ELECTRONIC SUPPLEMENTARY MATERIAL 2 (ESM 2)

Inspection of Measurement Invariance of the QUVA-p factorial structure across sex

Measurement bias was examined by investigating if the determined factor structure for the QUVA-p varied across sex. Sex invariance was tested in the framework of multiple-group Confirmatory Factor Analysis (CFA). Precisely, the hypotheses of configural, metric, scalar, and strict measurement invariance have been verified.

The invariance test of the factorial structure was verified through a series of steps (Byrne et al., 2009). Firstly, two confirmatory factorial analyses were carried out, one for boys and one for girls, in order to verify if the model was suitable for each group. Secondly, the measurement model was estimated simultaneously for the two groups by leaving all the parameters free in order to test for configural invariance. Then, the metric invariance was tested by constraining factors loadings to equality across groups. Scalar invariance was then tested by additionally constraining the thresholds of the observed variables to equality across groups. Subsequently, the hypothesis of strict measurement invariance was tested by constraining the residual variances of the observed variables to the equality between boys and girls.

Configural invariance has been considered verified if CFI, TLI were greater than .90, and RMSEA was minor than .06 (Brown, 2015). According to Scott-Lennox and Scott-Lennox (1995), metric, scalar, and strict invariance hypotheses were considered verified across groups, if, for each model compared, the difference in CFI was not greater than 0.010, *and* the chi-square difference test was not significant.

Table S2.1

Measurement invariance of the QUVA-p structure across sex

Models	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	TLI	RMSEA	90% CI	ΔCFI
Boys	2146.77	1371	-	-	.921	.917	.048	.044 –.053	-
Girls	2004.63	1371	-	-	.912	.908	.047	.043–.051	-
Configural Invariance	4151.40	2742	-	-	.912	.908	.050	.047 –.053	-
Metric Invariance	4156.40	2792	47.33^{ns}	50	.914	.912	.049	.046 – .052	-.002
Scalar Invariance	4168.93	2842	54.11^{ns}	50	.917	.916	.048	.045 – .051	-.003
Strict Scalar Invariance	4090.75	2896	74.60*	54	.925	.926	.045	.042 – .048	-.008
Strict Scalar Invariance p1	4073.60	2895	61.71^{ns}	53	.926	.927	.045	.041 – .048	-.009
Factorial var/cov invariance	3811.16	2905	65.83^{ns}	63	.943	.944	.039	.036 – .042	-.017*
Factorial var/cov invariance p1	3888.54	2902	56.06^{ns}	60	.938	.939	.041	.037 – .044	.005
Factorial Means invariance	10539.49	2904	1938.54***	62	.521	.528	.114	.111 – .116	.417*
Factorial Means invariance p1	5521.19	2902	568.14**	60	.836	.838	.067	.064 – .069	.102*

Note. df = degree of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; p1 = partial; ns = non-significant.

* $p < .05$; ** $p < .01$; *** $p < .001$

As seen in Table S2.1, the QUVA-p factorial structure (Model 3) shows a good fit to empirical data in boys and girls. The hypotheses of configural, metric, and scalar invariance are verified ($\Delta\chi^2 \approx 50.72$, *p*s > .05; ΔCFI < .010), thus the number of factors and the pattern of factor-indicators are equal across groups, as well as the factor loadings and the thresholds (see Table S2.1).

Concerning strict measurement invariance, the χ^2 difference between the strict *versus* scalar invariance model was significant ($\Delta\chi^2 = 74.60$, $\Delta df = 50$), $p > .05$). Therefore, strict measurement

invariance was not verified. This means there are some differences in items residuals dispersion, between boys and girls. Thus, the hypothesis of partial strict measurement invariance was tested by removing the constraint referred to item 24 (i.e. "Gets picked on by other children") – as suggested by the Modification Indices – whose residual variance was higher in boys. Finally, the new estimated model shows a non-significant variation in global fit ($\Delta\chi^2 = 61.71$, $\Delta df = 53$; $p > .05$); thus, the hypothesis of a partial invariance of the factorial structure between boys and girls is verified.

In addition, the hypotheses of factorial variances and covariances invariance and factors mean invariance were tested, by introducing constraints on factorial variance and covariances and, subsequently, on factors mean, to be equal across groups.

The hypothesis of factorial variances and covariance invariance is partially verified. Since the *CFI* of the factorial variances and covariances invariance model shows a decrease greater than 0.010, the parameter referred to the covariance between *Cognitive* and *Behavioural Dysregulation* ($\psi_{girls} = 1.506$; $\psi_{boys} = 1.437$), as indicated by the Modification Indices, has been freed. The removal of this constraint caused a non-significant variation in the model fit; thus, the hypothesis of a partial invariance of the factorial variances and covariances is verified across groups. In the same way, also the two latent dimensions mean varies across groups. Even releasing one parameter at a time, the chi-square and CFI difference result significant, thus the factorial mean invariance cannot be verified ($\Delta\chi^2 = 568.14$ $\Delta df = 60$; $p > .05$; $\Delta CFI = .102$) across groups. Girls show significantly lower means at the *behavioral dysregulation* and *socio-relational and adaptive difficulties* subscales (respectively, $M_{boys} - M_{girls} = 1.58$, $p = .001$; $M_{boys} - M_{girls} = 0.97$, $p = .019$) than boys.

In conclusion, the results of the invariance test confirm that the construct validity of the questionnaire is maintained across sex. Relevant sex differences emerge only at the level of factorial means and covariances. In line with the literature (Beaman et al., 2007), teachers have observed, on average, higher frequency of problem behaviours and socio-relational difficulties in

boys. Furthermore, symptoms of cognitive dysregulation were more associated with behavioural dysregulation issues in boys compared to girls.

Inspection of the QUVA-p power to detect EC differences in preschool population

In order to investigate the power of the QUVA-p to detect differences in EC performance between children with and without SR problems (Objective 5, see the main manuscript), n = 45 questionnaires with scores for *Global Self-Regulation Difficulty* greater than the 85th percentile rank were selected (boys = 23, girls = 22, $M = 54.44$, $SD = 9.62$) from the total sample to form a case-study group of at-risk children. From the remaining 371 questionnaires, n = 47 (mdn = 49th percentile) were selected by stratified random sampling to form a control-study group (boys = 21, girls = 26, $M = 56.77$, $SD = 8.70$). MANOVA, Mann-Whitney and chi-square statistics were used to confirm that the two groups are statistically equal in socio-demographic characteristics. A multivariate analysis of variance (MANOVA) was conducted with EC task performance as the dependent variable and the case *versus* the control-study group as the between factor.

The Mann-Whitney and chi-square statistics confirmed substantial equality between the case- and control-study groups for the socio-economic ($U_{\text{mothers}} = 1317.50$, $p = .353$; $U_{\text{fathers}} = 1326.00$, $p = .869$) and socio-cultural status ($U_{\text{mothers}} = 1203.50$, $p = .070$; $U_{\text{fathers}} = 1193.00$, $p = .154$) of families, children's age (χ^2 (2, $N = 92$) = 2.59, $p = .273$) and children's sex (χ^2 (1, $N = 92$) = 0.07, $p = .778$).

The results of the one-way MANOVA with age, sex and family status as covariates showed significant differences in inhibitory control (ST), $F(1, 84) = 6{,}559$, $p = .012$, short-term (FDS) and working memory (BDS), $F(1, 84) = 4.32$, $p = .041$; $F(1, 84) = 4.02$, $p = .048$, respectively, and cognitive flexibility (DCCS), $F(1, 84) = 5.46$, $p = .022$), between the at-risk and control groups. Table S2.2 shows the effect size of these differences, which were moderate (Cohen's d [0.434 – 0.547], Wilks' $\lambda = .805$). With regard to abstract reasoning, no differences were found for the CPM between the groups, $F(1, 84) = 0.70$, $p = .402$). This means that even when controlling for the

effects of sex, age or status, children with severe SR problems – classified as *at-risk* by the QUVA-p – showed poorer EC than typically-developing children. According to these results, the instrument seems to be able to detect EC differences in preschool children.

Table S2.2

Descriptive statistics and Effect Size of the at-risk and the control groups.

Executive Tasks	At-risk group (n = 45)		Control group (n = 47)		Effect Size
	M	*SD*	*M*	*SD*	Cohen's *d*
Forward Digit Span (Range: 0-16)	4.16	1.88	5.21	2.03	0.444
Backward Digit Span (Range: 0-16)	0.98	1.61	1.91	2.08	0.434
Statue (Errors, Range: 0-45)	9.04	10.19	4.06	5.55	0.547
Dimensional Change Card Sort (Range: 0-24)	16.27	5.82	19.49	4.22	0.499
Colored Raven Matrices (Age-corrected percentile rank: 0-99)	56.51	27.73	61.00	24.15	0.171

Note. Cohen's $d = (M_2 - M_1) / SD_{pooled}$