

# Determinants of Elite Fitness: A Factor Analysis of the CrossFit Games

Alex Whitworth

Department of Statistics  
University of California, Los Angeles  
[whitworth.alex@gmail.com](mailto:whitworth.alex@gmail.com)

December 12, 2013

# Outline

## 1 Overview of CrossFit Games

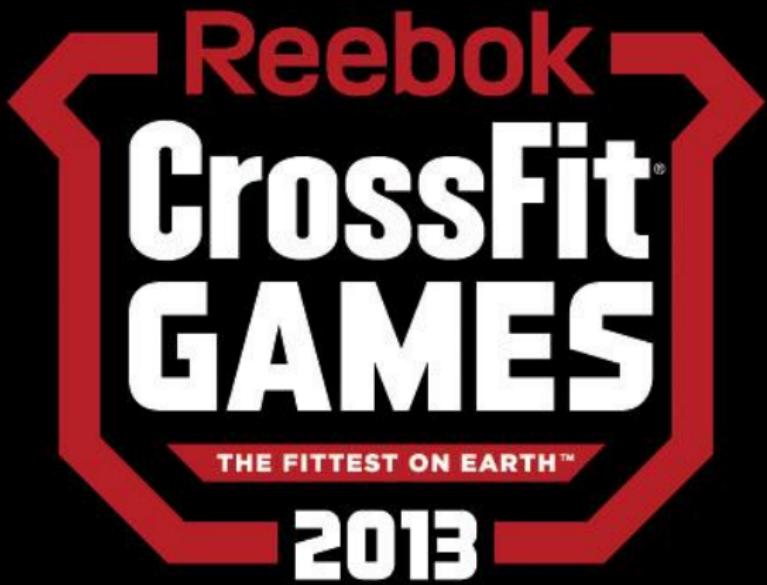
## 2 Physiological Background (factor interpretations)

## 3 Factor Analysis

- Data
- Factor Model
- Missing Data Imputation
- Model Uncertainty and Model Averaging

## 4 Results and Discussion

- Summary Information
- Factor Models
- Factor Discrimination































HANNAHAYWORTHPHOTOGRAPHY

# Outline

- 1 Overview of CrossFit Games
- 2 Physiological Background (factor interpretations)
- 3 Factor Analysis
  - Data
  - Factor Model
  - Missing Data Imputation
  - Model Uncertainty and Model Averaging
- 4 Results and Discussion
  - Summary Information
  - Factor Models
  - Factor Discrimination

# Definitions of Fitness

- CrossFit defines three standards for evaluating fitness
  - ① Ten general athletic skills
  - ② Performance across constantly varied athletic tasks
  - ③ Metabolic pathways, or energy systems, of human movement
- Bompa is more specific: strength, speed, endurance, and their combinations "create crucial physical athletic qualities" and are the key determinants of athletic performance.
- Understanding the links between athletic abilities and the energy systems is critical to understanding athletic performance.

# Outline

1 Overview of CrossFit Games

2 Physiological Background (factor interpretations)

3 Factor Analysis

- Data
- Factor Model
- Missing Data Imputation
- Model Uncertainty and Model Averaging

4 Results and Discussion

- Summary Information
- Factor Models
- Factor Discrimination

# Data

- The data are taken from the CrossFit Games website archive.
- 12 events, 44 male athletes, 43 female athletes
- All athletes compete in first 10 events, top 30 male / female athletes advance to final 2 events
  - Leads to missing data problem (discussed later)

# Factor Model

- Consider the  $n \times p$  data matrix  $X = \{X_1, \dots, X_p\}$  where  $X_j = \{x_{1j}, \dots, x_{nj}\}'$  is an  $n \times 1$  vector of values for the  $j^{\text{th}}$  variable  $j = 1, 2, \dots, p$ .
- $X' = L^k F^k + \epsilon$ 
  - where  $X$  is an  $n \times p$  data matrix
  - $L^k$  is a  $p \times k$  matrix of  $k$  factor loadings
  - $F^k$  is a  $k \times n$  matrix consisting of  $k$  underlying factors
  - $\epsilon$  is the  $p \times n$  matrix of stochastic errors.
- covariance matrix:  $\Sigma_k = L^k L^{k' \prime} + \Psi$ 
  - As is typical, I term the  $\Psi_i^2$ ,  $i = 1, 2, \dots, p$  as 'uniquenesses'.

# Missing Data Imputation

- 12 events, where only 30 athletes advance to final 2 events
- 30 complete observations with 14 (male) and 13 (female) observations missing data
- Ignoring missing observations can lead to biased estimates and even misspecified models
  - ① Estimate factor loadings for the 10 events with complete information
  - ② Impute missing values via OLS - estimate 12 event factor loadings
  - ③ Impute missing values via KNN - estimate 12 event factor loadings
- Results in three candidate models for interpretation

# Model Uncertainty and Model Averaging

- The typical procedure is to single out the "best" model. This approach neglects uncertainty about model choice and may bias estimates.
- compromise estimator:  $\widehat{\theta} = \sum_{i=1}^k w_i \widehat{\theta}_i$ 
  - where  $w_i$  is the weight applied to the estimators  $\widehat{\theta}_i$
  - $w_i = \frac{e^{-1/2AIC_i}}{\sum_{i=1}^k e^{-1/2AIC_i}}$
  - $AIC_k = -2 * logL(k) + 2p$
  - $L^k = \sum_{i=1}^k w_i L^k$  and  $\Psi_i^2 = \sum_{i=1}^k w_i \Psi_i^2$
- Guarantees the incorporation of model selection uncertainty; requires using the same rotation for all candidate models.

# Outline

- 1 Overview of CrossFit Games
- 2 Physiological Background (factor interpretations)
- 3 Factor Analysis
  - Data
  - Factor Model
  - Missing Data Imputation
  - Model Uncertainty and Model Averaging
- 4 Results and Discussion
  - Summary Information
  - Factor Models
  - Factor Discrimination

# Male Event Summary

Event	Athletes	Average	Std. Dev.	Min	Max	Min(Std. Dev)	Max(Std. Dev)
The Pool	44	13.78	2.43	10.10	20.49	-1.52	2.80
Row 1	44	6.83	0.20	6.35	7.59	-2.37	3.77
Row 2	44	83.65	3.85	78.04	98.81	-1.51	3.99
Burden Run	44	33.65	2.80	28.97	40.10	-1.65	2.07
Zig-Zag Sprint	44	11.71	0.41	10.90	12.60	-1.92	2.18
Legless	44	8.27	1.40	6.12	10.50	-1.54	1.74
Naughty Nancy	44	19.42	0.98	16.53	20.45	-2.96	1.05
Clean & Jerk	44	320.57	18.62	275.00	355.00	-1.85	2.45
2007	44	10.68	1.06	8.56	13.46	-2.00	2.62
Sprint Chipper	44	3.31	0.46	2.47	4.36	-1.82	2.28
Cinco 1	30	6.08	1.02	4.09	7.47	-1.95	1.37
Cinco 2	30	6.84	0.54	4.68	7.22	-4.02	0.70

**Table:** Summary of scores for all events of the 2013 CrossFit Games for male athletes.

# Female Event Summary

Event	Athletes	Average	Std. Dev.	Min	Max	Min(Std. Dev)	Max(Std. Dev)
The Pool	43	18.16	2.99	13.32	25.03	-1.60	2.16
Row 1	43	7.88	0.30	7.28	8.47	-1.98	1.96
Row 2	43	94.78	3.94	87.80	105.10	-1.80	2.65
Burden Run	43	36.02	3.85	28.07	40.12	-2.09	1.04
Zig-Zag Sprint	43	12.97	0.66	11.90	16.00	-1.62	4.61
Legless	43	10.49	0.35	9.56	10.92	-2.68	1.24
Naughty Nancy	43	19.43	0.99	16.77	20.47	-2.69	1.04
Clean & Jerk	43	205.00	17.25	180.00	240.00	-2.03	1.45
2007	43	12.35	1.53	9.81	15.30	-1.66	1.93
Sprint Chipper	43	3.32	0.49	2.59	4.35	-1.48	2.09
Cinco 1	30	6.24	1.04	3.54	7.78	-2.59	1.48
Cinco 2	30	7.13	0.17	6.79	7.45	-2.04	1.93

**Table:** Summary of scores for all events of the 2013 CrossFit Games for female athletes.

# Male Athletes - Final Factor Model

Event	Factor1	Factor2	Factor3	$\psi^2$
The Pool	0.36	0.06	0.03	0.83
Row 1	0.80	-0.12	-0.15	0.32
Row 2	0.54	0.22	0.12	0.68
Burden Run	0.58	-0.00	0.30	0.43
Zig-Zag Sprint (time)	0.12	-0.24	-0.26	0.87
Legless	0.21	-0.10	0.21	0.81
Naughty Nancy	-0.06	-0.03	0.88	0.24
Clean & Jerk	0.01	1.00	0.01	0.00
2007	0.28	0.25	0.42	0.69
Sprint Chipper	0.15	-0.64	0.20	0.40
Cinco 1	0.02	-0.20	0.12	0.81
Cinco 2	0.03	-0.08	0.30	0.69

**Table:** Matrix of Loadings and Uniquenesses, Male Model Averages.

# Female Athletes - Final Factor Model

Event	Factor1	Factor2	$\Psi^2$
The Pool	0.27	0.34	0.71
Row 1	0.88	0.01	0.19
Row 2	0.75	-0.02	0.43
Burden Run	0.78	0.07	0.34
Zig-Zag Sprint (time)	0.09	0.01	0.97
Legless	0.14	0.26	0.88
Naughty Nancy	0.53	0.05	0.69
Clean & Jerk	0.15	-0.69	0.56
2007	-0.00	0.47	0.67
Sprint Chipper	0.18	0.82	0.16
Cinco 1	0.06	0.21	0.84
Cinco 2	0.07	0.23	0.74

**Table:** Matrix of Loadings and Uniquenesses, Female Model Averages.

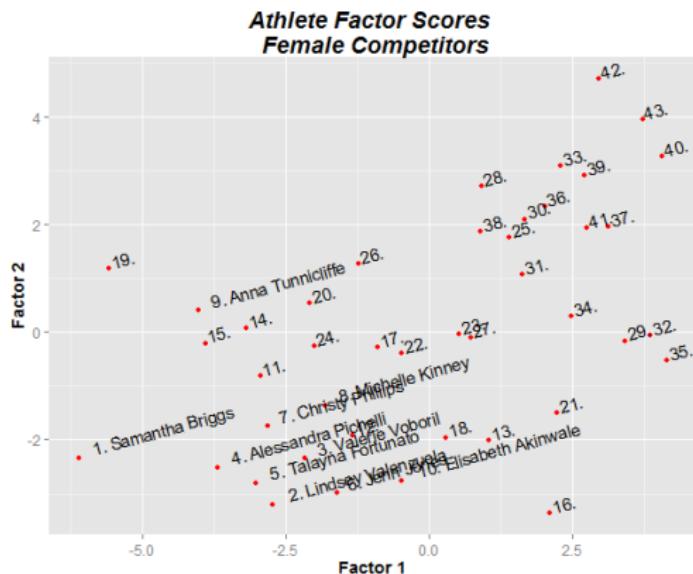
## Factor Interpretations - Male

- Factor 1: Based on the prior physiological discussion, this may be interpreted as loading highly on the aerobic system and muscular-endurance of medium and long duration. This could also be interpreted as a measure of general CrossFit fitness.
- Factor 2: This may be interpreted as the anaerobic alactic system and anaerobic lactic system; or a pure power and power-endurance component.
- Factor 3: loading highly on Naughty Nancy and moderately on 2007. Both of these events required extended use of the anaerobic lactic system in 1-3 minute sessions with the aerobic system highly active throughout. They also required good event strategy to properly balance the timing of these systems.
- Uniquenesses: Of high uniqueness events, The Pool, Zig-Zag Sprint and Legless represent the most "non-standard CrossFit" events so this result is unsurprising. Cinco 1 and Cinco 2 may represent a last-event-effect but may also represent an effect of the model averaging process, indicating that results for these two events should be interpreted with caution.

## Factor Interpretations - Female

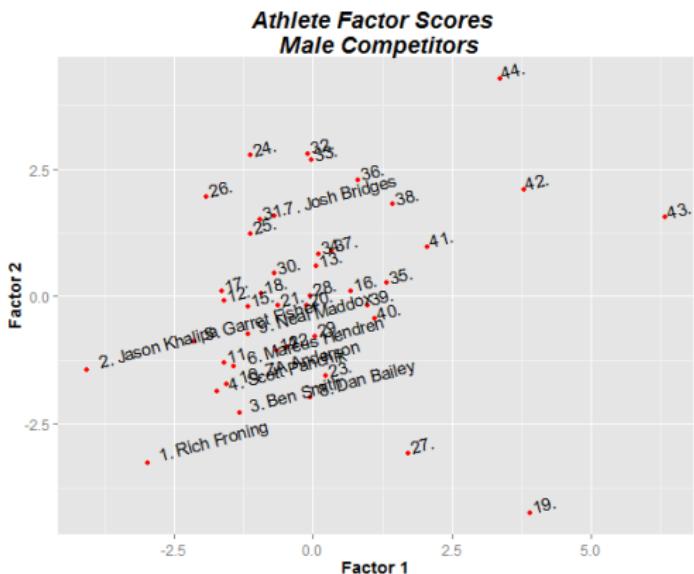
- Factor 1: Similar interpretation to Male Factor 1.
- Factor 2: This is nearly the opposite rotation to the male Factor 2 and may be interpreted somewhat similarly. I clarify that Factor 2 for the females appears to be more highly loaded on the anaerobic lactic system than the anaerobic alactic system.
- Uniquenesses: Similar interpretation to Male uniquenesses.
- Extremely interesting that the 2-factor model for the women is able to explain approximately the same amount of variance as the 3-factor model for the men.
  - ① More men take a very strong interest in athletic competition than women.
  - ② CrossFit requires large amounts of upper body strength; and, biologically, men have more upper body strength than women.
  - ③ Spurious noise in the data or small sample size. I urge caution, and a healthy dose of skepticism, in accepting any single explanation.

# Women's Factor Scores



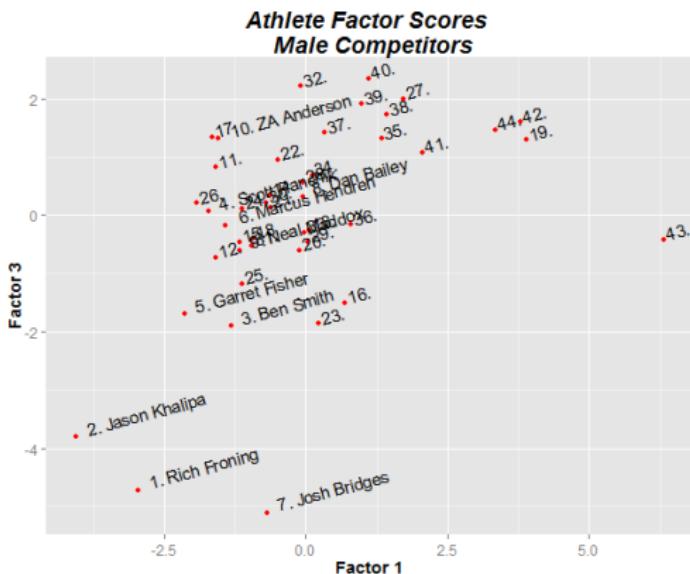
**Figure:** Female Factor scores. Points are labeled for each athlete's placing. Only the top 10 athletes names are shown to improve readability.

# Men's Factor Scores - Factor 1 vs Factor 2



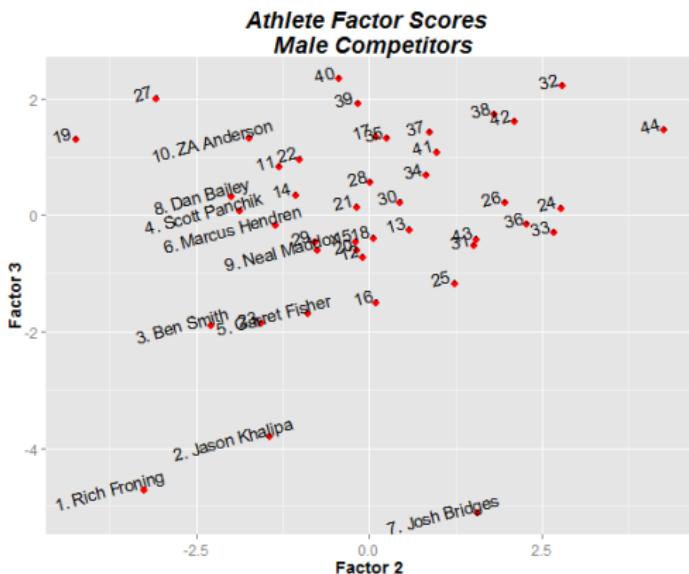
**Figure:** Male Factor scores. Points are labeled for each athlete's placing. Only the top 10 athletes names are shown to improve readability. Factor 2 has been inverted such that negative scores indicate better performance.

# Men's Factor Scores - Factor 1 vs Factor 3



**Figure:** Male Factor scores. Points are labeled for each athlete's placing. Only the top 10 athletes names are shown to improve readability.

## Men's Factor Scores - Factor 2 vs Factor 3



**Figure:** Male Factor scores. Points are labeled for each athlete's placing. Only the top 10 athletes names are shown to improve readability. Factor 2 has been inverted such that negative scores indicate better performance.