Digital Life

Journal of Digital Life 2025,5,3

Article

A practical evaluation method using item response theory to evaluate children's form of Jumping-over and crawling-under

Yasufumi Ohyama [1]* and Osamu Aoyagi [2]

[1] National Institute of Technology (KOSEN), Sasebo College, 1-1, Okishin-machi, Sasebo-shi, Nagasaki, 857-1193, Japan
[2] Faculty of Health and Sports Science, Fukuoka University, 8-19-10, Nanakuma, Jyonan-ku, Fukuoka, 814-0180, Japan

Abstract

In order to develop the test battery evaluating the movement of Jumping-over and crawling-under for young children, we measured and videotaped it performed by 350 kindergarten children. After that, 16 movements by body part were assess using three categories of "possible," "no idea" and "impossible." Since the change of eigenvalues derived from this data represented a one-dimensional structure having a high homogeneous each other, successively, the parameters of step difficulty and samples were computed using IRT. Among various IRT models, this study used Partial Credit model because obtained data is ordinal and the sample size is few. Then we found the outcomes as follows: 1) The correlation between the two sets of parameters of step difficulty computed from two sets of randomly divided samples was high, so that it can be concluded that the difficulty parameters are not depended on any ability level of samples. Again, the correlation between the two sets of sample parameters calculated from two groups of randomly grouped item parameters was also significant. This fact allowed us to conclude the obtained sample parameters did not depend on any difficulty level of items. 2) As significant difference between ages in obtained thetas was found, it is considered that thetas reflect motor ability that advances with maturity. There was also high correlation between thetas and measurements of Jumping-over and crawling-under. 3) Judging from information function consisting of 16 items, this test is suitable to determine the ability with from middle to a little low level because the information reached around there. 4) A practical estimation and evaluation sheet utilizing thetas by total score based on Zhu and Cole (1996) was developed and the application examples were represented. 5) As a result that the relationship between strictly computed indexes of fitness and the number of aberrant patterns detected in the practical method in this study was examined, the high correlation was found. Judging from this fact, it can be concluded that IRT is useful to estimate and evaluate the movement of Jumping-over and crawling-under.

Keywords: Partial Credit model; young child

1. Introduction

As young children's motor ability is immature, they cannot perform their adult-like movements that serve the purpose such as "jumping further" or "running fast." On that condition, it is meaningless that children's motor ability is evaluated based on measurements of performance (the records measured by the units of seconds, cm, kg and the like), but how close to the efficient movement (i.e. adult motor performance) children perform is important (Matsuura, 1975, 1982; Malina and Bouchard, 1991). In this context, the attempts to measure the motor development of children have been conducted by classifying children's motor patterns into several typical mature and immature ones (DeOreo, 1980a, 1980b; Gallahue, 1982; Kim and Matsuura, 1988; Miyamaru and Hirakoba, 1982; Nakamura, 2001; Nakamura et al., 1987; Wild, 1938). However, many instances that cannot group into the typical patterns were detected (Nakamura, 2001; Nakamura et al., 1987). In such cases, after rating whether individual movement at various body

Received: 17 October 2024, Revised: 26 November 2024, Accepted: 23 December 2024, Published: 21 February 2025
* Correspondence: yasufumi@sasebo.ac.jp

parts can be accomplished or not, a method estimating individual motor ability is used by scaling the rating results (Aoyagi, 1999, 2003).

Now, when scaling, if the parameter of item difficulties and estimated sample abilities can put on the same scale, it can be predicted whether an individual with a definite level of ability can perform a given motor task or not. This advantage provides useful information when instructing physical education or motor exercise. For this reason, item response theory (IRT) can be applied into the problem like this. However, since ability parameters (θs) in an IRT model is estimated by using a nonlinear optimization method, it is unfeasible to utilize the IRT in a kindergarten or nursery school because kindergarten or nursery school teachers are unfamiliar with statistical knowledge or programming tasks. For this reason, Zhu and Cole (1996) developed a practical calculation sheet that estimates children's fundamental motor skills (θs in IRT), understands relative evaluation of them and comprehends the relationship between estimated motor ability (θ) and their pass or fail of motor tasks using only a pencil and ruler.

Thus, taking Zhu and Cole's practical calculation sheet of children's motor skills (1996) as reference, this study aimed to propose a simple and feasible subjective rating sheet for the movements of "Jumping-over and crawling-under (Katsube, 1971; Research Center in Physical Education, 1976, 1981)" which is a combined movement required for coordination and is much popular in Japan. The IRT used to implement this goal has additional advantage that it can estimate the item parameters not depending on the ability level of sample and directly compare with both difficulty of motor tasks (difficulty parameters) and subject's ability (sample parameters) on the same scale (Hambleton and Swaminathan, 1985).

Now, Zhu and Cole's practical calculation sheet of children's motor skills (1996) had the following characteristics: 1) Sample parameters (θs) can be estimated from scores (the sum of possible motor tasks) using the relation that the score corresponds to the θ in the Rasch model. 2) Relative evaluation is conducted while referring to the distributions of θs by gender and age. 3) The pass or fail of a child who has a certain θ can be predicted by the positional relation between the difficulty of a motor task and the child's θ located in the same scale.

Now, their procedure is based on the Guttman scale that is a preferable response pattern in classical test theory but not in IRT. In the Guttman scale, the pass and fail responses are perfectly separated in the response pattern chart sorted by scores of both samples and motor tasks. However, in IRT, an adequate randomness in response patterns is regarded to be desirable not because it is redundant by duplicating the information of other motor tasks (Wright and Masters, 1982). Namely, in IRT, the estimation of pass or fail has been conducted with a certain range. So, when predicting a child's pass or fail in a motor task with his/her sample parameter (θ), the range of judgement is not utilized with a standard error of motor task but only that of sample parameter. However, some inconsistencies will be found in the judgement because several indexes of fitness have been proposed until now. Thus, this study investigated the relationship between the practical method proposed in this study and ordinal indexes of fitness as well as the validity of it.

2. Methods

2.1 Sample and evaluation items

Three-hundred and fifty young children in S-kindergarten and S-nursery school of F-city were asked to perform Jumping-over and crawling-under. Figure 1 shows a schematic exhibition of the measurement scene of Jumping-over and crawling-under. In the measurement of Jumping-over and crawling-under, children are asked to jump over a 35 cm high rubber string with both feet and then crawl under the rubber string while changing the direction after landing as soon as possible. The required time for a 5-time repeated movement is measured. At the same time, children's performance was video-taped obliquely from the front side. While playing back the VTR, their actions were subjectively rated as "possible," "no idea" and "impossible" using 16 items.

Based on the general description of the development of fundamental motor skills such as running, jumping and throwing by DeOreo (1980a, 1980b), Galahue (1982), Kim and Matsuura (1988), Miyamaru and Hirakoba (1982), Nakamura (2001), Nakamura et al. (1987) and Wickstrom (1977), 16 subjective rating items were developed after categorizing the movement of Jumping-over and crawling-under by body part and phase. Subjective rating items as well as abbreviated item names, body parts, motor phases and objectivity coefficients are shown in Table 1. Objectivity coefficients was computed using Goodman-Kruscal rank order correlation coefficient (Ikeda, 1989) between 2 raters who subjectively rated 30 children.

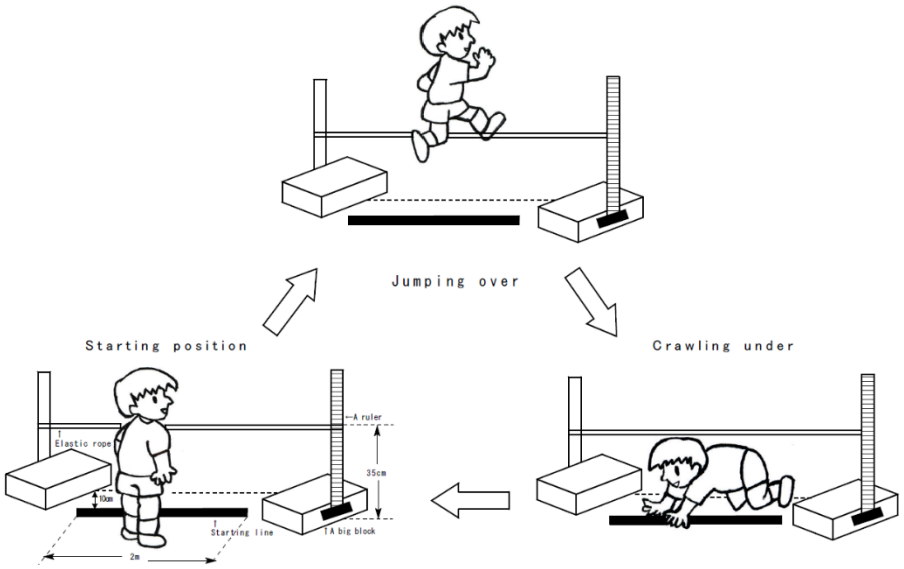

Fig.1. Jumping Over and Crawling Under performance.

Table 1. Subjective rating items and objectivity coefficients

No.	Full description of item	Abbreviated description	Body parts used	Phase	Objectivity coefficient †)
1	A child can swing both arms forward from backswing with a full range of motion.	Use of back swing	Arm	Before jumping	0.719
2	A child can use both arms before jumping (i.e., does not leave arms hanging at side).	Use of both arms before jumping			0.536
3	A child does not elevate arms backward while jumping.	Elevation of arms backward while jumping		While jumping	0.832
4	A child does not elevate arms sideward with stiffened shoulders while jumping.	Elevation of arms sideward while jumping			0.963
5	Child's feet are parallel and apart appropriately.	Feet location before jumping	Leg	Before jumping	0.492
6	Child's knees face forward (not sideways).	Direction of knees when jumping		While jumping	0.766
7	A child can jump with both feet but not one foot.	Jumping with two feet			0.945
8	A child does not land on the knees but on the feet.	Not landing on the knees		On landing	0.721
9	A child can land with both feet but not one foot	Landing with both feet			0.926
10	A child can complete the fifth landing successfully.	Complete success of landing			0.852
11	Child's trunk faces front before jumping.	Direction of trunk before jumping	Trunk	Before jumping	0.551
12	While jumping, child stretches trunk out fully.	Trunk extension while jumping		While jumping	0.965
13	A child can complete 5 jumps.	Complete success of jumping 5 times	Whole body	While jumping	0.935
14	When landing, child prepares for the next movement by twisting.	A succession of turning back		On landing	0.865
15	When crawling under, child goes through straight but not like a crab.	Direction of crawling		While crawling	0.820
16	When crawling under, child is not caught by the elastic because hip was not lowered.	Complete success of crawling movement			0.671

†) Goodman and Kruscal's ordinal correlation coefficient between two raters

2.2 Partial Credit model

Taking the evaluation is 3-categorized rank order into account, difficulty parameters were obtained using a Partial Credit model, one of IRT models, in which a relatively stable parameter can computed even in the situation of few samples (Masters,1982). The Partial Credit model has a trait that focusing on the probability between adjacent categories and defines the conditional probability that an individual prefers category (k) to category (k-1). In this

model, the conditional probability that an individual (j) prefers category (k) in the item (i) is expressed category characteristic function shown in formula (1).

$$\frac{P_k}{P_k + P_{k-1}} = \frac{\exp(\theta - \delta_k)}{1 + \exp(\theta - \delta_k)} \quad (1)$$

Figure 2 displays the category characteristic curves which has four categories and their difficulty parameters monotonically increase with the delta of category 1 (delta1) = -2.0, delta2 = 0.0 and delta3 = 1.8. In particular, the intersection point between adjacent category characteristic curves is called "step difficulty," in which the conditional probability which category includes an ability parameter (θ) out of the adjacent two ones is 50%. Therefore, this step difficulty is, what is called, the difficulty of two categories. In addition, as this model is fundamentally based on the Rasch model, it has a merit to take advantage of a much practical estimation method which can be estimated sample parameters by score computed by adding the number of passing each step difficulty, i.e. possible=2, no idea=1 and impossible =0 because scores are sufficient statistics (Harris et al., 1988; Masters, 1984; Wilson et al., 1988).

However, this step difficulty does not always magnify monotonically with categories advance. Figure 3 exemplifies the case which step difficulties do not increase monotonically (delta1 = 0.0, delta2 = -1.t and delta3 = 2.0). The non-monotonicity like this occurs when all individuals can reach at category 3 if only they accomplish category 1 because category 2 is much easy.

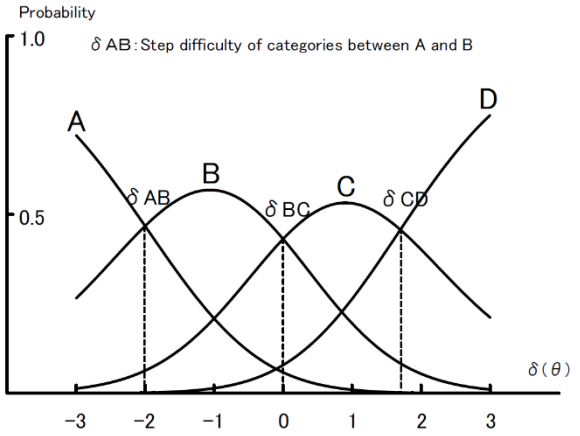

Fig. 2. Graphical representation of the category characteristic curve of the Partial Credit model.

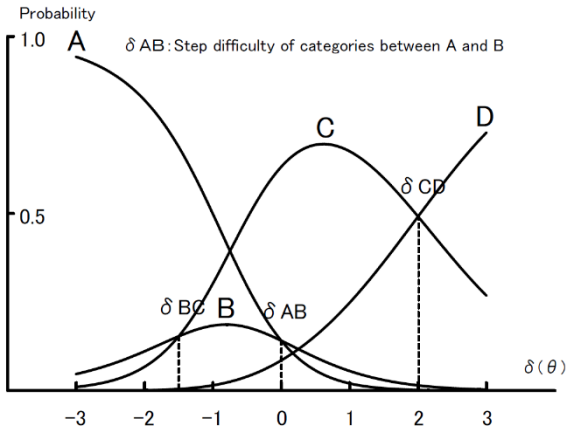

Fig. 3. Graphical representation of step difficulties that do not increase monotonically.

Step difficulties and sample parameters were computed based on Master's algorism (1982).

2.3 Indexes of fitness

In Rasch model, based on standardized residual z which standardize the difference between expectation and actual evaluation, whether obtained θs fit to the model or not is statistically tested. Besides this, there are 1) (simple) mean square u (formula 2) which is the average of sums of squared standardized residuals, 2) weighted mean square (formula 3) which is the average of sums of squared standardized residuals weighted by variance (w), and 3) standardized weighted mean square (formula 4) which makes approximate it normal distribution by adjusting with variance and kurtosis (Wright and Masters, 1982). The condition that is unfitted to the model with random variations is judged as "under fit." In contrast, if there are much reluctant and little own information because it has duplicated information with other items, it is concluded as "over fitted." However, although the latter is interpreted as "useless" in IRT, on the contrary it is regarded to be desired in classic test theory as it is perfectly fitted to the Guttman scale. The cases that "an individual was able actually to accomplish a motor task though it was expected to not from the θ estimated by IRT" or "one was not able really to do it though it was presumed to be able to implement it judging from obtained a θ" are called "aberrant patterns" in this study. The validity of the number of aberrant pattern derived from the practical method in this study was investigated based on these indexes of fitness computed strictly.

$$P_k = \exp(\theta - \delta_k) P_{k-1} \quad (2)$$

$$P_k = \exp(\theta - \delta_k) \exp(\theta - \delta_{k-1}) \exp(\theta - \delta_{k-2}) \cdots \exp(\theta - \delta_1) P_0 \quad (3)$$

$$P_0 + P_1 + \cdots + P_m = 1 \quad (4)$$

2.4 Ethical considerations

We explained in writing to the parents through kindergartens the purpose of this study and that the video footage would not be used for any purpose other than research and academic study and obtained their permission. After the subjective evaluation, the video footage was erased, and only the age, gender, and evaluation data were saved as Excel data and analyzed. The data was managed on a desktop computer on campus that was protected by a password.

3. Results and discussion

3.1 Applicability of IRT

As motor skills are trying to be identified using a scale, i.e. a θ, whether all items are thought to measure a potential ability or not is investigated. After computing a correlation matrix using Goodman-Kruscal rank order correlation coefficient taking an ordinal evaluation such as "possible," "no idea" and "impossible" into account, principal component analysis was applied to it. Figure 4 displayed the change of eigenvalues in the descending order of the amount of them. The eigenvalue of the first component with 7.58 is much greater than the second one with 1.77. This fact allows us that these items are highly homogeneous and have a one-dimensionality.

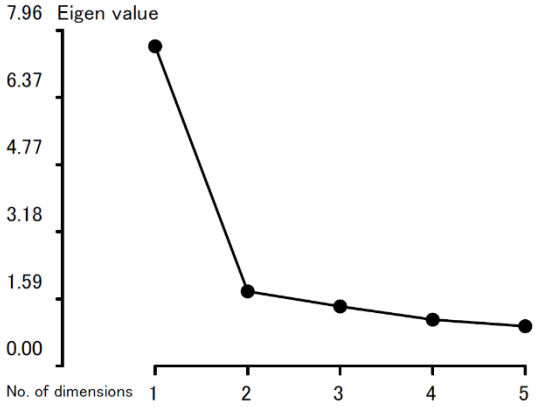

Fig. 4. Change in Eigen value by number of dimensions.

Next, whether difficulty parameters depend on sample or sample parameters relay on the easiness of a given item or not is inspected. After randomly separating all samples into the two groups, the correlation coefficient between two groups of difficulty parameters derived from each group of sample parameters. Similarly, after randomly categorizing all items into the two sets of items, the correlation coefficient between sample parameters computed using the two difficulty parameters was calculated. The correlation coefficient between two sets of difficulty parameters obtained from randomly separated two subgroups was 0.999. Likewise, that of two groups of sample computed from randomly divided into two groups of items was 0.928. That is to say, it is thought that since no remarkable difference is found in two sets of difficulty parameters even if it is computed from any group of samples, they do not depend on the ability level of sample. Alike, it is considered that as there is no noticeable gap between in two sets of sample parameters even if calculated from any group of items, they are independent of the difficulty level of items.

In addition, estimated sample parameters (θs) has a significant correlation coefficient of 0.139 ($p < 0.05$) with age. In other words, the sample parameters are regarded to measure children's motor skills that will be supposed to develop with advancing age. Again, there was a significant association of -0.158 ($p < 0.05$) with the measurement of Jumping-over and crawling-under. Namely, this fact induced us that the finer the form of Jumping-over and crawling-under, the quicker its performance.

3.2 Step difficulty

Table 2 shows obtained step difficulty parameters and their standard errors in two intersections from "impossible" to "no idea" and from "no idea" to "possible." In "Jumping with two feet," as the step difficulty of the former and latter is not monotonically going up, relative easy and accurate evaluation is thought to make the judgement of "no idea" few. Furthermore, judging from the two sets of step difficulty in each item, "Use of back swing," "Complete success of crawling movement" and "Landing with both feet" were difficult to perform because their parameters of step difficulty are large. In contrast, "Direction of trunk before jumping," "Trunk extension while jumping" and "Direction of crawling" were easy to perform because of their small values of step difficulty parameters.

3.3 Estimates of sample parameters by score

In order to obtain individual ability parameter θ from these items, in usual item response theory or graded response model, a nonlinear optimization method is used by pattern of pass or fail. It is not practical because actually it can be

done only using a computer but not hand-calculating. Additionally, θs corresponding to all combinations of pass-or-fail patterns cannot be computed in advance because the number of patterns is almost innumerable. However, Partial

Table 2. Step difficulties and standard errors

No.	Item †)	"Failed" to "Indiscernable"		"Indiscernible" to "Successful"	
		Step difficulty	Standard error	Step difficulty	Standard error
11	Direction of trunk before jumping	-3.095	0.505	-0.342	0.119
12	Trunk extension while jumping	-3.400	0.505	0.230	0.112
15	Direction of crawling	-2.633	0.415	-0.418	0.120
8	Not landing on the knees	-1.161	0.221	-0.474	0.119
4	Elevation of arms sideward while jumping	-2.613	0.296	1.166	0.117
10	Complete success of landing	-1.588	0.220	0.296	0.112
5	Feet location before jumping	-1.578	0.212	0.481	0.112
13	Complete success of jumping 5 times	-0.414	0.162	-0.275	0.115
7	Jumping with two feet††)	0.439	0.140	-0.713	0.117
3	Elevation of arms backward while jumping	-1.810	0.200	1.638	0.129
6	Direction of knees when jumping	-0.928	0.155	1.051	0.119
2	Use of both arms before jumping	-1.121	0.160	1.380	0.125
14	A succession of turning back	-0.554	0.141	0.913	0.118
9	Landing with both feet	-0.147	0.132	0.629	0.117
16	Complete success of crawling movement	-0.533	0.130	1.962	0.149
1	Use of back swing	-0.705	0.134	2.162	0.156

†) Items are arranged according the mean of step difficulties.
††) Difficulties for this item do not increase monotonically.

Credit model, one of the Rasch models, can have already yielded the θs by a total score by using the characteristics that a total score is a sufficient statistic. By doing this, only if all θs have already computed once, after that, a required θ can be easily obtained directly from a total score. After tallying even almost innumerable patterns by total score, as the number of the pattern of total scores decreases drastically, finding a θ from all total score patterns comes to be easy. Table 3 indicates the corresponding table between θs and total scores. The total score is obtained by summing each score (possible = 2, no idea = 1 and impossible = 0) up. However, since maximum likelihood method is utilized, both cases of full mark and 0 point are removed in the table because it is impossible to compute. Figure 5 displays the relationship between total scores and estimated sample parameters. It indicates that their relationship is monotonous but not linear, i.e. although sample parameters linearly go up with advancing total scores in the average level of sample parameters, but suddenly increase in highest score level and decline in lowest score level.

3.4 Information function

Figure 6 shows the test information function comprising of all items. Information function indicates the accuracy of estimation for θs, i.e. the larger the information function, the more accurately θs are estimated in the spot of a scale. In this scale, information function reaches around at θ = -0.240 with 6.642 of the maximum amount of information. The amount of information of 6.642 is equivalent to 0.869 of reliability coefficient in classical test theory (Toyoda, 2002, pp. 124 – 126). When we see it conversely, the reliability coefficient of 0.85 corresponds to the amount of information of 5.66. As in this scale, the interval covering this value is approximately from -1.0 to +0.5, it seems to be possible to estimate θs in this interval with the accuracy that reliability coefficient is more than 0.85. In short, this scale can more accurately estimate θs of individuals with from middle to a little poor ability. Thus, it is thought to be effective to discriminate the individual difference in this ability level.

3.5 Practical estimation and evaluation sheet for the movement of Jumping-over and crawling-under

In the evaluation using IRT, a computer is necessary because it should be done by nonlinear optimization method or an adapted test (Aoyagi, 2005; Koch and Dodd, 1989) which conducts on each time the test proceeds using only test items making information highest but not all items is used. However, since Partial Credit model is belonged into Rasch model, more simple and practical estimation of θs can be implemented using a paper-and-pencil method without any computer based on the result found in this study. From such a perspective, Figure 7 is the practical estimation and evaluation sheet for the movement of Jumping-over and crawl-under developed based on Zhu and Cole (1996). This sheet is drawn using a Visual Basic program while accurately reflecting the values computed by the program. The lower part of this Figure indicates the chart which is exchanged an abscissa and ordinate in Figure 5 and added the band of standard error of estimate in Table 4 in the both side of the curve. The middle part of the Figure displays the step difficulty parameters plotted on the same scale. An open triangle mark is a step difficulty from "impossible" to

"no idea" and an open circle mark one from "no idea" to "possible." However, a filled triangle presents the values over the range of plus/minus 3.0. The upper part of the Figure shows the normal distribution of with the average and

Table 3. Conversion table of θ by score

Score†)	θ††)	Standard error
1	-2.190	1.064
2	-1.762	0.792
3	-1.487	0.676
4	-1.275	0.606
5	-1.100	0.558
6	-0.949	0.521
7	-0.816	0.492
8	-0.697	0.469
9	-0.587	0.449
10	-0.487	0.433
11	-0.392	0.420
12	-0.303	0.409
13	-0.218	0.401
14	-0.136	0.395
15	-0.056	0.391
16	0.023	0.388
17	0.101	0.388
18	0.179	0.390
19	0.259	0.394
20	0.341	0.400
21	0.425	0.408
22	0.514	0.419
23	0.608	0.433
24	0.709	0.450
25	0.819	0.471
26	0.941	0.498
27	1.078	0.533
28	1.238	0.580
29	1.433	0.650
30	1.690	0.768
31	2.097	1.045

†) The score is the sum of all items in cases in which "failure" = 0, "indiscernible" = 1, and "successful" = 2.

††) 0 and full marks (32 points in this study) cannot be estimated because of the use of the maximum likelihood

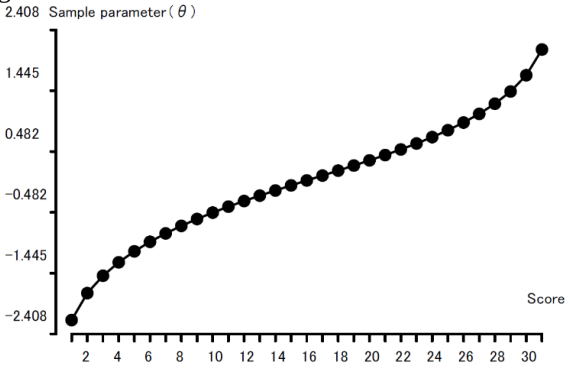

Fig. 5. The relationship between scores and θ.

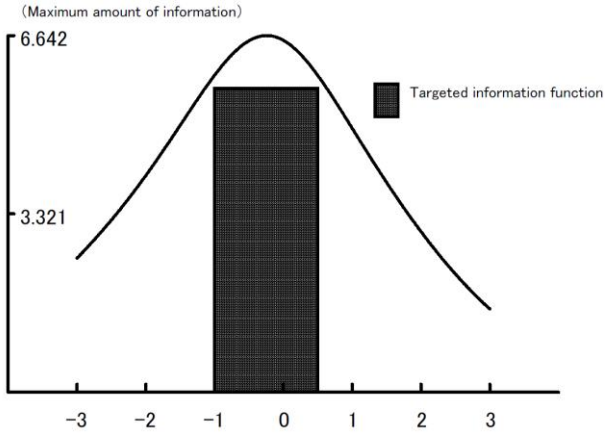

Fig. 6. Test information function.

standard deviation of computed θs by gender and age group. It is divided into three categories of A (excellent), B (normal) and C (poor) with the borderlines of the average plus/minus half of standard deviation.

First, in the evaluation item in the middle of the practical estimation and evaluation sheet, a checkmark is given to only an open triangle in the case of "no idea" or to both an open triangle and circle if applicable to "possible" (filled in the Figure). After all items are checked, the number of checkmarks, which becomes the total score, is counted. Next, after finding the corresponding total test score in the vertical axis of the lower part of the sheet, a horizontal straight line is drew from the spot and intersection points between the straight and three curves in the graph of the lower part of the sheet is detected. Then, three vertical straight lines are traced from the three intersection points to the normal distribution by gender and age in the upper part of the sheet. The two side lines out of three vertical lines indicates the band of standard error of estimation, which can be regarded an allowable range of judgment. In the parameters of step difficulty in the middle part of the sheet, it can be estimated as "possible" in the left side of the vertical lines and "impossible" in the right side of them within an allowable range. Among normal distributions by gender and age in the upper part of the sheet, the shift of approximate 0.2 in θ is found with advancing age and the gender difference of round 0.4 in θ is detected while boys are superior to girls in all age range. Additionally, 5-year-old girl group had larger variance than other gender and age groups. Taking these conditions of gender and age difference into consideration, relative evaluation can be conducted using this sheet.

3.6 Evaluation instance

Figures 8 and 9 demonstrates the examples of estimation and evaluation of 5-year-old boys with the total score of 20 points using this sheet. In case of Figure 8, there were two filled marks (attained) and three open mark (not attained) among three vertical lines. When considering that step difficulty is the spot two adjacent categories arise with the probability of 50%, it can be regarded to be a usual pattern to perform skills. However,
"Jumping with two feet" is an unusual pattern because although it should be expected to be attainable, it could not in reality ("Possible" is an open mark but not filled.). However, it is likely that it is not necessarily aberrant because parameters of this item themselves are not monotonous and a practical reference point for evaluation should be ambiguously put on around the middle spot.

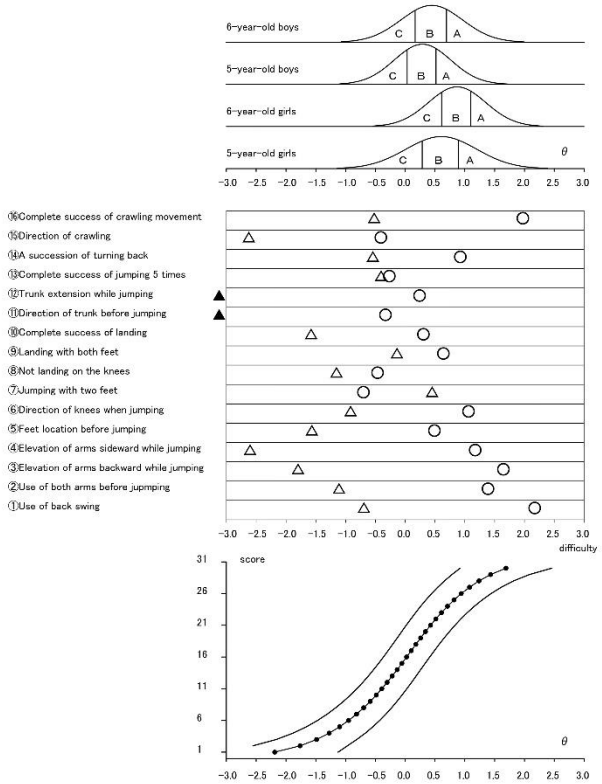

Fig. 7. Diagonal sheet.

In Figure 9, almost filled marks were located among or on the left side of the three vertical lines. Only "Not landing on the knees" is opened (not attained) but not filled even though it is on the left side of three lines. Although a child with $\theta = 0.341$ corresponding to a total score = 20 is naturally expected to be evaluated as "possible," he/she was not actually accomplished. Thus, it should be regarded to be aberrant due to some sort of cause. As the reason for this, it seems that the knowledge of fine movements and the experiences of these efficient movements unexpectedly happened to be insufficient in spite of sufficient readiness from the viewpoint of maturity. If experience of fine movements is inadequate, the instructions to improve their form and to increase the experience of these movements are needed. In such a case, it is induced that he/she can be possible to attain it easily.

3.7 Investigation of the result of practical evaluation

The fact that the number of aberrant patterns is in this practical method few indicates that a series of typical movement patterns which are related each other exist in mature (or immature) movements. In contrast, if there are many aberrant patterns, we will find various individual differences. Figure 10 shows the frequencies by number of aberrant pattern after actually counting them by each individual. Whereas 32 patterns (= 16 item x 2 categories) in total is capable of causing in theory, three per individual is the most and that of 12 or more was not found. In other words, approximately 10% of the aberrant patterns existed. This result allowed us that the traditional approach to understand motor development by investigating the changes from typical immature to mature patterns after categorizing them, and that there was about 10% of cases that the traditional procedure cannot cover.

Three kinds of the indexes of fitness (simple mean square, weighted mean square and standardized mean square) as well as the number of aberrant pattern produced by the practical method were computed. The result showed that no over-fitted values (0.7 or less) nor no under-fitted values (1.3 or more) was detected (Bond and Fox, 2007). As a result of computing the correlation coefficients of the number of aberrant patterns in the practical method to three kinds of the indexes of fitness, they were 0.798, 0.760 and 0.678, respectively. Inspecting the scatter plots (Figures 11 to 13) between the numbers of aberrant patterns derived from the practical method and three kinds of indexes of fitness confirmed that because of no outliers, the numbers of aberrant patterns roughly reflected the restrictedly computed indexes of fitness.

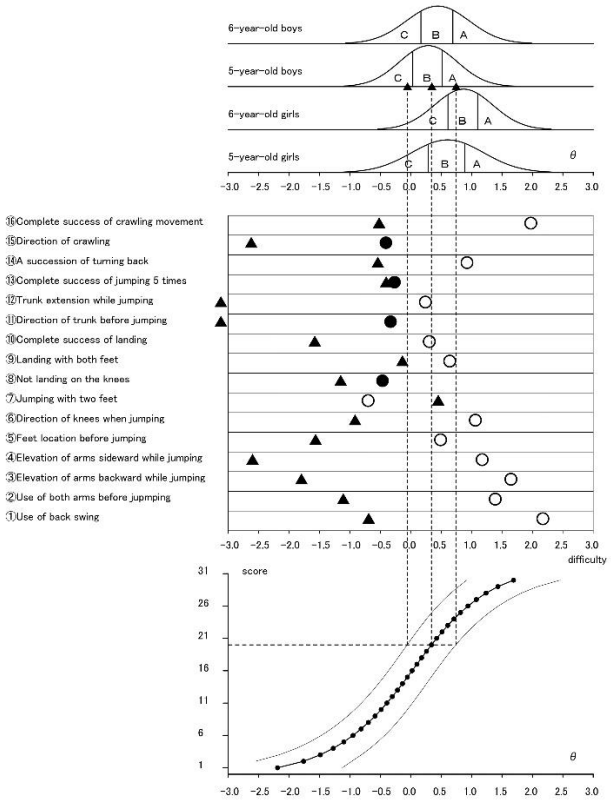

Fig. 8. Example of a diagonal sheet of a 5-year-old boy with a score of 20.

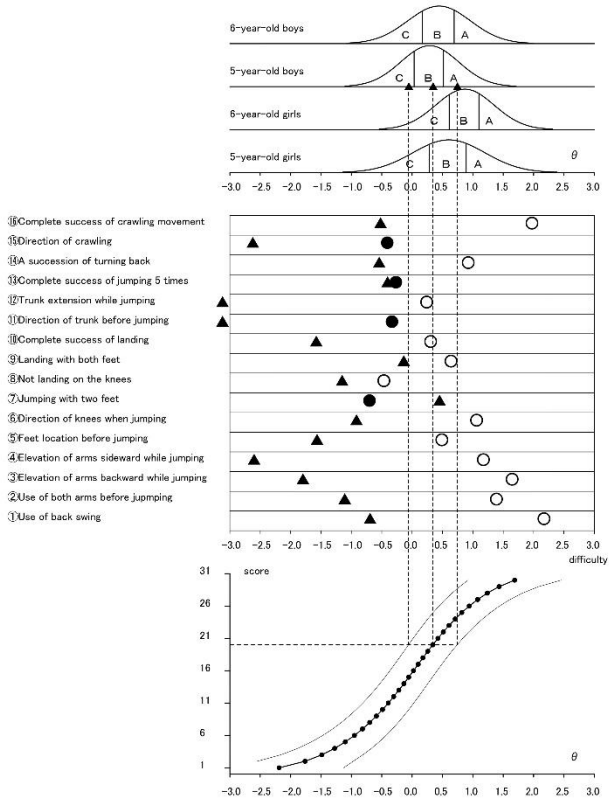

Fig. 9. Example of a diagonal sheet of a 5-year-old boy with a score of 20.

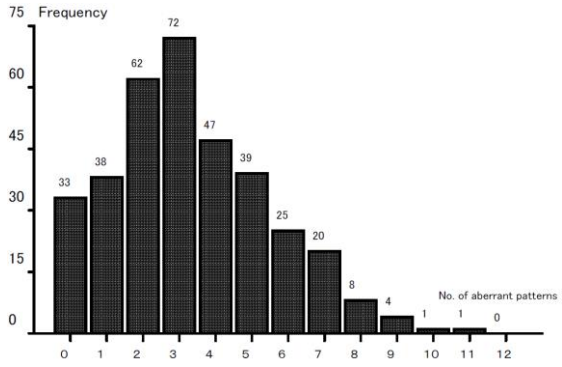

Fig. 10. Frequency of aberrant pattern.

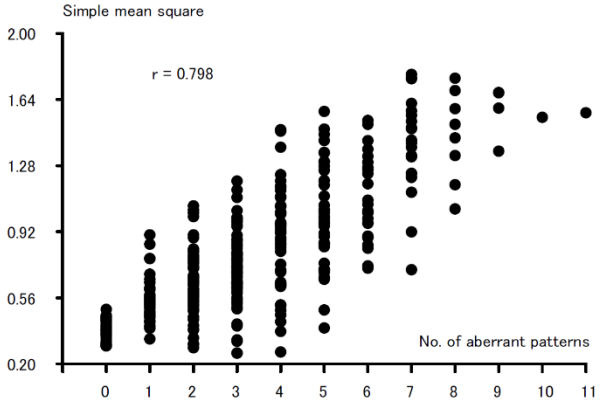

Fig. 11. Scattergram of the frequency of aberrant patterns and simple mean squares.

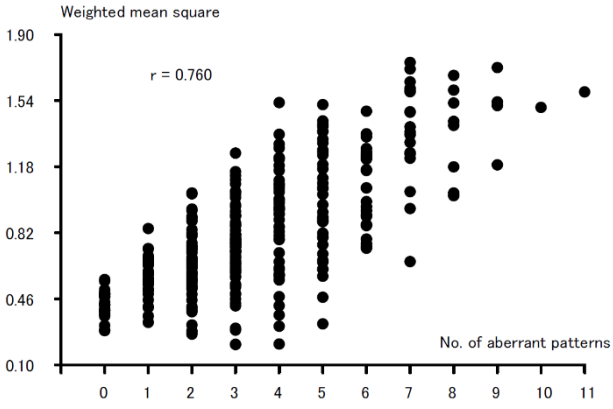

Fig. 12. Scattergram of the frequency of aberrant patterns and weighted mean squares.

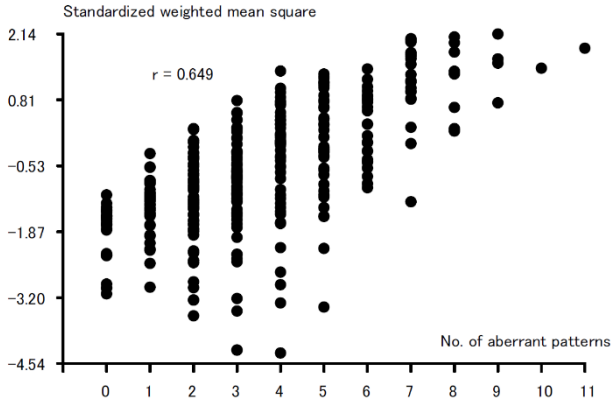

Fig. 13. Scattergram of the frequency of aberrant patterns and standardized weighted mean squares.

Author Contributions

Conceptualization, Y.O. and O.A.; methodology, Y.O. and O.A.; validation, Y.O. and O.A.; writing—original draft preparation, Y.O. and O.A.; writing—review and editing, Y.O. and O.A.; supervision, Y.O. and O.A.; project administration, O.A.; funding acquisition, Y.O. and O.A.; All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by grants-in-aid for scientific research from the JSPS Numbers 22K11695.

References: (APA Style)

Aoyagi, O. (1999). Investigation of subjective rating for jumping in childhood. *The reports presented at the 49th Japanese Sport and Physical Education Congress*: 424.

Aoyagi, O. (2003). A subjective rating of a 25m run in childhood. *Kyushu Journal of Physical Education and Sport, 17(1)*:1-8.

Aoyagi, O. (2005). Measurement of motor ability using Item Response Theory. Touka-Shobo: Fukuoka, pp.105-117.

Bond, T. G. and Fox, C. M. (2007). Applying the Rasch model fundamental measurement in the human sciences, 2nd edition, Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, pp. 309-314.

Coordination Committee (1976). Test manual and norm of coordination. *Report of Research Center in Physical Education, 4*: 207-217.

Coordination Committee (1981). Test manual and norm of coordination in Report of Research Center in Physical Education in final version: Report of coordination committee. *Report of Research Center in Physical Education, 9*: 207-212.

DeOreo, K. (1980a). Refining locomotor skills. In Corbin, C. B. (Ed.), A textbook of motor development, 2nd edition, Wm. C. Brown Company Publishers: Dubuque, IA, pp.59-66.

DeOreo, K. (1980b). Refining nonlocomotor skills. In Corbin, C. B. (Ed), A textbook of motor development, 2nd edition, Wm. C. Brown Company Publishers: Dubuque, IA, pp.67-70.

Gallahue, D. L. (1982). Understanding motor development in children, John Wiley & Sons, Inc.: New York, pp.207-210.

Hambleton, R. K. and Swaminathan, H. (1985). Item Response Theory: principle and applications. Kluwer Nijhoff Publishing: Boston, pp.10-13.

Harris, J., Laan, S. and Mossenson, L. (1988). Applying Partial Credit analysis to the construction of narrative writing tests. *Applied Measurement in Education, 1(4)*: 335-346. https://doi.org/10.1207/s15324818ame0104_5

Ikeda, H. (1989). Guidebook of statistics, Shin'yosha: Tokyo, p.101.

Katsube, A. (1971). Theory and practice of child physical education: Kyorin-shoin: Tokyo, pp.40-47.

Kim, S. and Matsuura, Y. (1988). A study on quantitative change and qualitative change of fundamental movement skills in children. *Japanese Journal of Physical Education, 33*: 27-38.

Koch, W. R. and Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using Partial Credit scoring. *Applied Measurement in Education, 2(4)*: 335-357. https://doi.org/10.1207/s15324818ame0204_5

Malina, R. M. and Bouchard, C. (1991). Growth, maturation, and physical activity. Human Kinetics Books: Champaign, IL, pp.178-183.

Masters, G. N. (1982). A Rasch model for Partial Credit scoring. *Psychometrika, 47*: 149-174.

Masters, G. N. (1984). Constructing an item bank using Partial Credit scoring. *Journal of Educational Measurement, 21(1)*: 19-32. https://doi.org/10.1111/j.1745-3984.1984.tb00218.x

Matsuura, Y. (1975). New physical education lecture 67: Developmental kinesiology. Shoyo-shoin: Tokyo, pp.141-142.

Matsuura, Y. (1982). Development of physical fitness. Asakura-shoin: Tokyo, pp.45-67.

Miyamaru, M. and Hirakoba, K. (1982). A study on development of motor coordination in ball-handling skill of young children (3): The development of throwing pattern and the effects of training on ball throwing. *Report of Research Center in Physical Education, 10*: 111-124.

Nakamura, K. (2001). Development of running movement in child by observational evaluation. Miyamaru, M. (Ed.) Development of running. Kyorin-shoin: Tokyo, pp.61-69.

Nakamura, K., Miyamaru, M. and Kuno, S. (2001). Development and evaluation of throwing pattern in young children. *Bulletin of Institute of Health and Sports Sciences, University of Tsukuba, 10*: 157-166.

Toyoda, H. (2002). Beginner's book of Item Response Theory: Science of test and measurement. Asakura-shoten: Tokyo, pp.124-126.

Wickstrom, R. L. (1977). Fundamental motor patterns, 2nd edition, Lea & Febiger: Philadelphia, pp.59-90.

Wild, M. R. (1938). The behavior pattern of throwing and some observations concerning its course of development in children. *Research Quarterly, 9*: 20-25. https://doi.org/10.1080/23267429.1938.11802445

Wilson, M. and Iventosch, L. (1988). Using the Partial Credit model to investigate responses to structured subtests. *Applied Measurement in Education, 1(4)*: 319-334. https://doi.org/10.1207/s15324818ame0104_4

Wright, B. D. and Masters, G. N. (1982). Rating scale analysis. Mesa Press: Chicago, pp.94-111.

Zhu, W. and Cole, E. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research Quarterly for Exercise and Sport, 67(1)*: 24-34. https://doi.org/10.1080/02701367.1996.10607922