent velocity — with eight candidate cut-points evaluated per variable, yielding 64 threshold-pair combinations per grid iteration. For each candidate pair, the complete eight-axis deduction rule set was applied to every calibration-subset attempt, and the combination minimizing the mean absolute error (MAE) between the resulting expert-system score and the reference E-score was retained.

Stage 2 (Independent Validation): The calibrated thresholds obtained in Stage 1 were then applied, without further adjustment, to the 14 holdout attempts that had not contributed to threshold selection, in order to verify that the observed accuracy reflected genuine generalization rather than overfitting to calibration-set idiosyncrasies (Hastie, Tibshirani, & Friedman, 2009). This train/holdout separation directly addresses a documented weakness of earlier rule-based sports-scoring attempts, in which threshold values were frequently set via visual inspection of the full dataset without any provision for out-of-sample verification.

3.4 Statistical Analysis

Agreement between expert-system scores and the mean judge score was quantified using mean absolute error (MAE) and Pearson's correlation coefficient, computed separately for the full sample and the independent holdout subset. A paired-samples t-test was used to test whether the mean signed difference between actual and expert-system scores differed significantly from zero, with Cohen's d reported as a standardized effect size. All analyses were conducted in Python 3 using pandas, NumPy, SciPy, and scikit-learn.

4. RESULTS

4.1 Effect of Calibration on Expert-System Accuracy

Table 2 compares expert-system performance using descriptive (uncalibrated) thresholds against performance using the grid-search calibrated thresholds.

Table 2. Expert-system agreement with judge scores before and after grid-search calibration

Metric	Descriptive thresholds	Calibrated thresholds
Mean absolute error, full sample (n = 70)	0.476	0.309
Pearson correlation, full sample (n = 70)	0.887	0.914
Mean absolute error, holdout subset (n = 14)	—	0.272
Pearson correlation, holdout subset (n = 14)	—	0.913

Calibration reduced full-sample MAE by approximately 35% (from 0.476 to 0.309) and increased the correlation with judge scores from 0.887 to 0.914. Critically, performance on the independent holdout subset (MAE = 0.272, r = 0.913) was equal to or better than full-sample performance, providing direct evidence that the calibrated thresholds capture a generalizable relationship rather than an artifact of the calibration sample.

Natural Resources for Human Health
eISSN: 2583-1194

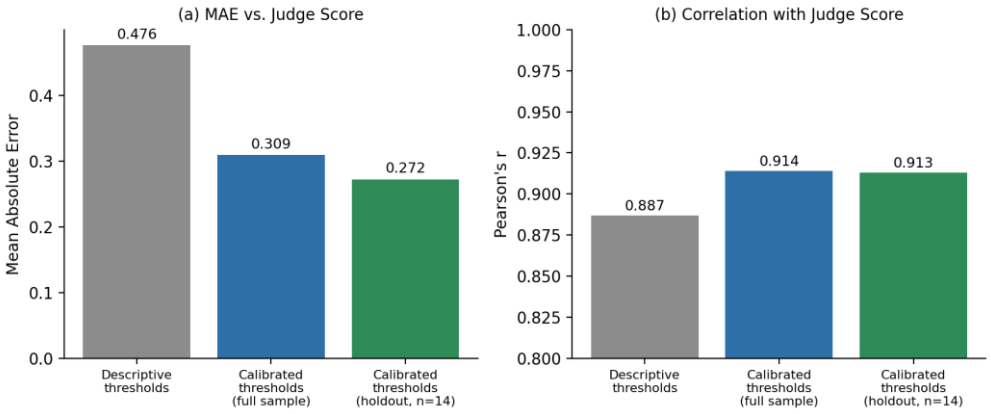

Figure 1. Expert-system agreement with judge scores across threshold conditions: (a) mean absolute error; (b) Pearson correlation.

4.2 Final Deduction Rule Set

Table 3. Final deduction axes, corresponding biomechanical variables, and threshold provenance

Deduction axis	Biomechanical variable	Threshold source	Deduction value
Knee flexion	Knee angle deviation from 180°	Descriptive (percentile-based)	0.10 / 0.30 / 0.50
Elbow flexion	Elbow angle deviation from 180°	Descriptive (percentile-based)	0.10 / 0.30 / 0.50
Trunk flexion	Trunk angle	FIG Code of Points, Art. 14.3	0.10 / 0.30 / 0.50
Leg separation	Inter-ankle distance	Grid-search calibrated (n = 56)	0.10 / 0.30 / 0.50
Asymmetry	Shoulder/hip left–right difference	Descriptive (cumulative)	0.10 / 0.20
Ascent stall / power loss	Minimum ascent velocity	Grid-search calibrated (n = 56)	0.10 / 0.30 / 0.50
Path deviation	Hip displacement from reference path	FIG Code of Points, Art. 14.3	0.10 / 0.30 / 0.50
Post-catch oscillation	Shoulder/hip oscillation after catch	FIG Code of Points, Art. 14.3	0.10 / 0.30

Three of the eight deduction axes (trunk flexion, path deviation, post-catch oscillation) were grounded directly in the published regulatory text, while the remaining five relied on either percentile-based descriptive thresholds or the grid-search calibration procedure described above.

4.3 Illustrative Worked Example

To illustrate the system's interpretability in practice, consider a representative attempt with the following measured deviations: trunk angle at the upper (major) severity band, triggering a 0.50 deduction under the FIG-grounded rule; post-catch shoulder oscillation at the medium band (0.30); inter-ankle distance within the minor band under the calibrated threshold (0.10); and all remaining axes within tolerance (0.00 each). The resulting total deduction is 0.90, yielding an expert-system score of 9.10. Unlike a neural-network prediction, which would return a single numerical estimate without further explanation, this output can be decomposed transparently into its three contributing axes and their respective rule sources, allowing a coach or review panel to verify — or contest — each component individually against the underlying kinematic measurement.

4.4 Statistical Assumptions Check

Prior to conducting the paired-samples t-test, the distribution of the 70 signed differences (actual minus expert-system score) was examined for normality using the Shapiro-Wilk test, a prerequisite for valid parametric inference. The test did not reject the null hypothesis of normality ($W = 0.985$, $p = .552$), supporting the appropriateness of a parametric paired-samples t-test over a non-parametric alternative (e.g., Wilcoxon signed-rank test) for this comparison.

4.5 Agreement with Human Judges: Paired-Samples Test

Table 4. Paired-samples t-test comparing actual (judge) and expert-system scores (n = 70)

Statistic	Value
Mean difference (actual − expert system)	-0.185
SD of difference	0.338
t (df = 69)	-4.564
p	< .001
95% CI of difference	[-0.264, -0.105]
Cohen's d	0.546 (medium)

The paired-samples t-test revealed a statistically significant systematic difference ($t(69) = -4.564$, $p < .001$, $d = 0.546$): the expert system tended to award scores approximately 0.185 points higher, on average, than the human judges. Despite reaching statistical significance, the effect size falls in the medium range and the absolute magnitude of the bias (< 0.2 points) is small relative to the overall scale (0–10).

5. DISCUSSION

The two-tier expert system achieved a strong correlation with the mean judge score ($r = 0.914$ overall; $r = 0.913$ on an independent holdout sample), supporting the feasibility of encoding a substantial share of FIG deduction logic as explicit, auditable rules rather than relying solely on opaque, data-driven prediction. The close correspondence between full-sample and holdout performance is a critical methodological safeguard: rule-based systems calibrated on a single dataset without independent validation are prone to reporting inflated, non-generalizable accuracy (James et al., 2021), a pitfall this design was explicitly structured to avoid.

The systematic, statistically significant positive bias identified via the paired t-test (mean difference = 0.185, $d = 0.546$) is, we argue, an expected finding given the system's design. The expert system's knowledge base is restricted to indicators that are directly and objectively measurable from kinematic data; it has no mechanism for penalizing artistic, aesthetic, or amplitude-related deficiencies that trained judges routinely factor into their scores despite these not being reducible to a single biomechanical variable. This asymmetry — the system being 'blind' to certain deduction-worthy qualities that lower human scores but leave measured kinematics unaffected — offers a parsimonious explanation for the direction of the observed bias.

A key strength of the proposed architecture, relative to black-box predictive models such as artificial neural networks, is complete interpretability: for any given attempt, the specific deduction axis (and corresponding biomechanical variable) responsible for a score reduction can be identified explicitly, a property increasingly emphasized in the explainable-AI literature as essential for high-stakes, human-facing automated decision systems (Molnar, 2022).

These findings extend the narrative-synthesis observation reported above, namely that none of the five directly comparable prior biomechanical studies of parallel-bar skills incorporated either a composite motion-error index or an AI-based scoring mechanism. By explicitly linking a small set of LASSO-selected kinematic indicators (Table 1) to FIG-consistent deduction values through a held-out-validated calibration procedure, the present study offers an initial, single-athlete attempt to narrow this gap while preserving full rule-level transparency — a design choice we consider preferable, for officiating contexts specifically, to marginal gains in raw predictive accuracy obtainable from opaque neural architectures.

From a measurement-theory perspective, the excellent inter-judge reliability of the reference E-score (ICC = 0.988) provides an unusually solid ground truth against which to evaluate the expert system; many prior automated-scoring feasibility studies have not reported, or have reported markedly lower, reliability for their human-judge reference standard, which complicates direct comparison of reported accuracy figures across studies. We recommend that future work in this area routinely report inter-rater reliability of the reference measure alongside model or system accuracy, to allow like-for-like comparison of automated-scoring performance across independent studies.

The magnitude of the systematic bias identified here (0.185 points on a 10-point scale, i.e., under 2% of the scale range) is small in absolute terms, and its practical significance for any specific application context — whether as a fully autonomous scoring aid or as a discrepancy flag for review — should be weighed against the intended use case rather than treated as a uniform threshold for acceptability. In a decision-support role, where the system's output is intended to prompt closer human review of specific deduction axes rather than to stand as a final score, even a somewhat larger bias would arguably remain acceptable provided its direction and approximate magnitude are known and can be communicated to end users.

5.1 Practical Implications for Judging Panels and Coaching Staff

Beyond its potential role in supporting judging panels, the axis-level transparency of the proposed system has direct applied value for coaching and athlete-development contexts. Because each deduction is traceable to a named, measurable kinematic quantity rather than an undifferentiated composite score, coaching staff can use the system's per-attempt output as a diagnostic tool, identifying which specific biomechanical fault (e.g., excessive post-catch shoulder oscillation versus insufficient trunk extension) is most consistently responsible for lost execution points across a training block, and targeting subsequent technical work accordingly. This diagnostic function is unavailable from an aggregate execution score alone, whether awarded by a human judge or predicted by an opaque statistical model, and represents a practical benefit of the rule-based design independent of its ultimate adoption — or non-adoption — within formal competition judging.

6. LIMITATIONS AND FUTURE RESEARCH

This study is subject to several limitations that constrain the generalizability of its findings. First, the sample comprises 70 attempts by a single elite gymnast; the calibrated thresholds and the observed bias pattern are specific to this individual's movement characteristics and cannot be assumed to transfer directly to other athletes without further validation. Second, three of the eight deduction axes relied on descriptive, percentile-derived thresholds rather than either an official regulatory source or an empirically validated calibration procedure, reflecting the partial availability of the regulatory text accessible to the authors; extending official-rule coverage remains an important direction for future refinement. Third, the linkage between the retained biomechanical variables and the qualitative phases of the skill (preparatory, main, and terminal) presented in this research program is conceptual rather than derived from frame-level, phase-segmented kinematic data, and should be tested quantitatively once such data become available. Future work should extend this calibration methodology to larger, multi-athlete samples and to additional apparatus and skills, enabling formal tests of cross-athlete generalizability and supporting the eventual use of such systems as decision-support tools for judging panels.

7. CONCLUSION

A two-tier, rule-based expert system combining officially documented FIG deduction thresholds with a rigorously validated grid-search calibration procedure achieved strong, holdout-confirmed agreement with human judge scores ($r = 0.914$ full sample; $r = 0.913$ independent holdout) for the Tippelt skill on parallel bars, while retaining full interpretability of its scoring logic. A small but statistically significant positive bias was identified and attributed to the system's inability to capture non-kinematic, artistic dimensions of execution quality. These findings support the potential of transparent, partially-calibrated expert systems as interpretable decision-support tools in gymnastics judging, complementary to — rather than a replacement for — qualified human officials.

Ethics Statement

Data collection procedures involving the human participant were conducted with informed consent and in accordance with the ethical standards of the institution overseeing this research program. No identifying personal information beyond performance-relevant kinematic and scoring data is reported.

Data Availability Statement

The biomechanical dataset and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

[1] Brüggemann, G. P. (1994). Biomechanics in gymnastics. International Journal of Sport Biomechanics, 10(3), 95–103.

[2] Fédération Internationale de Gymnastique (FIG). (2025–2028). FIG Judges' Rules: Specific rules for men's artistic gymnastics (Version 2.0). Lausanne, Switzerland: FIG.

[3] Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[6] Hindawi, H. H. (2021). Measurement statistics: Philosophy and application (1st ed.). Iraq: Dar Al-Dhiyaa Printing.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

[8] Knudson, D. (2020). Fundamentals of biomechanics (3rd ed.). Springer.

[9] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163.

[10] McKenzie, J. E., & Brennan, S. E. (2019). Chapter 12: Synthesizing and presenting findings using other methods. In J. P. T. Higgins, J. Thomas, J. Chandler, et al. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions (Version 6.5). Cochrane.

[11] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.).

[12] Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme.