

# Rasch Calibration of Perceived Weights of Different Sports Games

Sang-Jo Kang

*Department of Physical Education  
Korea National Sport University*

Minsoo Kang

*Department of Health and Human Performance  
Middle Tennessee State University*

In many countries, an athlete's performance at sporting competitions is often used as part of the selection criteria for entry into college. These criteria could be biased depending upon the procedures utilized by the authorities in a particular country. The purpose of this study was to calibrate, by using the Rasch rating scale model, the judge-based weights associated with the different quality of games. The participants ( $N = 202$ ) were asked to weight the quality of the gold medal of 10 sports games using a scale ranging from 1 to 100. In this study the judges' weights on the quality of the games differed among the sports games. The Olympic Games had the highest quality measure (logit = 7.69) whereas the National Sport Games had the lowest quality measure (logit = -3.73). The severity level (i.e., a tendency of a judge to rate higher or lower than other judges) was also different among judges, ranging from -6.73 (*less severe*) to 3.82 (*more severe*) logits. The results from a Rasch model analysis make it possible for a college board to identify the relative quality of games and to select from high school athletes more objectively.

Key words: weights, Rasch model, judge, sports games

In many countries, an athlete's performance at sports competitions/games is often used as part of the selection criteria for entry into college. The selection committees of college entrance boards have recognized the sensitive nature of using the

outcomes of various athletic performances as a criterion for acceptance into college. The selection of high school athletes, who have achieved success at different quality of games, has been controversial. The quality of the athlete's performance (i.e., winning first place, second place, etc.) is related to the game in which the successful performance occurred.<sup>1</sup> This has forced college entrance examination boards and sport-related organizations to employ various criteria for evaluating athletes' performances. As a result, many sport-related institutions and organizations (Canadian Academy of Sport Medicine, 2002; Federation Internationale de Football Association [FIFA], 2003; Federation Internationale de Volleyball [FIVB], 2003; International Badminton Federation [IBF], 2003; Korea National University of Physical Education, 2001, 2002) have suggested criteria for evaluating the quality of athletes' performances at different games.

Assigning different "weights" to athletic successes based on the quality of the games has been used as a criterion in many institutions and organizations. The Korea National University of Physical Education (2001, 2002) uses a 1-to-100 scale for the weights of different sports competitions with higher weight indicating greater quality (e.g., first place at the Olympic Games is given a weight of 100). The Canadian Academy of Sport Medicine (2002) suggested that greater weight should be given to success at an international game (15 points) than national/provincial (10 points) or local (5 points) games, and greater weight should be given to major games than a single competition, tournament, or tour.

An athlete's past performances (e.g., first place at the Asian Games) have been judged based on the weights assigned to a particular level of competition. For example, if an athlete who has had successful performances (e.g., first place finish at two local competitions and one international competition), the athlete would have a total weighted score of 25 (e.g.,  $2 \times 5$  points for local level +  $1 \times 15$  points for international level) using the system recommended by the Canadian Academy of Sport Medicine. This weighted score of 25 will be compared against the scores of the other applicants. The current practice is that the athlete with a total higher weighted score has a greater chance of acceptance into the university.

Though a wide variety of sports organizations and colleges assign different weights to the games, they all share similarities in the development process. The typical practice for establishing these weights for success at the games relies heavily on a few experts in the organizations or colleges. These content experts set the weights given to success at the various games. Expert judgment has played an important role in many circumstances, including test development (Zhu, Ennis, & Chen, 1998), content-related evidence of validity (American Educational Research Association, American Psychological Association, & Na-

---

<sup>1</sup>A game in this study refers to a sports-competition event such as the Olympics, not to a particular area of the competition such as skating or swimming.

tional Council on Measurement in Education, 1999), and setting standards (Glass, 1978). This is also true in determining weights for the different sports games.

Expert ratings typically have been analyzed by traditional approaches (e.g., the average of all experts' ratings). Unfortunately, this practice may suffer from some psychometric disadvantages and could lead to incorrect interpretations. First, expert's ratings are ordinal data but often incorrectly assumed to be interval data. Because ordinal scale data are not additive (e.g., the distance between rating scores of 4 and 5 may differ from the distance between scores of 9 and 10), the traditional approach is no longer a useful way to analyze the data.

In addition, experts' severity of rating and inconsistency in ratings in the decision-making process may threaten the validity of the judgments from experts. The term *severity* is used to describe how experts often rate higher or lower compared to other experts. Experts' severity may become a major threat when only a few experts are involved in determining standards or weights, where the same performance from an athlete may result in weights that differ considerably across experts. This is the current situation at many of the universities. The inconsistency in ratings from the experts is also a problem. For example, when a judge rates a certain performance a 5 and then rates an equal performance a 3 or a 7, this inconsistency becomes problematic (Lunz & Stahl, 1993; Zhu et al., 1998).

These psychometric issues, including the nonadditive feature of ordinal data as well as severity of raters and inconsistency in ratings, can be addressed in a Rasch model analysis (Rasch, 1960, 1980). Certain advantages are possible from the Rasch model approach over traditional approaches. A relatively simple Rasch model, known as a two-facet model, characterizes the relation between an examinee's underlying trait and a test item. The item and examinee parameter estimates are calibrated and placed along a continuum with a common metric (i.e., *logit*), which allows examining the relative positions of items and examinees (Zhu & Cole, 1996). A logit is a nonlinear function of the probability of obtaining a certain rating for an examinee for a given trait.

In the present application of the two-facet Rasch model, *item* refers to the various games and *examinee* refers to the judges. Item scores are the weights that the judges assign to the games. Because the Rasch model takes both facets into consideration at the same time in the logit, accounting for the item scores (i.e., weight of games), it is possible to identify the judges' inconsistency (i.e., a deviation from the judging pattern expected by the model) and judge-by-item "bias" (i.e., there is a pattern to unexpected responses).

A Rasch model has been successfully used with experts' judgment data. Lunz, Wright, and Linacre (1990) examined judge severity with three facets (examinee performances, items, and judges) and calibrated them using a many-faceted Rasch model; this is an extension of the Rasch model, which includes more than two fac-

ets. Looney (1997) applied the many-faceted Rasch model to analyze the judges' ratings from Olympic figure-skating competition. Zhu et al. (1998) used the many-faceted Rasch model to examine the severity and consistency of the experts' judgments in the Value Orientation Inventory-2, which assesses the value of physical education curriculum goals. Kang and Ahn (1999) examined the judges' severity in rating the quality of athletic performances in gymnastic competition using three facets: place (i.e., rank order of finish within a game), game, and judge. More research is needed to determine other applications of this statistical model. This study was designed to calibrate the judge-based weights for the different quality games using the Rasch model.

## METHOD

### Participants and Data Collection

A total of 202 participants from Korea were asked to serve as judges for this study. The participants consisted of 75 university athletes and 127 specialists. Athletes and specialists had participated on a regular basis from the following sports federations affiliated with the Korea Amateur Sport Association: archery, badminton, basketball, bowling, boxing, canoe, cycling, fencing, field hockey, gymnastic, handball, horse riding, judo, modern pentathlon, regatta, shooting, ski, swimming, table tennis, taekwondo, tennis, track and field, weight lifting, wrestling, and yacht. Prior to the data collection, verbal consent of the participants was obtained according to the local university research policy. The number of participants by sport is presented in Table 1. All the participants had experience as an athlete or coach at national or international levels.

The following 10 national and international games were selected for the participants to rate (a) Olympic Games, (b) World Cup Games, (c) Asian Games, (d) World Junior Competition, (e) Asian Championship Competition, (f) Universiade Games, (g) East Asian Games, (h) Asian Junior Competition, (i) International Sport Games delegated by each federation, and (j) National Sport Games. The participants were asked to rate the 10 national and international games; specifically, the participants rated the quality of the gold medal for each of the games. The gold medal or first-place finish represented the peak quality. A scale ranging from 1 (*lowest quality*) to 100 (*highest quality*) was used to rate the gold-medal performance at the various games. Though some sports are not included in each of the games (e.g., bowling is not included in the Olympic Games), judges rated each of the 10 national and international games listed. The participants' ratings were obtained at the Korea Amateur Sport Association Training Center over a 2-week period during the summer.

TABLE 1  
Numbers of Judges by Sports Federations

<i>Sports Federations</i>	<i>Judges</i>		<i>Total</i>
	<i>Athletes</i>	<i>Specialists</i>	
Archery	1	6	7
Badminton	3	4	7
Basketball	0	8	8
Bowling	0	3	3
Boxing	2	4	6
Canoe	2	5	7
Cycling	0	5	5
Fencing	1	6	7
Field hockey	2	6	8
Gymnastic	3	10	13
Handball	0	9	9
Horse riding	4	0	4
Judo	3	10	13
Modern pentathlon	3	2	5
Regatta	3	4	7
Shooting	0	4	4
Ski	3	1	4
Swimming	1	11	12
Table tennis	4	5	9
Taekwondo	4	7	11
Tennis	1	5	6
Track and field	25	0	25
Weight lifting	2	5	7
Wrestling	4	7	11
Yacht	4	0	4
Total	75	127	202

Data Analyses

The FACETS program (Linacre, 2002a) was used to analyze the judges’ weights for the 10 games. Two facets, games and judges, were evaluated in the analysis and the Rasch rating scale model was defined as

$$\ln [P_{nj k} / 1 - P_{nj (k-1)}] = D_n - C_j - F_k \tag{1}$$

where  $P_{nj k}$  is the probability of quality game  $n$  being awarded  $k$  category by judge  $j$ ;  $P_{nj (k-1)}$  is the probability of quality game  $n$  being awarded  $k-1$  category by judge  $j$ ;  $D_n$  is the level of quality game  $n$ ;  $C_j$  is the severity of judge  $j$ , and  $F_k$  is the difficulty of category step  $k$ . The higher the quality of games and the more severe ratings of

the judges, the higher the logit score. Because the FACETS program has a technical limitation whereby the program will not run with a range of scores for ratings greater than 100 or with ratings that include a decimal point, each judge's rating score was divided by 10 and rounded to the nearest integer (e.g.,  $87 / 10 = 8.7$ , which became 9).

First, a map of the distribution of games and judges was illustrated. The map displayed the relative position of judges over the 10 games. Then, the proper functioning of the rating scale was evaluated. Three evaluation criteria were used: (a) Was mean square residual appropriate for each category? (b) Did the average measure (i.e., a mean of logit measures in category) per category increase as the category score increased? and (c) Were category thresholds (i.e., boundaries between categories, also stated as the step difficulty values) ordered (Linacre, 1999, 2002b)? If the rating categories did not function properly, then that may suggest that one or more experts experienced problems when using the rating scale (Myford & Wolfe, 2003).

The model–data fit was evaluated by Infit and Outfit statistics for each game and judge in the Rasch analysis. Infit represents the information-weighted mean square residuals between observed and expected responses, and Outfit is similar to Infit statistics but is more sensitive to the outliers. Linacre (2002c) and Lunz et al. (1990) proposed a criterion for deciding acceptable and unacceptable fit whereby Infit and Outfit statistics with a value close to 1 are considered satisfactory model–data fit, and values greater than 1.5 or less than 0.5 are considered a misfit of model and data. Values greater than 1.5 indicated inconsistent performance (e.g., unexpectedly high or low quality ratings for a particular game across judges or for a particular judge across games), and values less than 0.5 showed too little variation (e.g., almost identical ratings for a particular game across judges or similar ratings by a judge across games).

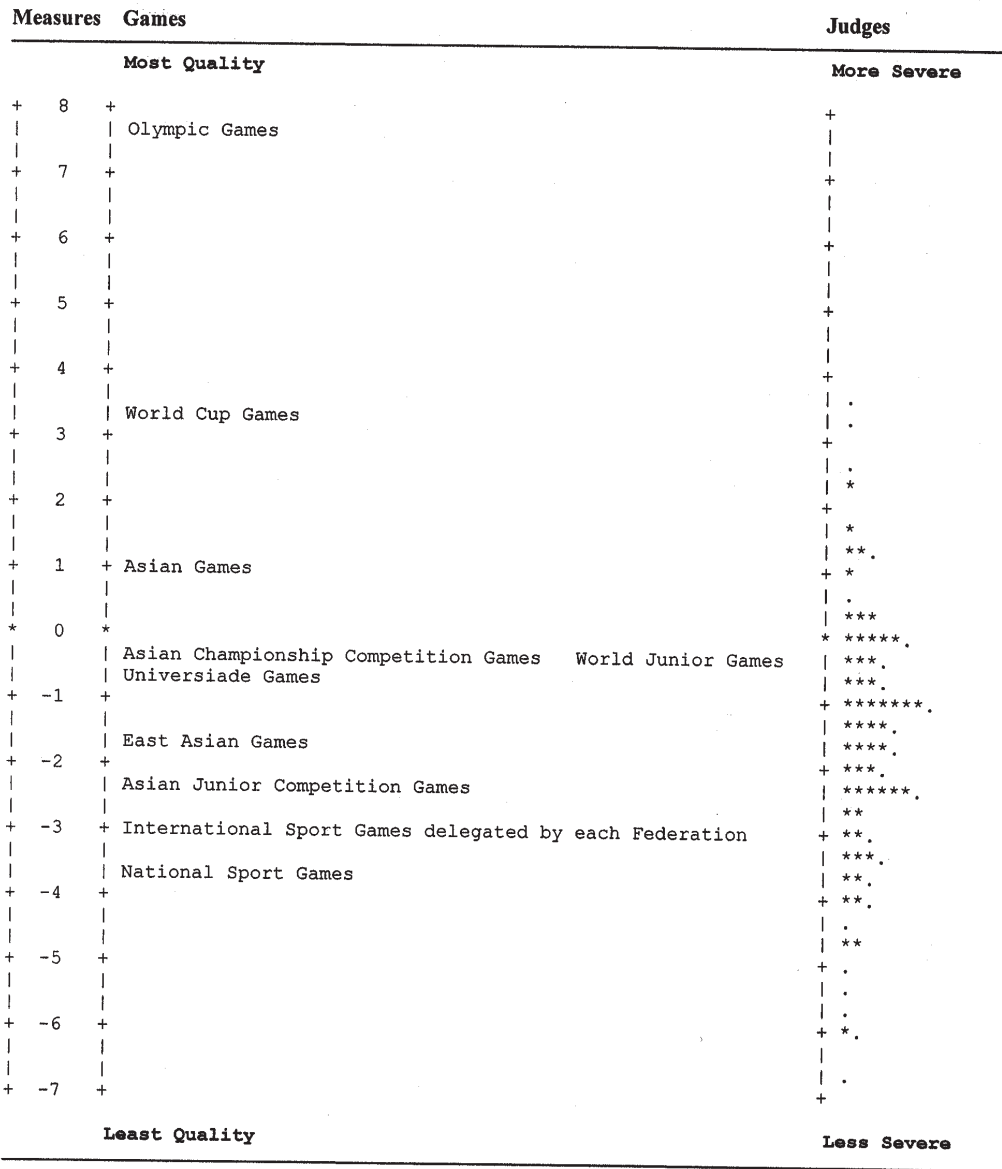
In addition to the model–data fit procedure, other statistics, including the separation and separation-reliability indexes and chi-square statistic were evaluated. The separation statistics indicate how well games and judges are spread along the measurement scale. Reliability of game separation indicates how confident one can be that the games would have the same respective order of calibrations within measurement error for another sample of judges from their respective sports affiliations. Reliability of judge separation indicates how confident one can be that the judges would separate along measurement scale into same strata within measurement error for another sample of games to which they assigned weights (Wright & Masters, 1982). A high separation index and separation-reliability statistics close to 1.00 indicate that there is a good discrimination for a facet along the measurement scale with a high degree of confidence (Fisher, 1992). The chi-square statistic was also used to identify whether the judges' ratings for the games were homogeneous (Linacre, 2001; Looney, 1997; Myford & Wolfe, 2003).

The judge's severity was estimated during the calibration process in logits. Because of the lack of sample size, the judge type (i.e., athletes and coaches/managers) and experts' affiliated sports federations were not treated as facets in the analysis, but rather, the judge's mean severity scores were compared by different judge type and experts' affiliated sports federations to examine how judge type and affiliation with sports federations are related to judge-severity measures. Finally, the comparison was made between the weights of the 10 games set by a few experts (Korea National University of Physical Education, 2002) and the Rasch analysis results. A comparison was made of the disparity between experts' weights and the weights from a Rasch analysis adjusted for the idiosyncrasies of 202 judges. The scale scores for the games from the Rasch analysis were converted to a 1-to-100 scale to compare to weights given by three experts associated with the Korea National University of Physical Education (2002) who used a 1-to-100 scale.

## RESULTS

A map of the distribution of judges and the 10 sports games is illustrated in Figure 1, which displays graphically the logit measures for the games and logit calibrations for the judges. The logit scale is shown on the left side of the map. On the middle of the map, games are located by their quality measures, and the histogram of asterisks and dots on the right side of the map represents the distribution of judge severity. The map reveals that the distribution of the games is well targeted to the judges. The majority of games fall within range of judge-severity calibrations. Both games and judges were dispersed across a broad spectrum of the logit measurement scale.

The data for the 202 judges were 100% complete. Out of 2,020 individual rating scores, 106 (5.2%) were flagged as being unexpected by the model. The unexpected marks were identified by examining the standardized residuals equal to or exceeding an absolute value of 2.00 (Looney, 2004). The World Cup Games had the most unexpected marks (22/106); 10 judges marked the World Cup Games higher than expected whereas another 12 judges marked these games lower than expected. The least unexpected marks (4/106) were found in the Asian Junior Competition Games; two judges marked the Asian Junior Competition Games higher than expected whereas the other two judges marked lower than expected. This indicates that the World Cup Games was the most troublesome out of the 10 sports games for the judges to weight. The unexpected mark for each judge helps to identify if there may be any judge bias for particular games. Though judges were flagged from 21 out of 25 federations, and more athletes ( $n = 40$ ) than specialists ( $n = 23$ ) were flagged with unexpected weights, there did not appear to be any obvious pattern in the unexpected marks.



Note: An asterisk represents 3 judges, while a dot denotes a single judge.

FIGURE 1 A map of the distribution of judges and the 10 sports games.

A summary of rating scale steps for 10 weight categories are reported in Table 2. Overall, the rating scale functioned well. The Outfit mean square residuals for all categories were within the range of 0.6 to 1.7. Considering that mean square values over 2.0 indicate more unexpected than expected randomness in choosing categories (Linacre, 2000b), the observed mean square residual for each category was appropriate. Because the higher categories are intended to reflect higher measures, a mean of logit measures in the category is expected to be increased. The average measure per category increased as the category score increased. Finally, step difficulty values (i.e., the category thresholds) were ordered as expected.

The judges' weight of the quality level of the 10 sports games, including calibrated logit scores with standard errors, Infit and Outfit statistics, and separation and separation-reliability statistics are reported in Table 3. Though the fit statistics were acceptable for all 10 games ( $M$  Infit  $MNSQ = 1.0 \pm 0.3$ ;  $M$  Outfit  $MNSQ = 0.9 \pm 0.3$ ), there was some unexpected noise related to the World Cup Games (Infit = 1.5) and the World Junior Competition (Infit = 1.5). This corresponds in part with some of the unexpected ratings mentioned earlier. The estimated quality level of the 10 games ranged from  $-3.73$  (lowest quality) to  $7.69$  (highest quality) logits, with a mean of  $0.00$  and standard deviation of  $3.23$ . The Olympic Games had the highest quality measure (logit =  $7.69$ ), followed by the World Cup Games (logit =  $3.41$ ) and the Asian Games (logit =  $1.13$ ). The National Sport Games had the lowest quality measure (logit =  $-3.73$ ). The game separation index and separation-reliability statistics were  $33.31$  and  $1.00$ , respectively. This was evidence for a good discrimination that the weights of the sports games displayed acceptable variability along the measurement scale with a high degree of confidence (separation reliability =  $1.00$ ) in replicating placement within measurement error for another sample of judges. The chi-square test of

TABLE 2  
Summary of Rating Scale Steps for 10 Weight Categories

Category Score	Counts Used	Average Measure	Outfit MNSQ	Category Thresholds
1	124	-3.72	1.2	low
2	113	-2.64	.9	-3.41
3	160	-1.89	.8	-2.47
4	178	-1.07	.6	-1.59
5	275	-.10	.8	-.80
6	221	.81	.9	.33
7	267	1.90	.9	1.22
8	272	3.15	1.0	2.53
9	195	5.54	1.0	4.54
10	209	8.91	1.7	7.14

Note. Average measure = a mean of logit measures in category; MNSQ = mean square residuals.

TABLE 3  
Summary of Quality Evaluation for 10 Sports Games

<i>Sports Games</i>	<i>Calibration Logit</i>	<i>SE Logit</i>	<i>Infit MNSQ</i>	<i>Outfit MNSQ</i>
Olympic Games	7.69	0.19	0.8	0.8
World Cup Games	3.41	0.11	1.5	1.3
Asian Games	1.13	0.09	1.2	1.2
World Junior Competition	-0.23	0.08	1.5	1.4
Asian Championship Competition	-0.49	0.08	0.8	0.8
Universiade Games	-0.78	0.08	1.0	1.0
East Asian Games	-1.66	0.07	0.6	0.6
Asian Junior Competition	-2.37	0.07	0.5	0.5
International Sport Games delegated by each federation	-2.97	0.07	0.8	0.8
National Sport Games	-3.73	0.08	1.1	1.1
<i>M</i>	0.00	0.09	1.0	0.9
<i>SD</i>	3.23	0.03	0.3	0.3

*Note.* Item separation = 33.31; Item separation reliability = 1.00; *SE* = standard error; *MNSQ* = mean square residual.

homogeneity,  $\chi^2(9) = 7,055.2$ ,  $p < .01$ , also indicated that the game quality differed among the 10 games; consequently, the judges weighted some of the games with higher quality than others.

The fit statistics for judges ( $M$  Infit  $MNSQ = 0.9 \pm 0.9$ ;  $M$  Outfit  $MNSQ = 0.9 \pm 0.9$ ) were not favorable. Thirty-eight percent of the judges (i.e., 77/202) had poor fit statistics. Most of the poor fit statistics were due to muting of responses (i.e., 54/77;  $MNSQ < .05$ ), which indicated similar ratings by a judge across games. The muted ratings do not alter the measurement continuum; however, they may result in underestimates of the standard errors, and thus, inflate the reliability estimates (Linacre, 2002c; Looney, 2004). The table of fit statistics for each of the 202 judges was not provided because of the lack of space. Judges' severities ranged from  $-6.73$  (*less severe*) to  $3.82$  (*more severe*) logits, with a mean of  $-1.57$  and standard deviation of 1.89. The judge separation index and separation-reliability statistics were 4.96 and 0.96, respectively. This indicated that the judges fell into seven distinct strata along the measurement scale with a high degree of confidence (separation reliability = 0.96) in replicating placement into these strata. The chi-square test of homogeneity,  $\chi^2(201) = 4,266.2$ ,  $p < .01$ , showed that there was a difference in the severity level among judges.

To examine the difference between judge's ratings and logit measures, the correlation coefficient between average judges' transformed raw scores on the 1-to-10 scale and logit measures was calculated and plotted in Figure 2. A strong negative

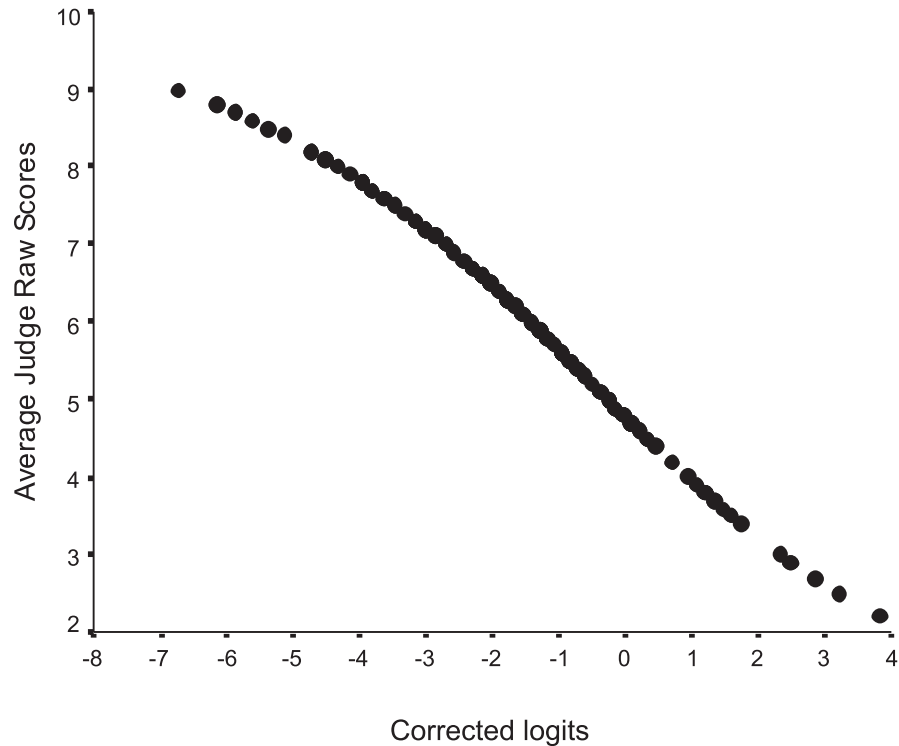


FIGURE 2 A plot of average judges' transformed raw scores (1–10 scale) against the logit measures.

correlation ( $r = -0.99$ ) was found between transformed raw scores ( $M \pm SD = 6.01 \pm 1.40$ ) and logit measures ( $M \pm SD = -1.57 \pm 1.90$ ). The plot shows that little variation of average judge raw scores was found corresponding to logit measures.

The judges' severity was compared by type of judge (i.e., athletes or coaches/managers) and experts' affiliated sports federations. The descriptive statistics for judges' severity scores by different judge type and experts' affiliated sports federations are summarized in Table 4. Athletes ( $M \text{ logit} = -1.54 \pm 1.90$ ) and specialists, including coaches and managers ( $M \text{ logit} = -1.63 \pm 1.89$ ), rated games in a similar manner. Regarding the experts' affiliated sports federations, the judges from the Badminton Federation gave the most severe ratings ( $M \text{ logit} = 0.38 \pm 2.13$ ), followed by the Bowling Federation ( $M \text{ logit} = -0.43 \pm 0.65$ ) and the Tennis Federation ( $M \text{ logit} = -0.48 \pm 1.37$ ). The judges from the Horse-Riding Federation gave the least severe ratings ( $M \text{ logit} = -3.22 \pm 2.06$ ), followed by the Regatta Federation ( $M \text{ logit} = -2.63 \pm 2.19$ ) and the Gymnastic Federation ( $M \text{ logit} = -2.47 \pm 1.90$ ). Note that there is a great deal of variability in judges' measures within federations.

TABLE 4  
Summary of Judges' Severity by Group

<i>Group</i>	<i>n</i>	<i>Judge Severity (in Logit)</i>	
		<i>M</i>	<i>SD</i>
Judge type			
Athletes	127	-1.54	1.90
Specialists	75	-1.63	1.89
Affiliated sports federations			
Archery	7	-1.06	1.85
Badminton	7	0.38	2.13
Basketball	8	-2.41	1.71
Bowling	3	-0.43	0.65
Boxing	6	-0.85	1.33
Canoe	7	-1.08	1.68
Cycling	5	-1.49	0.37
Fencing	7	-2.32	1.96
Field hockey	8	-2.05	1.54
Gymnastic	13	-2.47	1.90
Handball	9	-1.58	2.55
Horse riding	4	-3.22	2.06
Judo	13	-0.88	1.81
Modern pentathlon	5	-1.92	3.05
Regatta	7	-2.63	2.19
Shooting	4	-1.86	0.66
Ski	3	-1.25	0.75
Swimming	12	-1.90	2.08
Table tennis	9	-1.20	1.50
Taekwondo	11	-0.94	2.40
Tennis	6	-0.48	1.37
Track and field	26	-1.93	1.93
Weight lifting	7	-2.32	1.66
Wrestling	11	-1.17	1.17
Yacht	4	-0.89	1.46

## DISCUSSION

The various international sports institutions and federations have provided different weights for the various games according to the sport (FIFA, 2003; FIVB, 2003; IBF, 2003; Korea National University of Physical Education, 2001, 2002). The weights of different games contribute to establishing a comparison across the various national and international games. However, there are differences in judges' severities with respect to comparing the quality of different games. Comparisons of the performances achieved from different games are difficult to make unless the

same scale is applied to evaluate the different performances. A major benefit of applying a Rasch model analysis to examine experts' judgment for the weight given to different sports games is that the weights can be adjusted for the judges' idiosyncrasies in assigning weights (Zhu et al., 1998). This study was designed to calibrate, by using the Rasch rating scale model, the judge-based weights associated with the different quality of games. The results from the Rasch model analysis showed that the judges' weights on the quality of the games differed among the sports games.

An unexpected result from this study was the number of misfit judges; 38% of judges had Infit and/or Outfit statistics lower than .5 or greater than 1.5. There are a few possible explanations for this large number of misfit judges. There could be systematic differences among the judges in the relative weights the judges gave to the various games. Another possibility is that the experts were using the rating scale or weights in a different way (e.g., response set or bias). Small sample size specifically with some sport federations (e.g.,  $n = 3$  for Bowling and Ski Federations) could have also contributed to the finding. Further investigation with larger sample size focused on the judges' characteristics and perceived importance of various games is needed for Rasch modeling studies.

The scale measures derived from Rasch analysis can be used to compare the quality of the various national and international games. The logit measures often are not convenient for the reporting of results to athletes, managers, and administrators. Wright and Stone (1979) suggested a useful transformation method, which is to take the high and low logit measures on the vertical scale and rescale the logit measures to measures that range from 1 to 100. In other words, the highest quality game is assigned 100 and the other games are transformed linearly. The weights of 10 games derived from the transformation method are summarized in Table 5. For more information regarding the computation, see Wright and Stone.

The weights of the 10 games assigned by faculty and administrators from the Korea National University of Physical Education (2002) and from the Rasch analysis were compared (see Table 6). Only a few specialists ( $n = 3$ ) from the Korea National University of Physical Education were involved in the process of setting weights of the 10 games. The results showed that the derived weights from the Rasch analysis were different from those of these few experts. The ordering and weights for the top three and the bottom two games were similar for the three experts from Korea National University of Physical Education (2002) and Rasch analysis results. Differences in ordering and weights occurred for the other five games. Three experts, for example, believed that the World Junior Competition Games is the sixth highest in quality out of 10 sports games, but the weights from the Rasch analysis indicated that this competition is the fourth highest. Because the derived weights from the Rasch analysis are adjusted for judge idiosyncrasies, they may be fairer (if model-data fit is satisfactory) than the current criteria employed in the Korea National University of Physical Education (2002) set by a few experts.

TABLE 5  
The Weights of 10 Sports Games (Highest 100 Point and Lowest 1 Point)

<i>Sports Games</i>	<i>Calibration Logit</i>	<i>Rescaled Weight</i>
Olympic Games	7.69	100
World Cup Games	3.41	63
Asian Games	1.13	43
World Junior Competition	-0.23	31
Asian Championship Competition	-0.49	28
Universiade Games	-0.78	26
East Asian Games	-1.66	18
Asian Junior Competition	-2.37	12
International Sport Games delegated by each federation	-2.97	7
National Sport Games	-3.73	1

*Note.* Rescaled weights were rounded to the nearest integer.

TABLE 6  
Comparison of Weights Determined by Three Experts and Rasch Analysis

<i>Sports Games</i>	<i>KNUPE<sup>a</sup></i>	<i>Rasch Analysis</i>
Olympic Games	100	100
World Cup Games	70	63
Asian Games	50	43
Universiade Games	50	26
Asian Championship Competition	30	28
World Junior Competition	20	31
East Asian Games	20	18
Asian Junior Competition	20	12
International Sport Games delegated by each federation	10	7
National Sport Games	3	1

*Note.* KNUPE = Korea National University of Physical Education.

<sup>a</sup>Weights employed in Korea National University of Physical Education (2002) set by three experts.

Though using perceived weights of past first-place success at sport games is one way of deciding which athletes will be admitted into college, the weighting method does, however, have some limitations. The weights of national and international games were based only on receiving a gold medal. The problem is that the weights for places (i.e., rank-order within a game, e.g., first, second, etc.) might be different among the raters (i.e., athletes and specialists) and even among different games. To further compare the athletes' performances at the different quality

games, researchers may need to examine the experts' ratings on various placements rather than just on the gold medal. For example, an athlete who finishes in sixth place at the Olympic Games probably should receive a higher score for that performance than an athlete who finishes in first place at a local game. Emphasis should also be focused on not only perceived quality but also actual athletic performance. Because a sprinting athlete who finishes in first place at a local game may be faster and have a lower time than an athlete who finishes in first place at a national game, the scores given to the slower athlete with the higher time at the perceived more important game would be lower. Relying on finishing in first place is heavily dependent on the sample of athletes. Further research should focus on incorporating perceived quality and actual physical performance to make a better decision on entrance into college.

In conclusion, the Rasch model provides a fairer and more interpretable scale for national and international games. Judges may differ significantly in their severity, and it is crucial to model this difference. Using the Rasch model, the judges' severity is modeled allowing for better evaluation of the games. The practical application of the Rasch model analysis of games provides college boards more information to better identify potential athletes for entrance into college based on their previous success at games.

## REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Canadian Academy of Sport Medicine. (2002). *Selection committee report*. Retrieved December 13, 2003, from <http://www.casm-acms.org/Committees/selectioncriteria.htm>
- Federation Internationale de Football Association. (2003). *The FIFA/Coca-Cola world ranking: Overview of basic principles and method of calculation*. Retrieved December 13, 2003, from <http://www.fifa.com/en/mens/statistics/rank/procedures.html>
- Federation Internationale de Volleyball. (2003). *FIVB world ranking*. Retrieved December 13, 2003, from <http://www.fivb.ch/EN/Volleyball/Rankings/Rankings.htm>
- Fisher, W. (1992). Reliability statistics. *Rasch Measurement Transactions*, 6, 238.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237–261.
- International Badminton Federation. (2003). *IBF world ranking system* (Appendix 6). Retrieved December 17, 2003, from <http://www.intbadfed.org/Portal/documents/app6wldrnk.doc>
- Kang, S. J., & Ahn, E. J. (1999). Objectivity of gymnastic performance judgment: Application of many-facet Rasch model. *The Korean Journal of Physical Education*, 38, 641–650.
- Korea National University of Physical Education. (2001). *Development plan of Korea National University of Physical Education*. Seoul, Korea: Author.
- Korea National University of Physical Education. (2002). *Athletic performance evaluation*. Seoul, Korea: Author.
- Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3, 103–122.

- Linacre, J. M. (2001). *A user's guide to FACETS: Rasch Measurement Computer Program* [Computer software manual]. Chicago: MESA Press.
- Linacre, J. M. (2002a). *FACETS* [Computer program, version 3.4]. Chicago, IL: MESA Press.
- Linacre, J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85–106.
- Linacre, J. M. (2002c). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.
- Looney, M. (1997). Objective measurement of figure skating performance. *Journal of Outcome Measurement*, 1, 143–163.
- Looney, M. (2004). Evaluating judge performance in sport. *Journal of Applied Measurement*, 5, 31–47.
- Lunz, M. E., & Stahl, J. A. (1993). The impact of rater severity on person ability measure: A Rasch model analysis. *American Journal of Occupational Therapy*, 47, 311–317.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity of examination scores. *Applied Measurement in Education*, 3, 331–345.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–421.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rasch, G. (1980). *Probabilistic models for some intelligence and achievement tests* (Expanded Ed.). Chicago: University of Chicago Press.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Zhu, W., & Cole, E. L. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research Quarterly for Exercise and Sport*, 67(1), 24–34.
- Zhu, W., Ennis, C. D., & Chen, A. (1998). Many-faceted Rasch modeling expert judgment in test development. *Measurement in Physical Education and Exercise Science*, 2, 21–39.