RESTQ-Sport 77 Athlete Self-Report Measure for Soccer Players :The Revisit using Item Response Theory

Nurul Syuhadah Khusaini
Universiti Teknologi MARA
Raja Mohammed Firhad Raja Azidin
Universiti Teknologi MARA
Mohd Hanif Mohd Ramli
Universiti Teknologi MARA
Norheliena Aziz
Universiti Teknologi MARA
Abdul Rahman Hemdi
Universiti Teknologi MARA
Zulkifli Mohamed
zulkifli127@uitm.edu.my
Universiti Teknologi MARA

Research Article

Keywords: RESTQ-Sport, Fatigue, Stress, Survey, Item Response Theory

Posted Date: March 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4129481/v1

Additional Declarations: No competing interests reported.

RESTQ-Sport 77 Athlete Self-Report Measure for Soccer Players : The Revisit using Item Response Theory

Nurul Syuhadah Khusaini[1,2],
Raja Mohammed Firhad Raja Azidin[3],
Mohd Hanif Mohd Ramli[1,2], Norheliena Aziz[1,2],
Abdul Rahman Hemdi[4], Zulkifli Mohamed[4,2*]

[1]School of Mechanical Engineering, College of Engineering, Universiti Teknologi MARA, Jalan Ilmu, Shah Alam, 40450, Selangor, Malaysia.
[2*]Sports Engineering & Artificial Intelligence Centre, College of Engineering, Universiti Teknologi MARA, Jalan Ilmu, Shah Alam, 40450, Selangor, Malaysia.
[3]Faculty of Sport Science and Tecreation, Universiti Teknologi MARA, Jalan Ilmu, Shah Alam, 40450, Selangor, Malaysia.
[4*]School of Mechanical Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang Kampus Permatang Pauh, Permatang Pauh, 13500, Pulau Pinang, Malaysia.

*Corresponding author(s). E-mail(s): zulkiflii127@uitm.edu.my;

Abstract

Monitoring athletes' well-being by using subjective or objective approach is one of the important aspects to ensure optimum performance for both technical and physical during match or training. Athlete Self Report Measure (ASRM) is one of the subjective approaches and it is used widely at the global level. However, most ASRM are still printed on paper RESTQ-Sport is one of the well-known ASRM and it is established using Classical Test Theory (CTT). This study aims to re-validate RESTQ-Sport via a different approach; by implementing Item Response Theory (IRT) which will convert ordinal data to interval data for a sound and prudent objective measurement. A total of 45 respondents volunteer to participate in the pilot study; and based on the interaction between respondents (person) and items, a reliability index of 0.95 and 0.93 is recorded for both person and item. The response for this pilot study yields an explained variance by measures of 53.2%, while 9.2% is the unexplained variance. Further scrutiny is made for improvement, and the test is re-analyzed. It is evident that the explained variance by measures improve tremendously to 56.9%; whilst the unexplained variance is reduced to 5.4%.

Keywords: RESTQ-Sport, Fatigue, Stress, Survey, Item Response Theory

1 Introduction

Soccer is widely acknowledged as one of the most universally recognized sports, with a pervasive global presence and widespread participation across various levels of competition. Due to the innate dynamism involved in this sport, it is crucial to exercise prudence when evaluating the training load, addressing injuries, and facilitating recuperation to optimize the performance capabilities of each athlete.

Two methodologies, objective and subjective measurement, are presently employed to gauge and track the performance of the athlete Thorpe et al (2017) . An athlete self-report measures (ASRM), which has garnered considerable interest among practitioners in recent times, represents a prominent method for subjective athlete monitoring Thorpe et al (2017).

A self-report measure is normally administered among a specific demographic to evaluate an individual's overall physical and mental well-being Saw et al (2015). Hence, it is conventional for the self-report measure to be administered utilizing a traditional pen-and-paper approach. Through the implementation of a paper-based system, the subsequent transfer of data to a dedicated database gives rise to concerns related to time optimization and redundant labor. The self-administered self-report approach is currently subject to advancements as a result of the rise of online technologies, which are presenting novel prospects. Remarkable advancements in technology have resulted in the emergence of cost-effective and immensely productive methodologies for data acquisition Griffis and Goldsby (2003) .

In a broad assessment, the implementation of computer-assisted data collection techniques has exhibited the capacity to enhance efficiency and drive cost savings in the conversion of data into a format suitable for analysis on software platforms. An example of this application would be the extraction of data from Google Form, which can be seamlessly stored in a Google Sheets database Fowler (2013). Field researchers may have found emerging technologies to be both convenient and effective, thereby generating intrigue within their professional domain. However, there are various factors that influence the effectiveness of the response in relation to any self-report instrument. In survey research, it is frequently observed that the integrity of responses can be impacted by several factors, such as the rate of response and the length of the measurement tool. This could potentially be attributed to the conclusions drawn from prior research, wherein it was indicated that the response rate quality can be influenced by the typical attention span of an adult, estimated to be approximately 20 minutes Revilla and Ochoa (2017).

Based on research findings, it is recommended that the response time for a comprehensive assessment should ideally be less than 13 minutes. This timeframe strikes a favorable balance between the need for promptness and the delivery of high-quality outcomes Menon and Muraleedharan (2020) . Therefore, in order to transition from a paper-based ASRM to a paperless ASRM, it is advisable to reduce the number of questions (items) included in the ASRM as much as feasible.

Regarding the topic of ASRM, it is worth noting that RESTQ-Sport stands out as a prominent example. RESTQ-Sport consists of a lengthy 77 items, and it explores factors particularly related to athlete's perceived physical, psychological, social well-being, which are query on a regular, often daily, basis. The tool is developed utilizing the principles of Classical Test Theory (CTT), a psychometric theory widely recognized for enhancing the standardization of assessments Guttman (1945) . Classical test theory is a traditional quantitative method for evaluating the validity and reliability of a scale based on its items Cappelleri et al (2014) . There are efforts done to further improve RESTQ-Sport, especially in simplification and reduction of items. For example, four shorter version of RESTQ-Sport were documented; where the 77 items are reduced to 52,53,36 and 33 respectively Kuan and Kueh (2015) Laux et al (2015) Nicolas et al (2019) Fronso et al (2021). Hence, we could say there are ongoing efforts to reduce the items for existing instrument. The approaches which are most used for this purpose, similar like other instruments, rely on statistical method related to Classical Test Theory (CTT). According to Prieto, while CTT may be the preferred method for validating an instrument, it has two major conceptual limitations: the lack of an explicit ordered continuum of items representing a unidimensional construct and the lack of additivity of rating scale data Prieto et al (2003) . Prieto also pointed that Item Response Theory (IRT) approach called Rasch Model provides an alternative scaling methodology. Based on previous research, any ASRM that are developed using Classical Test Theory is commonly comprised of ordinal level items on a Likert response rating. The fact that the observed scores are linearly regressed on the latent constructs is the primary problem with the CTT and all its generalizations. If the latent trait is assumed to be on an interval scale, then the researchers will treat the observed scores or the sum scores as if they were interval scaled as well. This is a major flaw in the assumption of a linear relationship and is one of its most problematic aspects. However, the data from the Likert rating contradicts this statement. Because the precise "distance" between the subjective criteria that are being evaluated cannot be determined, hence it is not possible to locate the mean for Likert rating. To put it in another way, while it is possible to find an average of the rating (for example, rating 1, 2, and 3), it is not possible to compute an average of the subjective measures varying from "Strongly Agree" to "Strongly Disagree". Perhaps a good example will help. Referring to Fig. 1, a rating scale containing the following response options are provided: Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), and Strongly Agree (SA). Common approach is to assign rating to all five-response option and then to compute the mean (μ). The process is shown in Table 1.

Considering that there are 30 respondents, fifteen of them responded to SD, one chose D, one chose N, one of them chose A and the remaining twelve respondents responded to SA. It is a common practice to find the mean (μ) for the subjective

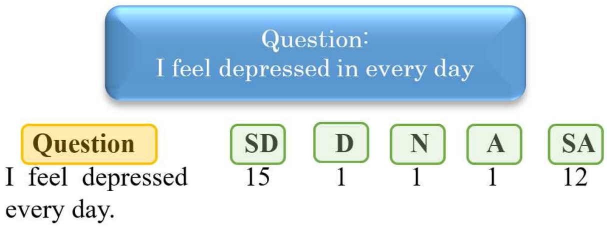

Fig. 1: Example of a common Likert Rating and its responses (n=30)

Table 1: Typical Likert Rating Calculation

Item	**Strongly Disagree (1)**	**Disagree (2)**	**Neutral (3)**	**Agree (4)**	**Strongly Agree (5)**
I feel depressed every day.	15	1	1	1	12
Finding mean (μ)					
i) 15 x 1 = 15					
ii) 1 x 2 = 2					
iii) 1 x 3 = 3					
iv) 1 x 4 = 4					
v) 12 x 5 = 60					
Total = 84					
Mean (μ) = 84 / 30 = **2.8**					

measures (for this example, the subjective measure is on depression level). Based on the calculation as shown above, the mean is 2.8. But what is 2.8 exactly? Is it the feeling of disagreement but somehow has a hint of neutral preference? Is it true that averagely the respondents' depression level lies between Disagree (not depressed) and Neutral even though most of them chose Strongly Disagree (SD) and Strongly Agree (SA)?

We can see from this elementary example that ordinal data analysis can be difficult because it does not allow results from different respondents to be compared with respect to the questions in the instrument, using the same unit of measurement.

As known to many, any number can be considered meaningless, if it is not associated with any unit of measurement. A number 5 means nothing but putting a unit of measurement to it (i.e USD5, 5 centimeters, 5 yards) will give a meaningful value to it. One way to get around this problem is to use the Rasch Measurement Model (RMM), a subclass of Item Response Theory (IRT) that evaluates the measurement characteristics of the fitness of item and instrument uni-dimensionality based on the individual's response to the items in question using a measurement unit called Logit.

A shorter ASRM is required to keep the attention of the average adult, as was stated earlier. If items on a construct are evaluating the same level of difficulty, the

Rasch Measurement Model (RMM) can tell you that they are psychometrically redundant. Such understanding can help in revising instruments to make them more efficient and effective.

Hence, this study aims to offer new paradigm of RESTQ-Sport instrument validation using the Item Response Theory (IRT) approach. To the author the best knowledge, this is the first study to analyze RESTQ- Sport using IRT.

2 Rasch Measurement Model (RMM)

One of the most well-known methods in item response theory (IRT) is the Rasch Measurement Model (RMM), which necessitates that response data be standardized to fit the measurement model. In honor of Danish mathematician Georg Rasch, the Rasch Measurement model bears his name. RMM has been used extensively in areas such as education Khusaini et al (2017) , psychology Chen et al (2021) and rehabilitation Post et al (2022) .

Rasch Model offers valuable insights at both the scale and item-level, enabling the customization of scales to suit specific research populations and contextual requirements. It is a prevalent practice among researchers in the field of social sciences to employ Classical Test Theory (CTT) as a means to analyze Likert scale responses, such as those gathered through survey instruments.

CTT is excellent for presenting an overall overview of descriptive data, but it is insufficient for analyzing a fully objective measurement. Most Likert rating responses are presented as ordinal numerical assignments. For a measurement to be legitimate, it is necessary to distinguish between ordinal and interval data. Stevens highlighted this concern is his statement: "when operations are available to determine only rank order, it is of questionable propriety to compute means and standard deviations... If we want to interpret the result of averaging a set of data as an arithmetic mean in the usual sense, we need to begin with more than an ordinal assignment of numerals" (p.29) Stevens et al (1959) . Treating a 'ranking-type' ordinal data as if they are interval data will cause misleading findings. Ordinal data only permits simple descriptive statistics such as count, relative frequency, percentage and median. To find mean or perhaps standard deviation, interval or ratio data are needed Royal (2010) . RMM could transform ordinal data to interval data, hence allows statistical analysis to be conducted since the data is now linear and have equal interval.

The RMM is a probabilistic model that gives each person and item a standard error Post et al (2022) . This method is different from CTT, where the standard error of measurement covers the whole range of people's abilities. This means that every person in the dataset has the same amount of error in their measurements. RMM stands out because of three key characteristics. First, the technique assumes that the construct being studied is regularly distributed (but not certainly the observable data being examined). Second, derived measures should be "test-free" and "person-free," according to this idea. The third idea is that the goal of Rasch measurement is to create a one-dimensional scale and then assess how well the data conforms to it Stemler and Naples (2021) .

The Rasch Measurement Model (RMM) considers the interconnectedness between individual responses and item difficulty, whereby the estimation of item difficulty is contingent upon the individual's ability. In summary, RMM operates on the foundation of these two fundamental principles Post et al (2022) :

i. An individual possessing higher competence is more likely to respond accurately to all items.

ii. A straightforward item is more inclined to elicit accurate responses from all individuals.

In regards to this particular study, both theorems have been reconstructed in a similar fashion:

i. An individual who is in optimal physical and mental condition is more likely to respond affirmatively to all questions.

ii. Simple practices have a higher chance of being successfully used by all participants.

3 Methodology

The methodology presented in this paper consist of validating the instrument through a pilot study. A total of 45 responses were obtained from the pilot study. The participants' demographic details (age, weight, height, BMI, cohort, and playing position) are shown in Table 2.

Table 2: Participants Demographic Attributes

Demographic Attributes	Total Sample (N=45)
Age	Mean, $\mu = 21.6$ years
Weight	Mean, $\mu = 67.74$ kg
Height	Mean, $\mu = 1.73$ meter
BMI	Normal (N=39)
	Overweight (N=6)
Cohort	Premier League (N=6)
	President Cup (U22) (N=22)
	Super League (N=17)
Playing Position	Premier League (N=6)
	President Cup (U22) (N=22)
	Super League (N=17)

3.1 RESTQ-Sport 77

The Recovery Stress Questionnaire for Athletes (RESTQ-Sport) is developed by Kallus & Kellman in year 2001 Kellmann and Kallus (2001) . The RESTQ-Sport is being used widely up until today. Most if not all research investigating RESTQ-Sport construct validation use the Classical Test Theory (CTT) as a means of analysis. This instrument consists of 77 items, segregated into four clusters namely General Stress Scales, General Recovery Scales, Sport-Specific Stress Scales and Sport-Specific Recovery Scales as shown in Table 3.

These clusters consist of nineteen (19) constructs; with four items in each construct. All the items are asked with a statement at the beginning of it which says, “In the past (3) days/nights...”. Though the items are seventy-seven (77) in total, only seventy-six (76) items are grouped into the mentioned clusters, since the first item is considered as “warming-up item”. However, for the purpose of analyzing the instrument validation via The RMM approach, all seventy-seven items will be analyzed without considering which cluster they belong to. Also, the RESTQ-Sport 77 consists of a combination of positive and negative worded questions.

Table 3: Scales in RESTQ-Sport 77

General Stress Scales	
• A – General Stress	• E – Fatigue
• B – Emotional Stress	• F – Lack of Energy
• C – Social Stress	• G – Physical Complaints
• D – Conflicts / Pressure	
General Recovery Scales	
• H – Success	• K – General Well-Being
• I – Social Recovery	• L – Sleep Quality
• J – Physical Recovery	
Sport-Specific Stress Scales	
• M – Disturbed Breaks	• O – Injury
• N – Emotional Exhaustion	
Sport-Specific Recovery Scales	
• P – Being in Shape	• R – Self Efficacy
• Q – Personal Accomplishment	• S – Self Regulation

3.2 Pilot Study

A pilot study was conducted, and the participants are soccer players from Super League, Premiere League and President Cup (U22). The pilot study approval is granted from the university Research Ethics Committee (REC/11/2020 (MR/392)).

As previously indicated, the questionnaire comprises items that are phrased in both positive and negative forms, accompanied by a seven-point Likert scale. The scale consists of rating options ranging from 0, representing ”Never,” to 6, representing ”Always.” The rating of negatively worded questions is recorded only after obtaining responses to facilitate subsequent RMM analysis.

The aim of conducting the pilot study is to present an alternative framework of data analysis. Classical Test Theory (CTT) is widely employed in the field of social sciences for the assessment of un-observable attributes. Therefore, it would be valuable to conduct an examination of the tool employing Item Response Theory (IRT).

4 Rasch Measurement Model Analysis

The RMM analysis is conducted to investigate the reliability, goodness of fit test as well as its uni-dimensionality.

4.1 Reliability Indices

The RMM exhibits a distinction from CTT due to its comprehensive examination of both the reliability of items and individuals. Table 4 displays the reliability coefficients of both the items and individuals, accompanied by the corresponding standard error values.

Table 4: Summary of Fit Statistics for RESTQ-Sport 77 for N = 45; i = 77

(N=45 ; I = 77)	**Item – Person Interaction**	
	Item	**Person**
Mean	0.00 logit	+0.64 logit
SD	0.64	0.39
MnSq (Infit & Outfit)	0.99 (Infit) 1.03 (Outfit)	1.03 (Infit) 1.03 (Outfit)
Z-Std (Infit & Outfit)	-0.11(Infit) 0.05 (Outfit)	-0.42 (Infit) -0.38 (Outfit)
Maximum Measure	1.62 logit	1.41 logit
Minimum Measure	-1.38 logit	-0.02 logit
Reliability Indices		
Separation	4.49	3.65
Reliability	0.95	0.93
Model Error	0.13	0.09
Cronbach Alpha	0.94	

A comprehensive analysis was conducted using the Winsteps Version 4.8.0.0 software to analyze 3465 data points collected from 45 participants who responded to a total of 77 items in the RESTQ-Sport 77 questionnaire. Utilizing the Winsteps software, Response Model Measurement (RMM) converts the raw scores of responses (representing ordinal data) into a ratio of success-to-failure probabilities, commonly referred to as odds. Following this, the data is further translated into its logarithmic form, also known as logit unit. The concept of item reliability pertains to the inclusion of an ample number of items in the measurement instrument that adequately capture a broad range of abilities or, in the context of this study, an individual player's level of well-being.

Based on the results outlined in Table 4, it can be observed that the item reliability demonstrates a strong value of 0.95 on a scale ranging from 0 to 1, utilizing the identical approach employed for interpreting Cronbach Alpha. According to the reliability indices provided, the item separation index indicates a positive range of items, with a value of 4.49 Fisher (2007) . In a different vein, a person's degree of dependability can be used to assess how consistently they will respond to a test with the same questions and a similar conceptual structure. In the present study, the individual reliability illustrates a commendable significance of 0.93, accompanied by a minimal model error of $\pm$ 0.09. Both the person and item demonstrate a significant level of conformity based on the MnSq and ZStd measures, indicating that the instrument effectively and accurately assesses the intended factors. The maximum value for the item measure is documented as 1.62 logit with a standard error of 0.26, while the maximum value for the person measure is slightly higher at 1.41 logit with a standard error of 0.12.

Conversely, an ample quantity of items is available, evidenced by the minimum item measure of -1.38 logit, in comparison to the minimum person measure of -0.02 logit.

According to the data presented in Table 4, the item reliability demonstrates a robust value of 0.95 when measured on a 0-1 scale, which aligns with the interpretation method utilized for Cronbach Alpha. Based on the reliability indices, the item separation index has been observed to be 4.49, indicating a favorable range of items Fisher (2007) .

Conversely, person reliability signifies the extent to which an individual's responses remain consistent when presented with the same test measuring the same construct. In the present study, the person reliability coefficient demonstrates an outstanding value of 0.93, accompanied by a negligible model error of $\pm$ 0.09. Both the individual's and the item's MnSq and ZStd values are close to 1 and 0, respectively. This indicates a strong level of fit and confirms that the instrument effectively measures the intended variables. The highest item measure recorded is 1.62 logit with a standard error of 0.26, whereas the highest person measure is marginally higher at 1.41 logit with a standard error of 0.12.

However, it should be noted that there is an adequate number of items available, as the minimum item measure of -1.38 logit is higher than the minimum person measure of -0.02 logit.

4.2 Item Fit Analysis

One effective approach to identifying a misfit item is through the analysis of the Point Measure Correlation (PMC). The occurrence of a negative PMC value indicates the presence of multidimensionality Abdullah et al (2013) . In order for an item to be deemed a misfit, it is important to take into account three (3) specific parameters Othman et al (2014) . The three parameters are:

i) Point Measure Correlation; $0.4 < x < 0.8$

ii) Infit Mean Squared; $0.5 < y < 1.5$

iii) Z-Standard; $02.0 < z < +2.0$

A comprehensive analysis is conducted on all 77 items to assess their level of compatibility. According to the findings in Table 5, there are 37 items that may be outliers based on their value of MnSq, ZStd and PMC. Notably, one particular item (Item M3: In the past (3) days/nights, too much was demanded of me during the breaks) displayed a negative Point Measure Correlation (PMC) value of -0.10. This result suggests that this specific item is not effectively capturing the intended measurement and should be considered for elimination from the analysis. Hence, only thirty-six items from Table 5 are valid for further scrutiny.

Next, out of the thirty-six (36) items in Table 5, thirty one (31) items have value between 0.01 and 0.40. Items exhibiting a PMC value below 0.4 can be deemed to demonstrate a low or weak level of correlation Othman et al (2014). As a result, these items are also discarded. Out of the total of thirty-five (35) items with low PMC value, four (4) items exhibited In fit MnSq and Infit ZStd values exceeding 1.5 and 2.0 respectively.

Table 5: Largest Standardize Residual Correlation for N=45; i=77

Item No	Measure	Model SE	Infit Mnsq	Infit Z-Std	Outfit MnSq	Outfit Z-Std	PMC	Item
28	-.59	0.15	1.63	1.91	2.24	3.00	.11	B3_Felt anxious or inhibited
66	-.31	0.13	1.56	2.02	2.19	3.36	-.10	M3_Too much during breaks
11	-.20	0.12	1.51	1.99	2.13	3.43	.14	F2_Difficulties concentrating
77	1.09	0.1	2.05	4.22	1.99	3.85	.11	Q4_Dealt emotional problems very calmly
63	-.27	0.13	1.61	2.22	1.85	2.64	.06	N2_Felt emotionally drained performance
60	1.62	0.13	1.53	1.92	1.74	2.34	.10	Q2_Dealt effectively teammates' problems
76	-.87	0.18	1.73	1.87	1.42	1.14	.34	N4_Felt frustrated by my sport
23	1.10	0.1	1.65	2.85	1.49	2.15	.28	I3_Visited close friends
40	-.37	0.13	1.05	.28	1.53	1.71	.15	F4_Put off making decisions
38	-.59	0.15	1.49	1.56	1.03	.20	.53	C3_Was upset
51	-.29	0.13	1.33	1.30	1.49	1.65	.29	M1_Could not get rest breaks
31	-.51	0.14	1.03	.21	1.41	1.30	.17	F3_Was lethargic
58	-.59	0.15	1.40	1.31	1.37	1.13	.21	M2_Impression too few breaks
16	-.31	0.13	1.09	.42	1.36	1.28	.03	E2_Tired from work
3	.97	0.1	1.33	1.66	1.29	1.41	.30	H1_Finished important tasks
68	-1.38	0.26	1.33	.83	1.33	.81	.28	N3_Felt that I wanted to quit sport
50	-.16	0.12	1.12	.60	1.30	1.16	.01	O1_Parts body aching
6	.48	0.1	1.20	1.10	1.29	1.47	.24	I1_I laughed

Continued on next page

Table 5 continued

Item No	Measure	Model SE	Infit Mnsq	Infit Z-Std	Outfit MnSq	Outfit Z-Std	PMC	Item
41	.99	0.1	1.17	.89	1.27	1.31	.30	H3_Made important decisions
54	-.51	0.14	1.05	.27	1.24	.84	.12	N1_Felt burned out by sport
2	-.14	0.12	.78	-.98	.86	-.51	.24	E1_Not enough sleep
4	-.43	0.14	.73	-1.03	.86	-.39	.30	F1_Unable to concentrate
42	-.37	0.13	.77	-.91	.83	-.53	.24	G4_Felt physically exhausted
72	-.75	0.17	.83	-.47	.79	-.55	.27	M4_Breaks not at the right time
37	-.75	0.17	.79	-.59	.74	-.72	.35	B4_Was annoyed
44	-.59	0.15	.76	-.81	.79	-.62	.32	D4_Felt under pressure
12	.07	0.11	.71	-1.51	.78	-1.01	.25	D1_Worried unresolved problems
57	-.09	0.11	.76	-1.12	.77	-.98	.02	O2_Muscles felt stiff or tense
61	.26	0.1	.64	-2.08	.76	-1.22	.71	P2_Good condition physically
48	-.69	0.16	.69	-1.01	.74	-.75	.37	C4_Angry with someone
7	-.72	0.17	.58	-1.46	.73	-.75	.36	G1_Felt physically bad
26	-.75	0.17	.61	-1.33	.71	-.84	.12	C2_Other people got on my nerves
69	.46	0.1	.66	-2.06	.68	-1.85	.83	P3_Felt very energetic
17	.61	0.1	.65	-2.17	.65	-2.10	.73	H2_Successful in what I did
21	-.67	0.16	.63	-1.32	.59	-1.36	.39	C1_Annoyed by others
75	.59	0.1	.61	-2.46	.63	-2.25	.74	P4_Body felt strong
8	-.43	0.14	.48	-2.35	.50	-2.02	.39	B2_In a bad mood

4.3 Uni-Dimensionality Test

The Rasch Model allows for the identification of redundant items, wherein the responses yield similar outcomes. One benefit of having awareness of items that overlap is the potential to streamline the questionnaire by eliminating redundancies. This ultimately leads to a shorter survey, which in turn has the potential to enhance the overall response rate. This factor holds significant importance in the process of transitioning a conventional self-report assessment into an online self-report assessment.

The RMM methodology encompasses the application of Principal Component Analysis (PCA) on the standardized residual. As indicated by the data presented in Table 6, the RMM accounted for 46.8% of the raw variance, a figure that closely aligns with the expected model variance of 47.8%. Therefore, it can be concluded that the degree of variance explained by the RMM is deemed acceptable. The unexplained variance attributable to the first construct stands at 9.2%, falling below the prescribed threshold of 15% Linacre (2007) . This variance is caused by "noise" from items in Table 7.

Table 6: Standardized Residual Variance (in eigenvalue units) for N=45; i=77

Description	Empirical (%)	Model
Variance explained by measures	53.2	52.2
Unexplained variance	46.8	47.8
Unexplained variance in 1st contrast	9.2	

This phenomenon of "noise" originated from two distinct components that exhibit a high level of inter-correlation (correlation > 0.7), suggesting that respondents perceived them as posing similar inquiries albeit with subtle variations in wording. Upon analysis of the provided sentences, it becomes apparent that there is a clear correlation between each pair of items. This correlation suggests that respondents perceive these items as inherently similar to one another. One possible course of action is to assess the options and determine the most suitable item, while disregarding the other item that may not be as suitable.

Furthermore, according to the data presented in Table 7, it is evident that there exist two (2) sets of items that exhibit a significant level of local dependence (with a correlation coefficient greater than 0.7, denoted in light grey), as well as six (6) sets of items that demonstrate a moderate level of local dependence (with a correlation coefficient between 0.5 and 0.7, denoted in dark grey). Additionally, items that are denoted in white are also observable. The items that have been marked in white have been previously deleted and therefore will be excluded from the scrutiny process.

Therefore, following a meticulous examination, certain elements have been deemed unsuitable for inclusion, specifically *F3_Was lethargic, O1_Parts body aching, A1_Felt down, G4_Felt physically exhausted, K2_Was in a good mood, and O2_Muscles felt stiff or tense.* The variance explained by measures goes from 53.2 to 56.9 after redundant items are removed from the instrument.

Table 7: Largest Standardize Residual Correlation for N=45; i=77

Correlation	Item	Item
0.72	F3_Was lethargic	E4_Was overtired
0.72	O1_Parts body aching	O3_Muscle pain after performance
0.69	A1_Felt down	C3_Was upset
0.69	E4_Was overtired	G4_Felt physically exhausted
0.69	K2_Was in a good mood	K3_Felt happy
0.68	B3_Felt anxious or inhibited	M3_Too much during breaks
0.68	E4_Was overtired	O1_Parts body aching
0.67	O2_Muscles felt stiff or tense	O3_Muscle pain after performance

4.4 Wright Map

The Wright Map (WM) is regarded as the most distinguished characteristic of RMM. The distribution of both the respondent and the item are illustrated by WM on a single continuum. In simpler terms, the concept of a weighted measurement can be likened to that of a ruler due to its attributes of having numerical values in a specific order, equal intervals, and a designated unit of measurement Khusaini et al (2014). The measurement utilized for WM is referred to as the "Logit" unit.

As depicted in Figure 2, the Wright Map (WM) comprises a single vertical dashed line that serves to distinguish the respondent (positioned on the left) from the item (located on the right). Regarding this study, the left side displays the chosen group or cohort for the RESTQ-Sport study, categorized based on their level of concurrence towards the items in all four scales of RESTQ-Sport. The athlete displaying the highest number of dissenting responses regarding their stress level is positioned at the highest rank, while the athlete with the lowest count is positioned at the lowest rank. The map displays the most agreeable RESTQ-Sport item at the uppermost position on the right side, while the least agreeable RESTQ-Sport item is positioned at the lowermost point on the same side.

The letter M, specifically marked at the dashed line, represents both the mean of the item and the mean of the person. While the letters S and T denote one and two standard deviations, respectively, from the mean. An item is deemed appropriately calibrated for a test-taker when the likelihood of them answering it correctly stands at 50%. In this scenario, the logit value can be calculated as ln $(50/50) = 0$, as both the likelihood of agreeing and disagreeing with each item is equal at 50% Stemler and Naples (2021) . Therefore, the average value of the item has been calibrated to zero on the Logit scale. The individual's score is significantly influenced by their responses to the items in the RESTQ-Sport questionnaire.

Self-report assessments allow for the inference that a person's level of stress is either (a) equal to, (b) higher than, or (c) lower than the stress level expressed in an item Wilson (2023) . In summary, the purpose of WM not only provides a visual

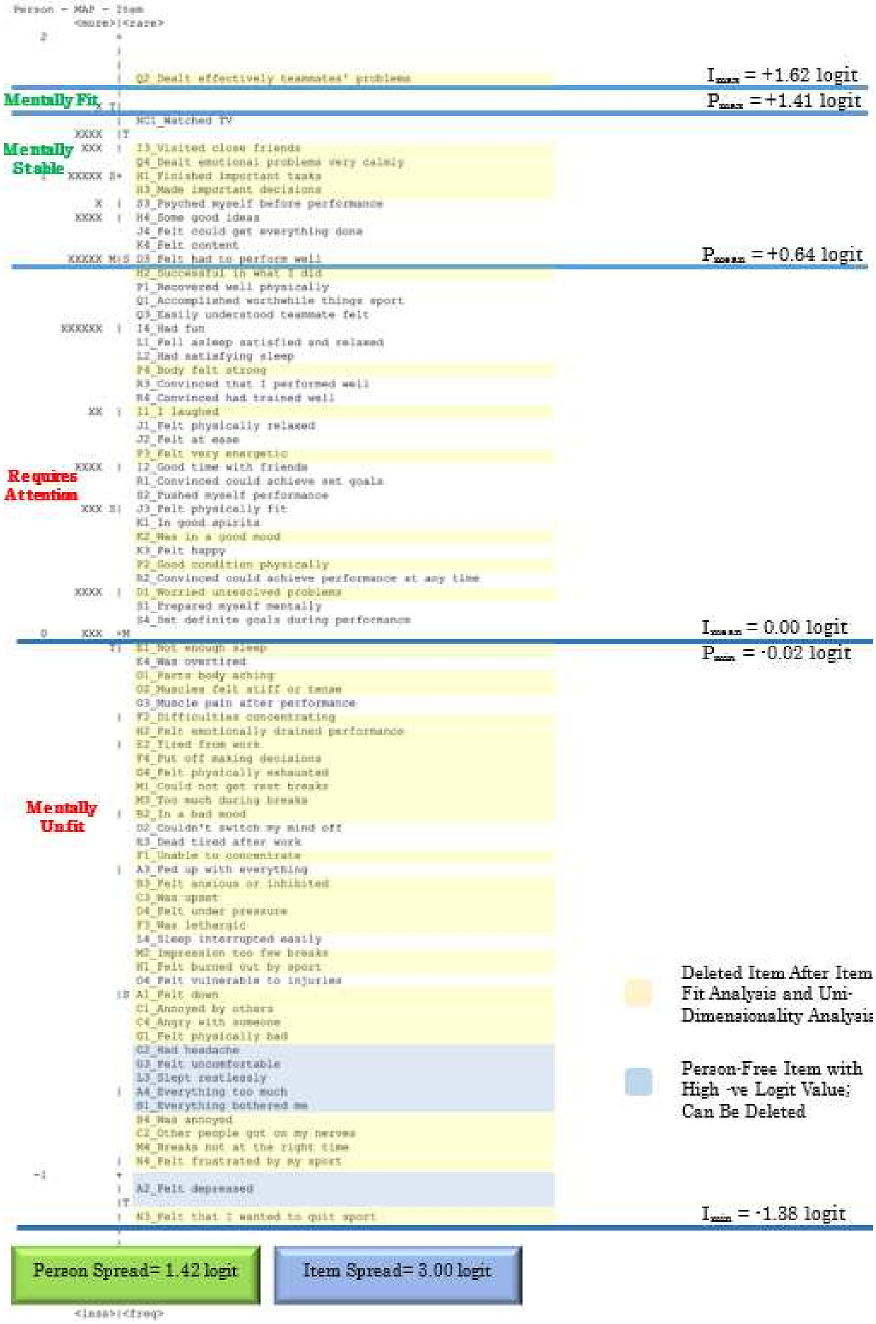

Fig. 2: Wright Map for RESTQ-Sport 77

representation of the concept but also establishes its validity through objective measurement. The left panel of the WM represents the participant's position in accordance with their proficiency, while the right panel indicates the item's position in relation to its level of difficulty.

According to the WM analysis, it can be observed that the minimum Person measure (Pmin) corresponds to -0.02 Logit, whereas the maximum Person measure (Pmax) is equivalent to 1.41 Logit. Therefore, the range of the Person measure extends up to 1.43 Logit. The average value of the Person's response (Pμ) is 0.64 Logit, which is near the average value of the Item (Iμ), 0.00 Logit, within this dataset. The item classified as having an average level of difficulty will be assigned a logit unit value of zero, as per the Rasch model which standardizes the mean item difficulty at zero Logit units Bond et al (2015) . Furthermore, the proximity of the mean score for the Person (0.64 Logit) to the mean score for the Item (0.00 Logit) indicates favorable alignment between the test and its items. This indicates that the selection of test items for the respondents is aligned with the characteristics and needs of the intended respondent group Boone (2016) .

In contrast, the Imin value for the minimum Item measure is -1.38 Logit, whereas the Imax value for the maximum Item measure is 1.62 Logit. Therefore, the Item measure exhibits a spread of 3.00 Logit. Based on the discrepancy between the Item measure spread (3.00 Logit) and Person measure spread (1.41 logit), it can be inferred that the items sufficiently capture the stress level of the target population.

After compiling and organizing all pertinent Logit measures, the process of categorizing the respondents can commence. According to the data presented in Figure 2, the participants can be classified into four distinct groups, characterized as follows: Mentally Fit (Group 1), Mentally Stable (Group 2), Requiring Attention (Group 3), and Mentally Unfit (Group 4). Therefore, this finding provides compelling evidence that athletes can be categorized based on their levels of stress, facilitating the development of targeted intervention strategies for individuals falling within Group 3 and Group 4.

Also, can be seen from Figure 2, items highlighted in yellow and blue. The items highlighted in yellow is the items deleted from previous section (Item Fit Analysis and Uni-Dimensionality Test). The total of items deleted so far is 39 items. On the other hand, for Group 4 (Mentally Unfit), there are a total of thirteen (13) items (excluding the items deleted which is highlighted in yellow) with high -ve Logit measures. Those items are shown in Table 8.

Also, all these thirteen (13) items are below Item's mean (0.00 Logit) and some of it can be deleted since these items are Person-Free items. Hence, six (6) items with high -ve measures are deleted (Items G2 to A2).

4.5 Cleaned Rasch Measurement Model Analysis

After data cleaning (significant item deletion), the variance explained by measures increase from 53.2 to 56.9; as seen in Table 9. This is due to the decrement of unexplained variance in 1^{st} contrast; whereby previously a total of 15.1 items strength are likely to have the sign of secondary dimension. One might question, what is the problem with having secondary dimension? Well, the answer is quite straightforward.

Table 8: Items with NegativeLogit Value for N=45; i=77 in Group 4(Mentally Unfit)

Item	Measure (Logit)
O3_Muscle pain after performance	-.09
E4_Was overtired	-.14
E3_Dead tired after work	-.41
D2_Couldn't switch my mind off	-.43
L4_Sleep interrupted easily	-.51
A3_Fed up with everything	-.57
O4_Felt vulnerable to injuries	-.57
G2_Had headache	-.62
G3_Felt uncomfortable	-.62
L3_Slept restlessly	-.64
B1_Everything bothered me	-.77
A4_Everything too much	-.80
A2_Felt depressed	-1.15

Objective measurement requires creation of uni dimensional measurement scales [31]; which means a study should describe and measure the attribute of the event under observation; one at a time. For instance, when a doctor needs to decide whether a patient is having a fever, the doctor will use a thermometer (which measures only one attribute (uni dimensional); regardless of age, weight or height.

Table 9: CleanedRESTQ-Sport Construct Properties

Description	Item	Person
PCA Variance explained by measures	**56.9** (53.2)	
Unexplained Variance in 1st contrast (Eigenvalue Observed)	**3.99** ; **5.4%** (15.1 ; 9.2%)	
Reliability	**0.93** (0.95)	**0.95** (0.93)
Separation	**3.79** (4.49)	**4.23** (3.65)
Infit MnSq SD	**0.36** (0.31)	**0.64** (0.50)
Mean Error	**0.12** (0.13)	**0.14** (0.09)

Thus, having a uni-dimensional instrument is important, so we could deduce clearly the outcome of the trait that is being measured. With the new unexplained variance in 1^{st} contrast of only 3.99 items strength, the reliability for both item and person maintains at a value of more than 0.9. Though a slight drop in item reliability, its power of person separation increases from 3.65 to 4.23; and the infit MnSq S.D and

S.E remains very small. In essence, with the new RESTQ-Sport 32 (which indicates 32 items), the instrument has been proven sufficient to measure the recovery and stress level of the Malaysian soccer players.

5 Conclusion and Recommendation

This study is intended to unfold the application of RESTQ-Sport 77 for Malaysian Soccer Players, as well as its analysis from Rasch Measurement Model point of view. Based on the objectives, clearly conducting the instrument's validation using Rasch Measurement Model offers empirical evidence for the fitness of item, as well as making sure the item is indeed uni-dimensional, which then permits a meaningful data interpretation and analysis. Also, it is worth noting that, in the case of Malaysian soccer players, 31 items are sufficient to measure what is intended to be measured using the RESTQ-Sport.

Future works should include the involvement of practitioners (coaches, physicians etc) to use the newly refined instrument for player's monitoring. The input from practitioners may further improve this instrument to provide meaningful data for each individual player.

References

Abdullah N, Masodi S, Aziz AA, et al (2013) Investigation of parallel resistive circuit pcut-40 instrument construct validity to assess engineering students' understanding in electricity. Res J Appl Sci 8(10):470–476

Bond TG, Yan Z, Heene M (2015) Applying the Rasch Model Fundamental Measurement in the Human Sciences, Third Edition. Routledge, New York

Boone WJ (2016) Rasch analysis for instrument development: why,when,and how? CBE Life Sci Educ 15(4)

Cappelleri JC, Lundy JJ, Hays RD (2014) Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 36(5):648–662

Chen W, Liang Y, Yin X, et al (2021) The factor structure and rasch analysis of the fear of covid-19 scale (fcv-19s) among chinese students. Front Psychol 12

Fisher WP (2007) Rating scale instrument quality criteria. Rasch Meas Trans 21(1)

Fowler FJ (2013)

Fronso SD, Robazza C, Montesano C, et al (2021) Initial validation of a 33-item recovery-stress questionnaire for italian athletes. Open Sports Sci J 14(1):43–50

Griffis SE, Goldsby TJ (2003) Web-based and mail surveys : a comparison of response, data and cost. J Bus Logist 24(2):237–258

Guttman L (1945) A basis for analyzing test-retest reliability. Psychometrika 10(4):255–282

Kellmann M, Kallus KW (2001) Recovery Stress Questionnaire for Athletes: User Manual. Human Kinetics Publisher, Inc, Illinois

Khusaini NS, Jaffar A, Yusoff N (2014) A survey on lean manufacturing tools implementation in malaysian food and beverages industry using rasch model. Adv Mater Res 845:642–646

Khusaini NS, Saad NH, Aziz N, et al (2017) Relationship between demographic attributes and model. Int Conf Eng Educ (ICEED) pp 117–121

Kuan G, Kueh YC (2015) Psychological skills during training and competition on recovery-stress state among adolescent state athletes. J Sport Res 2(4):122–130

Laux P, Krumm B, Diers M, et al (2015) Recovery-stress balance and injury risk in professional football players: a prospective study. J Sports Sci 33(20):2140–2148

Linacre JM (2007) A User's Guide to WINSTEPS Rasch Model Computer Programs. MESA Press, Chicago

Menon V, Muraleedharan A (2020) Internet-based surveys: relevance, methodological considerations and troubleshooting strategies. Gen Psychiatry 33(5):1–3

Nicolas M, Vacher P, Martinent G, et al (2019) Monitoring stress and recovery states: structural and external stages of the short version of the restq sport in elite swimmers before championships. J Sport Heal Sci 8(1):77–88

Othman N, Salleh SM, Hussin H, et al (2014) Assessing construct validity and reliability of competitiveness scale using rasch model approach. 2014 WEI Int Acad Conf Proc pp 113–120

Post MWM, Fellinghauer CS, Charlifue S, et al (2022) Rasch analysis of the international quality of life basic data set version 2.0. Archives of Physical Medicine and Rehabilitation

Prieto L, Alonso J, Lamarca R (2003) Classical test theory versus rasch analysis for quality of life questionnaire reduction. Health Qual Life Outcomes pp 1–1

Revilla M, Ochoa C (2017) Ideal and maximum length for a web survey. Int J Mark Res 59(5):557–565

Royal KD (2010) Making meaningful measurement in survey research: a demonstration of the utility of the rasch model. IR Appl 28:1–16

Saw AE, Main LC, Gastin PB (2015) Role of a self-report measure in athlete preparation. J Strength Cond Res 29(3):685–691

Stemler SE, Naples A (2021) Rasch measurement v. item response theory: knowing when to cross the line. Pract Assessment, Res Eval 26:1–16

Stevens SS, Measurements P, Utility (1959) Meas Defin Theor

Thorpe RT, Atkinson G, Drust B, et al (2017) Monitoring fatigue status in elite team-sport athletes: implications for practice. Int J Sports Physiol Perform 12:27–34

Wilson M (2023) Constructing Measures : An Item Response Modeling Approach, second edition edn. Routledge Taylor & Francis Group