Part 7: The Rasch Model in Practice

Example Analysis

Example Analysis using the RM

Study: Investigation of the ZAREKI-R (Aster et.al., 2006)

Neuropsychological Test Battery for Number Processing and Calculation in Children

test for the assessment of dyscalculia for children of age 6 – 10
was developed using classical testtheory techniques
based on a sample of children from Germany and Switzerland
also versions in English, Portuguese, Greek, French, Turkish and Chinese

contains 16 Subtests: e.g., mental arithmetic: addition, subtraction, multiplication, number line, estimate quantity, count forward and backward etc.

Study description

Koller & Alexandrowicz (in press) check the assumptions of the Rasch model with a sample of 341 second to fourth grade children from five primary schools

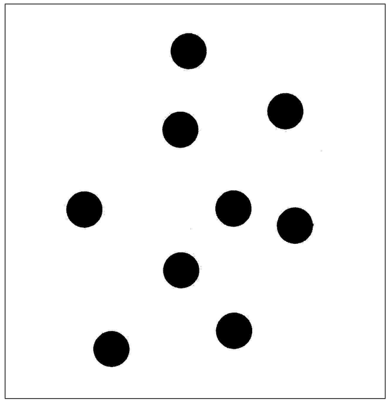
Sample:

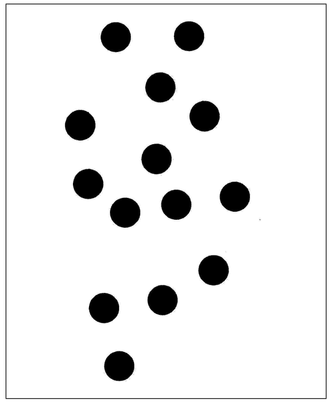
		grade 2	grade 3	grade 4
Gender	male	71	55	59
	female	51	44	61
Age	Median	8	9	10

we present some results on the subscale:
Perceptive Quantity Estimation

Perceptive Quantity Estimation (5 Items)

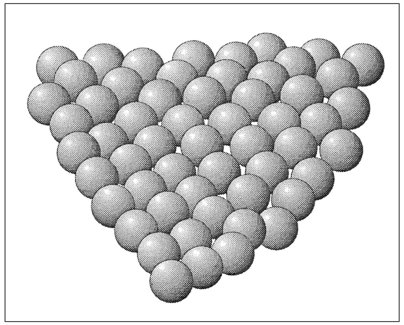
Item 1 and Item 2 (each presented for 2 seconds)





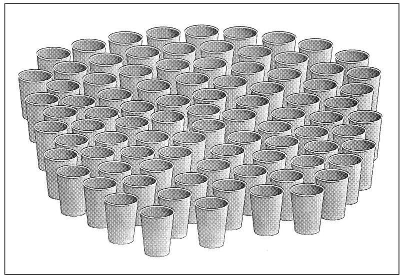
Perceptive Quantity Estimation (5 Items)

Item 3: Balls (presented for 5 seconds)



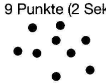
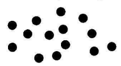
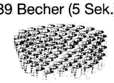
Perceptive Quantity Estimation (5 Items)

Item 4: Cups (presented for 5 seconds)



Item 5: “Were there more balls or more cups?”

Protokoll:[*)]

Nr.	Stimulus	Antwort		Scores
9.1	9 Punkte (2 Sek.) 	7 X ←······→ X 11		0 / 1
9.2	14 Punkte (2 Sek.) 	11 X ←······→ X 19		0 / 1
9.3	57 Bälle (5 Sek.) 	25 X ←······→ X 80		0 / 1
9.4	89 Becher (5 Sek.) 	35 X ←······→ X 125		0 / 1
9.5	Waren es mehr Bälle oder mehr Becher?			0 / 1
			Summe:	

Perceptive Quantity Estimation (5 Items)

Qualitative analysis by Koller & Alexandrowicz (in press) shows a few problems for this subscale:

- presentation time: differences particularly with Item 1: faster children can count parts of the points and estimate the rest, slower children can estimate only
- Item 5 has different structure than other items latent trait is ability to compare and not ability to estimate quantitites
- Additionally: the question (Item 5) irritates the children some think they have given a wrong answer to previous items

Rasch Analysis

data:

```
> library(foreign)
> dat <- read.spss("zareki.sav", to.data.frame = T, use.value.labels = F)
> head(dat)
  id class time sex item1 item2 item3 item4 item5 score
1  1     2   39   1     1     0     1     1     1     4
2  2     2   37   1     1     1     1     0     1     4
3  3     2   34   1     1     1     1     1     1     5
4  4     2   28   1     1     1     0     1     1     4
5  5     2   32   1     1     1     1     1     1     5
6  6     2   27   1     1     0     1     1     1     4
```

extract the items

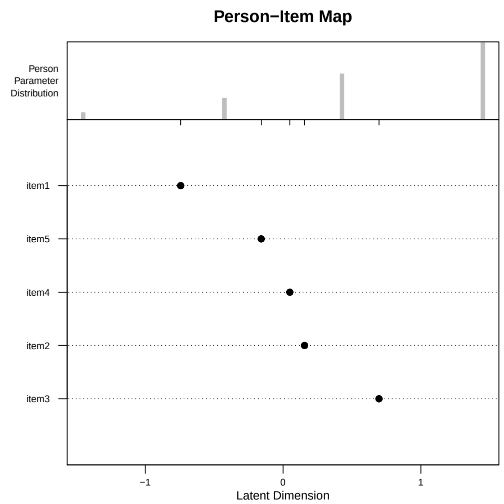
```
> data <- dat[, 5:9]
```


fit the model

```
> res <- RM(data)
> res
Results of RM estimation:

Call:  RM(X = data)

Conditional log-likelihood: -340.2692
Number of iterations: 12
Number of parameters: 4

Item (Category) Difficulty Parameters (eta):
             item2     item3      item4      item5
Estimate 0.1560360 0.6961884 0.04932307 -0.1586715
Std.Err  0.1315110 0.1237928 0.13396856  0.1396862
```



```
> summary(res)
Results of RM estimation:

Call:  RM(X = data)

Conditional log-likelihood: -340.2692
Number of iterations: 12
Number of parameters: 4

Item (Category) Difficulty Parameters (eta) with 0.95 CI:
      Estimate Std. Error lower CI upper CI
item2    0.156      0.132   -0.102    0.414
item3    0.696      0.124    0.454    0.939
item4    0.049      0.134   -0.213    0.312
item5   -0.159      0.140   -0.432    0.115

Item Easiness Parameters (beta) with 0.95 CI:
           Estimate Std. Error lower CI upper CI
beta item1    0.743      0.163    0.424    1.062
beta item2   -0.156      0.132   -0.414    0.102
beta item3   -0.696      0.124   -0.939   -0.454
beta item4   -0.049      0.134   -0.312    0.213
beta item5    0.159      0.140   -0.115    0.432
```

Item Fit Statistics

first estimate the person parameters

```
> pers <- person.parameter(res)
> itemfit(pers)
Itemfit Statistics:
         Chisq  df p-value Outfit MSQ Infit MSQ Outfit t Infit t
item1 150.825 189   0.981      0.794     0.862    -2.11   -1.85
item2 197.634 189   0.319      1.040     1.025     0.79    0.58
item3 184.359 189   0.582      0.970     0.955    -0.76   -1.33
item4 148.917 189   0.986      0.784     0.828    -4.01   -3.74
item5 242.900 189   0.005      1.278     1.199     3.95    3.54
```

Item 5 is significant, slight underfit for Infit and Outfit

Items 1 and 4 slightly overfitting

Person Fit Statistics

```
> personfit(pers)
```

extract and display some of the persons

```
> pfit <- print(personfit(pers), visible = FALSE)
> head(pfit)
   Chisq df p-value Outfit MSQ Infit MSQ Outfit t Infit t
P1 4.676  4   0.322      0.935     1.027     0.13    0.28
P2 5.115  4   0.276      1.023     1.072     0.30    0.38
P4 2.958  4   0.565      0.592     0.756    -0.63   -0.37
P6 4.676  4   0.322      0.935     1.027     0.13    0.28
P7 4.676  4   0.322      0.935     1.027     0.13    0.28
P8 5.416  4   0.247      1.083     1.117     0.51    0.74
```

information criteria

```
> IC(pers)
Information Criteria:
                      value npar       AIC      BIC     cAIC
joint log-lik     -519.2466    8 1054.4932 1080.469 1088.469
marginal log-lik  -768.1714    4 1544.3428 1559.670 1563.670
conditional log-lik -340.2692  4  688.5385  703.866  707.866
```


Parametric Approach

Likelihoodratio- and Wald Tests

for the criteria:

- gender (female . . . 1, male. . . 2)
- high/low ability (median person raw score)
- grade

Gender:

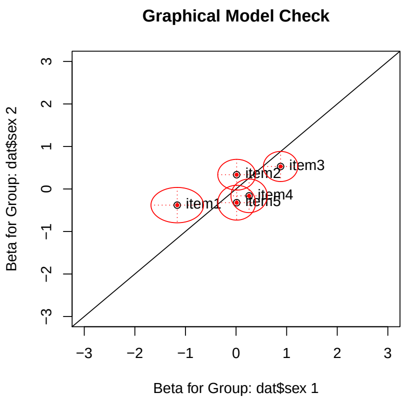
LR Test:

```
> lrt1 <- LRtest(res, splitcr = dat$sex, se = TRUE)
> lrt1
Andersen LR-test:
LR-value: 9.857
Chi-square df: 4
p-value:  0.043
```



```
> summary(lrt1)
Andersen LR-test:
LR-value: 9.857
Chi-square df: 4
p-value:  0.043


Subject subgroup dat$sex 1:
Log-likelihood:  -171.4529

Beta Parameters:
         beta item1  beta item2 beta item3 beta item4  beta item5
Estimate  1.1624804 -0.01219534 -0.8800760 -0.2580137 -0.01219534
Std.Err.  0.2648616  0.19076698  0.1736495  0.1830662  0.19076698


Subject subgroup dat$sex 2:
Log-likelihood:  -163.8879

Beta Parameters:
         beta item1 beta item2 beta item3 beta item4 beta item5
Estimate  0.3800989 -0.3349843 -0.5290260  0.1612697  0.3226418
Std.Err.  0.2122085  0.1846469  0.1806321  0.2015569  0.2092184
```

```
> plotGOF(lrt1, conf = list())
```



Item 1 not well fitting: easier for girls

Wald Test

```
> wald1 <- Waldtest(res, splitcr = dat$sex)
> wald1
Wald test on item level (z-values):

            z-statistic p-value
beta item1        2.305   0.021
beta item2        1.216   0.224
beta item3       -1.401   0.161
beta item4       -1.540   0.124
beta item5       -1.183   0.237
```

as seen from the plot: significant difference for Item 1.

GRADE: (3 groups)

LR Test

```
> lrt2 <- LRtest(res, splitcr = dat$class)
> lrt2
Andersen LR-test:
LR-value: 9.115
Chi-square df: 8
p-value:  0.333
```

Wald Test (only for 2 groups)

ABILITY: (2 groups)

split according to median rawscore (default)

LR Test

```
> lrt3 <- LRtest(res)
```

produces error:

```
Error in LRtest.Rm(res) :
No items with appropriate response patterns left to perform LR-test!
```

reason: median is 4 (there are only 5 items)

```
> table(dat$score, ifelse(dat$score <= median(dat$score),
+     "<=med", ">med"))
    <=med >med
  1     3    0
  2    24    0
  3    59    0
  4   104    0
  5     0  151
```

better use another split value, e.g., rawscore < 4

```
> split4 <- ifelse(dat$score < 4, 1, 2)

> lrt3 <- LRtest(res, splitcr = split4, se = TRUE)
> lrt3
Andersen LR-test:
LR-value: 31.059
Chi-square df: 4
p-value:  0
```

is highly significant

items have different properties for better and worse performing children

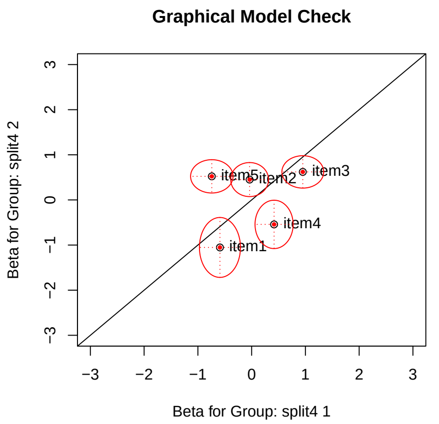
the Wald test gives

```
> Waldtest(res, splitcr = split4)
Wald test on item level (z-values):

           z-statistic p-value
beta item1      -1.185   0.236
beta item2       1.857   0.063
beta item3      -1.224   0.221
beta item4      -2.927   0.003
beta item5       4.584   0.000
```

Item 5: easier for less able group

Item 4: is much easier for better performing children


```
> plotGOF(lrt3, conf = list())
```



Conclusion: (for the parametric procedures)

ITEMFIT STATISTICS (Infit and Outfit):
indicate model deviation for Item 5.

DIF:

Gender:
significant LR Test Item 1 easier for girls (Wald Test)

Grade:
no differences

Ability:
highly significant LR Test
Item 5: easier for less able group
Item 4: is much easier for better performing children

→ scale does not seem to be Rasch homogeneous

Nonparametric Tests

Stochastic Dependence

it is assumed that items 1,2 and items 3,4 are dependent:

the black points have the same theoretical structure
the balls and cups have the same theoretical structure, but different from the points

Global Test

```
> t11 <- NPtest(rmat1, method = "T11")
> t11
Nonparametric RM model test: T11 (global test - local dependence)
    (sum of deviations between observed and expected inter-item correlations)

Number of sampled matrices: 1000
one-sided p-value: 0
```


All Itempairs

```
> t1 <- NPtest(rmat2, method = "T1")
> print(t1, alpha = 0.05)
Nonparametric RM model test: T1 (local dependence - increased
     inter-item correlations)
    (counting cases with equal responses on both items)

Number of sampled matrices: 1000

Number of Item-Pairs tested: 10

Item-Pairs with one-sided p < 0.05

(1,2) (3,4)
    0     0
```


Item specific

```
> t21 <- NPtest(rmat3, method = "T2", idx = 1:2)
> t21
Nonparametric RM model test: T2 (local dependence - model
     deviating subscales)
    (dispersion of subscale person rawscores)

Number of sampled matrices: 1000
Items in subscale: 1 2
Statistic: variance
one-sided p-value: 0.002

> t22 <- NPtest(rmat3, method = "T2", idx = c(3, 4))
> t22
Nonparametric RM model test: T2 (local dependence - model
     deviating subscales)
    (dispersion of subscale person rawscores)

Number of sampled matrices: 1000
Items in subscale: 3 4
Statistic: variance
one-sided p-value: 0
```


Stochastic dependence: Conclusion

it was assumed that the itemgroups 1,2 and 3,4 are dependent

– the global test T1 gives significant results

– more specific test for both item groups on item level further exhibit this pattern

Different Discrimination

Lower Discrimination is assumed for items 1 and 5:
– item 1, because faster children start to count
– item 5, different task

```
> t7a <- NPtest(rmat5, method = "T7a", idx = 1:5)
> t7a
Nonparametric RM model test: T7a (different discrimination - 2PL)
    (counting cases with response 1 on more difficult and 0 on easier item)

Number of sampled matrices: 1000
Item Scores:
  1   2   3   4   5
309 275 246 280 289

Item-Pairs: (i>j ... i easier than j)

(1>5) (1>4) (1>2) (1>3) (5>4) (5>2) (5>3) (4>2) (4>3) (2>3)
0.461 0.590 0.999 0.114 0.029 0.042 0.001 0.201 1.000 0.163
```

⟶ only Item 5 shows different discrimination compared to 2,3,4

Differential Item functioning

DIF hypothesised for gender and ability

High/Low Ability (1. . . low, 2. . . high)

Global Test

```
> t101 <- NPtest(rmat6, method = "T10")
> t101
Nonparametric RM model test: T10 (global test - subgroup-invariance)

Number of sampled matrices: 1000
Split: median
Group 1: n =  151   Group 2: n = 190
one-sided p-value: 1
```


Item-level Tests: Item groups 1,2 and 3,4

```
> t41 <- NPtest(rmat7, method = "T4", idx = c(1, 2), group = split4 ==
+     1, alternative = "low")
> t41
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting low raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 1 2
Group: split4 == 1   n = 86
one-sided p-value: 0.785

> t42 <- NPtest(rmat8, method = "T4", idx = c(3, 4), group = split4 ==
+     1, alternative = "low")
> t42
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting low raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 3 4
Group: split4 == 1   n = 86
one-sided p-value: 0
```


Item-level Test: Item 5

```
> t44 <- NPtest(rmat10, method = "T4", idx = c(5), group = split4 ==
+     1, alternative = "high")
> t44
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting high raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 5
Group: split4 == 1   n = 86
one-sided p-value: 0
```

Conclusion: DIF for Ability

- no differences for items 1,2
- for (the more difficult) items 3,4:
  more difficult for children with a lower ability
- item 5: easier for children with low ability

Interpretation: DIF for ability

- developmental psychology:
  low ability means little understanding of quantities
- children with a high/low abilities apply different strategies:
  counting vs. estimating the quantities
- item 5: different task – irritates children
  they suspect having given a wrong answer to items 3 and 4 and correct it
  wrong answer becomes correct and vice versa

DIF for Gender (1...female, 2...male)

Global Test

```
> t102 <- NPtest(rmat11, method = "T10", splitcr = dat$sex)
> t102
Nonparametric RM model test: T10 (global test - subgroup-invariance)

Number of sampled matrices: 1000
Split: dat$sex
Group 1: n =  185   Group 2: n = 156
one-sided p-value: 0.039
```


Item-level Tests: Item groups 1,2 and 3,4

```
> t46 <- NPtest(rmat12, method = "T4", idx = c(1, 2), group = dat$sex ==
+     2, alternative = "low")
> t46
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting low raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 1 2
Group: dat$sex == 2   n = 156
one-sided p-value: 0.002

> t47 <- NPtest(rmat13, method = "T4", idx = c(3, 4), group = dat$sex ==
+     2, alternative = "high")
> t47
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting high raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 3 4
Group: dat$sex == 2   n = 156
one-sided p-value: 0.018
```


Item-level Test: Item 5

```
> t48 <- NPtest(rmat14, method = "T4", idx = c(5), group = dat$sex ==
+     2, alternative = "high")
> t48
Nonparametric RM model test: T4 (Group anomalies - DIF)
    (counting high raw scores on item(s) for specified group)

Number of sampled matrices: 1000
Items in Subscale: 5
Group: dat$sex == 2   n = 156
one-sided p-value: 0.205
```

CONCLUSION: DIF for Gender

- significant gender differences
- items 1,2 easier for female children
- for (the more difficult) items 3,4:
  more difficult for female children
- item 5: no gender differences
  'irritation' depends on ability rather than on gender

Interpretation:

- well-known phenomenon in developmental psychology:
  – typical gender differences for numerical tasks
  – boys better than girls

- also different strategies assumed

General Conclusion

data do not conform with RM assumptions

subscale Perceptive Quantity Estimation is not an adequate measurement instrument for diagnosing dyscalculia

specific violations:

- local dependence for item groups (1,2) and (3,4)
- worse discrimination for item 5
- DIF for high/low ability and gender: – items 3,4 more difficult for females and lower ability – item 5: reversed solving probabilities for high/low ability

items provoke different problem solving strategies
exhibit gender differences even in very young children (supported by various studies)