Baltic Journal of Health and Physical Activity 2018; 10(4): 34-48
Journal of Gdansk University of Physical Education and Sport
e-ISSN 2080-9999

ORIGINAL

doi: 10.29359/BJHPA.10.4.03

Development and standardization of a rating scale designed for floorball skills diagnostics of young school-age children

Authors' Contribution:
A Study Design
B Data Collection
C Statistical Analysis
D Data Interpretation
E Manuscript Preparation
F Literature Search
G Funds Collection

Zuzanna Dragounova

Faculty of Physical Education and Sport, Charles University in Prague, the Chech Republic

abstract

Background: The purpose of the study is to develop a standardized diagnostic tool designed to predict the level of the tested floorball skills in young school-age children that is necessary for future game performance.

Material and methods: For the construction of the Guttman-type scale, the Rasch model was applied. The methodology employed the procedures for standardization by Stochl & Musalek, fit functions to determine the fit of the data model, KR-20 coefficient for the reliability calculation, Fleiss' kappa coefficient to determine the inter-rater agreement, and PCA of residuals to determine the unidimensionality.

Results: Only 9 items out of a total of 30 were selected and retained in the developed rating scale. However, the items covered the continuity of the diagnosed feature very well, and the standardization procedure has been successful – the Rasch model fit the data, three criteria of unidimensionality were met, the reliability value of the rating scale was 0.81 and the inter-rater agreement reached 98.5%.

Conclusions: The developed rating scale includes 9 items suited to assess ball handling, ball controlling and passing techniques. Unfortunately, items containing shooting were not selected; they were too difficult and misfit the Rasch model.

Key words: floorball, Guttman scale, Rasch model.

article details

Article statistics: **Word count:** 5,709; **Tables:** 11; **Figures:** 2; **References:** 42
Received: May 2018; **Accepted:** October 2018; **Published:** December 2018

Full-text PDF: http://www.balticsportscience.com

Indexation: Celdes, Clarivate Analytics Emerging Sources Citation Index (ESCI), CNKI Scholar (China National Knowledge Infrastructure), CNPIEC, De Gruyter - IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences), De Gruyter - IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences), DOAJ, EBSCO - Central & Eastern European Academic Source, EBSCO - SPORTDiscus, EBSCO Discovery Service, Google Scholar, Index Copernicus, J-Gate, Naviga (Softweco, Primo Central (ExLibris), ProQuest - Family Health, ProQuest - Health & Medical Complete, ProQuest - Illustrata: Health Sciences, ProQuest - Nursing & Allied Health Source, Summon (Serials Solutions/ProQuest, TDOne (TDNet), Ulrich's Periodicals Directory/ulrichsweb, WorldCat (OCLC)

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interests: Author has declared that no competing interest exists.

Corresponding author: Corresponding author: Zuzanna Dragounova, Universita Karlova Fakulta telesne vychovy a sportu Ringgold standard institution, José Martího 269/31 Prague 6 , Praha 162 52; Czech Republic; phone no.: +420 220 172 003; e-mail: dragounova@ftvs.cuni.cz.

INTRODUCTION

Many sports specialists try to use different diagnostic tools to control the training process, but in the vast majority, these are just experience-based approaches. The same situation is currently in floorball. In floorball, due to its short history, there are not fully standardized diagnostic tools (tests, scales) for children that would allow the coaches to objectively assess the level of acquired floorball skills.

Young school age can be divided into two periods: childhood and prepubescence. The first period is characterized by a lower level of the quality of movement; child's movements lack economy, and every action is done with a number of additional moves. Later, the prepubescent period can be described as a stage with a good quality of movement and, therefore, as a favorable period for motor development and new skills acquirement [1]. We chose to create a floorball skill diagnostic tool for a young school-age category because of the unequal level of motor development that requires different demands on the content and difficulty of tests and assessment scales.

Floorball players' aim is to score more goals than the opponent, that is, to control the ball at a level to be conveyed to the opponent's goal. Manipulating the ball, in spite of the opponent's defense, and placing the ball into the goalkeeper's goal is almost impossible without the corresponding technical skills [2–5]. Technique means an effective way of dealing with a movement action that is in line with an individual's abilities, biomechanical patterns of movement, and is based on neurophysiological mechanisms of motion control [6]. During floorball game, as players are confronted with a changing environment, cognitive and perceptual skills are important determinants of technical skill performance and playing ability. There are a number of ways, from simple to complex tests (mental concentration, perceptual, anticipation and psychomotor tests) in which these aspects of team game are monitored [7–11]. There are also several tests and scales intended to assess the technique of different sports specialties described in the literature [12–19], using various diagnostic tools including scaling techniques.

The aim of our work is to design a standardized diagnostic tool for the young school-age category that will testify to the level of acquired floorball skills (controlling the ball with the floorball stick) that is necessary for future game performance with the Guttman-type scale designed through Rasch's analysis [20–23].

MATERIAL AND METHODS

Participants

The research sample was composed of 212 floorball players (197 boys and 15 girls) from the Central Bohemia Region, divided into three different age categories (Table 1). The pilot sample included 25 players of younger school age, and 29 players of the research sample were evaluated by four raters to establish inter-rater agreement. The raters were experienced youth floorball trainers from the Central Bohemia Region with 15, 9, 6 and 3 years of training experience. The trainers with varying lengths of training experience were selected due to the possible future use of the scale by both beginner and experienced trainers. The research sample included players with unequal

levels of acquired floorball skills (Table 2). In order to create a motor scale, it is desirable that the sample is not homogeneous [23]. Participants were players from competitive and non-competitive teams (non-competitive participants were floorball players who did not participate in league matches).

Table 1. Vitamin D supplementation based on the concentration of 25(OH)D in serum

Age	6–8 years	9–10 years	11–12 years	Total
Research sample	42	141	29	212
Pilot sample	5	12	8	25

Table 2. The research sample – level groups

Level	Non-competitive team	Non-competitive team	Non-competitive team	Competitive team	Competitive team	Competitive team
Years of training	1 or less	1–2	3–4	1–2	3–4	5 or more
Practice sessions per week	1x or 2x	1x or 2x	1x or 2x	3x	3x	3x or 4 x
Research sample	36	48	39	38	36	15
Pilot sample	0	3	5	5	8	4

Guttman scale and Rasch analysis

The Guttman scale is a set of items arranged from the easiest to the most difficult item. The tested person should complete the block of items from the beginning of the scale to the critical point, which indicates the maximum possible level of the personal latent feature [23]. This is a "cumulative" scale; it means that the critical point item guarantees successful completion of all previous items. The skill level is evaluated dichotomically as 1–0 or correct–incorrect.

The basic assumption for the Guttman scale is unidimensionality, which means that all of the scale items diagnose the same latent feature. "The unidimensionality of items is a limiting factor for the design of the Guttman-type scale" [23]. Latent features, or latent variables, are variables that are not directly observed and can be attributed to a general characteristic such as a physical ability or a movement skill [24]. For latent variable modelling used in the analysis of test results of a binary and generally categorial type, the item response theory is used [25].

The Guttman scale is a theoretical and mathematical ideal, and although it is an ordinal scale, it carries no information about intervals between items or about intervals between persons. Measurement variability translates to errors from confounding variables in the Guttman scale. IRT models exploit these errors as a means to estimate interval scales from ordinal scores assigned to observations [26]. One of the basic models of the item response theory is the Rasch model and it is offered as a suitable tool for constructing a perfect scale [20, 22].

Using the Rasch model, we will try to explain a relationship between the theoretical property represented by a latent variable and the empirical property represented by a manifest variable. The latent variable is the latent feature of the floorball skill of controlling the ball with the floorball stick. The manifest variable is the answer to the dichotomous item (1 corresponds to a correct motor task, 0 means an incorrect motor task).

PROCEDURE FOR DESIGNING AND STANDARIZING A MOTOR SCALE

We used findings from the design of the perfect scale for motor skills diagnostics developed by Cepicka [22, 23] and recommended procedures for the standardization of motor tests by Stochl & Musalek [27]. To design a motor scale, we followed steps shown in Figure 1.

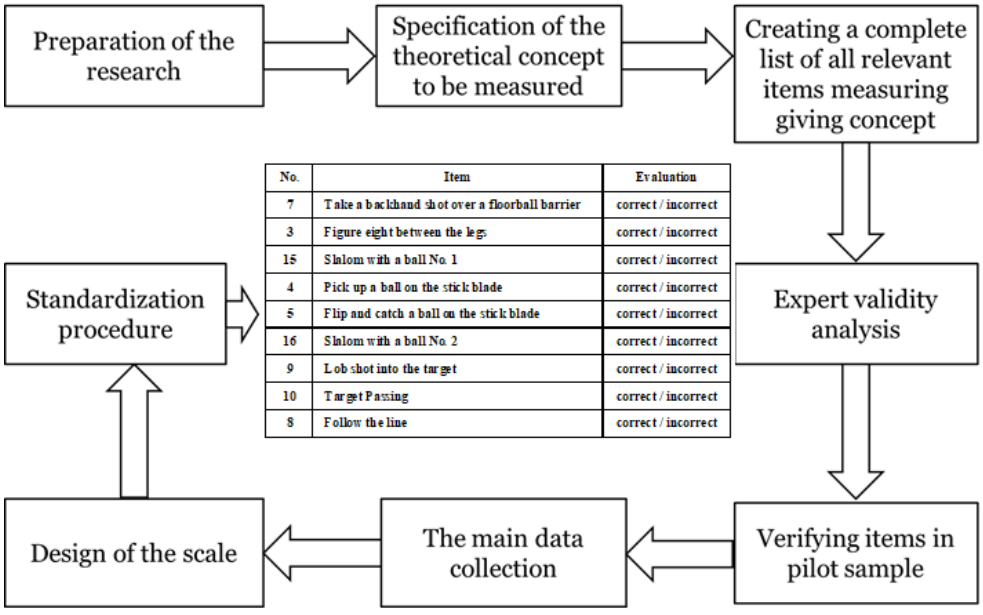

Figure 1. The design of the scale for motor skills diagnostics

EXPERT VALIDITY ANALYSIS

Expert analysis was applied as a technique to study the content validity of each item. The content validity index (CVR) according to Lawshe [28] was used to select appropriate items measuring the given concept. The coefficient CVR can range from 1, when none of the experts has indicated that the item complies with the theoretical concept to be measured, to +1, when all the experts accept the item corresponding to the measured concept.

DESIGN OF THE SCALE

The scale was designed in three steps. First of all, we analyzed the fit statistics for each item, then we created a graphical form of the scale according to the difficulty parameter of each item and, finally, we evaluated the scale values:

1. Removing misfit items

The suitability of items is assessed on the basis of their unidimensionality and it must be evaluated relative to the model. Infit and outfit statistics for each item were used, and values between 0.5 and 1.5 were considered acceptable [29]. Also the difficulty parameters were evaluated in this phase. Item difficulty parameter values should range from 3 logits to +3 logits [23]. Items with the difficulty parameter < -3 are too simple for a sample for which the scale is determined (young school age children) [23]. Most players are evaluated correctly on this item and it does not distinguish between a low and a high level of the latent feature. Items with the difficulty parameter > +3 are too difficult for this sample. Items outside this range and misfit items were removed.

2. Graphical form of the scale and the distribution of scale values
Information about the difficulty parameters of the items is obtained from the raw scores by Rasch analysis. We sorted items by difficulty, from the easiest items to the most difficult ones, and we have created a graphical form of the scale to see the distribution of the scale values. Quantification of the latent feature takes place on the same scale; the latent feature value corresponds to the difficulty parameter of the item.

3. Evaluating the scale distribution
The distribution of the scale values needs to be examined in terms of distance, which should not be too large due to the loss of discrimination. Items should also cover the continuity of the diagnosed feature within a sufficient range. The value of the difficulty of an item should be in the range of -3 logits to +3 logits with a probability of 95% [23]. If the distribution of the scale values does not meet the above requirements, the scale should be supplemented by the missing items.

Standarization procedure

1. Validity and unidimensionality of the scale
In considering validity and unidimensionality of the scale, the fit diagnosis of Rasch analysis was used. We had to evaluate infit and outfit MNSQ and infit and outfit ZSTD values to consider if the data fit the Rasch model well.

MNSQ (Mean–square) value is the chi-square statistic divided by its degrees of freedom, and its expected value is close to 1.0. Values greater than 1.0 indicate unmodeled noise and degrade measurement; values less than 1.0 indicate that the model predicts the data too well. It is less productive for measurement, but not degrading. ZSTD (Z-Standardized) value reports probability of MNSQ statistics occurring by chance when the data fit the Rasch model. They are also called t-statistics reported with infinite degrees of freedom and 0.0 are their expected values. Values less than 0.0 indicate too predictable measurement, and values more than 0.0 lack predictability in measurement. [22, 23, 29, 30]

There are two indicators of misfit: infit means sensitive to unexpected responses to items near a person's ability level, and outfit is more sensitive to unexpected observations by persons on items that are relatively very easy or very difficult for them [29, 30].

The general principles of fit diagnosis, according to Linacre [29], are:
1. investigate outfit before infit statistics,
2. investigate MNSQ before ZSTD values,
3. investigate high values before low or negative values,
4. if MNSQ values are acceptable (between 0.5 and 1.5), then ZSTD values can be ignored.

There was also a two-step process for judging unidimensionality [30, 31, 32]. First, we used the Rasch model – if the data fit the Rasch model, we can confirm the assumption of unidimensionality [22, 23, 29]. Second, a principal component analysis of the standardized residuals (PCA) was used [33]. We used three judging criteria for assessing unidimensionality [29, 30, 34]:
1. The Rasch dimension explains at least 50% of the variance in the data.
2. The largest secondary dimension, the first principal component of the residuals, explains no more than 5% of the variance, or the eigenvalue in the first contrast is less than 2.

3. There is a minimum ratio of 3:1 for the variance explained by the items compared to the variance of the first principal component of the residuals.

2. Reliability of the scale

Reliability of the scale was calculated using the KR-20 coefficient. Cronbach's alpha is a general version of the Kuder-Richardson coefficient of equivalence. The KR-20 coefficient applies only to dichotomous answers, whereas Cronbach's alpha applies to any set of items regardless of the response scale [35, 36]. We interpreted the KR-20 coefficient values according to Tavakol & Dennick [37]: excellent ($\alpha \geq 0.9$), good ($0.9 > \alpha \geq 0.8$), acceptable ($0.8 > \alpha \geq 0.7$), questionable ($0.7 > \alpha \geq 0.6$), poor ($0.6 > \alpha \geq 0.5$) and unacceptable ($0.5 > \alpha$).

Rasch's reliability is calculated for persons and for items. "Person reliability" is equivalent to the traditional "test" reliability, and "item reliability" has no traditional equivalent. "Person reliability" chiefly depends on sample ability variance, length of the test and sample-item targeting. "Item reliability" depends on item difficulty variance and person sample size [29].

3. The inter-rater agreement of the scale

The inter-rater agreement of the scale was calculated using Fleiss' kappa coefficient [38] to determine the agreement between the raters. "Item-by-item inter-rater agreement analysis" [39] was used to evaluate inter-rater agreement for each item of the resulting scale separately. We also calculated the average Fleiss' kappa value from the item values, as well as the total percentage agreement of examiners for each item [40] and the average value of all items. We interpreted results according to Landis & Koch [41]: almost perfect agreement (0.99–0.81), moderate agreement (0.80–0.61), substantial agreement (0.60–0.41), fair agreement (0.40–0.21), slight agreement (0.20–0.01) and poor agreement (< 0.00).

Statistical analysis

For the purposes of our research, software Winsteps (version 4.0) [29] designed for Rasch analysis and the Kappa Calculator software [42] to calculate Fleiss's kappa were used.

Results

Expert validity analysis

Five floorball experts evaluated 30 items in terms of content validity. Coefficient CVR according to Lawshe [28] was calculated for each item. All items were ranked in the order from the highest to the lowest values (Table 3). 18 items with the three highest coefficients 1, 0.6 and 0.2 were chosen for the next steps. We removed items 2, 13, 15, 20 and 30 with negative coefficients -1, -0.6 and -0.2 and also items 5, 12, 17, 18, 24, 27 and 28 with higher coefficients for various reasons (content similarity with another chosen item or impact of other factors on a successful solution).

Table 3. Coefficient CVR, according to Lawshe [28]

Item	16	22	25	3	4	6	8	9	10	11	12	28	29	1	5
CVR	1	1	1	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.2	0.2

Item	7	14	17	18	19	21	23	24	26	27	13	20	2	15	30
CVR	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	-0.2	-0.2	-0.6	-1	-1

Pilot testing

We tested 25 children of young school age on 18 items in terms of adequate difficulty and fitness level for this age group and also in terms of simple organization of each test. In terms of difficulty, we rated all items as appropriate for this age category. Items 3 and 4 were evaluated as unsatisfactory for testing due to the difficulty with the assessment procedure. The content of the items is the evaluation of two different types of dribbling over a wide line – "floorball dribbling" (item 3) and "hockey dribbling" (item 4). It was very difficult to divide the execution of the movement task into two categories – correct and incorrect. Some participants were able to accomplish a movement task, they were able to control the ball and move it over the wide line, but the technical execution of the movement was in fundamental contradiction with the proper floorball technique [2–5].

Design of the scale

Based on the pilot testing, we removed 2 items from the set of items, and for the main testing we used the remaining 16 items that we renumber as shown in Table 4.

Table 4. List of the numbered items

Item 1	Forehand spin with a ball
Item 2	Backhand spin with a ball
Item 3	Figure eight between the legs
Item 4	Pick up a ball on the stick blade
Item 5	Flip and catch a ball on the stick blade
Item 6	Carry a ball on the stick blade over a floorball barrier
Item 7	Take a backhand shot over a floorball barrier
Item 8	Follow the line
Item 9	Lob pass into a target
Item 10	Target passing
Item 11	Target passing while running
Item 12	Target shooting
Item 13	Target shooting while running
Item 14	Slalom with a ball No. 1
Item 15	Slalom with a ball No. 2
Item 16	Slalom with a ball No. 3

First of all, we analyzed the fit statistics for each of the 16 items (Table 5). The MNSQ outfit values of items 13, 12, 8, 10, 14, 5 and 2 were outside the required interval (0.5, 1.5). Items 13, 12, and 2 also have the value of the parameter of difficulty outside the interval (-3, +3), and they seem unsuitable for the final scale. On the other hand, items 14 and 5 almost met the interval

value 0.5 and had the parameter of difficulty inside the interval (-3, +3). Items 10 and 8 have the values of their parameters of difficulty suitably matched to the distances between the items in the scale, but their outfit MNSQ values were too high. Linacre [29] claims that high outfit MNSQ values may be the result of a few random responses by low performers. He recommends removing these performers when doing item analyses. We tried to identify these misfit performers and removed them from the research sample. We removed three misfit performers on item 10 and one misfit performer on item 8. The Rasch analysis after removing four performers is shown in Table 6.

Table 5. Rasch analysis of 212 performers

ITEM	DIFFICULTY PARAMETER	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
13	4.34	0.80	-0.9	0.25	-1.6
11	3.68	0.94	-0.3	1.12	0.4
12	3.38	0.80	-1.3	0.31	-1.6
8	2.80	1.49	3.2	1.56	1.1
10	2.13	1.27	2.1	2.15	2.1
9	1.59	1.00	0.0	0.82	-0.3
14	1.51	0.77	-2.2	0.48	-1.5
16	0.42	1.11	1.1	0.93	-0.1
5	-0.12	0.71	-3.3	0.48	-2.3
4	-0.88	0.82	-1.9	0.55	-1.7
15	-1.77	1.20	1.8	1.20	0.6
3	-2.39	1.00	0.1	1.11	0.4
7	-2.78	0.92	-0.6	0.58	-0.9
6	-3.11	1.00	0.1	0.82	-0.2
2	-3.22	1.05	0.4	2.42	2.1
1	-5.59	0.92	-0.2	1.03	0.3

Table 6. Rasch analysis of 208 performers

ITEM	DIFFICULTY PARAMETER	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
13	4.34	0.80	-0.9	0.25	-1.6
11	3.68	0.94	-0.3	1.12	0.4
12	3.38	0.80	-1.3	0.31	-1.6
8	2.80	1.49	3.2	1.56	1.1
10	2.13	1.27	2.1	2.15	2.1
9	1.59	1.00	0.0	0.82	-0.3
14	1.51	0.77	-2.2	0.48	-1.5
16	0.42	1.11	1.1	0.93	-0.1
5	-0.12	0.71	-3.3	0.48	-2.3
4	-0.88	0.82	-1.9	0.55	-1.7
15	-1.77	1.20	1.8	1.20	0.6
3	-2.39	1.00	0.1	1.11	0.4
7	-2.78	0.92	-0.6	0.58	-0.9
6	-3.11	1.00	0.1	0.82	-0.2
2	-3.22	1.05	0.4	2.42	2.1
1	-5.59	0.92	-0.2	1.03	0.3

difficulty and fitness level for this age group and also in terms of simple organization of each test. In terms of difficulty, we rated all items as appropriate for this age category. Items 3 and 4 were evaluated as unsatisfactory for testing due to the difficulty with the assessment procedure. The content of the items is the evaluation of two different types of dribbling over a wide line – "floorball dribbling" (item 3) and "hockey dribbling" (item 4). It was very difficult to divide the execution of the movement task into two categories – correct and incorrect. Some participants were able to accomplish a movement task, they were able to control the ball and move it over the wide line, but the technical execution of the movement was in fundamental contradiction with the proper floorball technique [2–5].

After Rasch analysis of the test results for the remaining 208 performers (Table 5), we decided to include items into the final scale due to recommendation of McCreary [30] and Linacre [29] and also due to the values of the difficulty parameters [16, 17] as follows:

Item 1: "Forehand spin with a ball" – The item was not selected for the final scale despite the appropriate values of fit statistics (infit MNSQ = 0.93; outfit MNSQ = 1.05; infit ZSTD = 0.1; outfit ZSTD = 0.3). The reason was that the item was too simple (the difficulty parameter was -5.61).

Item 2: "Backhand spin with a ball" – The item was not included into the scale because of the high value of the MNSQ outfit function (2.49) that degrades the measurement system, the high outfit value of the ZSTD (2.2) that indicates lack of predictability and also because of the difficulty parameter (-3.20) that was out of the interval (-3, 3).

Item 3: "Figure eight between the legs" – The item was chosen for the final scale, the value of the difficulty parameter -2.38 was inside the interval (-3, 3), and fit statistics had the required values (infit MNSQ = 1.01; outfit MNSQ = 1.13; infit ZSTD = 0.1; outfit ZSTD = 0.4).

Item 4: "Pick up a ball on the stick blade" – The item was chosen for the scale. The value of the difficulty parameter was -0.93, which was inside the interval (-3, 3). Infit and outfit MNSQ values were also inside the required intervals (infit MNSQ = 0.85; outfit MNSQ = 0.57). Infit and outfit ZSTD values were both -1.6. These negative values indicated higher predictability of the data, but they were not a threat to the validity of the scale.

Item 5: "Flip and catch a ball on the stick blade" – The item was selected for the final scale. The value of the difficulty parameter -0.15 was close to the center of the scale. The MNSQ infit value 0.73 was in the required interval, but the MNSQ outfit value 0.49 slightly exceeded the interval. However, we decided to accept this value for further processing. Too low infit and outfit values of ZSTD statistics (infit ZSTD = -3.0; outfit ZSTD = -2.1) indicated higher predictability of data as in item 4.

Item 6: "Carry a ball on the stick blade over a floorball barrier" – The item was not included in the final scale despite perfect values of fit statistics (infit MNSQ = 1.0; outfit MNSQ = 0.87; infit ZSTD = 0.0; outfit ZSTD = -0.1). The reason was an unsatisfactory parameter of the difficulty (-3.26), the item was too simple.

Item 7: "Take a backhand shot over a floorball barrier" – Item was placed in the final scale. The value of the difficulty parameter was -2.85 that was inside the interval (-3, 3) and infit and outfit MNSQ were found in the required interval (infit MNSQ = 0.92, outfit MNSQ = 0.58). ZSTD infit and outfit values were -0.6 and -0.8 (lack of predictability), and the values of statistics were not a reason to remove this item from the final scale (these values were not a threat to the validity).

Item 8: "Follow the line" – The item was retained in the final scale, the value of the difficulty parameter was 2.90 and was close to the extreme value of the scale (+3). Infit and outfit MNSQ values and outfit ZSTD were inside the required interval (infit MNSQ = 1.48, outfit MNSQ = 1.14, outfit ZSTD = 0.4). Infit ZSTD reached a high value of 3.1. This value indicated lack of predictability but did not decrease the value of the scale.

Item 9: "Lob pass into a target" – The item was selected for the final scale, the value of the difficulty parameter 1.61 was inside the interval (-3, 3). The infit and outfit values of MNSQ and ZSTD were very good (infit MNSQ = 1.01, outfit MNSQ = 0.86, infit ZSTD = 0.2, outfit ZSTD = -0.2).

Item 10 "Target passing" – Item was chosen for the final scale, the value of the difficulty parameter 2.30 was in the interval (-3, 3). The infit and outfit MNSQ values and outfit ZSTD value were in the required intervals (infit MNSQ = 1.18; outfit MNSQ = 1.04; outfit ZSTD = 0.2); the infit ZSTD value reached 1.4. This value indicated some lack of predictability of data.

Item 11: "Target passing while running" – The item was not included in the final scale despite the relatively appropriate values of the fit functions (infit MNSQ = 0.95; outfit MNSQ = 1.19; infit ZSTD = -0.2, ZSTD outfit = 0.5). The reason was an unsatisfactory difficulty parameter (3.72); the item was too difficult.

Item 12: "Target shooting" – The item was not included in the final score due to the low MNSQ outfit (0.31) and too high value of the difficulty parameter (3.41). Low MNSQ outfit values do not degrade the measurement tool, but may produce misleadingly good reliability and separation values.

Item 13: "Target shooting while running" – The item was not included in the final scale due to the low MNSQ outfit (0.25) as low as at item 12. This item was also too difficult (the difficulty parameter = 4.39).

Item 14: "Slalom with a ball No. 1" – The item was removed despite a suitable difficulty parameter (1.52) and relatively acceptable values of fit statistics (infit MNSQ = 0.78; outfit MNSQ = 0.49; infit ZSTD = -2.1; outfit ZSTD = -1.4). The reason was that this item had almost the same difficulty parameter as item 9. For the final scale we preferred item 9 because of the content ("lob shot" instead of another slalom as in items 15 and 16).

Item 15: "Slalom with a ball No. 2" – The item was chosen for the final scale. The value of the difficulty parameter -1.87 was in the required interval (-3, 3). Infit and outfit MNSQ values and outfit ZSTD value fell within the correct interval (infit MNSQ = 1.25; outfit MNSQ = 1.26; outfit ZSTD = 0.7). Infit ZSTD value reached 2.1, referring to the lack of predictability. This value does not degrade the final scale.

Item 16: "Slalom with a ball No. 3" – The item was placed in the final scale; the value of the difficulty parameter was 0.41 and was inside the required interval (-3, 3). Values infit and outfit MNSQ and outfit ZSTD were good (infit MNSQ = 1.14; outfit MNSQ = 0.97; outfit ZSTD = 0.0). Only the infit value of ZSTD (1.14) was increased as in the previous item.

The second step of the scale development was the creation of a graphical form of the scale (Figure 2) according to the difficulty parameter of each chosen item shown in Table 7 (the difficulty parameter values were recalculated for nine item scale).

Finally we evaluated the scale values. Nine items were chosen for the final scale (Table 8) with the items very well covering the continuity of the diagnosed feature within a range from -2.85 logits to +3.01 logits. The final scale optimally covers the continuum of the diagnosed feature; slightly greater is the distance between items 16 and 9 (1.19 logits) and between items 15 and 4 (0.98 logits). Despite these distances, the final scale appears to be a quality latent feature diagnostic tool due to the distribution of items and their difficulty parameters.

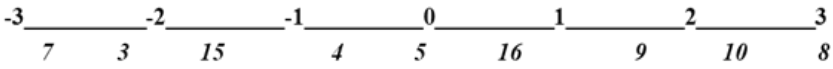

Figure 2. The graphical form of the scale

Table 7. The difficulty parameter

ITEM	7	3	15	4	5	16	9	10	8
DIFFICULTY PARAMETER	-2.85	-2.34	-1.79	-0.81	0.01	0.57	1.76	2.43	3.01

Table 8. The items of the final scales

Item 7	Take a backhand shot over a floorball barrier
Item 3	Figure eight between the legs
Item 15	Slalom with a ball No. 2
Item 4	Pick up a ball on the stick blade
Item 5	Flip and catch a ball on the stick blade
Item 16	Slalom with a ball No. 3
Item 9	Lob pass into a target
Item 10	Target passing
Item 8	Follow the line

STANDARIZATION PROCEDURE – VALIDITY AND UNDIMENSIONALITY

The Rasch analysis, specifically fit diagnosis, gave us information on the validity and unidimensionality of the scale. All the infit and outfit MNSQ values (Table 9) were in the required interval (0.5, 1.5), according to Linacre [29]. ZSTD values were ignored, because MNSQ values were acceptable [29]. The data fit the Rasch model well and the other three PCA criteria used for judging unidimensionality [29, 30, 34] were met (Table 10):

1. The variance explained by the measure was 54.1 % (more than the required 50%).

2. The variance explained by the first principal component of the residuals was 8.8%, just over the criterion of 5 %, but the eigenvalue in the first contrast was 1.7225 (less than the required 2). Exceeding one of these two criteria means that this condition for unidimensionality was met.

3. The ratio for the variance explained by the items compared to the variance of the first principal component of the residuals was 3.171: 1 (more than the required 3:1 ratio).

Taken together, the fit statistics and the PCA indicate that the scale is valid and unidimensional.

Table 9. Standardized residual variance

ITEM	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
8	1.28	2.0	1.07	0.3
10	1.04	0.4	0.95	0.1
9	0.98	-0.2	0.93	-0.1
16	1.07	0.7	0.81	-0.8
5	0.73	-3.0	0.53	-2.7
4	0.89	-1.1	0.66	-1.4
15	1.17	1.4	1.13	0.4
3	1.10	0.8	1.04	0.3
7	0.93	-0.5	0.65	-0.5

Table 10. Standardized residual variance

	EIGEN VALUE	Observed variance	Expected variance
Raw variance explained by measures	10.6188	54.1%	54.4%
Raw variance explained by persons	5.1400	26.2%	26.3%
Raw variance explained by items	5.4788	27.9%	28.1%
Unexplained variance in 1st contrast	1.7225	8.8%	19.1%
Unexplained variance in 2nd contrast	1.4461	7.4%	16.1%
Unexplained variance in 3rd contrast	1.2083	6.2%	13.4%
Unexplained variance in 4th contrast	1.1030	5.6%	12.3%
Unexplained variance in 5th contrast	1.0289	5.2%	11.4%

STANDARIZATION PROCEDURE – RELIABILITY

The reliability was calculated using the KR-20 coefficient. The reliability of the final scale was 0.81, which was a good result according to Tavakola & Dennick [37]. The Rasch reliability was also calculated for the tested persons and the items, and the results showed the “real” and “model” reliability. The “real” person reliability reached lower values; the value of 0.75 (“model” value = 0.78) was an acceptable result according to Tavakola & Dennick [37]. Item ”real” reliability was 0.99 (“model” value = 0.99), which was a very good result and very high reliability.

STANDARIZATION PROCEDURE – THE INTER-RATE AGREEMENT

The inter-rater agreement of the scale was evaluated by four raters with varying lengths of training experience. The inter-rater agreement reached 0.985313% on all nine items and the average Fleiss' kappa value reached 0.936887. This is a very good result and almost perfect agreement according to Landis & Koch [41]. The inter-rater agreement on each of nine items is shown in Table 11.

Table 11. Inter-rater agreement

ITEM	Inter-rater agreement %	Fleiss' kappa	Interpretation [41]
8	0.982759	0.963046	almost perfect agreement
10	1.00000	1.00000	almost perfect agreement
9	1.00000	1.00000	almost perfect agreement
16	1.00000	1.00000	almost perfect agreement
5	0.942529	0.865741	almost perfect agreement
4	0.977011	0.765657	moderate agreement
15	1.00000	1.00000	almost perfect agreement
3	0.965517	0.837535	almost perfect agreement
7	1.00000	1.00000	almost perfect agreement

DISCUSSION

The design of the Guttman-type assessment scale includes nine items that measure the level of acquired floorball skills. The scientific standardization procedure of the final nine items of the rating scale has been successful and the results have shown that the scale is a valid, reliable and objective diagnostic tool.

The Rasch model fit the data well – infit and outfit MNSQ values were in the required interval according to Linacre [29]. Exploring the model fit of the scale was also the first step in assessing the unidimensionality of the scale, the basic assumption of the Guttman scale. The second step was the principal component analysis of the standardized residuals in which we assessed the three criteria that were met – the first component explained 54.1% of the total variance in the data (more than the required 50%); the eigenvalue in the first contrast of the residuals was 1.7225 (less than the required 2), and the ratio between the variance explained by the items and the variance explained by the first contrast was 3.171: 1 (more than the required 3:1 ratio). The only value that did not meet the required criteria was the value of the first contrast of the residuals that did not exceed 5% but reached 8.8%. This result could indicate the existence of a second dimension [29], but since the above-mentioned eigenvalue in the first contrast of the residuals is low, the existence of the second dimension should be avoided and unidimensionality confirmed.

The inter-rater agreement (Fleiss kappa) reached almost perfect agreement 98.5% and the reliability value of the rating scale was 0.81.

All the results reached high criteria except the person reliability. The result of the person reliability (0.75) was not interpreted as excellent or good but was still acceptable [29]. Interpretation of this result [29] suggests that there

were not enough performers in the research sample with a sufficiently large range of the floorball skill distribution, probably missing a sufficient number of tested players with an extreme (high or low) level of tested skill.

Nine items of the final scale cover the continuity of the diagnosed feature with the gradual increase in the difficulty from the easiest to the most difficult items (from -2.85 logits to +3.01 logits). Final nine chosen items include skills as passing, slalom running or a ball manipulation, but we miss the "shooting" items (items 12 and 13) in the final rating scale. The "shooting" items were too difficult for the research sample. Although we have tried to ensure equal representation of all levels of the diagnosed feature in the research sample, there were probably more performers with a moderate or lower level of acquired floorball skills for which "shooting" items were too difficult. From this point of view, it would be appropriate to test players with a higher level of the diagnosed feature with 16 original items to shift the difficulty of the scale towards more difficult items. Another option would be to simplify the "shooting" items and retest them again together with 9 chosen items or 16 original items. This should be the next step in further development of this diagnostic tool.

CONCLUSIONS

Despite the absence of "shooting" items, we consider the rating scale as a high-quality diagnostic tool that evaluates the level of acquired floorball skills.

Relatively simple content of the items also allow practical applications. The nine items scale can be easily used in the training or teaching process. The results of our research will, therefore, be for coaches of youth categories in floorball and physical education teachers.

Finally, we would like to highlight the well-defined procedure for constructing the assessment scale as a contribution to the sport science. Diagnostic tool construction can also be used in other sports specializations to create a high quality and standardized rating scale designed to test the level of technical skills of a particular sport.

REFERENCES

[1] Peric T. Sportovni priprava deti [Sport preparation of children]. Prague: Grada; 2012. Czech.
[2] Kysel J. Florbal – kompletni pruvodce [Floorball – a complete guide]. Prague: CFbU; 2010. Czech.
[3] Martinkova Z. Florbal – prakticky pruvodce treninkem mladeze [Floorball – a practical guide to youth training]. Prague: CFbU; 2009. Czech.
[4] Skruzny Z. Florbal [Floorball]. Prague: Grada; 2005. Czech.
[5] Zlatnik D. Florbalovy trenink v praxi – herni cinnosti jednotlivce [Floorball training in practice – individual playing skills]. Prague: CFbU; 2004. Czech.
[6] Dovalil J. Vykon a trenink ve sportu [Performance and training in sport]. Prague: Olympia; 2012. Czech.
[7] Ali A. Measuring soccer skill performance: A review. Scand J Med Sci Sport. 2011;21(2):170-183.
[8] Ilic I. Structures and differences of the cognitive abilities of top handball, volleyball, basketball and soccer players. Facta Universitatis: Physical Education and Sport. 2015;13(3):403-410.
[9] Kioumourtzoglou E, Derri V, Tzetzis G, Theodorakis Y. Cognitive, perceptual, and motor abilities in skilled basketball performance. Percept Motor Skill. 1998;86(3):771-786.
[10] Huijgen BCH. Cognitive functions in elite and sub-elite youth soccer players aged 13 to 17 years. PloS One. 2015; 10(12).
[11] MacDonald LA, Minahan CL. Indices of cognitive function measured in rugby union players using a computer-based test battery. J Sport Sci. 2016; 34(17): 1669-1674.
[12] Baumgartner TA, Jackson AS, Mahar MT, Rowe AD. Measurement for evaluation in physical education and exercise science. Boston: McGraw Hill; 2003.

[13] Cepicka L. Modely teorie polozkovych odpovedi v diagnostice motoriky cloveka [Models of item response theory in human motor diagnostics]. Pilsen: Zapadoceska univerzita; 2002. Czech.

[14] Jansa P. Mnohorozmerove skalovani v telesne vychove a sportu [Multidimensional scaling in physical education and sport]. Teorie a praxe TV. 2012; 37(2). Czech.

[15] Knudson DV, Morrison CS. Qualitative analysis of human movement. Champaign: Human Kinetics; 2002.

[16] Mekota K, Blahus P. Motoricke testy v telesne vychove [Motor tests in physical education]. Prague: SPN; 1983. Czech.

[17] Mekota K, Cuberek R. Pohybove dovednosti, cinnosti, vykony [Movement skills, activities, performances]. Olomouc: Univerzita Palackeho; 2007. Czech.

[18] Morrow JR, Jackson AW, Disch JG, Mood DP. Measurement and evaluation in human performance. Champaign: Human Kinetics; 2005.

[19] Thomas JR, Nelson JK, Silverman SJ. Research methods in physical activity. Champaign: Human Kinetics; 2005.

[20] Andrich D. Rasch models for measurement. Newbury park: Sage Publications; 1988.

[21] Brichacek V. Uvod do psychologickeho skalovani [Introduction to psychological scaling]. Bratislava: Psychodiagnosticke a didakticke testy; 1978. Czech.

[22] Cepicka L. Konstrukce perfektni skaly v diagnostice motorickych dovednosti [Construction of a perfect scale in the diagnosis of motor skills]. Ceska Kinantropologie. 2003;7(1):7-18. Czech.

[23] Cepicka L. Prispevek k unidimenzionalnimu skalovani motorickych predpokladu [Contribution to unidimensional scaling of motor assumptions]. Prague: FTVS UK; 2005. Czech.

[24] Blahus P. K systemovemu pojeti statistickych metod v metodologii empirickeho vyzkumu chovani [Systemic concept of statistical methods in the methodology of empirical research of behavior]. Prague: Karolinum; 1996. Czech.

[25] Hendl J. Prehled statistickych metod: analyza a metaanalyza dat [Overview of statistical methods: data analysis and meta-analysis]. Prague: Portal; 2009. Czech.

[26] Massof W. Understanding Rasch and item response theory models: Applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. Ophthal Epidemiol. 2011;18(1):1-19.

[27] Stochl J, Musalek M. A practical guide to pilot standardization of tests. Acta Universitatis Carolinae Kinanthropologicae. 2009; 45(2): 5-15.

[28] Lawshe CH. A quantitative approach to content validity. Personel Psychol. 1975;28:563-575.

[29] Linacre JM. Winsteps® Rasch measurement computer program. Beaverton: Winsteps; 2017.

[30] McCreary LL. Using the Rasch Measurement Model in psychometric analysis of the family effectiveness measure. Nurs Res. 2013;62(3):149-159.

[31] Donovan NJ. Adding meaning to measurement: Initial Rasch analysis of the ASHA FACS Social Communication Subtest. Aphasiology. 2006;20:362-373.

[32] Wu TY. Rasch analysis of the general self-efficacy scale in workers with traumatic limb injuries. J Occup Rehabil. 2016;26:332-339.

[33] Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Measur. 2002;3:205-231.

[34] Basílio ML. Cross-cultural validity of the Brazilian version of the Abilhand questionnaire for chronic stroke individuals, based on Rasch analyses. J Rehabil Med. 2016;48:6-13.

[35] Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98-104.

[36] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.

[37] Tavakol M, Dennick R. Making sense of Cronbach's alpha. International Journal of Medical Education. 2011; 2:53-55.

[38] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378-82.

[39] Turner-Stokes L. The work-ability support scale: Evaluation of scoring accuracy and rater reliability. J Occup Rehabil. 2014; 24: 511-524.

[40] Welch V. Open access systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for a health equity plausibility algorithm. Med Res Methodol. 2012;12:187.

[41] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

[42] Randolph JJ. Online Kappa calculator computer software [Available at http://justus.randolph.name/kappa] [Accessed on 20 January, 2018].

Cite this article as:

Dragounova Z.
Development and standardization of a rating scale designed for floorball skills diagnostics of young school-age children.
Balt J Health Phys Act. 2018;10(4):34-48.
doi: 10.29359/BJHPA.10.4.03