Standardized Measure for Performance Assessment of Athletes in The CrossFit Open: Theoretical Structuring and Item Response Theory

Rafael da Silva Fernandes (✉ rafasfer2@ufra.edu.br)
Federal Rural University of the Amazon https://orcid.org/0000-0002-3035-8025

Bruna Gabriele Biffe
Centro Universitário Católico Salesiano Auxilium https://orcid.org/0000-0001-9650-5713

Mário Jefferson Quirino Louzada
Centro Universitário Católico Salesiano Auxilium https://orcid.org/0000-0002-5744-2235

Antônio Cézar Bornia
Universidade Federal de Santa Catarina https://orcid.org/0000-0003-3468-7536

Dalton Francisco de Andrade
Universidade Federal de Santa Catarina https://orcid.org/0000-0002-4403-980X

Research Article

Keywords: Fitness, Latent trait theory, Metrics, Physical conditioning, Sports

Posted Date: April 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1308148/v3

Standardized Measure for Performance Assessment of Athletes in The CrossFit Open: Theoretical Structuring and Item Response Theory

Rafael da Silva Fernandes [1,3*], Bruna Gabriele Biffe [2], Mário Jefferson Quirino Louzada [2], Antônio Cezar Bornia [3], Dalton Francisco de Andrade [3]

[1] Campus Parauapebas, Federal Rural University of the Amazon, Parauapebas, PA, 68.515-000, Brasil

[2] Department of Medicine, Salesian Catholic University Center Auxilium, Araçatuba, SP, 16.016-500, Brasil

[3] Postgraduate Program in Production Engineering, Federal University of Santa Catarina, Florianópolis, SC, 88040-900 Brasil

[*]CONTACT Rafael S.Fernandes. Resing rafasfer2@ufra.edu.br Parauapebas campus, Federal Rural University of Amazonas.

Absctract

In its competitive form, CrossFit® intends on assessing the performance of athletes in a wide variety of aspects that determine their conditioning. CrossFit Games is an official competition and intends to recognize the best conditioned athlete in the world to each class. Thus, measuring an athlete's Physical Conditioning is, in a sense, assigning a set of performance outputs that can determine the efficiency and efficacy of the athlete, discriminating the performance of one or more athletes. Since the scores obtained by the athletes in the various workouts are directly identifiable, the conditioning can be seen through the performance in a competition, it is, therefore, a result of the interpretation and scope of the workouts in measuring the performance. This work analyzed data form "CrossFit Open" and has as an objective to propose a new theoretical arrangement to the sport discipline while being of the Item Response Theory which is capable of providing data such as, discriminatory capacity and difficulty level of the workouts, as well as, an assessment of the competition. In other words, intends to describe the probability of an athlete performing a workout and obtaining a score, given his/her physical conditioning. Analysis of the main indications that refer to a good quality of the measurement tool indicates it is a high quality competition. Lastly, this work accomplished both objectives proposed, methodological as well as practical, and recognizes the limitations derived from the reduced amount of qualitative data on the topic and the little use of applied probability models.

Keywords: Fitness, Latent trait theory, Metrics, Physical conditioning, Sports.

Introduction

CrossFit®, sport discipline that has been growing worldwide, has become a popular sport with more than 15.000 members all over the world. Such an increase may be highlighted when compared to the number of athletes subscribed in its annual competition, called "The CrossFit Open", in which approximately 26.000 athletes participated in 2011, reaching 572.653 subscribed athletes in 2019 (CrossFit; Glassman, 2004).

In its official website, CrossFit®, is defined as:

> *"CrossFit is a lifestyle characterized by safe, effective exercise and sound nutrition. CrossFit can be used to accomplish any goal, from improved health to weight loss to better performance. The program works for everyone — people who are just starting out and people who have trained for years."*

Usually, every sport discipline, specially, CrossFit® – which presents multiple physical requirements – needs to identify effective techniques to analyze the performance through a smaller number of influential variables, thus facilitating the analysis and the development of training programs to enhance relevant physical skills. Due to its practical nature, this performance enhancement usually happens in an evolutive and adaptative manner (Gómez-Landero & Frías-Menacho, 2020).

Most of the studies dedicated to CrossFit® have been directed towards understanding physiological and nutritional factors, training strategies, physical and psychological recovery and other aspects that may directly influence the performance of the athletes (Claudino et al., 2018; Mangine, Stratton, et al., 2020; Mangine, Tankersley, et al., 2020; Schlegel, 2020).

Typically, such studies vary depend on whether one is analyzing beginners, athletes with longer sport experience, athletes who are focused on maintaining their health, high-level competitors and several classes relating to age group and sex.

Regarding functional limitations, sporting performance is regulated via different factors that go through the ability of efficiently repeating the contractile motor activity, however it is limited by the progression of the fatigue – characterized for a decrease in the strength or production of musculoskeletal energy causing the reduction in the capacity of keeping the intensity of the exercise, so that greater fatigue leads to better performance (García-Pinillos et al., 2019; Hargreaves & Spriet, 2020; Khassetarash et al., 2021; Potvin & Fuglevand, 2017; Taylor et al., 2016; Wan et al., 2017).

is the reduced variability in the performance of athletes when they are position in a specific class. Thus, after considering decisive performance aspects, strategies design a better adaptive response or responses to induce the best gain to the athlete's output (Hanin & Hanina, 2009; Silva-Grigoletto et al., 2013).

Currently, the athlete's performance evaluation criteria is provided by the score obtained in the execution of a workout. So, performance can be conceptualized, in the CrossFit® context, as an output or score, presented in time, number of repetitions or pounds, originated from the execution of a workout by an athlete, being possible to distinguish between the efficiency and efficacy of the execution.

Based on this concept, three points can be highlighted:

i) Performance, based on the output, it is a tool to measure physical conditioning.

ii) Performance ascertains the athlete's physical conditioning, in other words, the efficiency and efficacy of the workout performed by the athlete.

iii) Performance allows the differentiation or distinction between the conditioning of two or more athletes.

It is important to clarify that the definition of conditioning is not directly identifiable and observable. What is directly observable and identifiable are the outputs obtained by the athletes in the various workouts performed. In other words, conditioning is perceived through the performance of the athletes in the workouts proposed in a competition, therefore, it is a resultant of the interpretation and scope of the workouts in measuring this execution.

For this purpose, evaluating or determining the athlete with the best conditioning based on a single workout is flawed, since it is insufficient to encompass the wide variety of exercises proposed by CrossFit® itself.

Despite the validation of the *"CrossFit Games"* as a measurement tool of the athlete's conditioning, principles of the Classical Test Theory (CTT) are applied to rank athletes on each workout. Thus, CTT determines the "final score" as a simple rating score, which in CrossFit® is the sum of the "ranks" obtained (Nunnally, 1975).

Due to the large diversity and number of athletes that can sign up for "CrossFit Games", it is expected that when applying CTT, subgroups of athletes are placed on the same rank, thus, information regarding performance on the different workouts is lost. Hence, Item respond theory (IRT) has been employed to measure latent traits and characteristics of the measurement(Bock et al., 1997; Fernandes, Luz, Reis, Luz, & Guimarães, 2022; Fernandes, Luz, Reis, Luz, Guimarães, et al., 2022).

IRT application contributes to provide information regarding the performance of each athlete in different workouts, moreover, it becomes possible to obtain a scale for measuring and interpretating the scores in the CrossFit® setting (Bonifay, 2019; Henninger & Meiser, 2020a, 2020b).

In this scope, the first objective of this study is methodological: it is proposed the application of a probabilistic model of the Item respond theory (IRT), called Graded-Response Model proposed by (Bock & Zimowski, 1997) to describe the probability of an athlete executing a workout obtaining a certain output, given its physical conditioning, the latent trait being measured. Thus, we have *"CrossFit Games"* as the measurement tool to assess performance (output) of the athletes in different workouts (items) and determine their latent traits (physical conditioning).

The second objective of this work is practical: the application of the model encompasses the performance analysis, mechanisms to provide additional information that identify execution characteristics per workout and data regarding the quality of the measurement tool as a criterion for performance evaluation. In particular, when analyzing the athlete's performance in various workouts it is possible to distinguish or differentiate the physical conditioning of two of more athletes.

Materials e methods

Measuring Instrument

"The CrossFit Open" is a qualifying event that, since 2012, is composed by 5 workouts that are completed by the athletes and mobilizes thousands of athletes around the world to compete in the biggest participative CrossFit® event, "CrossFit Games" (CrossFit; G. CrossFit).

For this purpose, we can describe workout as a group or repetitive series of exercises that require some combination of strength, cardiopulmonary ability and/or gymnastic, to be performed within a specific time frame (Time Cap).

In most cases, there are two ways of determining the stop criterion: first, after a specified number of completed repetitions during a predetermined time frame or time cap, which is called truncated by repetitions, the score is given based on the execution time. In the second case, after a specified time, the athletes complete the maximum number of repetitions, this is called truncated by time, and the score in given by the number of repetitions, workouts that do not utilize the metrics of the number of repetitions and/or time may eventually appear, the most common amongst them being the one set by strength movement, in which the athlete is assessed (score) through the maximum load executed within a specified time frame. It is worth highlighting the cases in which there is a repetition sectioning, a case where the athlete could not complete the execution before the time cap the score is given by the number of completed repetitions.

Therefore, it is possible to consider that a workout has its complexity defined by the number of repetitions to be executed, the time frame defined for execution, the number of types of exercises, and the complexity of their execution, the complexity being able to differentiate the various workouts. Thus, we have the following specifications that specify its complexity:

- **Number of Repetitions:** referring to the total amount of repetitions to be completed (for repetition sectioning) or number of completed repetitions (for time sectioning), usually, it is divided in amount or repetitions per exercises or number of rounds.
- **Execution time or Time Cap:** referring to the time-limit available for the athlete to execute all or the maximum number of constant repetitions in a workout.
- **Number of types of exercises:** it is the number of different exercises proposed in a workout. The difference among the exercises can be determined by the increase in complexity and/or the increase of the imposed load.
- **Complexity of execution on each exercise:** referring to the categorization of the exercises as to type, exercises that include gymnastic elements, Olympic weightlifting, or aerobic conditioning.

In general, workouts are defined to consider the diversity of athletes and could be classified in two classes: types or division.

The first class aims on differentiating beginner athletes to the ones with larger experience in sports practice and are known as: Rx'd and Scaled. It is worth highlighting the assumption that athletes who have an extended time of practice tend to be able to execute more complex workouts or have better skills, whereas beginners need workouts with adapted movements and reduced load.

The second class is consisted of factors such as sex and age range and takes into consideration physical and biological characteristics. The workout is specified with varying complexity, according to these characteristics.

As shown in Figure 1 there is a distribution of the athletes in the various divisions and types for 2019 and in Figure 2 for 2020.

Data Set

For this study, data obtained at CrossFit Games (G. CrossFit) website referring the years 2019 and 2020 of "The CrossFit Open" were used. Data were subdivided, according to **Table 1**, by year, category and division.

Table 1*. Frequency Table subdivided by year, category and division.*

The Open 2019			**The Open 2020**		
Category	**Division**	**Count**	**Category**	**Division**	**Count**
Scaled	Men (18-34)	14.638	Scaled	Men (18-34)	13.865
	Men (35-39)	2.758		Men (35-39)	2.619
	Men (40-44)	2.340		Men (40-44)	2.261
	Men (45-49)	1.954		Men (45-49)	1.797
	Men (50-54)	1.409		Men (50-54)	1.289
	Men (55-59)	658		Men (55-59)	422
	Men (60+)	871		Men (60+)	557
	Women (18-34)	26.324		Women (18-34)	17.727
	Women (35-39)	4.732		Women (35-39)	3.274

	Women (40-44)	3.691		Women (40-44)	2.434
	Women (45-49)	2.869		Women (45-49)	1.796
	Women (50-54)	2.095		Women (50-54)	1.280
	Women (55-59)	891		Women (55-59)	659
	Women (60+)	884		Women (60+)	636
	Sub-Total Scaled	**66.114**		**Sub-Total Scaled**	**50.616**
Rx’d	Men (18-34)	195.512	Rx’d	Men (18-34)	133.874
	Men (35-39)	39.490		Men (35-39)	27.108
	Men (40-44)	26.044		Men (40-44)	18.664
	Men (45-49)	15.940		Men (45-49)	11.389
	Men (50-54)	8.067		Men (50-54)	6.174
	Men (55-59)	4.589		Men (55-59)	3.848
	Men (60+)	2.961		Men (60+)	2.452
	Women (18-34)	146.363		Women (18-34)	94.157
	Women (35-39)	27.202		Women (35-39)	17791
	Women (40-44)	17.761		Women (40-44)	11.942
	Women (45-49)	10.763		Women (45-49)	7.054
	Women (50-54)	5.877		Women (50-54)	3.968
	Women (55-59)	3.669		Women (55-59)	2.692
	Women (60+)	2.301		Women (60+)	1.806
	Sub-Total Rx’d	**506.539**		**Sub-Total Rx’d**	**342.919**
	Total	**572653**		**Total**	**393535**

Data showed the relevance relative to the variety of the participants that corelates age range and sex. A greater interest of the athletes is observed in competing in the same “Rx’d” type, the male sex in larger number compared to the female sex, regarding the 18 to 34 age range. In the “Rx’d” type, athletes with extended practice time are expected to have longer practice time.

Besides, the “Scaled” class is an indication of a smaller amount of practice and/or non-competitive objectives, in other words, beginner athletes may or may not seek the practice for health purposes and face competition only as personal challenge.

Item Response Theory

In the CrossFit® context, the outputs obtained in a set of workouts have been traditionally used as an assessment and selection process to find the most conditioned athlete. However, due to the complexity of each workout, those may benefit athletes with abilities on certain movements. In order to avoid this situation, and even as a premise of physical conditioning, the set of workouts need to have a wide variety of requirements.

Particularly in “The CrossFit Open”, the score obtained in a specific workout serve as criterion to rank the athletes, and after the 5 workouts, the general ranking is calculated through the sum of the rankings in each workout. Consequently, an athlete who is placed in a lower ranking position in each workout indicates a greater contribution in the final sum, this athlete with lower score being the most conditioned one.

This method of evaluation relies on the specific set of workouts that composes the competition; thus, analysis and interpretation are always associated to the competition as a whole, which is the main characteristic of the Classical Test Theory. Therefore, it is made unfeasible the comparison between people that were not subjected to the same competition, or at least to what are called parallel methods of evaluation (Andrade et al., 2000; Mangine, Tankersley, et al., 2020).

Contextualizing, the Item Response Theory – IRT refers to the set of probabilistic models that intend to represent the probability of an athlete obtaining a specific score in a workout as a function of the characteristic parameters of the workout and the athlete's physical conditioning. This relation is always expressed in a way that the better the physical conditioning higher is the probability of obtaining a greater score in a workout (Andrade et al., 2000).

From the concept of Performance, the main characteristics of a workout are its complexity and ability of distinguishing two or more athletes. As a result, from the IRT point of view, the two-parameters logistic model is an adequate model to this context (Chalmers, 2012; Hori et al., 2020a, 2020b).

In a context of expansion of the dichotomous model, the polytomous models can handle items with three or more sorted or unsorted classes. Particularly, (Bock & Zimowski, 1997)proposed the Graded-Response Model – GRM as an extension of the Two-parameter model. Thus, in the context of the CrossFit, it is intended to describe the probability of an athlete fitting certain group, based on his/her physical conditioning, hence, it is expected that a better conditioned athlete will have an increased probability of obtaining improved performances in a set of workouts, thereby obtaining better outputs. It means that the sectioning of the athletes may be done gradually and orderly. Athletes with better performance are assigned to the primary groups and, as the score decreases, they are assigned to the last groups.

However, an issue appears when establishing criteria to define the sectioning of the athletes, mostly due to the continuous property of the scores, regarding the time unit, or the discrete property, relating to the number of valid repetitions, which results in estimative precision biases. Particularly, it is possible to make an empiric comparison between the IRT models that encompasses characteristic parameters of the evaluator when presenting a notation that refers to the data from the performance evaluation as well as a discussion regarding the common characteristics amongst evaluators (Ueno & Okamoto, 2008; Uto & Ueno, 2016, 2018). In this project, our aim is to discuss the rater biases on types. Usual rater characteristics on which the accuracy depends are as follow:

- **Severity:** the tendency of ranking with lower positions that what is justifiable by the results.
- **Consistency:** the point to which the evaluator classifies similarly the results from similar quality.
- **Range restriction:** the tendency to overuse some classes from restricted sections.

In practice, the consistency bias is disregarded, once the evaluator does not attribute a result to the athlete, thus, it does not represent a bias.

It will be implied, for the purpose of this project, that only an evaluator is going to determine the age restriction, also being described as a specialist or professional in the field.

Thereunder, the sectioning of the athletes may be completed in one of two ways: based on the score obtained, called grouping by score or based in the raking of the athlete, called grouping by rank. Thus, given that the athletes performed a specific workout and that the number of groups and age restriction are pre-established, we can define it as:

- **Grouping by score** refers to the distribution of the athletes through their obtained score. This grouping has a discriminatory nature and aims to compare athletes in classes, consequently, results in an inference about athlete's common characteristics and predictor factors and, due to subjectivity, it is reasonable considering it as an intuitive process that must be done by specialists or professionals from this field.
- **Grouping by rank** refers to the distribution of athletes through their obtained classification. This grouping has a qualifying nature.

Whether by score or rank, the grouping criteria also relies on the type of truncated and the tiebreaker criterions. Besides, it is still necessary to introduce the premises of **growing grouping** per range and that represents the classification of the athlete according to his/her competitive objectives, presuming that better conditioned athletes will tend to perform a bigger number of repetitions and be placed in primary groups.

To exemplify the process of grouping by score, let's first look at **Figure 3**, in which the construction of the frequency histogram of work frequency is carried out as a function of the number of repetitions of the work 19.1. Note that this training is characterized by a truncation by time and it is reasonable to think that the grouping should be done according to the athlete's performance. Furthermore, it is possible to observe in Figure 3 a figure of normal curvature and in the Figure 4 inclusion of the assumption of increasing clustering.

The premises is that the grouping per specialist will tend to be rising, given that it is expected that as the person becomes more competitive, less people would be interested in dedicating time and effort, and consequently, fitting the primary groups.

Also, it is possible to group by score, for workouts truncated by repetitions, as is the case of Workout 20.1 in which it has the characteristic of being truncated by repetitions, that is, it means that after a Time Cap, the athlete interrupts the execution and the score is given by the number of repetitions, not by the shortest time. Thus, as a graphic example of Figure 5, after the time of 900 seconds, the time is adjusted as follows:

Regarding the section scores, workout have as a characteristic the repetition grouping, in other words, after a Time Cap the athlete interrupts the execution and the score is given based on the execution time, and not for the maximum number of repetitions. Therefore, time may be adjusted as follows:

$$time_{fit} = (reps_{max} - reps_{exec}) \times \left(\frac{time_{cap}}{reps_{exec}}\right) + time_{cap} \qquad (1)$$

in which, $reps_{max}$ represents the number of the repetitions to be done within the $time_{cap}$ and $reps_{exec}$ represents the number of repetitions the athlete was able to perform.

Thus, if the athlete has finished all the repetitions within the time cap, their time will be kept. Otherwise the time would be the Time Cap plus the average time of execution of the constant repetitions.

It is worth highlighting that in practice, in the case of time cap being a really big number, the athletes would tend to take longer in the completion of the remaining repetitions, given that they are spending more body energetic resources.

In this way, **Figure 5** demonstrates the grouping by score performed by worktout truncated by time, and thus, facilitates the grouping by the specialist, according to **Figure 6**, adapting to the Gradual Response Model, with ordered categories, in a single dimension and taking into account consider the assumption of increasing clustering.

Finally, to exemplify the grouping by rank, it is simply a matter of grouping according to the frequencies or number of athletes of interest. As can be seen in **Figure 7** the value described in each column of the figure needs to be determined according to interest. This grouping is important, especially when it is necessary to define the first places in a competition, as is the case of "CrossFit Games – The Open Stage".

Graded-Response Model

The Graded-Response Model by (Samejima, 1968, 1969) assumes that the classification of the response to an item may be sorted with each other. This model obtains more information from people's answers than simply if they have given yes or no answers (Andrade et al., 2000; Bonifay, 2019; Uto & Ueno, 2016).

In the CrossFit context, we assume the grouping (classes) by score, representing the output of the athlete's performance in a workout (item), may be sorted amongst each other, thus the Graded-Response Model may be applied. Furthermore, the GRM is useful and allows an estimative of the probability of an athlete obtaining a score in a workout given his/her physical conditioning. In other words, it means that the athletes can be classified gradually and in an orderly manner, with the best performing athletes in the primary groups and, as their score worsens, they are placed in the final groups.

For instance, assuming that the scores of the workout classes are arranged in order, from lowest to highest, and denoted by $k = 0,1, \dots, m_i$, where $(m_i + 1)$ it is the same number of classes of the *i-th* workout. The probability of and athlete j being placed in a certain group, or a higher one of the *i* workout is given by the extension of the Two-parameter logistic model

$$P_{i,k}^{+} = \frac{1}{1 + e^{a_i(\theta_j - b_{i,k})}} \tag{2}$$

with$i = 1,2, \dots, I$, $j = 1,2, \dots, n$ and $k = 0,1, \dots, m_i$, where $b_{i,k}$ is the parameter of difficulty of the *k-th* class of the *i* workout and θ_j represents the physical conditioning (latent trait) of the j-th athlete.

Regarding the models for dichotomous items, the slope parameters a_i is the item discrimination. However, regarding models for non-dichotomous items, the discrimination of a specific class depends on the slope parameter, common to all the item classes, as well as the distance from adjacent difficulty classes.

Thus, the probability of a person *j* receiving a score *k* in the *i* item is given by the expression:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{a_i(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{a_i(\theta_j - b_{i,k+1})}} \tag{3}$$

Notice that if we have a test with *i* items, each one with $(m_i + 1)$ output classes, then we shall have $[\sum_{i=1}^{I} m_i + I]$ parameters to be estimated.

Results

Preliminary Analysis of the Data Set

In Figure 4, we can see the grouping by score performed in Workout 19.1. Thus, as presented in [3], the description of this worktout is described in Figure 8:

We initiate our analysis focusing on the workout 19.1, because it is a time grouping, the score is given by the number of executed repetitions until $T_{cap} = 900\ seconds$. Therefore, it is necessary to establish new values to represent the output of the athletes, thus, if the athlete fits a specific group, for example, Group 1, it means that he or she obtained a score 5, Group 2 with a score of four and so on. Thus, characterizing and sorting the workout 19.1, the data is summarized in **Table 2**.

Table 2. *Characterization and sorting of workout 19.1.*

workout	Group	Score	Truncated	Frequency	Inferior Limit	Upper limit

19.1	01	5	by times	2224	342	418
19.1	02	4	by times	13006	304	342
19.1	03	3	by times	20194	281	304
19.1	04	2	by times	27549	258	281
19.1	05	1	by times	42950	228	258
19.1	06	0	by times	58140	0	228
19.1	n/a	0	n/a	19193	n/a	n/a

n/a = not assessed by the measurement tool.

The parameters to each of the workouts are estimated assuming that the distribution of θ_j follows a normal distribution with $\mu = 0\ e\ \sigma = 1$. Values for $a < 1$ indicate the item has little discrimination capacity. Values for $a \geq 1$ mean the item discriminates well. It is possible to observe on **Table 3** that all the workouts present values for $a \geq 1$.

Table 3. *Estimates and standard error (SE) of the workout parameters of Graded-Response Mode on scale (0,1).*

workout	$\hat{a}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$
19.1	2.26 (0.009)	0.25 (0.004)	0.51 (0.004)	1.11 (0.005)	1.80 (0.006)	3.05 (0.013)
19.2	4.60 (0.025)	0.46 (0.003)	1.23 (0.004)	2.23 (0.007)	3.13 (0.014)	3.76 (0.027)
19.3	3.43 (0.015)	0.05 (0.003)	0.61 (0.003)	1.30 (0.004)	2.35 (0.008)	3.66 (0.023)
19.4	4.41 (0.021)	0.26 (0.003)	0.88 (0.003)	1.44 (0.004)	2.04 (0.006)	2.64 (0.009)
19.5	4.48 (0.022)	0.31 (0.003)	0.73 (0.003)	1.15 (0.004)	1.70 (0.005)	2.76 (0.01)
20.1	3.32 (0.017)	-0.14 (0.004)	0.49 (0.004)	1.06 (0.005)	1.65 (0.007)	2.54 (0.011)
20.2	4.45 (0.024)	0.25 (0.004)	1.04 (0.005)	1.692 (0.006)	2.51 (0.010)	3.30 (0.020)
20.3	3.66 (0.024)	0.39 (0.004)	1.45 (0.006)	3.23 (0.019)	3.98 (0.039)	4.62 (0.089)
20.4	3.49 (0.019)	0.14 (0.004)	0.79 (0.005)	1.32 (0.006)	2.32 (0.010)	3.66 (0.025)
20.5	3.43 (0.018)	0.10 (0.004)	0.62 (0.005)	1.24 (0.006)	1.85 (0.007)	2.61 (0.011)

Furthermore, as shown in **Figure 9**. Workout 19.1 characteristic curve., it is observed that the peak of the curves referring to each group is greater than a value of 30% and this is positive evidence in relation to the information generated by the Workout under analysis.

Finally, we may still evaluate the Test Information Function (TIF) and the Standard Error of Measurement (SEM) presented on Figure 10. So, we can verify the degree of precision of the workout set to several scale ranges (0,1), and as can be seen, SEM presents lower values, better precision, in the interval [0,4].

General Analysis of the Data Set

Preliminary analysis is important to describe the process of analyzing and evaluating a set of workouts. However, it is important to understand all the information generated by the competition with their respective numbers of registered athletes. In this sense, **Table 1** presents a set of 2 years of competition, 2 types of categories and 14 types of division (gender and age group), thus making it necessary to analyze 56 subsets of data and/or scenarios.

The general analysis, then, consists of an analysis of the main indicators that refer to a good quality of the measurement instrument, namely: the analysis of the parameter a when providing information regarding the discrimination power of the workouts; the frequency of respondents, which in this context, refers to the number of athletes included in the groups; the analysis of the worktous characteristic curves, which leads to the idea that flat curves or low probability peaks generate little information and, finally; analysis of the FRT and EPM curves.

Initially, all estimates of parameters a were analyzed and presented in **Figure 11**. It is a multidimensional graph, varying the value of the estimation of parameter a (y axis), with the 14 divisions (gender and age group), the two categories and the two years under analysis represented by colors and the sizes of the points representing the number of athletes in each scenario. Finally, a dotted line was drawn informing the value of interest, typically being $a \geq 1$. Therefore, it is noted that in all scenarios the value of interest was reached.

In a second stage, Figure 12 evaluates the occurrence of athletes in each situation, in which a point in every grid represents a specific workout in analysis. Hence, we are interested in verifying, at first, the smallest points, and later in which analyses scenario it fits. In order to do that, frequencies greater than 200 were transformed in 200, in the intent of better presenting (visually) the situations with low frequency.

To guide the analysis process, for example, we can fixate vertically the section “Men (18-34)” and observe a lower frequency of athletes that fit Group 01, workout 03, “Rx’d” type in the year 2020.

A strategy to avert this situation is to unite groups 1 and 2 and estimate the interest parameters. In particular, the estimation of parameters after this regrouping did not present a significative difference.

Discussion

Typically, in world level sports the quality of a competition in determining the best athletes or teams is given by their acceptance in recognizing and validating this competition. Another way of recognizing a competition is based on the rules and norms that a bigger authority states and in turn, exerts validation of the competitions around the world. As an example, the soccer regulation by FIFA.

Regarding CrossFit, the validation of the competitions is exerted by the entity itself, which sanctions world events and function as qualifying stages for the final competition called *CrossFit Games*. However, due to the proportions reached by the sport and its growing pace, several competitions, classified as amateurs, aim at determining the fittest athletes, those being validated by their own competitors.

In this section, we present an analysis of the events and an assessment of the quality of the *"The CrossFit Open"* as a mechanism for measuring physical conditioning.

Due to the size of the data set, a preliminary analysis was done regarding the "Rx'd" type, section "Men (18-34)", 2019.

Conclusions

We focused our work in presenting a conceptual arrangement of the main definitions, terms and expressions applied in the CrossFit context and that were pointed towards Samejima's Graded-Response Model of the Item Response Theory (Samejima, 1969). In this sense, we were able to accomplish the primary objective of the production by applying GRM to the respective context, and as a result, describing the probability of an athlete performing a Workout and obtaining a score, given his/her physical conditioning.

On the other hand, given the adjustment of the IRT to the context presented, it was possible to achieve the second objective and incorporate to the athletes' performance analyses useful data that seek to identify performance characteristics and positively evaluate the quality of the CrossFit Open.

Commonly, the evaluation process of a measurement tool requires an interactive process of regrouping the results, that in a sense, intend to improve the quality and validation of the measurement instrument. In the Item Response Theory this process means that the classes of responses are to assessing well enough the analyzed item, and it is necessary a grouping, in other words, in the context of the CrossFit and this work, it means that the grouping of the athletes, in face of their results, cannot contain all the data that the Model would be able to collect, and with the regrouping, it could offer more data about the measurement tool.

Considering this and the 56 studied scenarios, individually evaluation each presented situation could be a costly work. Besides the regrouping analysis, it is still necessary to answer the following questions: Did the score grouping, designed by a specialist, allocated an adequate number of athletes or did it in the best way? Is the number of specified groups sufficiently adequate? Given that the grouping elaborated by the specialist and done per Workout is based on a single scenario, could the remaining scenarios be better regrouped?

Those are the complex answers that, besides considering the premises and characteristics of the referred context, require a greater discussion on how the optimization process must be completed. However, this work did not focus on this optimization process.

In addition, the Item Response Theory provides us other tools that make possible to infer and measure, qualitatively, predictor factors of performance, for example, a subject of great relevance in the field of Sports Science and Exercise Physiology.

Studies that merge the Item Response Theory and the Sports Science and Exercise Physiology context in the presented manner were not found in literature. There are, on the other hand, studies that merge Sports Science and Psychology and encompasses psychometric assessments, among those, ones that employ the Item Response Theory.

In conclusion, this work accomplished both objectives proposed, methodological as well as practical, and recognizes the limitations derived from the reduced amount of qualitative data on the topic and the little use of applied probability models.

Future research is needed to build a standardized performance measurement scale that allows for a contextual and practical interpretation of the performance metrics obtained.

Acknowledgments: This study was fulfilled employing data collected form CrossFit® official website, which describes its privacy policy that the publicly available information may contain personal information. However, this work is restricted to the generic use of the data, not using or disclosing athlete’s personal data. The authors thank CrossFit® for making the data available.

Authors’ contributions: All the authors contributed equally to this assignment.

Ethics Approval: Not applicable.

Competing Interests: The authors declare that they have no competing interests.

References

Andrade, D. F. d., Tavares, H. R., & Valle, R. d. C. (2000). *Teoria da Resposta ao Item: Conceitos e Aplicações*. Associação Brasileira de Estatística.

Bock, R. D., Thissen, D., Steinberg, L., Andersen, E. B., Samejima, F., Masters, G. N., Wright, B. D., Verhelst, N., Glas, C. A. W., de Vries, H. H., Tutz, G., & Muraki, E. (1997). *In Handbook of Modern Item Response Theory*. Springer-Verlag.

Bock, R. D., & Zimowski, M. F. (1997). Multiple Group IRT. In (pp. 433-448). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_25

Bonifay, W. (2019). *Multidimensional item response theory* (Vol. 183). SAGE Publications.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*(6), 1-29. https://doi.org/https://doi.org/10.18637/jss.v048.i06

Claudino, J. G., Gabbett, T. J., Bourgeois, F., de Sá Souza, H., Miranda, R. C., Mezêncio, B., Soncin, R., Cardoso Filho, C. A., Bottaro, M., & Hernandez, A. J. (2018). CrossFit overview: systematic review and meta-analysis. *Sports medicine-open*, *4*(1), 1-14. https://doi.org/https://doi.org/10.1186/s40798-018-0124-5

CrossFit.). *What is CrossFit*. https://www.crossfit.com

CrossFit, G.). *Games CrossFit*. CrossFit. https://games.crossfit.com/

Fernandes, R. d. S., Luz, R. M. d. N., Reis, D. C. d., Luz, M. A. L. d., & Guimarães, G. V. (2022). Elaboration of quality perception instrument of remote teaching amidst COVID-19 pandemics in a University of Northern Brazil. *Research Square*. https://doi.org/https://doi.org/10.21203/rs.3.rs-1308160/v1

Fernandes, R. d. S., Luz, R. M. d. N., Reis, D. C. d., Luz, M. A. L. d., Guimarães, G. V., Bornia, A. C., & Andrade, D. F. d. (2022). Construction and validation of the remote teaching quality perception scale in the COVID-19 pandemic: an exploratory factor analysis and item response theory approach. *Research Square*. https://doi.org/https://doi.org/10.21203/rs.3.rs-1269691/v1

García-Pinillos, F., Molina-Molina, A., Párraga-Montilla, J. A., & Latorre-Román, P. A. (2019). Kinematic alterations after two high-intensity intermittent training protocols in endurance runners. *Journal of sport and health science*, *8*(5), 442-449. https://doi.org/https://doi.org/10.1016/j.jshs.2016.11.003

Glassman, G. (2004). What is crossfit. *The CrossFit Journal*, *56*, 1-7.

Gómez-Landero, L. A., & Frías-Menacho, J. M. (2020). Analysis of Morphofunctional Variables Associated with Performance in Crossfit® Competitors. *Journal of Human Kinetics*, *73*(1), 83-91. https://doi.org/https://doi.org/10.2478/hukin-2019-0134

Hanin, Y., & Hanina, M. (2009). Optimization of Performance in Top-Level Athletes: An Action-Focused Coping Approach. *International Journal of Sports Science & Coaching*, *4*(1), 47-91. https://doi.org/10.1260/1747-9541.4.1.47

Hargreaves, M., & Spriet, L. L. (2020). Skeletal muscle energy metabolism during exercise. *Nature Metabolism*, *2*(9), 817-828. https://doi.org/https://doi.org/10.1038/s42255-020-0251-4

Henninger, M., & Meiser, T. (2020a). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological methods*, *25*(5), 560-576. https://doi.org/https://doi.org/10.1037/met0000249

Henninger, M., & Meiser, T. (2020b). Different approaches to modeling response styles in divide-by-total item response theory models (part 2): Applications and novel extensions. *Psychological methods*, *25*(5), 577-595. https://doi.org/https://doi.org/10.1037/met0000268

Hori, K., Fukuhara, H., & Yamada, T. (2020a). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1531. https://doi.org/https://doi.org/https://doi.org/10.1002/wics.1531

Hori, K., Fukuhara, H., & Yamada, T. (2020b). Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1543. https://doi.org/https://doi.org/10.1002/wics.1543

Khassetarash, A., Vernillo, G., Krüger, R. L., Edwards, W. B., & Millet, G. Y. (2021). Neuromuscular, biomechanical, and energetic adjustments following repeated bouts of downhill running.

Journal of sport and health science. https://doi.org/https://doi.org/10.1016/j.jshs.2021.06.001

Mangine, G. T., Stratton, M. T., Almeda, C. G., Roberts, M. D., Esmat, T. A., VanDusseldorp, T. A., & Feito, Y. (2020). Physiological differences between advanced CrossFit athletes, recreational CrossFit participants, and physically-active adults. *PLoS One*, *15*(4), e0223548. https://doi.org/https://doi.org/10.1371/journal.pone.0223548

Mangine, G. T., Tankersley, J. E., McDougle, J. M., Velazquez, N., Roberts, M. D., Esmat, T. A., VanDusseldorp, T. A., & Feito, Y. (2020). Predictors of CrossFit Open Performance. *Sports*, *8*(7), 102. https://doi.org/https://doi.org/10.3390/sports8070102

Nunnally, J. C. (1975). Psychometric Theory— 25 Years Ago and Now. *Educational Researcher*, *4*(10), 7-21. https://doi.org/https://doi.org/10.3102%2F0013189X004010007

Potvin, J. R., & Fuglevand, A. J. (2017). A motor unit-based model of muscle fatigue. *PLoS computational biology*, *13*(6), e1005581. https://doi.org/https://doi.org/10.1371/journal.pcbi.1005581

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores1. *ETS Research Bulletin Series*, *1968*(1), i-169. https://doi.org/https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, *34*(4, Pt. 2), 100-100.

Schlegel, P. (2020). CrossFit® Training Strategies from the Perspective of Concurrent Training: A Systematic Review. *Journal of Sports Science & Medicine*, *19*(4), 670-680. https://pubmed.ncbi.nlm.nih.gov/33239940
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7675627/

Silva-Grigoletto, M. E. d., Valverde-Esteve, T., Brito, C. J., & García-Manso, J. M. (2013). Ability to repeat strength: effects of recovery between repetitions. *Revista Brasileira de Educação Física e Esporte*, *27*(4), 689-705. https://doi.org/https://doi.org/10.1590/S1807-55092013005000016

Taylor, J. L., Amann, M., Duchateau, J., Meeusen, R., & Rice, C. L. (2016). Neural Contributions to Muscle Fatigue: From the Brain to the Muscle and Back Again. *Medicine and science in sports and exercise*, *48*(11), 2294-2306. https://doi.org/10.1249/MSS.0000000000000923

Ueno, M., & Okamoto, T. (2008, 1-5 July 2008). Item Response Theory for Peer Assessment. 2008 Eighth IEEE International Conference on Advanced Learning Technologies,

Uto, M., & Ueno, M. (2016). Item Response Theory for Peer Assessment. *IEEE transactions on learning technologies*, *9*(2), 157-170. https://doi.org/10.1109/TLT.2015.2476806

Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon*, *4*(5), e00622. https://doi.org/https://doi.org/10.1016/j.heliyon.2018.e00622

Wan, J.-j., Qin, Z., Wang, P.-y., Sun, Y., & Liu, X. (2017). Muscle fatigue: general understanding and treatment. *Experimental & molecular medicine*, *49*(10), e384-e384. https://doi.org/10.1038/emm.2017.194

Figures

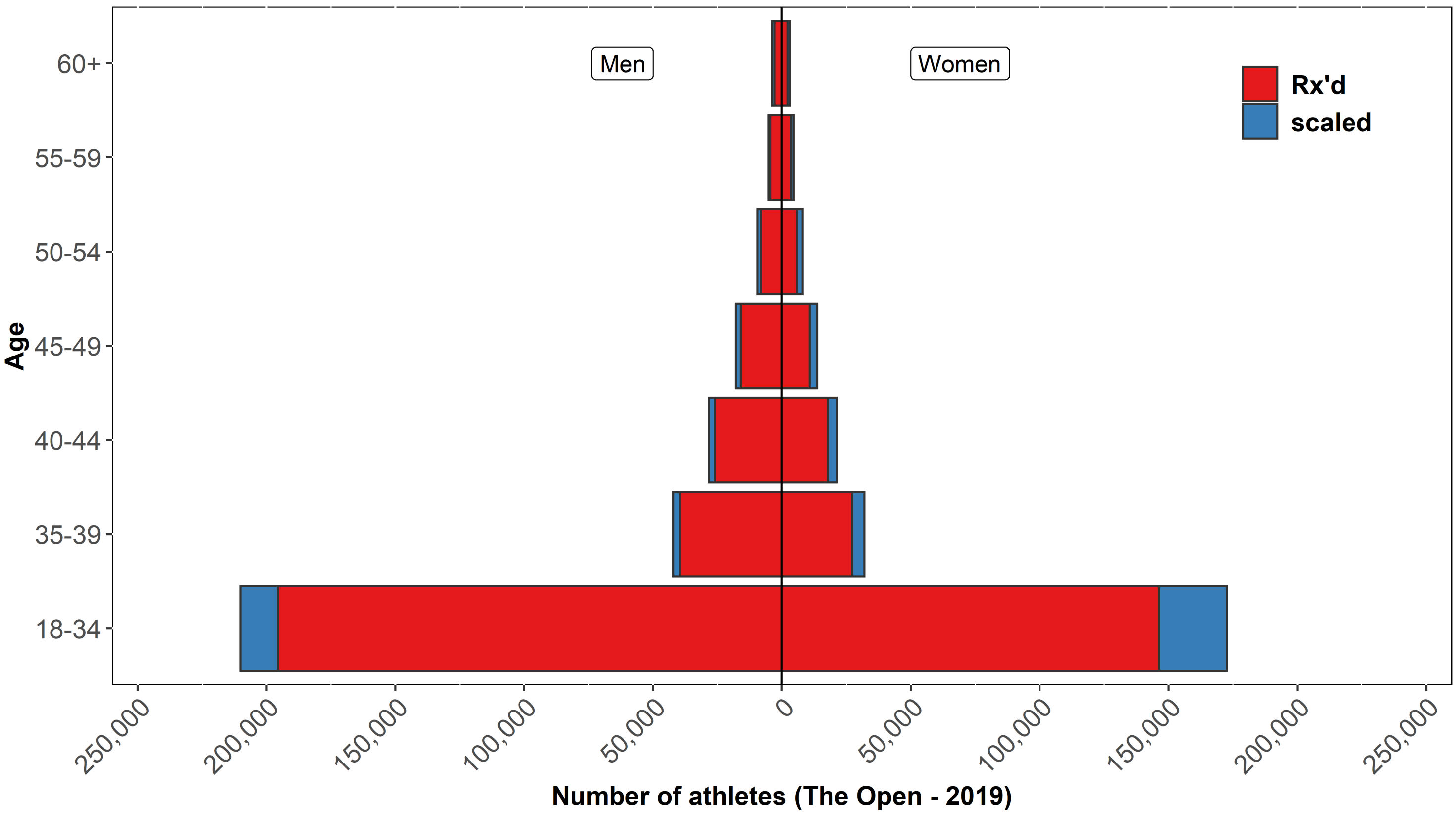


Figure 1

Number of athletes per division (bar stack) – The Open 2019.

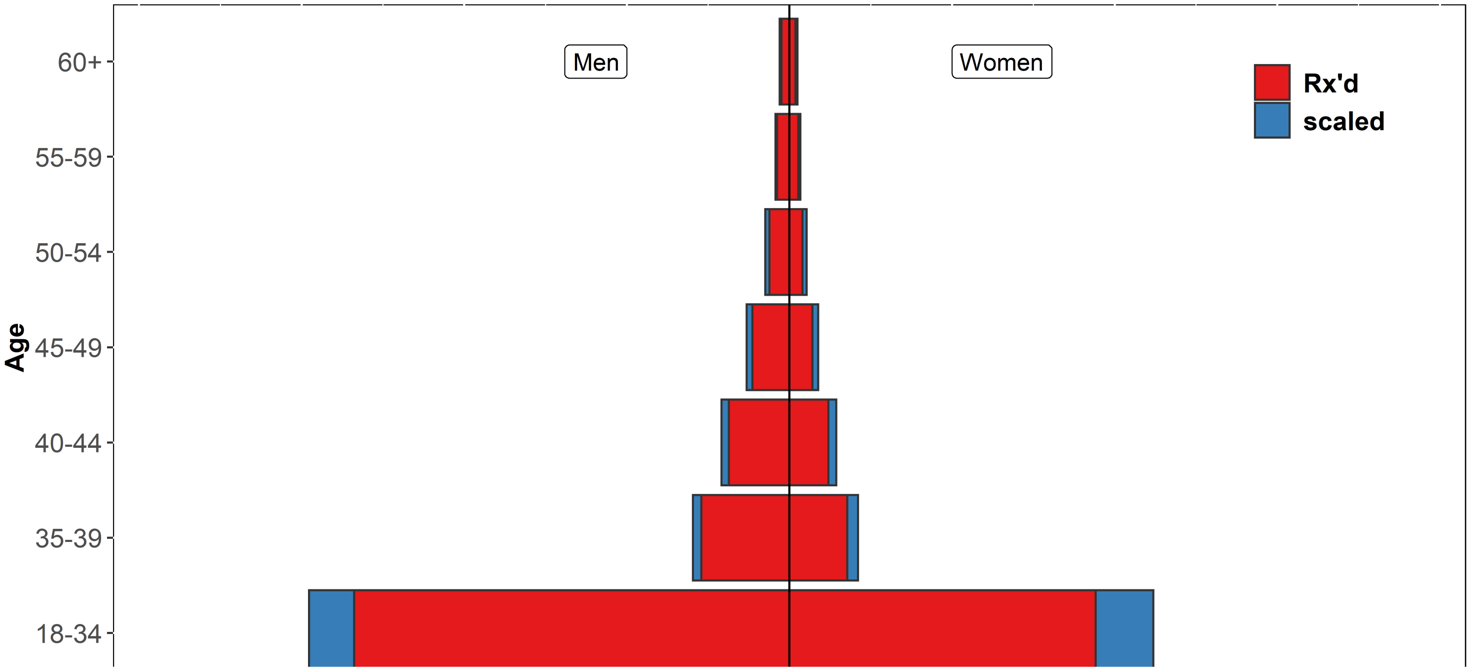


Figure 2

Number of athletes per division (bar stack) – The Open 2020.

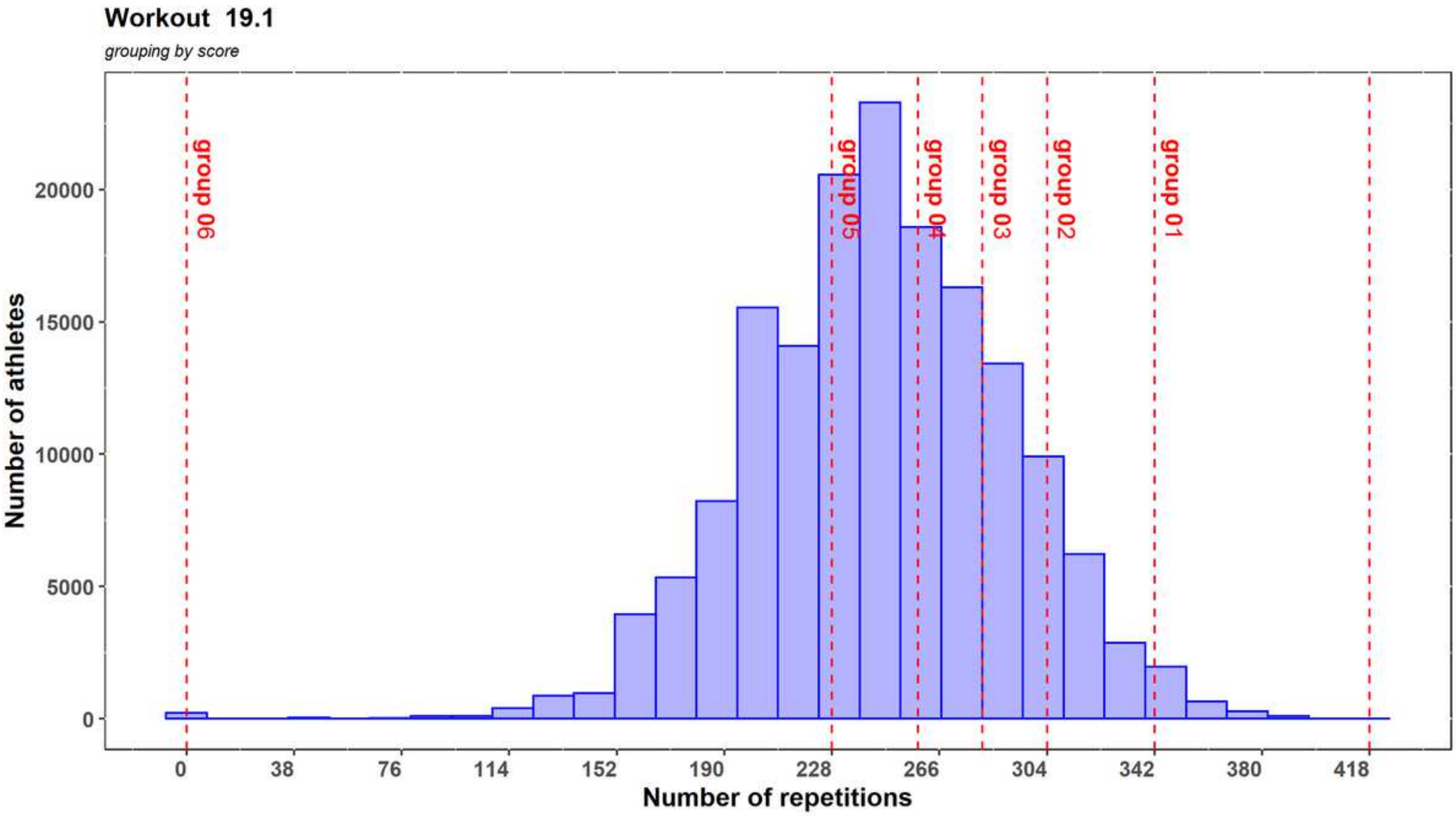


Figure 3

Histogram as a function of the number of repetitions performed. Workout 19.1.

Figure 4

Grouping by Workout 19.1 score, truncated by time and performed by a specialist.

Figure 5

Grouping by Workout 20.1 score, truncated by repetitions and performed by a specialist.

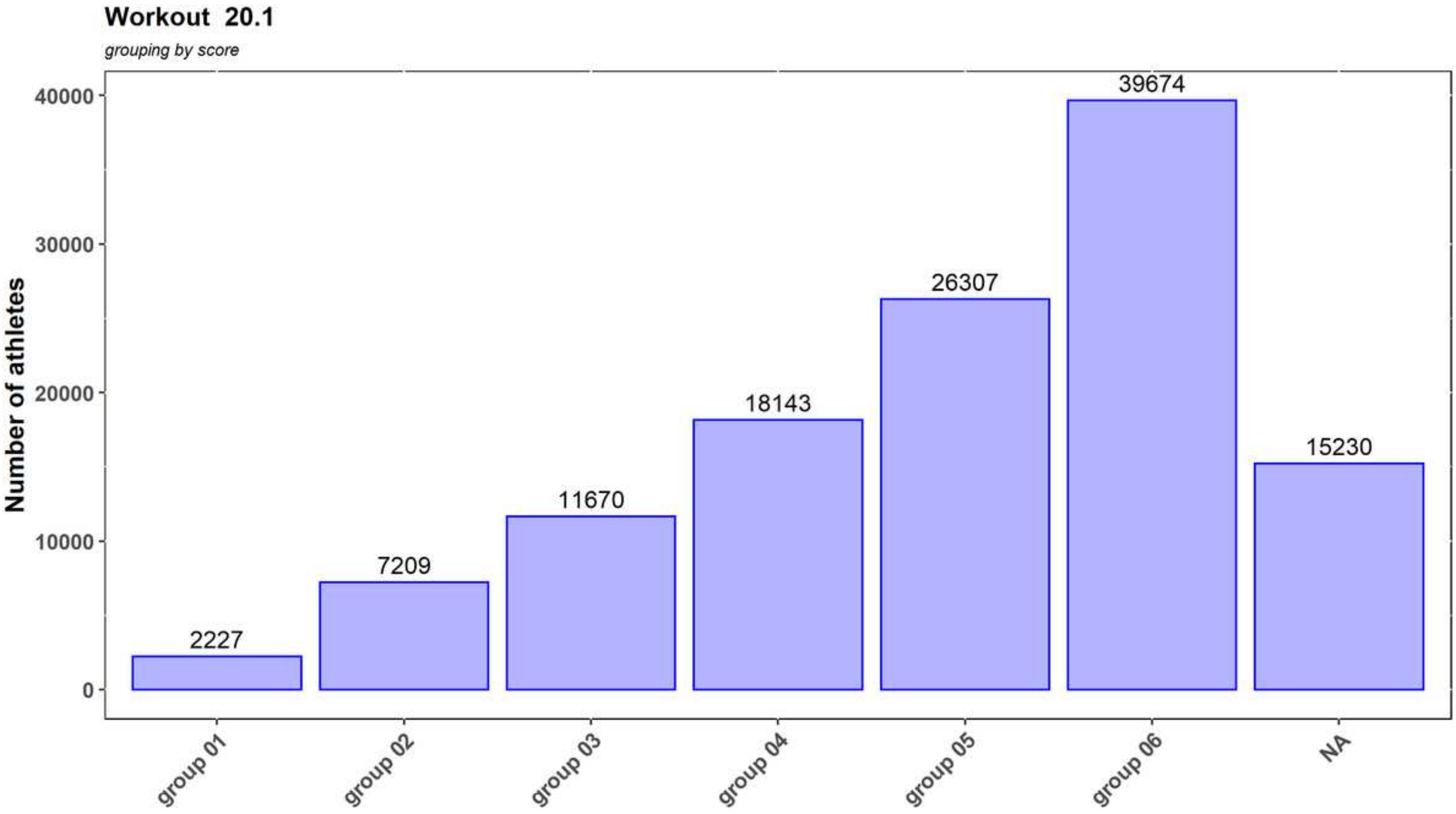


Figure 6

Grouping by Workout 20.1 score, truncated by time and performed by a specialist.

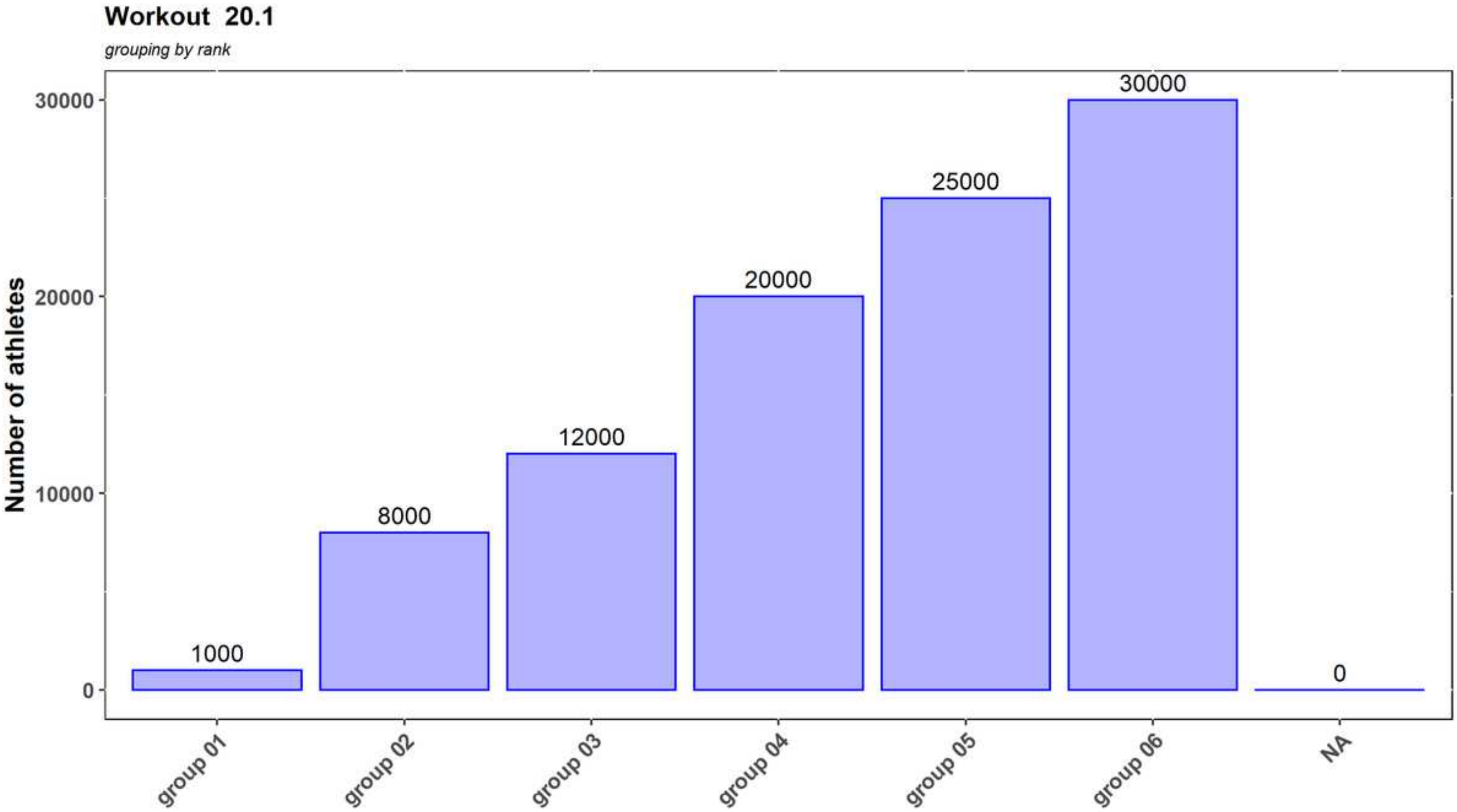
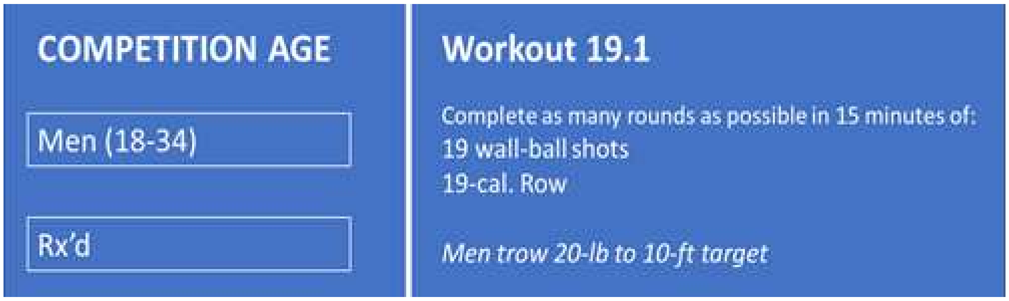


Figure 7

Grouping by Workout 20.1 rank, truncated by time and based on frequency.



COMPETITION AGE

Men (18-34)

Rx’d

Workout 19.1

Complete as many rounds as possible in 15 minutes of:
19 wall-ball shots
19-cal. Row

Men trow 20-lb to 10-ft target

Figure 8

Workout Description 19.1. Image taken from the site (G. CrossFit).

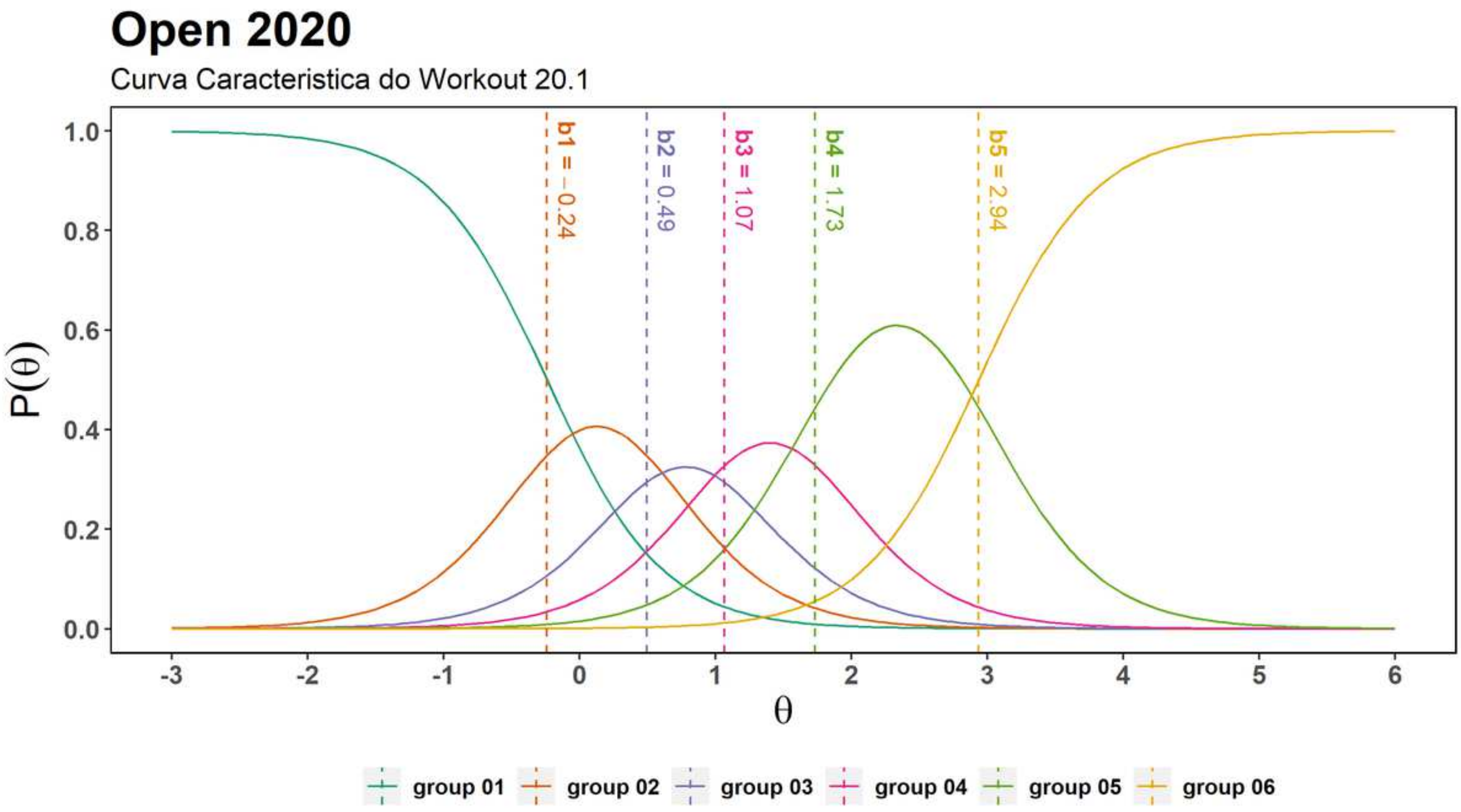


Figure 9

Workout 19.1 characteristic curve.

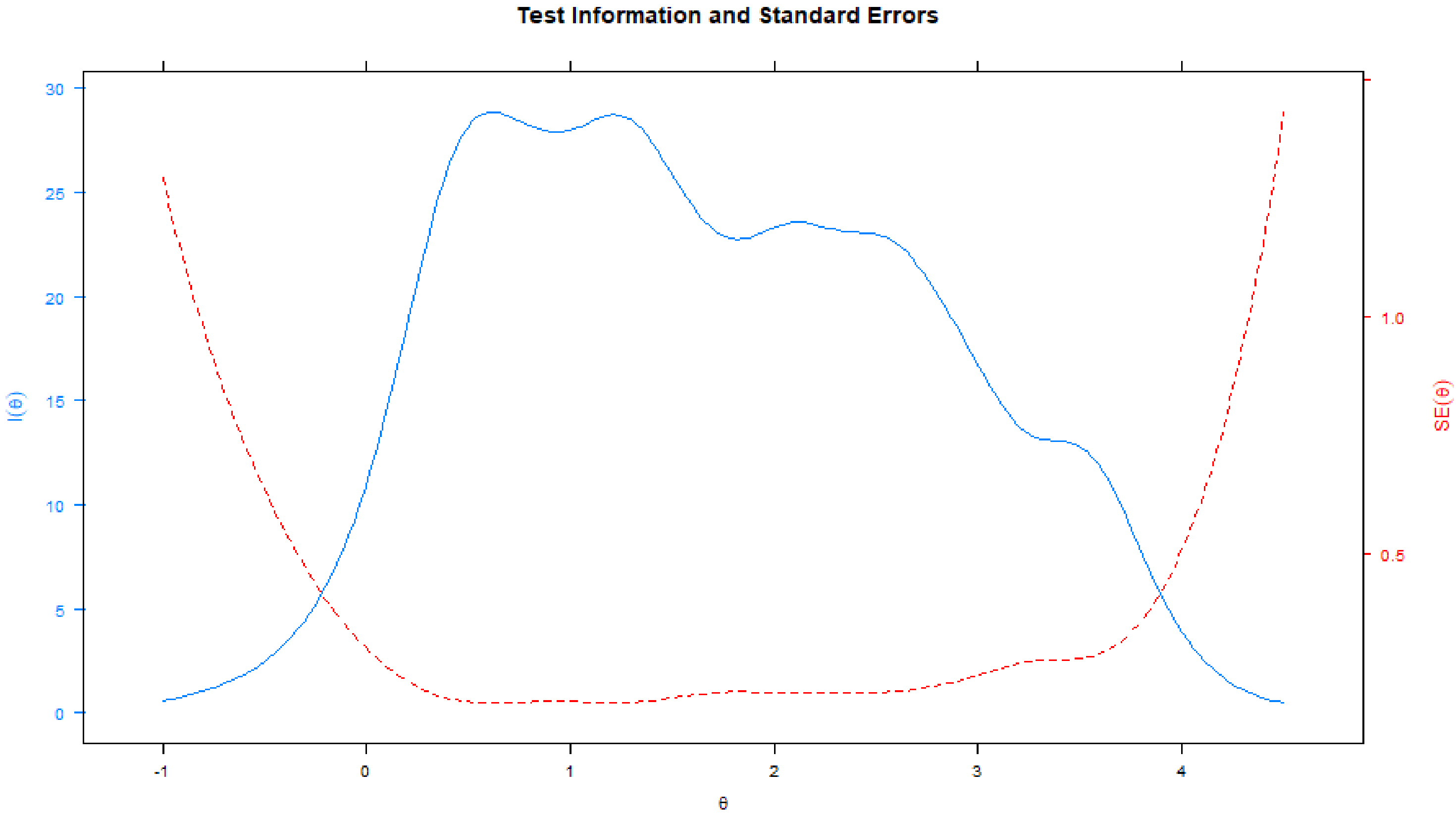


Figure 10

Test Information Function (TIF – continuous blue line) and Standard Error of Measurement (SEM – dotted red line).

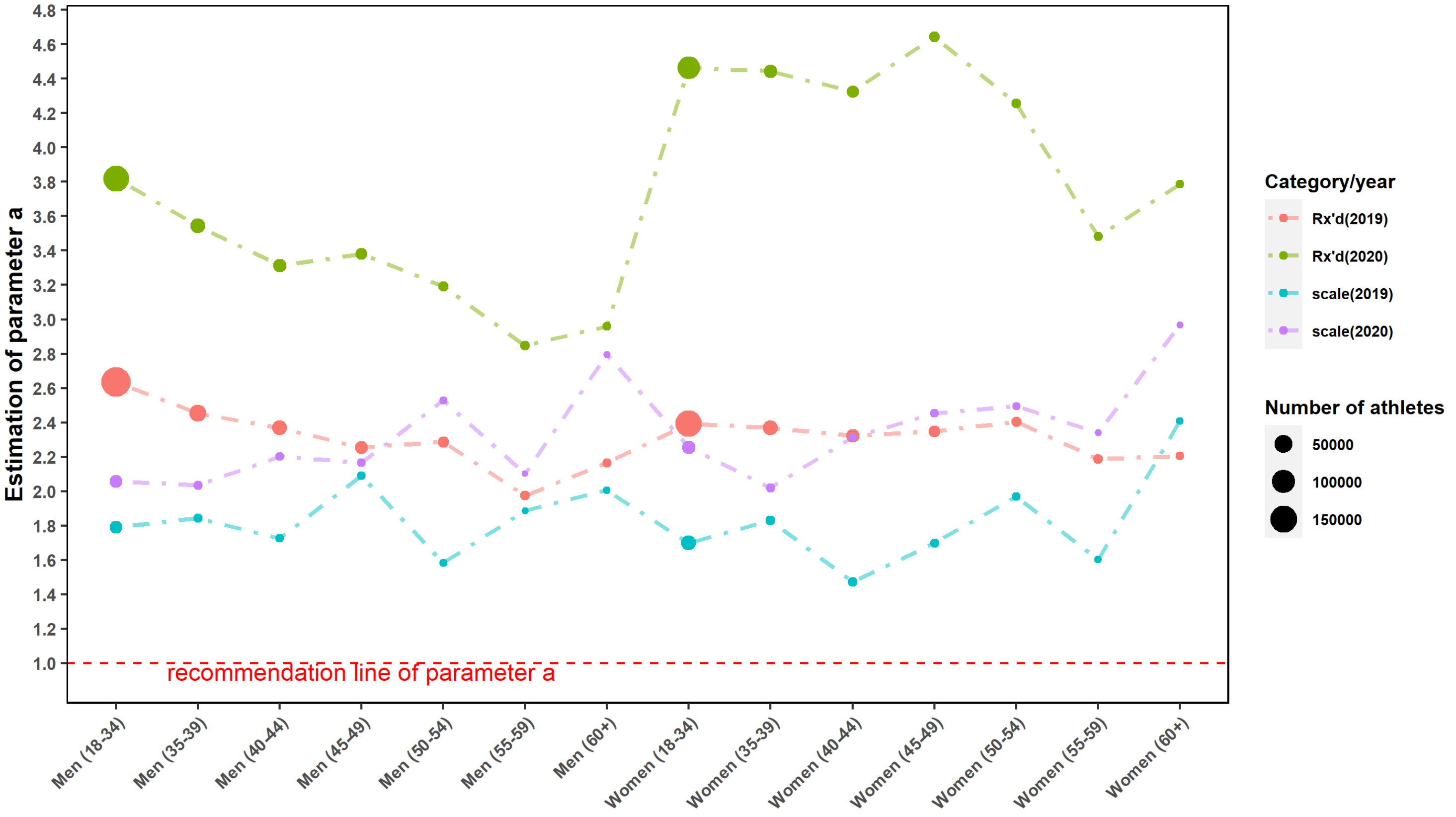


Figure 11

Estimates of the discrimination parameters.

Figure 12

Multidimensional analysis of the scenarios, groups, worktous and frequency of athletas in each situation. “NA” representes the group of athletes that did not compute their scores.