UNIVERSITY
OF AMSTERDAM

A novel predictive CAT method for screening instruments

Anastasios (Tasos) Psychogyiopoulos, Niels Smits, L. Andries van der Ark

Research Institute of Child Development and Education
University of Amsterdam

Prerequesites

- HR-QoL: Health Related - Quality of Life context
- IRT: Item Response Theory
- CAT: Computerized Adaptive Testing

Some challenges with *screening*

- screening: short ⟹ further assessment or intervention (Greenhalgh, 2009; Marshall et al., 2006)

QoL: **PHQ-9 questionnaire; depression screening** (Kroenke et al., 2001)

Some challenges with *screening*

- screening: short ⟹ further assessment or intervention (Greenhalgh, 2009; Marshall et al., 2006)

QoL: **PHQ-9 questionnaire; depression screening** (Kroenke et al., 2001)

- long questionnaires ⟹ boredom (e.g., Nelson et al., 2015) ⟹ infrequent use (e.g., Morris et al., 1997)

Some challenges with *screening*

- screening: short ⟹ further assessment or intervention (Greenhalgh, 2009; Marshall et al., 2006)

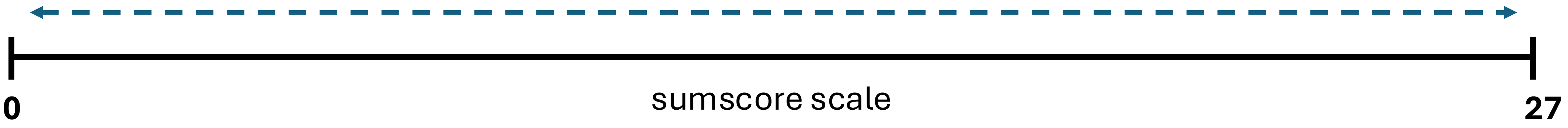
QoL:	**PHQ-9 questionnaire; depression screening** (Kroenke et al., 2001)

- long questionnaires ⟹ boredom (e.g., Nelson et al., 2015) ⟹ infrequent use (e.g., Morris et al., 1997)

- Efficient screening = predictive validity + administration time

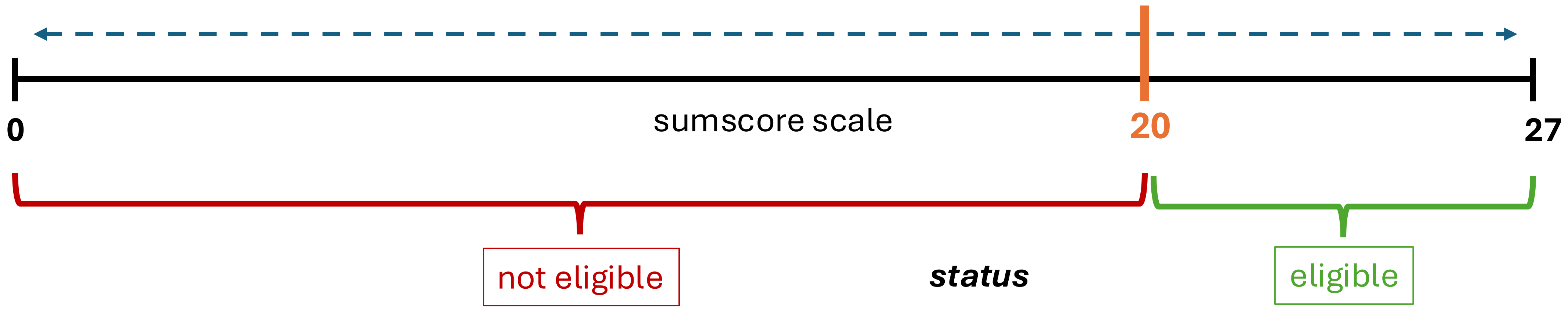
Screening as *prediction*

QoL: **PHQ-9: 9 items of 4 categories [0,1,2,3]**



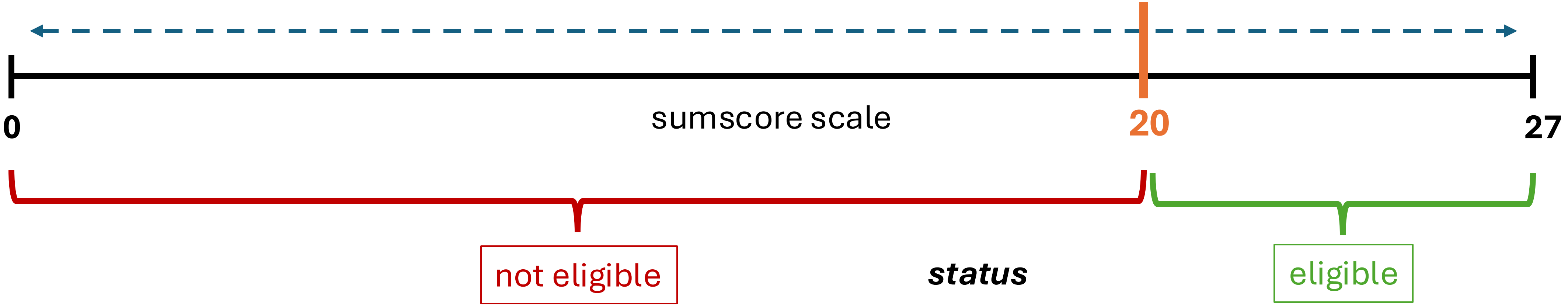
Screening as prediction

QoL: **PHQ-9: 9 items of 4 categories [0,1,2,3]**



Screening as *prediction*

QoL: **PHQ-9: 9 items of 4 categories [0,1,2,3]**



- [a screener] must be highly reliable around the cut-off value
- Can we choose the items with the highest predictive validity?

CAT as a solution

- Computer adaptive testing (CAT) methods: improving efficiency while ensuring accuracy and precision of **measurement**.

CAT as a solution

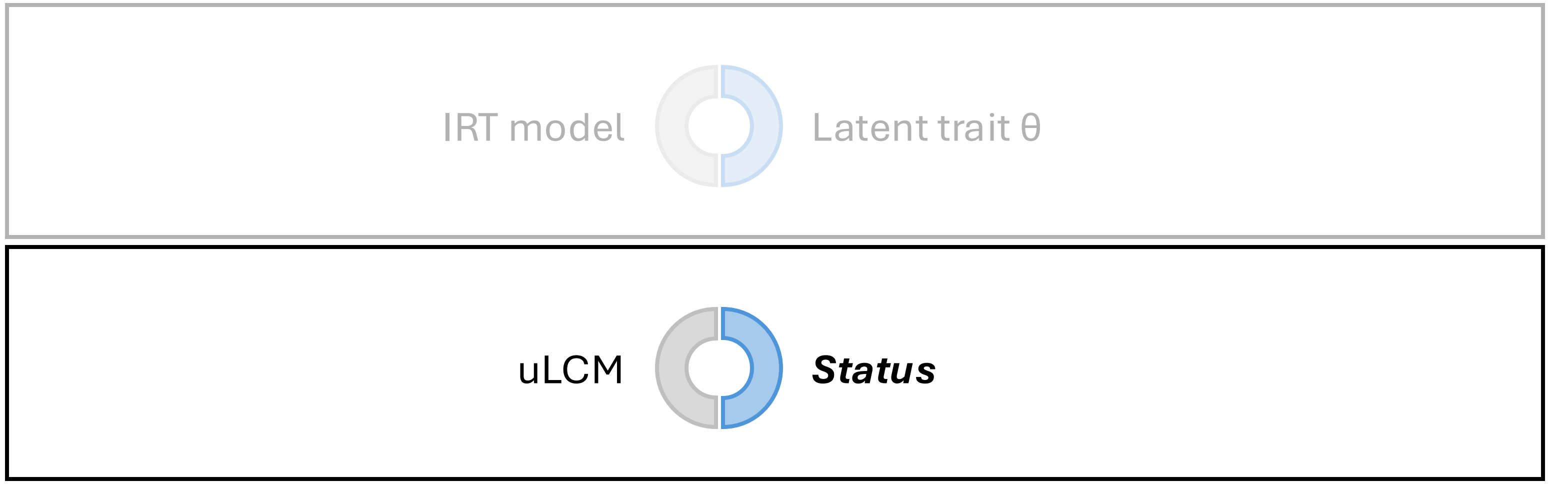
- Computer adaptive testing (CAT) methods: improving efficiency while ensuring accuracy and precision of **measurement**.

- Some disadvantages of ’traditional’ CAT,
  - IRT for measurement (Gibbons et al., 2016; Smits et al., 2018)
  - IRT assumptions are not always satisfied (e.g., Fayers, 2007)
  - Using sumscore instead of θ -> better communication
  - …

LSCAT: A novel approach

- **L**atent-class **S**umscore **C**omputerized **A**daptive **T**esting

- under FlexCAT framework: *engine* and *score* (Van der Ark & Smits, 2023)



engine

For 9 items with 4 answer categories,
there are $R = 4^9 = 262{,}144$ response patterns:

$\mathbf{y}_1, \dots, \mathbf{y}_R$: All possible response patterns

$\boldsymbol{\pi} = P(\mathbf{y}_1, \dots, \mathbf{y}_R)$: joint item score density

$$\mathbf{R} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_R \end{pmatrix} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} .009 \\ .012 \\ \vdots \\ .070 \end{pmatrix}$$

engine

$\mathbf{y}_1, \ldots, \mathbf{y}_R$: All possible response patterns

$\boldsymbol{\pi} = P(\mathbf{y}_1, \ldots, \mathbf{y}_R)$: joint item score density

Unrestricted latent class model (ULCM) can estimate

$$P(\mathbf{y}_r) = \sum_k P(\Lambda = k) \prod_j P(Y_j = y_{rj} | \Lambda = k)$$

$$\boldsymbol{\pi} = \prod_r P(\mathbf{y}_r)$$

- It uses the ULCM as a density estimator (e.g., Vermunt & Magidson, 2016; Linzer & Lewis, 2011)
- BIC generally provided the most accurate estimates of $\boldsymbol{\pi}$ (Psychogyiopoulos et al, 2025a)

For 9 items with 4 answer categories, there are $R = 4^9 = 262{,}144$ response patterns:

$$\mathbf{R} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_R \end{pmatrix} = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

$$\hat{\boldsymbol{\pi}} = \begin{pmatrix} .009 \\ .012 \\ \vdots \\ .070 \end{pmatrix}$$

score : The variable used to communicate the respondent's value.

I want to predict a *status*

e.g., eligible (sumscore > 20)
or not eligible (sumscore ≤ 20)

$$\underset{(4^9 \times 9)}{\mathbf{R}} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad \mathbf{r}_+ = \mathbf{R} \cdot 1 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 27 \end{pmatrix}$$

score : The variable used to communicate the respondent's value.

I want to predict a *status*

e.g., eligible (sumscore > 20)
or not eligible (sumscore ≤ 20)

$$\underset{(4^9 \times 9)}{\mathbf{R}} = \begin{pmatrix} \boxed{0 \quad 0 \quad \dots \quad 0} \\ \boxed{0 \quad 0 \quad \cdots \quad 1} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ \boxed{1 \quad 1 \quad \cdots \quad 1} \end{pmatrix} \qquad \mathbf{r}_+ = \mathbf{R} \cdot 1 = \begin{pmatrix} \boxed{0} \\ \boxed{1} \\ \vdots \\ \boxed{27} \end{pmatrix} \qquad \mathbf{Q} = \begin{pmatrix} \boxed{1 \quad 0} \\ \boxed{1 \quad 0} \\ \vdots \quad \vdots \\ \boxed{0 \quad 1} \end{pmatrix}$$

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} .009 \\ .012 \\ \vdots \\ .070 \end{pmatrix}$$

we calculate

$$\boxed{\widehat{\boldsymbol{\pi}}_S = \mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\pi}} = \begin{pmatrix} \boxed{.911} \\ \boxed{.088} \end{pmatrix}}$$

LSCAT steps

Calibration

- Full-test to a large sample
- Define the sample $\boldsymbol{\pi}_{\mathbf{status}}$

$$\hat{\boldsymbol{\pi}}_S = \mathbf{Q}^{\mathrm{T}}\hat{\boldsymbol{\pi}} = \begin{pmatrix} .911 \\ .088 \end{pmatrix} \begin{array}{l} \text{not eligible} \\ \text{eligible} \end{array}$$

LSCAT steps

Calibration

- Full-test to a large sample
- Define the sample $\boldsymbol{\pi}_{\mathbf{status}}$

$$\widehat{\boldsymbol{\pi}}_{\mathrm{S}} = \mathbf{Q}^{\mathrm{T}}\widehat{\boldsymbol{\pi}} = \begin{pmatrix} .911 \\ .088 \end{pmatrix} \begin{array}{l} \text{not eligible} \\ \text{eligible} \end{array}$$

Administration

- Use the initial $\boldsymbol{\pi}_{\mathbf{status}}$ as a **starting point**
- and a pre-defined **stopping rule**
  e.g., $c = .95 = 95\%$
- Start administer **the most informative items***
- Until the stopping rule is met
- Assign the repondent to the status with the highest probability

Put LSCAT into practice

Two Studies to demonstrate the potential of LSCAT for screening in HR-QoL

Study 1

Dataset: PHQ-9 data from a sample of $20{,}685$ individuals from the National Health and Nutrition Examination Survey (NHANES)

Methodology 1:

- Post-hoc simulation methodology using existing test data: **Use the full test score as the *true* status**
- Data split into *calibration* and *validation* sets ($N_v = 10{,}342$)

Study 1

Dataset: PHQ-9 data from a sample of 20,685 individuals from the National Health and Nutrition Examination Survey (NHANES)

Methodology 1:

- Post-hoc simulation methodology using existing test data: **Use the full test score as the *true* status**
- Data split into *calibration* and *validation* sets ($N_v = \ 10{,}342$)

- Dependent variables: **Predictive validity** was assessed by comparing predicted and *true* status (eligible/not eligible) using $\text{Type I ER} = \frac{FP}{TN+FP}$, $\text{Type II ER} = \frac{FN}{TP+FN}$, $\text{Accuracy} = \frac{TP+FN}{TN+FP+TP+FN}$
- Independent variables: Two Stopping criteria $c \ = \ .95$ and $c \ = \ .99$

Study 1 Results

Stopping criterion	Efficiency		Predictive Validity		
	M (SD)	Range	Type I ER	Type II ER	Accuracy
$\mathrm{c} = .95$	1.789(1.714)	1-9	0.001	0.128	0.989
$\mathrm{c} = .99$	3.045(1.919)	2-9	0.000	0.024	0.998

Note. ER = error rate

Study 2 : Benchmarking

Study 2 : Benchmarking

Objective: Compare LSCAT to other *test-shortening* methods

- **Stochastic Curtailment (SC)** (Smits & Finkelman, 2015; Finkelman et al. 2011)
- **CAT using Decision Trees (DTCAT)** (Yan et al. 2004; Gibbons et al. 2023;2013)

Study 2 : Benchmarking

Objective: Compare LSCAT to other *test-shortening* methods

- **Stochastic Curtailment (SC)** (Smits & Finkelman, 2015; Finkelman et al. 2011)
- **CAT using Decision Trees (DTCAT)** (Yan et al. 2004; Gibbons et al. 2023;2013)

Hypothesis: LSCAT would perform better because it employs dynamic item selection

Methodology 2:

- Same Post-hoc simulation methodology
- Efficiency (**average administered items**) was fixed across methods

Study B - Results

Stopping criterion	Method	Efficiency	
		M (SD)	Range
$c = .99$	LSCAT	3.045(1.919)	2-9
	SC	3.002(1.836)	2-9
	DTCAT	3.000(0.296)	2-4

Note. ER = error rate; LSCAT = Latent-class sum score computerized adaptive testing; SC = Stochastic curtailment; DTCAT = Decision tree based computer adaptive testing.

Study B - Results

Stopping criterion	Method	Efficiency		Predictive Validity		
		M (SD)	Range	Type I ER	Type II ER	Accuracy
	LSCAT	3.045(1.919)	2-9	**0.000**	**0.024**	**0.998**
$c = .99$	SC	3.002(1.836)	2-9	0.001	0.050	0.995
	DTCAT	3.000(0.296)	2-4	0.023	0.225	0.960

Note. ER = error rate; LSCAT = Latent-class sum score computerized adaptive testing; SC = Stochastic curtailment; DTCAT = Decision tree based computer adaptive testing.

Conclusion

Psychogyiopoulos, A., Smits, N., & Van der Ark, L. A. (2025). A novel CAT method for QoL screening: proof-of-principle study with comparisons to standard methods. *Quality of Life Research*, 1–9. https://doi.org/10.1007/s11136-025-04035-5

- **proof-of-concept** study
- **LSCAT consistently outperformed SC and DTCAT**
- High accuracy for all methods

- **SC similar performance to LSCAT:** The first items on the sequence were the most informative ones

Future Directions and Challenges

Next Steps:

- Large scale simulation studies needed to optimize settings
- Developing LSCAT further for larger item pools

Technical Challenges:

- Handling tests with > 20 binary items
- Addressing the curse of dimensionality

(e.g., for PHQ-9 all possible response patterns: $4^9 = 262{,}144$, $4^{10} = 1{,}048{,}576$

- Speed needs improvement

- Make LSCAT widely available

Reference list

Finkelman, M. D., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine*, *30*(16), 1989–2004. https://doi.org/10.1002/sim.4231

Gibbons, R. D. (2013). *The Computerized Adaptive Diagnostic Test for Major Depressive Disorder (CAD-MDD): A Screening Tool for Depression*. Psychiatrist.com; Primary Care Companion for CNS Disorders. https://www.psychiatrist.com/jcp/computerized-adaptive-diagnostic-test-major-depressive/

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2015). Computerized Adaptive Diagnosis and Testing of Mental Health Disorders. *Annual Review of Clinical Psychology*, *12*(1), 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Kroenke, K., Spitzer, R. L., & Janet. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Linzer, D. A., Lewis, (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis*, *19*(2), 173–187. https://doi.org/10.1093/pan/mpr006

Psychogyiopoulos, A., Smits, N., & Van der Ark, L. A. (2025b). A novel CAT method for QoL screening: proof-of-principle study with comparisons to standard methods. *Quality of Life Research*, 1–9. https://doi.org/10.1007/s11136-025-04035-5

Psychogyiopoulos, A., Smits, N., & Van der Ark, L.A. (2025a). Estimating the joint item-score density using an unrestricted latent class model: advancing flexibility in computerized adaptive testing. *Journal of Computerized Adaptive Testing, 12*(3), 136–164. https://doi.org/10.7333/2507-1203136

Smits, N., & Finkelman, M. D. (2015). Shortening the PHQ-9: a proof-of-principle study of utilizing Stochastic Curtailment as a method for constructing ultrashort screening instruments. *General Hospital Psychiatry*, *37*(5), 464–469. https://doi.org/10.1016/j.genhosppsych.2015.04.011

Van der Ark, L.A., & Smits, N. (2023). Computerized Adaptive Testing Without IRT for Flexible Measurement and Prediction. In: van der Ark, L.A., Emons, W.H.M., Meijer, R.R. (eds) Essays on Contemporary Psychometrics. Methodology of Educational Measurement and Assessment, 369–388. https://doi.org/10.1007/978-3-031-10370-4_19

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 549–553). Sage.

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive Testing With Regression Trees in the Presence of Multidimensionality. *Journal of Educational and Behavioral Statistics*, *29*(3), 293–316. https://doi.org/10.3102/10769986029003293

Thank you! ☺

Questions, Suggestions

a.psychogyiopoulos@uva.nl