Sample Distribution and Research Design Are Methodological Dilemmas When Identifying Selection and Using Relative Age as an Explanation of Results

thesportjournal.org/article/sample-distribution-and-research-design/

U.S. Sports Academy March 29, 2018

Authors: Torsten Buhre and Oscar Tschernij

Corresponding Author:
Torsten Buhre, PhD
Department of Sport Sciences
Malmö University
20506 Malmö
Sweden
+46-40-665845
torsten.buhre@mah.se

Torsten Buhre is the senior physiologist at the Department of Sport Sciences at Malmö University

Sample distribution and research design are methodological dilemmas when identifying selection and using relative age as an explanation of results

ABSTRACT

The use of a statistical test, such as the chi-squared test, to determine if selection has occurred within a sport has been used frequently in research. The assumed distribution of a sample could influence the occurrence of significant outcomes. The occurrence of significance is generally interpreted as RAE and explained as a result of selection within the sport. Most studies in this field have been done using a cross-sectional design. Therefore, the purpose of this study was to examine the influence of using different types of distribution when testing for significance, in swimming as an example, over a nine-year period of six cohorts in age by gender groups.

Results show that using either an assumed uniformed distribution or a proportional distribution of the national population distribution will lead to an increased number of significant results, in comparison to using either a distribution of the actual sample of the specific age by gender group or the distribution of the previous year within the age by gender group. In addition, when using a longitudinal design over a nine-year period, the occurrence of significance decreased over time. In order to interpret significant results as a consequence of selection within a sport the use of a sport specific and age by gender distribution and a longitudinal design is proposed.

Keywords: Chi-squared test, selection, relative age effect, distribution, research design

INTRODUCTION

The problems arising from grouping children into age groups, with specific cut-off dates has been studied in both the school setting [1, 2] and in sports [3, 4]. Depending upon the setting, the age cohorts vary in age-span [5, 6]. However, the use of a single age-group depending upon the cut of year set by the sport federation is often used in sports [7].

In sports the term relative age difference has been defined as the relative difference in age (months) within an age-group [8, 9]. The effect of this relative difference in age results in an asymmetrical distribution in the number of participants or player over the four different quartiles in specific age-groups. This is known as the relative age effect (RAE). Generally, a chi-squared test is used to see if the probability of the sample distribution shows significance [10].

Cobley et al. [4] concluded that the use of an assumed uniformed distribution (AUD) would not impact the possibility of detecting a RAE in a sample. However, samples tested are often drawn from a national population that does not have a uniformed distribution. Delorme and Raspaud [11], investigating the French basketball population in 2006, showed that the national population distribution (NPD) deviated from the AUD. In their study of age cohorts, age 7 through 17 among both girls and boys (n=22), 15 cohorts had the highest number of born children in the third quartile and 19 cohorts had a higher number being born in the fourth quartile in comparison to the first quartile, in the national population.

One of the possible explanations of RAE is a selection bias of more physically mature children within sporting systems due to the competitive nature of sports [12]. In order to verify this explanation numerous studies have measured differences in both anthropometric variables, such as height and weight [7, 13, 14, 15, 16] and also performance variables [7, 14, 15, 16], with contradicting outcomes. In addition, most RAE studies using selection bias as an explanation for RAE [3, 4, 5, 7, 13 14, 15, 16] have used a cross-sectional design. When using this type of design, there is an assumption that the distribution of number of athletes within each age-group is uniformed across all age-groups within the sport [4], which is not necessarily true.

Selection to a national team, regional teams, or the next year´s team is created by selecting a sample from the population of athletes within a sport. Therefore, as Delorme and Raspaud, [11] and Delorme et al. [17] pointed out, it is important to identify the population distribution of athletes within the specific sport and age group, in order to detect if a bias in

the selection has occurred. If this distribution already is asymmetrical, using either an AUD or NPD could impact the statistical outcome. Thus, leading to a hasty conclusion that a selection bias has occurred.

In Swedish sports in general, 10-year age groups have the highest number of participants in sports, for both boys and girls. Delorme and Raspaud [11] reported expected frequencies in different quartiles from the age of 7 to 17, based on information from the National Institute of Statistics and Economics (INSEE). Using this information and additional information from INSEE on the total number of children born in France during each of the investigated years, the proportion of the national population within an age group was determined. During the ages of 7 through 10, the relative proportion of participants increased from 1.45% to 3.65%. From age 10 through age 14 it was relatively stable, varying between 3.30% and 3.83%, at which point it declined to 1.97% of the national population at age 17. Thus, there is a temporal variation in the size of the specific sport population (SSP) in relation to the national population. If the intent is to discover if selection bias occurs it is important to identify if the sample is drawn from an increasing population or a decreasing population, in order to draw the right conclusion. When the population is increasing the recruitment into the sport is higher than the elimination, from the sport, and vice versa. A point of saturation of a population occurs when the rate of recruitment and elimination is equal, creating the highest number of participants within a sport specific age-group in relation to the national population. The age when this occurs can be defined as the age of saturated specific sport population (ASP). This takes into account that athletes eliminated in selection procedures, could possibly still participate in a sport, and thus has a chance to be selected in the future. Therefore, the notion of identifying bias selection also has a temporal aspect.

The distribution of ASP, is the distribution in the number of participants in quartiles of the specific year. The researchers term this distribution as the saturated sport distribution (SSD). This distribution can deviate both the NPD as well as from AUD. In order to identify ASP of an age-group in a particular sport, a longitudinal design must be used. Prior to ASP the rate of selection or self-elimination is surpassed by the rate of recruitment and engagement into the sport. For post-ASP, the mechanism is reversed.

Therefore, the purpose of this study compared the occurrence of significance pre- and post-the year of ASP, based on participation in competition when using four different types of distribution of a sample, i.e. assumed uniformed distribution (AUD), national population distribution (NPD), saturated sport distribution (SSD) and the distribution of the previous year within the age-group distribution (PYD) over a nine-year period between the ages of 8 and 16 using a database from the Swedish Swimming Federation as an example.

METHODS

Swedish swimming has had an online registration system of swimming results since early 2000. This has allowed the federation to file all competitive results from local, regional, and national competitions in databases. These databases are connected to individual swimmer ´s license number. The license number is personalized and contain information about birthdate in year, month, and day, and could therefore be used to identify what quartile the swimmer was born. The authors acquired access to databases from age cohorts born in 1998, 1999, and 2000 from the Swedish Swimming Federation. Each line in the database

included variables describing unique individual performances containing: license number, date of performance, event swum, race-time, and location. The variable of interest was license number only. Some of the license numbers were not identified with numbers but with XX-XX-XX. These could not be labelled to a specific quartile, thus, a number of individuals had to be excluded from the population database. Since exclusion occurred, the analysis is based on a sample of the population for each age group-cohort. This made it possible to use inferential statistics on each sample [10]. The cohorts were dived into gender and coded into birth quartiles. The following age-group by gender sample abbreviations were used; girls born 1998 (G98), born 1999 (G99), born 2000 (G00), boys born 1998 (B98), born 1999 (B), and born 2000 (B00). To gain a longitudinal perspective the statistical analysis included results from the age of 8 to the age of 16 for all age by gender groups. The span in years will probably include ASP. Thus, making it possible to compare three specific gender groups at the same time over a nine-year span during part of a competitive career. The number of participants in the total population was elusive, since no differentiation could be made between the individuals labeled XX-XX-XX. The exclusion from the population in order to create a sample was thus interpreted as a random sample. For each age-group sample by gender between 106,262 (B98) and 173,384 (G99) competitive results were coded. Deleted results, as represented by an unidentifiable birth data ranged between 18.0% and 25.3% for the different age by gender groups. For a complete overview of number of participants included at different stages in the samples and results excluded from the samples, see Table 1.

Table 1: Descriptive data for Swedish swimming in age by gender groups

	G98	G99	G00	B98	B99	B00
Total results (n)	145,901	173,384	168,395	106,262	119,341	110,301
Excluded results (%)	25.3	22.0	18.0	22.9	23.5	19.3
# of different participants	760	903	981	521	598	681
Year one (n)	105	117	147	71	68	96
ASP	Age 12	Age 12	Age 12	Age 11	Age 13	Age 12
(n) at ASP	460	554	611	291	329	367
% at ASP	60.5	61.4	62.2	55.9	55.0	54.9
Year nine (n)	157	236	247	164	187	189

Total results (n) are all results related to both an identifiable an unidentifiable ID
Proportion of results excluded from dataset, because they were unidentifiable.
Number of different athletes with unique identifiable athletes within an age by gender group
Year one (n) is the number of athletes included in the first-year age by gender group
Age of ASP based on sample
Number of participants at ASP
Proportion of participants at ASP in relation to the total number of athletes identified in the database
Year nine (n) is the number of athletes included in the ninth-year age by gender group

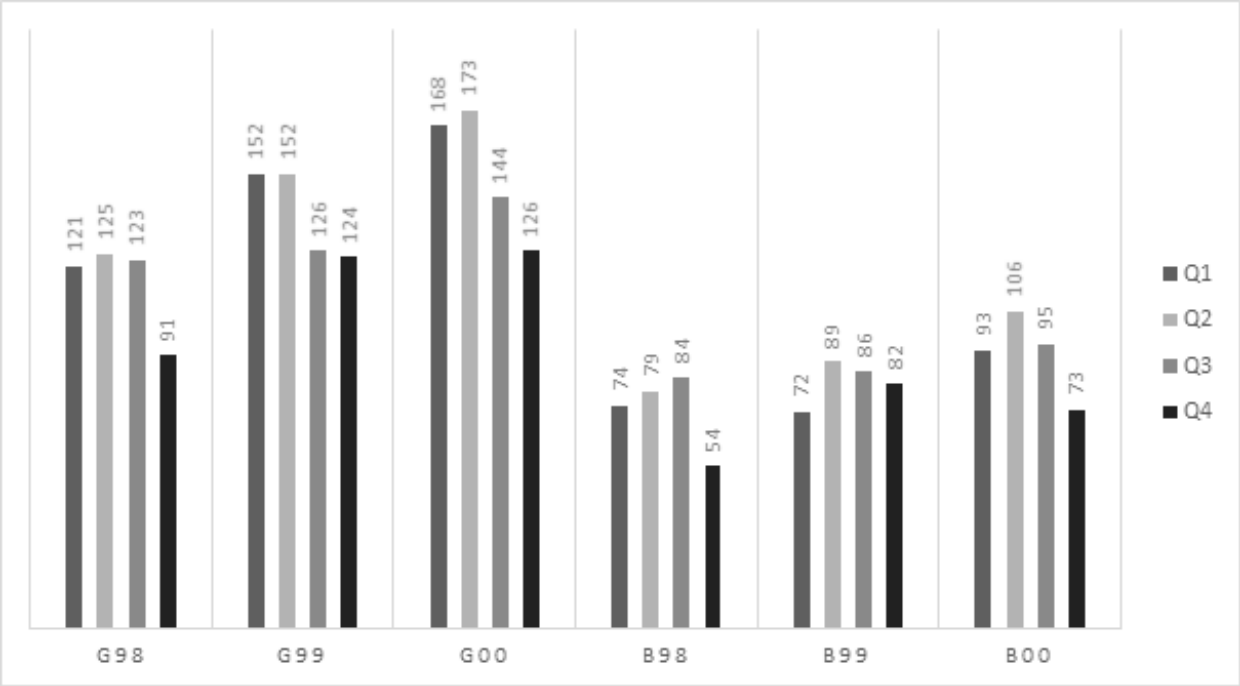
For calculation of the NPD, information was collected from the Swedish Census Bureau (SCB). The ASP was identified as the year with the highest number of registered identifiable swimmers with competitive results. Since some swimmers were excluded from this group, the construct of ASP was replaced with age of saturated sport specific sample

ASS. The chi-squared test, using goodness of fit was used to analyze each measure, i.e. the different distributions for each of the nine years, resulting in testing four different measures (AUD, NPD, SSD and PYD) each competitive year, in each age by gender sample, in total 216 statistical tests were performed. Level of significance was set a priori to $p<0,05$.

RESULTS

Of the six age-group samples investigated, four had an ASS at 12 years of age (G98, G99, G00, and B00). B98 reached ASS at the age of 11, and B99 at the age of 13. The distribution of the age-group by gender samples is depicted in Figure 1.

Figure 1: Distribution of participants in quartiles in age by gender groups at the age of saturation of the population (ASP).



For G98 the AUD was 115 and NPD in quartiles was 116 (Q1), 123 (Q2), 123 (Q3) and 98(Q4)
For G99 the AUD was 139 and NPD in quartiles was 140 (Q1), 149 (Q2), 149 (Q3) and 118(Q4)
For G00 the AUD was 153 and NPD in quartiles was 168 (Q1), 173 (Q2), 144 (Q3) and 126 (Q4)
For B98 the AUD was 73 and NPD in quartiles was 73 (Q1), 76 (Q2), 77 (Q3) and 64 (Q4)
For B99 the AUD was 82 and NPD in quartiles was 83 (Q1), 90 (Q2), 85 (Q3) and 72 (Q4)
For B00 the AUD was 92 and NPD in quartiles was 93 (Q1), 98 (Q2), 94 (Q3) and 83 (Q4)

The results are reported by age-group sample and divided into two parts, pre- and post-ASS. For G98 pre-ASS (age 12) the AUD and NPD showed significance for the first four years, SSD showed significance for the first three years, but for PYD it only showed significance for the second year. The sample distribution was in quartiles, thus deviated using three different distributions, AUD, NPD, and SSD, from age 8 through 11. However, the actual distribution did not deviate from PYD during the third, fourth, and fifth year, when the number of identifiable swimmers increased from the second year (253) to the fifth year

(450). The additional swimmers were distributed proportionally in the four quartiles in relation to PYD. Post-ASS, from year five through nine no significance was found using either of the four different distributions.

For G99 a similar pattern occurred with AUD and NPD showing significance the first three years of competition participation. SSD and PYD showed significance only the second year. The age by gender group reached ASS year five (age 12), increasing from 117 (age 8) to 554 (age 12), and then decreasing to 236 (age 16).

G00 showed significance for AUD and NPD the first five years, including ASS (age 12). Thus, SSD deviated from both AUD and NPD at ASS. SSD showed significance the second year and PYD showed significance the third year. Post-ASS no significance was found as the sample decreased in size from 511 (age 12) swimmers to 247 at age 16.

For B98, ASS was reached at age 11. AUD showed significance at all ages (8, 9, and 10 years of age) prior to ASS. NPD showed significance at ages 8 and 9, but not age 10. Both SSD and PYD did not show significance at any of these ages, including age 11, the year of ASS. Post-ASS, AUD showed significance in year nine, when the number of participants had decrease from 296 (year four) to 164 (in year nine). However, the distribution was skewed towards quartile three and not quartile one.

B99 only had one measure of significance, and this was in year one, when examining SSD. ASS was reached at age 13 (n=329) and decreased to 187 in year nine (age 16).

Similar to G98, G00, and B98, B00 showed significance on the measure of AUD on all years leading up to ASS, which occurred at age 12. The measure of NPD also showed significance in the years leading up to SSD, with the exception of year four (age 11), the year prior to ASS. SSD showed significance years one and two (age 8 and 9), whereas PYD only showed significance year two. The sample decreased from 374 (age 12) to 189 participants (age 16).

DISCUSSION

The purpose of this study compared the occurrence of significance when using different types of distributions. When utilizing theoretical distribution such as AUD and NPD, the possibility of getting significant outcomes using the chi-squared test goodness of fit increased in comparison to using distributions of the actual sample. Out of 24 tests done of each distribution prior to ASS, AUD had 18 outcomes, NPD had 16 outcomes, SSD had 8 outcomes and PYD had 4 outcomes that showed significance ($p<0.05$). In all of the outcomes where both SSD and PYD showed significance, both AUD and NPD also showed significance. The researchers interpret this as using a none-specific or assumed distribution could lead to the conclusion that selection occurred. Since the prevalence of significance in test using these two distributions out-number (18 and 16 versus 8 and 4) the occurrence of significant outcomes decreases when applying both to an actual distribution of the sample and the temporal aspect. Thus, allowing changes over time to influence how the sample was distributed over the four quartiles in each gender by age group. The assumption that participation in a sport (based on age by gender) is equal between quartiles, or is proportional in relation to the national population of the age by gender group is therefore questioned [11, 17]. The assumed distributions of AUD and NPD, do not take

into account the specific sample in question and does not allow for the flexibility to let time influence the changes within the participation rate of the sample in relation to the sport and level under investigation. This is strengthened by the fact that there were five incidences where both the measure of AUD and NPD were significant when neither the SSD nor PYD showed significance. These incidences occurred in a random pattern. These incidences occurred at the following year of participation in these age by gender groups, G00-year five, G98 and G00-year four and in addition G99 and B00-year three (see Table 2). Thus, the occurrence of significance in AUD and NPD can thus be interpreted as random events that probably occurred due to difference between the assumed distribution of the sample and the actual distribution of participants in the sample (SSD and PYD). At early ages, children ´s sport participation is influenced by their parent(s). They chose for their child when to start participating in the competitive sports, based on their collective perception of the child´s readiness to do so. This is only one factor that seems to be of importance. There are several other factors influencing parents to make the choice for their children to participate in competitive sports [18] on a micro level. On a meso-level a factor of importance is the clubs view on when a child should be allowed to enter competition. This factor is also based on the perception of the child's readiness, but from the different perspective. Table 2: Occurrence of significant outcomes on dataset, using for different types of distributions from the first year of competition (age 8) until the fifth year (age 12) when ASS was reached

	AUD					NPD					SSD					PYD			
Age by gender	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th
G98	*	*	*	*	ASS	*	*	*	*	ASS	*	*	*		ASS		*		
G99	*	*	*		ASS	*	*	*		ASS		*			ASS		*		
G00	*	*	*	*	ASS*	*	*	*	*	ASS*		*			ASS			*	
B98	*	*	*	ASS		*	*		ASS					ASS					ASS
B99 #											*								
B00	*	*	*	*	ASS	*	*	*		ASS	*	*			ASS		*		

*significant outcome ($p<0.05$) ASS= age of saturated sample of the age by gender group based on identifiable data points. # ASS was reached in B99 at the age of 13, the sixth year of competition The assumption that selection is the explanation for the occurrence of the significant result could be a sampling error if the population has not reached ASP. Selection or self-elimination (reduction in the number of participants) that might occur during these early years prior to ASP, is not surpassed by the increase in number of participants being recruited or chose themselves to start participating in the sport. In relation to using selection as an explanation for the asymmetrical distribution of a sample, the authors would encourage the researcher to identify the age of saturation of the sport specific population by gender and use the distribution at this age, in order to not draw hasty conclusion based on significant results. Changes in the sample are affected by changes over time and these changes are uniquely related to the population or sample, gender, level of proficiency, and sport under investigation. Assuming that a significance is equivalent to RAE and implying that selection has occurred within the specific age by gender group without identifying the distribution of ASP/ASS of the specific sample could lead to hasty conclusions about the reason for the statistical outcome. In the years, post-ASS, only one measure showed significance in three incidences and this was when utilizing AUD (B00, n=2 and B98, n=1). The randomness of occurrence strengthens the argument of using a longitudinal design in age by gender groups. It is further strengthened by

examining the use of PYD. Using this distribution, the occurrence of significance within an age by gender group was reduced to one incidence in four different age by gender groups (G98, G99, G00 and B00) over a nine-year period. These incidences occurred at age 9 (three times) and at age 10 (once), thus all four occurred in the phase where the sample size was increasing. Also, a selection-bias cannot be assumed based on significance, since the discrepancy between selection and self-elimination could not be identified in this sample. Utilizing PYD takes into account the uniqueness of the age by gender distribution, and could be used detect the possible unfairness in selection to different levels of performance within the age by gender group. An additional argument for the longitudinal design is based on the results of measure SSD. This measure showed no significant outcomes post-ASS. We interpret this as the selection or self-elimination away from the sport was evenly distributed in relation to the distribution at ASS. Identifying ASP and utilizing the distribution of the age when the population within a sport specific age by gender group is saturated is important. This in conjunction with using the PYD distribution is probably the best way to detect if a selection-bias has occurred over time in the age by gender sport specific sample. As reported earlier in the example of French basketball [11], the reduction in participation rate decreased in relation to NPD after the age of 14 to the age of 17. Thus, detecting ASS and using PYD, also increases the probability of detecting if a sport has a constant year effect [5] over time or not. **CONCLUSION**

Using the case of Swedish swimming, we have shown that the use of assumed distributions, either a uniformed or the national population, can increase the occurrence of significant results within a sample. The occurrence of significance has often lead to the conclusion that selection has occurred. We propose the use of distributions based on the actual population or sample, such as SSD and PYD and investigating how temporal aspects influences the changes in these distributions, by using a longitudinal design. Using actual distributions of the population that the sample represents, and repeating the sampling over time increases the possibility of detecting if a selection mechanism is in place within the specific sport and age by gender group.

APPLICATIONS IN SPORT

The method of detecting selection through the use of the chi-squared test goodness of fit test can be used on age by gender groups within the specific sport in a country. The utilization of PYD or SSD, in conjunction with an identification of ASP, takes into account the variations that occur within an age by gender group and are not specifically due to selection. When investigating selection to regional or national teams it is important to identify ASP and its distribution in order to draw conclusions if selections has occurred based on RAE. With today´s technological development it is more manageable for national federations in different sports to collect the data necessary to track the increase and decrease in numbers of the specific age by gender group populations. If such tracking is done continuously, it would also allow the national federations to monitor what is happening on the club level or at different proficiency levels of selection (i.e. local all-star team, regional, teams, etc.) of a specific age by gender group.

LIMITATIONS

The primary limitation of this study was the inability to identify all subjects in the data base in relation to results registered to determine the actual population distribution of each age

by gender group. The researchers assumed that the exclusion of unidentifiable ID registrations occurred randomly.

The researchers result, in this case, could have been impacted by the specific sport of swimming. Most RAE of studies have team sports. The researchers found an additional two studies related to swimming that showed contradictive results. Baxter-Jones [19] showed significance using a Kolmogorov Smirnov one sample test of uniformity in a small sample (n=54) representing individuals who were identified as elite. Costa et al. [20] found significance utilizing AUD in cross-sectional samples from seven age by gender group samples. This study utilized the top 50 performances in the respective age-group. In other words, they used performance as a criterion for inclusion in the sample, i.e. a selection of what individuals were included in the sample. Costa et al. [20] found a difference in distribution of competitors over the four quartiles. However, from performance perspective based on birth-quartile, they concluded that there is mostly no effect that could be attributed to what quartile the swimmers were born under.

ACKNOWLEDGMENTS
Data has been accessed with the aid and permission of the Swedish Swimming Federation.

REFERENCES
1. Dickinson, D. J. & Larson, J. D. (1963) The effect of chronological age in months on school achievement. J. Educ. Res., 56, 492-493.

2. Bedard, K & Dhuey, E. (2006) The persistence of early childhood maturity: International evidence of long-run age effects. Quart. J. Econ., 121, 1437-1472.

3. Barnsley, R. H. &Thompson, A. H. (1992) Family planning: Football style. The RAE in football. Int. Rev. Soc. Sport., 27, 77-88.

4. Cobley, S., Baker, J., Wattie, N. & McKenna, J. (2009) Annual age-grouping and athlete development: A meta analytical review of relative age effects in sport. Sport Med., 39, 235-256.

5. Schorer, J., Wattie, N. & Baker, J. R. (2013) A new dimension to relative age effects: Constant year effects in German youth handball. PLoS ONE., 8(4) e60336. doi:10.1371/journal.pone.006033.

6. Hollings, S. C., Hume, P. A. & Hopkins, W. G. (2014) Relative age effect on competition outcomes at world youth and world junior athletics championship. Eur. J. Sport Sci., 14, S456-S461.

7. Carling, C., le Gall, F., Reilly, T & Williams, A. M. (2009). Do anthropometric and fitness characteristics vary according to birth date distribution in elite youth academy soccer players? Scand. J. Med. Sci Sports., 19, 3-9.

8. Barnsley, R. H., Thompson, A. H. & Barnsley, P. E. (1985) Hockey success and birthdate: the RAE. Can. Assoc. Health, Phys Educ. Rec. J., 51, 23-28.

9. Barnsley, R. H.; Thompson, A. H. (1988) Birthdate and success in minor hockey: the key to NHL. Can. J. Beh. Sci., 20, 167-76.

10. Gibbs, B. J., Shafer, K. & Dufur, M. J. (2015) Why infer? The use and misuse of population data in sport research. Int. Rev. Soc. Sport., 50, 115-121.

11. Delorme, N. & Raspaud, M. (2009) The relative age effect in young French basketball players a study of the whole population. Scand. J. Med. Sci. Sports., 19, 235-242.

12. Musch, J. & Grondin, S. (2001) Unequal competition as an impediment to personal development: A review of the relative age effect in sport. Dev. Rev., 21, 147-167.

13. Carvalho, H. M., Coehlo-e-Silva, M. J., Gonclaves, C. E., Phiippaerts, R. M., Catsagna, C. & Malina, R. M. (2011) Age related variation of anaerobic power after controlling for size and maturation in adolescent basketball players. Ann. Hum. Bio., 38, 721-727.

14. Gorski, T., Rosser, T.;,Hoppeler, H. & Vogt, M. (2016) Relative age effect in young Swiss alpine skiers from 2004-2011. Int. J. Sports Phys. Per., 11, 455-463.

15. Skorski, S., Skorski, S., Faude, O., Hammes, D. & Meyer, T. (2016) The relative age effect in elite German youth soccer: Implications for a successful career. Int. J. Sports Phys. Per., 11, 370-376.

16. Robertson, S., Woods, C. & Gastin, P. (2014) Predicting higher selection in elite junior Australian rules football: The influence of physical performance and anthropometric attributes. J. Sci. Med. Sport., 18, 601-606.

17. Delorme, N., Boiche, J. & Raspaud, M. (2010) Relative age effect in elite sports: methodological bias or real discrimination. Eur. J. Sport Sci., 10, 91-96.

18. Fredricks, J. A. & Eccles, J. S. (2004) Parental influences on youth involvement in sports. In Developmental Sport and exercise physiology – a life span perspective; Weiss, M. R. Ed. Fitness Information Technology, Morgantown, WV, U.S.A.

19. Baxter-Jones, A. D. (1995) Growth and development of young athletes. Should competition levels be age related? Sports Med., 20, 59-64.

20. Costa, A. M., Marques, M. C., Louro, H., Ferreira, S. S. & Marinho, D. A. (2013) The relative age effect among elite youth competitive swimmers. Eur. J. Sport Sci., 13, 437-444.