Letter to the editor

Title: Pre-training inter-rater reliability of clinical instruments in an international psychosis research project.

Running title: Pre-training inter-rater reliability.

Total word count: 999.

Key words: Pre-training inter-rater reliability, psychosis instruments, assessor selection.

Authors: Steven Berendsen[1]*, Pim Kapitein[1], Frederike Schirmbeck[1], Mirjam J. van Tricht[1], Philip McGuire[2], Craig Morgan[3], Charlotte Gayer-Anderson[3], Matthew J Kempton[2], Lucia Valmaggia[4], Diego Quattrone[2], Marta di Forti[2], Mark van der Gaag[5], James B. Kirkbride[6], Hannah E Jongsma[6,7], Peter B Jones[7], Maria Parellada[8], Celso Arango[8], Manuel Arrojo[9], Miguel Bernardo[10], Julio Sanjuán[11], José Luis Santos[14], Andrei Szöke[15], Andrea Tortelli[16], Pierre-Michel Llorca[17], Ilaria Tarricone[18], Giada Tripoli[19], Laura Ferraro[19], Caterina La Cascia[19], Antonio Lasalvia[20], Sarah Tosato[20], Paulo Rossi Menezes[21], Cristina Marta Del-Ben[22], Barnaby Nelson[23,24], Anita Riecher-Rössler[25], Rodrigo Bressan[26], Neus Barrantes-Vidal[27,28], Marie-Odile Krebs[29], Merete Nordentoft[41], Stephan Ruhrmann[30], Gabriele Sachs[31], Bart P. F. Rutten[32], Jim van Os[2,32,33], EU-GEI High Risk Study, Eva Velthorst[1,34,35], Lieuwe de Haan[1,36].

EU-GEI High Risk Study Group not mentioned in main author list: Maria Calem[2], Stefania Tognin[2], Gemma Modinos[2], Sara Pisani[2], Tamar C. Kraan[1], Daniella S. van Dam[1], Nadine Burger[44], Patrick McGorry[23], G Paul Amminger[23], Athena Politis[23], Joanne Goodall[23], Stefan Borgwardt[37], Erich Studerus[37], Ary Gadelha[26], Elisa Brietzke[38], Graccielle Asevedo[38], Elson Asevedo[38], Andre Zugman[38], Tecelli Domínguez-Martínez[39], Manel Monsonet[28], Lidia Hinojosa [28], Paula Cristóbal-Narváez[28], Anna Racioppi[28], Thomas R. Kwapil[40], Mathilde Kazes[29], Claire Daban[29], Julie Bourgin[29], Olivier Gay[29], Célia Mam-Lam-Fook[29], Dorte Nordholm[41], Lasse Randers[41], Kristine Krakauer[41], Louise Birkedal Glenthøj[41], Birte Glenthøj[42], Dominika Gebhard[30], Julia Arnhold[43], Joachim Klosterkötter[30], Iris Lasser[31], Bernadette Winklbaur[31], Philippe A Delespaul[32].

*Corresponding author: Steven Berendsen, MD, UMC Amsterdam, Location AMC. Meibergdreef 9, 1105 AZ Amsterdam. Telephone: +3120 566 9111, email: s.berendsen@amsterdamumc.nl.

[1] Department of Psychiatry, Amsterdam UMC, Amsterdam, the Netherlands.

[2] Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, England.

[3] Department of Health Service and Population Research, Institute of Psychiatry, King's College London, London, UK.

[4] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[5] Amsterdam Public Mental Health Research Institute, Department of Clinical Psychology, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[6] PsyLife Group, Division of Psychiatry, UCL, London, England.

[7] Department of Psychiatry, University of Cambridge, Cambridge, England

[8] Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón, School of Medicine, Universidad Complutense, Instituto de Investigación Sanitaria Gregorio Marañón (IiSGM), Spanish Mental Health Research Network (CIBERSAM), Madrid, Spain

[9] Department of Psychiatry, Instituto de Investigación Sanitaria (IDIS), Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain

[10] Barcelona Clinic Schizophrenia Unit, Neuroscience Institute, Hospital Clinic of Barcelona, University of Barcelona, Barcelona, August Pi I Sunyer Biomedical Research Institute (IDIBAPS), Spanish Mental Health Research Network (CIBERSAM), Spain

[11]Department of Psychiatry, Hospital Clínico Universitario de Valencia, School of Medicine, Universidad de Valencia, Valencia, Spain

[12]Department of Psychiatry, University of Oviedo, Spanish Mental Health Research Network (CIBERSAM), Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Mental Health Services of Principado de Asturias (SESPA), Oviedo, Spain

[13]Neurobiological Research Group, Institute of Technology, Universidad de Castilla-La Mancha, Cuenca, Spain

[14]Department of Psychiatry, Hospital "Virgen de la Luz", Cuenca, Spanish Mental Health Research Network (CIBERSAM), SpainINSERM, U955, Créteil, France

[15]Etablissement Public de Santé Maison Blanche, Paris, France

[16]Centre Hospitalier Universitaire de Clermont-Ferrand, Clermont-Ferrand, France

[17]Department of Medical and Surgical Sciences, Psychiatry Unit, Alma Mater Studiorum Università di Bologna, Bologna, Italy

[18]Department of Biomedicine, Neuroscience and advanced Diagnostics, University of Palermo, Palermo, Italy

[19]Section of Psychiatry, Department of Neuroscience, Biomedicine and Movement, University of Verona, Verona, Italy

[20]Department of Preventive Medicine, Faculdade de Medicina, Universidade of São Paulo, São Paulo, Brazil

[21]Ribeirão Preto Medical School, University of São Paulo, Brazil

[22]Orygen, Parkville, Victoria, Australia

[23]Centre for Youth Mental Health, The University of Melbourne, Parkville, Victoria, Australia

[24]Medizinische Fakultät, Universität Basel, Basel, Switzerland

[25]LiNC-Lab Interdisciplinar Neurociências Clínicas, Depto Psiquiatria, Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil

[26]Departament de Psicologia Clínica i de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain

[27]Fundació Sanitària Sant Pere Claver, Spanish Mental Health Research Network (CIBERSAM), Spain

[28]University of Paris, GHU-Paris, Sainte-Anne, C'JAAD, Inserm U1266, Institut de Psychiatrie (CNRS 3557), Paris, France

[29]Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany

[30]Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria

[31]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Centre, Maastricht, the Netherlands
[32]Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[33]Seaver Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, NY, USA
[34]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, NY, USA
[35]Arkin Institute for Mental Health, Amsterdam, the Netherlands
[36]University of Basel, Basel, Switzerland
[37]Department of Psychiatry, Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
[38]CONACYT-Dirección de Investigaciones Epidemiológicas y Psicosociales, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, México
[39]Department of Psychology, University of Illinois at Urbana-Champaign, IL, USA
[40]Mental Health Center Copenhagen and Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research, CINS, Mental Health Center Glostrup, Mental Health Services in the Capital Region of Copenhagen, University of Copenhagen, Copenhagen, Denmark
[41]Centre for Neuropsychiatric Schizophrenia Research (CNSR) & Centre for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS), Mental Health Centre Glostrup, University of Copenhagen, Glostrup, Denmark
[42]Psyberlin, Berlin, Germany
[43]Parnassia Psychiatric Institute, Department of Psychosis Research, Zoutkeetsingel 40, 2512 HN The Hague, The Netherlands.

Dear Editor,

Inter-rater reliability (IRR) is an important component of methodology to establish valid results and prevent large measurement errors. However, only a minority of reports in psychiatric research present information concerning assessor training or reliability of applied instruments. For example, a recent study found that IRR coefficients and training procedures were strongly underreported in double-blind RCTs of antipsychotic medication(Berendsen et al., 2020).

IRR scores without training of raters are typically low, with only four studies investigated pre-training IRR (Muller and Dragicevic, 2003, Muller and Wetzel, 1998, Rosen et al., 2008, Loevdahl and Friis, 1996). ROSEN VERWIJDEREN The authors reported that the IRR scores of the PANSS, HAM-D or GAF [full titles available in the supplement] before training were generally moderate to poor, other observational instruments were not investigated. On the other hand, the authors reported significant improvement of the IRR after assessors were trained.

Selection of assessors based on their clinical backgrounds and assessment experience may also lead to improved pre-training IRR scores. However, merely three studies addressed the topic of assessor selection and pre-training reliability. The first study of Kobak et al. provided evidence that assessors with a PhD or medical degree showed significantly higher HAM-D clinical assessment skills necessary to conduct reliable assessments compared to assessors with lower educational degrees (Kobak et al., 2005). In contrast, Loevdahl et al. and Kollias et al. found no differences in pre-training reliability of the GAF or the CAARMS between psychiatrists, residents, psychologists and nurses (Loevdahl and Friis, 1996, Kollias et al., 2015).

This raises the question whether acceptable IRR scores can be achieved without assessor training or selection. Therefore, we aimed to determine the pre-training IRR of seven observational instruments that capture different aspects of psychosis in a large international multi-center research project by scoring video-taped interviews. In addition, we investigated the effect of assessor characteristics on pre-training IRR scores.

Assessors of the large multi-center study EU-GEI were instructed to rate participants on seven instruments via an online training platform (van Os et al., 2014). These instruments were chosen to measure predictors and outcome in psychosis. Ratings were based on videotaped assessments of interviews with actors playing the role of the patient. Demographic characteristics (age and gender), professional background (psychiatrists, psychologists, medical doctors or research assistants) and assessment experience (in months) of assessors were collected. The pre-training IRR of the following instruments were evaluated: CAARMS, SIS-R, LoTE, BQ, CECA, OPCRIT and GAF.

Pre-training IRR was calculated by Krippendorff's alpha (K-alpha) (Hayes and Krippendorff, 2007). According to interpretation guidelines, K-alpha values of >0.8 were considered high, 0.67 - 0.8 moderate, and <0.67 low (Krippendorff, 2011). For each K-alpha 95% confidence intervals were computed based on 10.000 bootstraps. Differences in age, assessment experience and IRR between different professional groups were analyzed for each assessment instrument using analysis of variance (ANOVA), followed by Bonferroni corrected pair-wise post-hoc comparisons.

**Table 1.*

In total 12 psychiatrists, 17 psychologists, 14 medical doctors and 13 research assistants participated in the online training platform. Mean age [30.18 years, F=13.43, p<0.001; see supplement table 1] and assessment experience (F=5,76, p=0.002; see supplement figure 1) were significantly higher for psychiatrists compared to medical doctors and research assistants, and at trend level compared to psychologists.

Observed pre-training IRR score was moderate for LoTE (K-alpha =0.67), low for GAF (K-alpha=0,45), BQ (K-alpha =0.47), SIS-R (K-alpha = 0.55), CAARMS (0,57), CECA (K-alpha =0.60) and OPCRIT (K-alpha =0.64).

IRR scores of subgroups are shown in Table 1. Overall mean IRR scores were significantly higher for psychiatrists compared to medical doctors (F=3,905, p= 0.0216). Comparisons for separate instruments showed significantly higher IRR scores for psychiatrists, psychologists and research assistants compared to medical doctors on the OPCRIT (F=18,38, p=<0.001), SIS-R (F=20,66, p=<0.001), GAF (F=12,53, p=<0.001) and CAARMS (F=13,34, p=<0.001). Additionally, medical doctors and research assistants scored significantly higher IRR scores compared to psychiatrists and psychologists on the BQ (F=16,75, p=<0.001). For detailed information on pair-wise comparisons of IRR scores between professionals and assessment experience see supplement figures 2a-2f.

Our study demonstrated that only one instrument showed moderate pre-training IRR, whereas the observed reliability scores of all other instruments were insufficient. Furthermore, medical doctors demonstrated significantly lower reliability scores compared to other professional subgroups in mean IRR ratings and several investigated instruments. These findings are important, particularly in light of previous research noting that rater training was strongly underreported and the impact of unreliability on study outcome (Mulsant et al., 2002, Kobak et al., 2007).

Our findings are in accordance with earlier results concerning insufficient pre-training IRR (Vatnaland et al., 2007, Muller and Dragicevic, 2003, Muller and Wetzel, 1998, Loevdahl and Friis, 1996). Differences in mean IRR scores between professions could be explained by the significantly higher assessment experience of psychiatrists compared to the other professions. However, observed IRR scores of separate instruments were also different between psychologists and research assistants compared to medical doctors, while the latter two subgroups did not significantly differ in assessments experience. Our hypothesis concerning the latter variation is that research assistants and psychologist probably received more training in psychopathology scales such as the CAARMS or SIS-R during their general education, in comparison to medical doctors.

Our findings concerning differences between professionals seem to contrast with previous literature, which found no significant differences in pre-training IRR of GAF scores between psychiatrists and psychologists, compared to psychiatric nurses (Loevdahl and Friis, 1996). Similarly, another study concerning the CAARMS provided evidence that psychiatry residents produced almost similar IRR scores compared to psychiatrists and psychologists (Kollias et al., 2015). Possible explanations for these inconsistent findings could be that psychiatry residents have more experience with observational instruments and psychiatric diagnosis compared to medical doctors.

Of note, we evaluated *pre-training* IRR in this report. All included researchers achieved high IRR scores after training before permitted to perform assessments. However, we should acknowledge an important limitation of our study: we do not have data concerning previous training or clinical background of raters.

In conclusion, our study emphasizes the importance of rater training and assessor selection for research in psychiatry. Without rater training, reliability is generally insufficient. This has potentially major implications for the interpretation of study-results because of decreased power and higher placebo-response[*see supplement] (Perkins et al., 2000, Kobak et al., 2010). Future research should focus on specific assessors characteristics that predict higher IRR scores after training. Finally, considering its importance, we propose training procedures and reliability coefficients should be reported in all studies.

References

BERENDSEN, S., VAN, H. L., VERDEGAAL, L. M. A., VAN TRICHT, M. J., BLANKERS, M. & DE HAAN, L. 2020. Burying Our Heads in the Sand: The Neglected Importance of Reporting Inter-Rater Reliability in Antipsychotic Medication Trials. *Schizophr Bull*.

HAYES, A. F. & KRIPPENDORFF, K. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. . *Communication Methods and Measures*.

KOBAK, K. A., KANE, J. M., THASE, M. E. & NIERENBERG, A. A. 2007. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol,* 27**,** 1-5.

KOBAK, K. A., LEUCHTER, A., DEBROTA, D., ENGELHARDT, N., WILLIAMS, J. B., COOK, I. A., LEON, A. C. & ALPERT, J. 2010. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. *J Clin Psychopharmacol,* 30**,** 193-7.

KOBAK, K. A., LIPSITZ, J. D., WILLIAMS, J. B., ENGELHARDT, N. & BELLEW, K. M. 2005. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol,* 25**,** 407-12.

KOLLIAS, C., KONTAXAKIS, V., HAVAKI-KONTAXAKI, B., SIMMONS, M. B., STEFANIS, N. & PAPAGEORGIOU, C. 2015. Inter-rater reliability of the Greek version of CAARMS among two groups of mental health professionals. *Psychiatriki,* 26**,** 217-22.

KRIPPENDORFF 2011. Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*.

LOEVDAHL, H. & FRIIS, S. 1996. Routine evaluation of mental health: reliable information or worthless "guesstimates'? *Acta Psychiatr Scand,* 93**,** 125-8.

MULLER, M. J. & DRAGICEVIC, A. 2003. Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *J Affect Disord,* 77**,** 65-9.

MULLER, M. J. & WETZEL, H. 1998. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. *Acta Psychiatr Scand,* 98**,** 135-9.

MULSANT, B. H., KASTANGO, K. B., ROSEN, J., STONE, R. A., MAZUMDAR, S. & POLLOCK, B. G. 2002. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry,* 159**,** 1598-600.

PERKINS, D. O., WYATT, R. J. & BARTKO, J. J. 2000. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry,* 47**,** 762-6.

ROSEN, J., MULSANT, B. H., MARINO, P., GROENING, C., YOUNG, R. C. & FOX, D. 2008. Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. *Psychiatry Res,* 161**,** 126-30.

VAN OS, J., RUTTEN, B. P., MYIN-GERMEYS, I., DELESPAUL, P., VIECHTBAUER, W., VAN ZELST, C., BRUGGEMAN, R., REININGHAUS, U., MORGAN, C., MURRAY, R. M., DI FORTI, M., MCGUIRE, P., VALMAGGIA, L. R., KEMPTON, M. J., GAYER-ANDERSON, C., HUBBARD, K., BEARDS, S., STILO, S. A., ONYEJIAKA, A., BOURQUE, F., MODINOS, G., TOGNIN, S., CALEM, M., O'DONOVAN, M. C., OWEN, M. J., HOLMANS, P., WILLIAMS, N., CRADDOCK, N., RICHARDS, A., HUMPHREYS, I., MEYER-LINDENBERG, A., LEWEKE, F. M., TOST, H., AKDENIZ, C., ROHLEDER, C., BUMB, J. M., SCHWARZ, E., ALPTEKIN, K., UCOK, A., SAKA, M. C., ATBASOGLU, E. C., GULOKSUZ, S., GUMUS-AKAY, G., CIHAN, B., KARADAG, H., SOYGUR, H., CANKURTARAN, E. S., ULUSOY, S., AKDEDE, B., BINBAY, T., AYER, A., NOYAN, H., KARADAYI, G., AKTURAN, E., ULAS, H., ARANGO, C., PARELLADA, M., BERNARDO, M., SANJUAN, J., BOBES, J., ARROJO, M., SANTOS, J. L., CUADRADO, P., RODRIGUEZ SOLANO, J. J., CARRACEDO, A., GARCIA BERNARDO, E., ROLDAN, L., LOPEZ, G., CABRERA, B., CRUZ, S., DIAZ MESA, E. M., POUSO, M., JIMENEZ, E., SANCHEZ, T., RAPADO, M., GONZALEZ, E., MARTINEZ, C., SANCHEZ, E., OLMEDA, M. S., DE HAAN, L., VELTHORST, E., VAN DER GAAG, M., SELTEN, J. P., VAN DAM, D., VAN DER VEN, E., VAN DER MEER, F., MESSCHAERT, E., KRAAN, T., BURGER, N., LEBOYER, M., SZOKE, A., SCHURHOFF, F., LLORCA, P. M., JAMAIN, S., TORTELLI, A., FRIJDA, F., VILAIN, J., GALLIOT, A. M., BAUDIN, G., FERCHIOU, A., et al. 2014. Identifying gene-environment interactions in schizophrenia: contemporary challenges for integrated, large-scale investigations. *Schizophr Bull,* 40**,** 729-36.

VATNALAND, T., VATNALAND, J., FRIIS, S. & OPJORDSMOEN, S. 2007. Are GAF scores reliable in routine clinical use? *Acta Psychiatr Scand,* 115**,** 326-30.