19.00 €

vol. LXVIII. September-December 2020

N° 289

3/3

Four-monthly Journal
ISSN 0006-6761

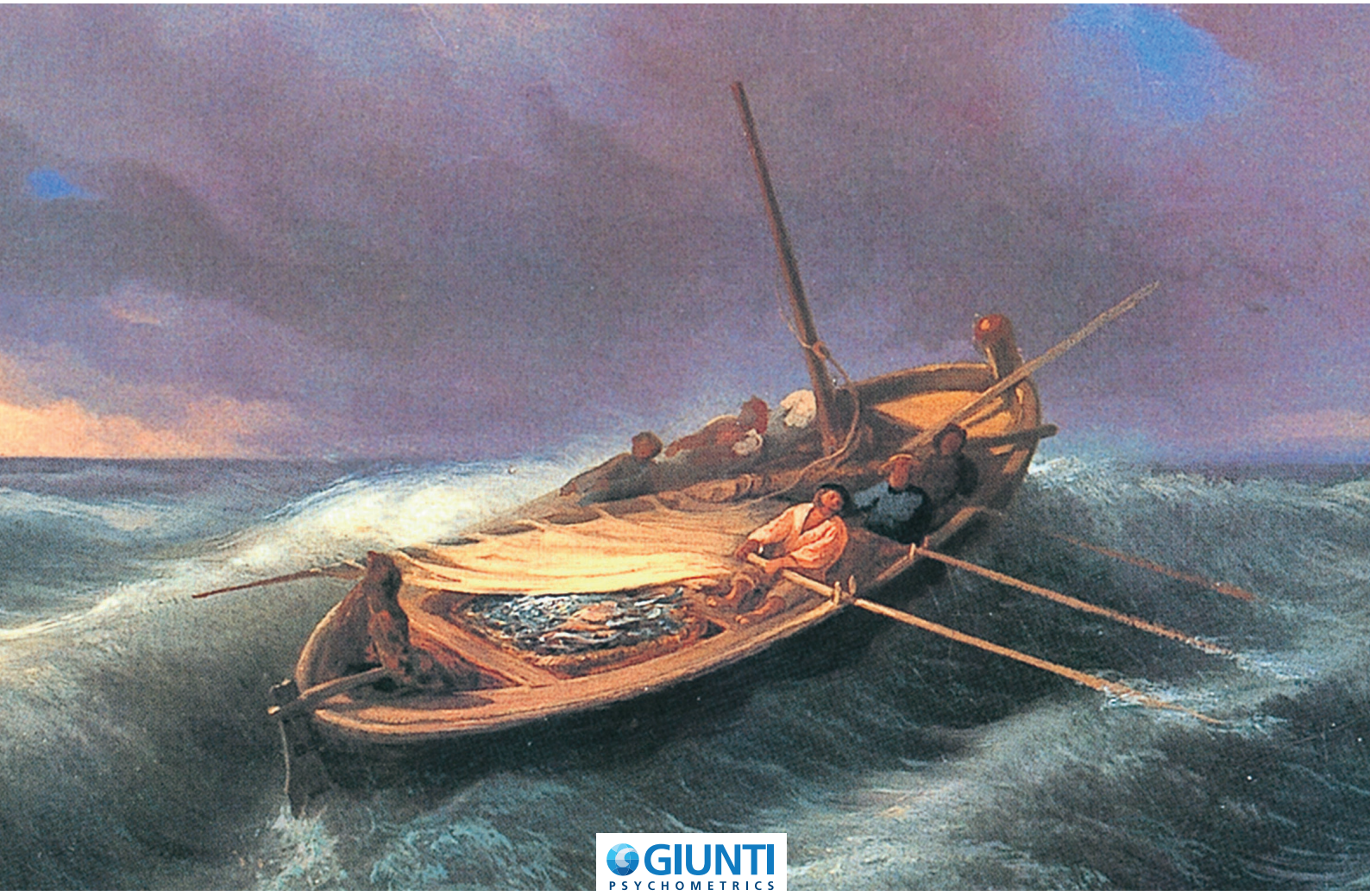
BPA

Bollettino di Psicologia Applicata

APPLIED PSYCHOLOGY BULLETIN

Indexed in PsycINFO® – Scopus Bibliographic Database

Research | Review | Experiences & Tools



GIUNTI PSYCHOMETRICS

CONTENTS

⬢ Research

■ Review

▲ Experiences & Tools

DOI: 10.26387/bpa.289.1

Depression at the time of the COVID-19 pandemic: A CES-D survey before and during the lockdown

Sabrina Rizzo, Luciano Giromini, Claudia Pignolo
Department of Psychology, University of Turin
claudia.pignolo@unito.it

ABSTRACT. Il 31 gennaio del 2020, in Italia, furono registrati i primi due casi di COVID-19; la portata e la rapidità della diffusione del virus costrinsero il governo italiano ad adottare delle misure di emergenza straordinarie per rallentare il contagio. L'obiettivo del presente studio è stato quello di indagare la sintomatologia depressiva sperimentata da un campione proveniente dalla popolazione generale, confrontando i dati raccolti prima e durante il lockdown da COVID-19. Sono stati utilizzati dati d'archivio che includevano dati relativi alla *Center for Epidemiologic Studies-Depression scale* (*CES-D*) somministrata prima (*n* = 151; gruppo Pre-COVID-19) e durante (*n* = 352; gruppo COVID-19) il primo lockdown italiano a un campione non-clinico. Le analisi si sono focalizzare sul confronto dei punteggi ottenuti alla CES-D nel gruppo Pre-COVID-19 e nel gruppo COVID-19. Inoltre, è stata valutata la possibile influenza di alcune variabili demografiche sui punteggi ottenuti alla CES-D prima e durante la pandemia da COVID-19. Inoltre, all'interno del gruppo COVID-19, sono state osservate delle differenze statisticamente significative tra i punteggi alla CES-D ottenuti da uomini e donne e una correlazione quasi significativa tra l'età dei partecipanti e i punteggi ottenuti alla CES-D. Lo studio ha rivelato che le misure restrittive e la pandemia stessa possono aver contribuito ad un incremento dei sintomi depressivi in un campione di individui non-clinici (e probabilmente nella popolazione generale italiana), specialmente nei giovani e nelle donne.

SUMMARY. On January 31, 2020, the first two cases of COVID-19 were detected in Italy; the extent and the rapidity of virus spread forced the Italian Government to take extraordinary measures to prevent contagion. In this study, we aimed to compare data collected before and during the COVID-19 pandemic on the depression symptomatology in a sample from the general population. We used archival data from a previous dataset we had access to, which included Center for Epidemiologic Studies-Depression scale (CES-D) data from non-clinical volunteers collected before (n = 151; Pre-COVID-19 group) and during (n = 352; COVID group) the pandemic. Statistical analyses compared CES-D scores yielded by the Pre-COVID-19 sample against those yielded by the COVID-19 sample. Additionally, the possible impact of demographic variables on CES-D scores before and during COVID-19 was tested. Moreover, in the COVID-19 group we found a statistically significant difference on the CES-D scores between men and women and a nearly significant relationship between age and CES-D scores. This study showed that the lockdown measures and the pandemic itself might have led to an increasing of the depressive symptoms in a non-clinical sample (and maybe in the Italian population), especially in women and youths.

Keywords: *COVID-19, Sars-Cov-2, Pandemic, Depression, Lockdown, Mental health, Women, Youths*

INTRODUCTION

Italy was the first European country to be hit by the COVID-19 pandemic, with about 200.000 confirmed cases and 30.000 deaths between March and May 2020 (https://covid19.who.int/region/euro/country/it). To mitigate virus diffusion, the Italian Government implemented emergency measures, based on the Chinese experience, including home confinement and limitation on movement in the entire country, except for justified work reasons and health needs. As such, lockdown was officially proclaimed on March 9th, 2020, and gradually extended until May 18th, 2020. The Italian lockdown was one of the most stringent ones in Europe, in terms of duration and intensity: it involved schools, universities, and almost all fields of business, the converting of many hospital wards or of whole hospitals into pandemic centers, social-distancing and self-isolation, and an unexpected drastic change of daily life. All these elements added fears to fears, and uncertainty to uncertainty, contributing to create an unprecedented situation in every aspects of life (Porcelli, 2020). Being constantly exposed to information about the pandemic, not having definite answers on its duration or effects, and feeling one's own balance threatened, can indeed affect individuals' mental health (Özdin & Bayrak Özdin, 2020).

As such, COVID-19 pandemic marked the beginning of a series of psychological processes and reactions that will interest clinicians and researchers for a long time. The most common individuals' psychological reactions to COVID-19 were depression, stress, anxiety, and sleep disorders (Ahmed et al., 2020; Choi, Hui & Wan, 2020; Huang & Zhao, 2020; Ozamiz-Etxebarria et al., 2020; Rossi et al., 2020a; Wang et al., 2020). These psychological reactions were stronger over time especially in individuals who were subjected to more restrictive measures of virus spread containment and who were exposed first to the pandemic (Choi et al., 2020; Ozamiz-Etxebarria, Dosil-Santamaria, Picaza-Gorrochategui & Idoiaga-Mondragon, 2020; Wang et al., 2020). Furthermore, post-traumatic stress and adjustment disorder symptoms were identified and correlated to measures of quarantine (Rossi, Socci, Pacitti et al., 2020; Rossi, Socci, Talevi et al., 2020). The level of stress was often associated with several COVID-19-related risk factors, such as losing jobs, having a loved one seriously threaten by the virus, being under quarantine, and the request to adapt to new way of working, studying, and communicating (Buonsenso, Cinicola, Raffaelli, Sollena & Iodice, 2020; Buzzi et al., 2020; Rossi, Socci, Talevi et al., 2020; Wang et al., 2020).

Referring to depression, different studies have found associations with demographic variables. In Italy, Mazza et al. (2020) assessed psychological distress variables in a sample from the Italian general population in March 2020 finding that 67.2% of the sample reported average levels of depression, whereas 32.8% reported high or very high levels of depression. In addition, they found that higher levels of depression were found in individuals with a lower level of education and in women, although they did not find any relationship with the age of the participants. In Italy, almost 50% of women had to renounce to their plans for the future because of the increased workload and 60% of them (versus 21% of men) had to manage alone family, children, and elders (https://alleyoop.ilsole24ore.com/2020/05/28/la-donna-tra-le-vittime-del-covid-una-su-due-rinuncia-ai-propri-progetti/). Furthermore, an Istat report revealed that on May 2020 more women than men lost their job (.7% *vs* .1%; https://www.istat.it/it/archivio/245093). In general, international studies have reported that women were more frequently associated with increased psychological distress during the pandemic (Qiu et al., 2020; Wang et al., 2020). Moreover, youths appeared to suffer more the psychological effects of the pandemic and lockdown compared to older people (Ahmed et al., 2020; Huang & Zhao, 2020; Odriozola-González, Planchuelo-Gómez, Irurtia & De Luis-García, 2020).

AIM

The aim of the current research was to compare data collected before and during the COVID-19 pandemic on the depression symptomatology in a sample from the general population. More specifically, the current cross-sectional study investigated the effects of isolation and social distancing on the onset of depressive symptomatology by comparing archival *Center for Epidemiologic Studies – Depression* (*CES-D*; Radloff, 1977) data collected before the spread of COVID-19 against those collected during the pandemic. Moreover, we also tested the extent to which demographic variables such as gender, age, and education were associated with the CES-D scores before and during the lockdown.

METHOD

Participants

Both the Pre-COVID-19 and COVID-19 groups were originally recruited with the snowball sampling technique to contribute to the study of the psychometric properties of the *Inventory of Problems – 29* (*IOP-29*; Viglione & Giromini, 2020; Viglione, Giromini & Landis, 2017), a recently introduced feigning measure. In addition to the IOP-29, all participants included in that sample were administered the CES-D and were asked to provide demographic information such as age, gender, and years of education. As data collection for that project occurred before and during the spread of COVID-19 pandemic, this dataset represents an optimal source of information for the goals of the current study.

The composition of both groups is reported in Table 1. The group recruited before the pandemic was composed of 151 adults, 57 men and 94 women, of Italian nationality, aged between 18 and 74 years old, and with an education level that ranged from 8 to 21 years. Among these participants, four did not provide any information about their education level. Moreover, geographical provenience was not reported. The group recruited during the COVID-19 pandemic was composed of 353 Italian adults, 114 men and 239 women, ranging in age between 18 and 60 years old, with a level of education ranging from 8 to 21 years. Most of the participants were native of the North-West (44.2%) and of the South of Italy (34.3%). Two participants did not report on their education level; ten did not disclose their geographical provenience.

Measures

The Italian version of the CES-D (Radloff, 1977; Italian version adapted by Fava, 1983) was administered through an online self-report survey to the participants, in order to detect depression symptoms before and during the COVID-19 pandemic. The CES-D was originally developed to measure depressive symptomatology in epidemiological studies about the general population (Radloff, 1977); however, it has also been used in primary care settings (Andresen, Malmgren, Carter & Patrick, 1994; Miller, Anton & Townson, 2008; Myers & Weissman, 1980; Vilagut, Forero, Barbaglia & Alonso, 2016). The questionnaire is a 20-item measure developed to explore the construct of depression through a 4-points Likert scale rating. The examinee is asked to specify the frequency with which each symptom was experienced over the last week (0 = Not at all or less than one day last week; 1 = It occurred a few times – one or two days last week; 2 = It occurred frequently – three to four days last week; 3 = It occurred always, or nearly always – five to seven days last week). CES-D items measure different depression symptomatic areas, i.e. negative affect, positive affect, and somatic symptoms (Al-Modallal, 2010), and they can be understood within the frame provided by Beck's Cognitive Theory of Depression (Beck, 1967; 1987; Zauszniewski & Graham, 2008). The CES-D total score has a possible range of 0-60, where a higher score suggests that more depression symptoms are experienced.

Research shows that CES-D scores possess good internal consistency, with alpha values ≥.85 (Spijker et al., 2004; Stockings et al., 2015; Tran et al., 2019; Zauszniewski & Graham, 2008), as well as a good test-retest reliability, construct validity, and concurrent validity (Spijker et al., 2004). Vilagut et al. (2016) inspected different possible cut-scores for the CES-D and found that a cut-off score ≥16 enhanced sensitivity (*Se* = .87; 95% CI .82-.91) over specificity (*Sp* = .70; 95% CI .65-.75); a cut-off score ≥20 produced *Se* = .83 (95% CI .75-.89) and S*p* = .78 (95% CI .71-.83); finally, a cut-off score ≥22 yielded similar results in both sensitivity (*Se* = .79; 95% CI .69-.85) and specificity (*Sp* = .80; 95% CI .75-.85). On the basis of Vilagut et al.'s (2016) findings, we chose to observe the trend of depression in the Italian population selecting the scores of 16, 20, and 22 as cut-off scores.

Procedure

All the participants gave their informed consent, and those who were not able to read and understand Italian fluently were excluded from the research. Additional exclusion criteria included having a history of severe psychiatric disorder, being younger than 18 years of age, not holding Italian citizenship, and not living in Italy during the lockdown period. The original research project received formal ethical approval by the Institutional Review Boards (approved November 19, 2019; Protocol Number 5072).

Table 1 – Demographic composition of the samples

	Pre-COVID-19 (*n* = 151)	**COVID-19 (*n* = 352)**
Gender		
Women	94 (62.3%)	238 (67.6%)
Men	57 (37.7%)	114 (32.4%)
Age		
M	30.97	34.67
SD	13.52	13.23
Education (yrs.)		
M	14.51	14.90
SD	3.03	2.57
Geographical provenience		
North-West	–	155 (44%)
North-East	–	11 (3.1%)
Centre	–	38 (10.8%)
South	–	121 (34.4%)
Islands	–	17 (4.8%)

Data analyses

To evaluate whether there were any differences between the CES-D scores before and during the COVID-19 pandemic, we computed a *t*-test for independent samples. Next, we evaluated whether the percentage of participants who scored above the CES-D clinical cut-off scores varied before and during the lockdown. To do so, we computed *Phi* coefficients applying the most commonly used cut-offs on the CES-D scores, i.e., ≥16, ≥20, and ≥22. Finally, we explored the relationship between the CES-D scores and demographic characteristics within each sample.

RESULTS

The COVID-19 group (M = 20.5, SD = 10.6) showed statistically significant higher CES-D scores compared to the Pre-COVID-19 group (M = 18.1, SD = 10.6; $t_{(501)} = -2.24$; $p = .025$; $d = .22$). Thus, the severity of depressive-related problems reported by the COVID-19 group was significantly greater than that reported by the Pre-COVID-19 group. Moreover, although the percentage of individuals who scored above the cut-off of 16 at the CES-D was significantly higher in the COVID-19 group compared to the Pre-COVID-19 group, we did not find any statistically

significant differences using the other two cut-scores, i.e., ≥20, and ≥22 (see Table 2).

Finally, considering the relationship between the CES-D scores and demographic characteristics, women showed higher CES-D scores compared to men during the COVID-19 lockdown only, with a small effect size; no gender differences were observed in the Pre-COVID-19 group (see Table 3). Furthermore, we correlated the CES-D scores of the two groups with the age of the participants. In the Pre-COVID-19 group, the correlation with age was not significant (r = .038; p = .646), whereas in the COVID-19 group the negative correlation between the depressive symptomatology and age was on the cut-off for statistical significance, with a small effect size (r = −.104; p = .051). As for the correlation between the CES-D scores and the years of education we did not find any statistically significant results in either group (Pre-COVID-19 group: r = −.022; p = .079; COVID-19 group: r = −.054; p = .311).

DISCUSSION

The main objective of the current study was to compare CES-D data collected before and during the COVID-19 lockdown in a sample from the general population. By comparing data collected in the fall 2019 with those gathered in the spring 2020 during the pandemic, we confirmed our expectations. We found a worsening of the depressive symptoms in the COVID-19 group, thus confirming the results of previous studies (Huang & Zhao, 2020; Lei et al., 2020; Özdin & Bayrak Özdin, 2020; Pappa et al., 2020). Another interesting result refers to the percentage of participants who scored above the screening cut-off score of 16 on the CES-D Total score during the lockdown. Indeed, while before the pandemic only 47% of the sample scored at or above the cut-off, during the lockdown 63% percent of the sample reported some depressive symptoms. This pattern of results, however, did not remain statistically

Table 2 – Percentage of above-threshold CES-D scores before and during COVID-19 pandemic

	Pre-COVID-19 (*n* = 151)	**COVID-19 (*n* = 352)**	***Phi***	***p***
CES-D Total ≥16	71 (47%)	223 (63%)	.152	.001
CES-D Total ≥20	61 (40%)	169 (48%)	.070	.116
CES-D Total ≥22	53 (35%)	152 (43%)	.075	.091

Table 3 – CES-D scores before and during COVID-19, divided by gender

	Men		**Women**					
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
Pre-COVID-19	17.65	9.83	18.52	12.11	−.46	149	.650	−.08
COVID-19	18.27	10.32	21.64	10.58	−2.82	350	.005	−.32

significant when we inspected the more conservative CES-D cut-off scores of 20 and 22, although a trend towards the same direction was observed (i.e., higher percentage of cases above threshold during COVID-19 pandemic). One explanation may rely on the nature of the sample: since they were non-clinical volunteers, the overall phenomenon could perhaps be observable only at a subclinical level. In other words, these participants seemed to have experienced an increased amount of depressive symptoms, but being healthy individuals, maybe they still had coping mechanisms to remain at a subclinical level. Nevertheless, these findings represent another confirmation of the consequences that a global health emergency can produce in a non-clinical population. The COVID-19 pandemic, indeed, being an extraordinary alarming situation that have threatened the health and freedom of the entire population, forced the population to cope with the existential concerns and the life changings that this situation has produced.

Our findings indicate that women and young adults were the most affected by the lockdown in terms of depression symptomatology. Indeed, young individuals may have suffered from the restrictions posed by the government more than older individuals did. A possible explanation for this finding is that youths are characterized by the need of relations with peer and of social gatherings, which were prevented by the lockdown. Furthermore, the pre-existing fear of the future that characterize the new generations (Buzzi et al., 2020) has been exacerbated by the economic crisis that COVID-19 pandemic produced. As for women, besides being more predisposed to depression (Maji, 2018; Noble, 2005; Thornton, McQueen, Rosser, Kneale & Dixon, 1997), they may have to face more serious social and economic consequences of the pandemic compared to men. Several organizations for human rights, like Amnesty International or United Nations (UN), launched appeals to politicians and citizens, driven by the worry about the actual destinies of women in almost every country in the world. In Italy, as mentioned before, women had to deal with the social and economic consequences of the pandemic much more than men, and this renews the inequalities already present in the society. An Istat report (https://www.istat.it/it/files//2020/05/Stat-today_Chiamate-numero-antiviolenza.pdf) showed the increase of domestic violence complaints during the quarantine period. In this frame, we can easily understand the gendered impact of COVID-19 (Wenham, Smith & Morgan, 2020) and the importance of a gendered approach in the crisis management. COVID-19 pandemic has perhaps further exacerbated the gender inequalities pre-existing in the Italian territory, and this can represent a risk factor for the increase of depressive symptoms in women.

Despite the interesting findings, we have some limitations to report. First, because the participants of the Pre-COVID-19 and COVID-19 groups are different, we could not evaluate directly differences in the experiencing of the depressive symptomatology, that is the limitation of cross-sectional studies. Second, we used only a self-report scale (i.e., the CES-D) to assess the depressive symptomatology in our sample without administering other psychological tools different in nature, such as clinical interviews, performance tests, or informant-reports. As such, we were able to assess only the subjective perspective of the participants on the matter, which depends on the self-awareness and insight of the participants. Third, given that the CES-D does not include validity scales, we did not evaluate the presence of negative impression management or response styles. As such, some participants may have adopted an intentional or unintentional response style, exaggerating or minimizing their experiences. Nevertheless, we have to face up a crisis that is upsetting the balances of the human life that we knew. For this reason, the mental health professionals should deeply analyze the short and long-term consequences that Sars-Cov-2 will bring with itself, with the aim of defining the best strategies to respond to the new individual and social needs, and of trying to deal with this dramatic situation in the best way as possible.

References

AHMED, M.Z., AHMED, O., AIBAO, Z., HANBIN, S., SIYU, L. & AHMAD, A. (2020). Epidemic of COVID-19 in China and associated psychological problems. *Asian Journal of Psychiatry*, 102092.

AL-MODALLAL, H. (2010). Screening depressive symptoms in Jordanian women: Evaluation of the Center for Epidemiologic Studies-Depression scale (CES-D). *Issues in Mental Health Nursing*, *31* (8), 537-544.

ANDRESEN, E.M., MALMGREN, J.A., CARTER, W.B. & PATRICK, D.L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventive Medicine*, *10* (2), 77-84.

BECK, A.T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper & Row.

BECK, A.T. (1987). Cognitive models of depression. *Journal of Cognitive Psychotherapy, An International Quarterly*, *1*, 5-37.

BUONSENSO, D., CINICOLA, B., RAFFAELLI, F., SOLLENA, P. & IODICE, F. (2020). Social consequences of COVID-19 in a low resource setting in Sierra Leone, West Africa. *International Journal of Infectious Diseases*, *97*, 23-26.

BUZZI, C., TUCCI, M., CIPRANDI, R., BRAMBILLA, I., CAIMMI, S., CIPRANDI, G. & MARSEGLIA, G.L. (2020). The psycho-social effects of COVID-19 on Italian adolescents' attitudes and behaviors. *Italian Journal of Pediatrics, 46*, 1-7.

CHOI, E.P.H., HUI, B.P.H. & WAN, E.Y.F. (2020). Depression and anxiety in Hong Kong during COVID-19. *International Journal of Environmental Research and Public Health*, *17* (10), 3740.

FAVA, G.A. (1983). Assessing depressive symptoms across cultures: Italian validation of the CES-D self-rating scale. *Journal of Clinical Psychology*, *39* (2), 249-251.

HUANG, Y. & ZHAO, N. (2020). Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: A web-based cross-sectional survey. *Psychiatry Research*, *288* (112954).

LEI, L., HUANG, X., ZHANG, S., YANG, J., YANG, L. & XU, M. (2020). Comparison of prevalence and associated factors of anxiety and depression among people affected by versus people unaffected by quarantine during the COVID-19 epidemic in southwestern China. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, *26*, e924609-1.

MAJI, S. (2018). Society and 'good woman': A critical review of gender difference in depression. *International Journal of Social Psychiatry*, *64* (4), 396-405.

MAZZA, C., RICCI, E., BIONDI, S., COLASANTI, M., FERRACUTI, S., NAPOLI, C. & ROMA, P. (2020). A nationwide survey of psychological distress among Italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors. *International Journal of Environmental Research and Public Health, 17* (9), 3165.

MILLER, W.C., ANTON, H.A. & TOWNSON, A.F. (2008). Measurement properties of the CES-D scale among individuals with spinal cord injury. *Spinal Cord*, *46* (4), 287-292.

MYERS, J.K. & WEISSMAN, M.M. (1980). Use of a self-report symptom scale to detect depression in a community sample. *The American Journal of Psychiatry*, *137* (9), 1081-1084.

NOBLE, R.E. (2005). Depression in women. *Metabolism*, *54* (5), 49-52.

ODRIOZOLA-GONZÁLEZ, P., PLANCHUELO-GÓMEZ, Á., IRURTIA, M.J. & de LUIS-GARCÍA, R. (2020). Psychological effects of the COVID-19 outbreak and lockdown among students and workers of a Spanish university. *Psychiatry Research, 290*, 113108.

OZAMIZ-ETXEBARRIA, N., DOSIL-SANTAMARIA, M., PICAZA-GORROCHATEGUI, M. & IDOIAGA-MONDRAGON, N. (2020). Stress, anxiety, and depression levels in the initial stage of the COVID-19 outbreak in a population sample in the northern Spain. *Cadernos de Saúde Pública*, *36*, e00054020.

ÖZDIN, S. & BAYRAK ÖZDIN, Ş. (2020). Levels and predictors of anxiety, depression and health anxiety during COVID-19 pandemic in Turkish society: The importance of gender. *International Journal of Social Psychiatry*, *66* (5), 504-511.

PAPPA, S., NTELLA, V., GIANNAKAS, T., GIANNAKOULIS, V.G., PAPOUTSI, E. & KATSAOUNOU, P. (2020). Prevalence of depression, anxiety, and insomnia among healthcare workers during the COVID-19 pandemic: A systematic review and meta-analysis. *Brain, Behavior & Immunity*, *88*, 901-907. doi:10.1016/j.bbi.2020.05.026

PORCELLI, P. (2020). Fear, anxiety and health-related consequences after the COVID-19 epidemic. *Clinical Neuropsychiatry*, *17* (2), 103-111.

QIU, J., SHEN, B., ZHAO, M., WANG, Z., XIE, B. & XU, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implication and policy recommendations. *General Psychiatry*, *33*, e100213.

RADLOFF, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1* (3), 385-401.

REYNOLDS, D.L., GARAY, J.R., DEAMOND, S.L., MORAN, M.K., GOLD, W. & STYRA, R. (2008). Understanding, compliance and psychological impact of the SARS quarantine experience. *Epidemiology & Infection*, *136* (7), 997-1007.

ROSSI, R., SOCCI, V., PACITTI, F., DI LORENZO, G., DI MARCO, A., SIRACUSANO, A. & ROSSI, A. (2020). Mental health outcomes among frontline and second-line health care workers during the coronavirus disease 2019 (COVID-19) pandemic in Italy. *JAMA Network Open*, *5*, e2010185-e2010185.

ROSSI, R., SOCCI, V., TALEVI, D., MENSI, S., NIOLU, C., PACITTI, F., ... & DI LORENZO, G. (2020). COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. An N= 18147 web-based survey. *medRxiv.* https://doi.org/10.1101/2020.04.09.20057802

SPIJKER, J., VAN DER WURFF, F.B., POORT, E.C., SMITS, C.H.M., VERHOEFF, A.P. & BEEKMAN, A.T.F. (2004). Depression in first generation labour migrants in Western Europe: The utility of the Center for Epidemiologic Studies Depression scale (CES-D). *International Journal of Geriatric Psychiatry*, *19* (6), 538-544.

STOCKINGS, E., DEGENHARDT, L., LEE, Y.Y., MIHALOPOULOS, C., LIU, A., HOBBS, M. & PATTON, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, *174*, 447-463.

THORNTON, E.W., McQUEEN, C., ROSSER, R., KNEALE, T. & DIXON, K. (1997). A prospective study of changes in negative mood states of women undergoing surgical hysterectomy: The relationship to cognitive predisposition and familial support. *Journal of Psychosomatic Obstetrics & Gynecology*, *18* (1), 22-30.

TRAN, T.D., KALIGIS, F., WIGUNA, T., WILLENBERG, L., NGUYEN, H.T.M., LUCHTERS, S., ... & FISHER, J. (2019). Screening for depressive and anxiety disorders among adolescents in Indonesia: Formal validation of the centre for epidemiologic studies depression scale–revised and the Kessler psychological distress scale. *Journal of Affective Disorders*, *246*, 189-194.

VIGLIONE, D.J. & GIROMINI, L. (2020). *Inventory of Problems-29: Professional manual*. Columbus, OH: IOP-Test, LLC.

VIGLIONE, D.J., GIROMINI, L. & LANDIS, P. (2017). The development of the Inventory of Problems-29: A brief self-administered measure for discriminating bona fide from feigned psychiatric and cognitive complaints. *Journal of Personality Assessment*, *99* (5), 534-544.

VILAGUT, G., FORERO, C.G., BARBAGLIA, G. & ALONSO, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): A systematic review with meta-analysis. *PloS One*, *11* (5), e0155431.

WANG, C., PAN, R., WAN, X., TAN, Y., XU, L., MCINTYRE, R.S., ... & HO, C. (2020). A longitudinal study on the mental health of general population during the COVID-19 epidemic in China. *Brain, Behavior & Immunity*, *87*, 40-48.

WENHAM, C., SMITH, J. & MORGAN, R. (2020). COVID-19: The gendered impacts of the outbreak. *The Lancet*, *395* (10227), 846-848.

ZAUSZNIEWSKI, J.A. & GRAHAM, G.C. (2008). Comparison of short scales to measure depressive symptoms in elders with diabetes. *Western Journal of Nursing Research*, *31* (2), 219-234.

DOI: 10.26387/bpa.289.2

The impact of extrapolation and examiner's judgment on Rorschach form quality coding: Interrater reliability and clinical implications

Davide Ghirardello[1], Francesca Ales[1], Valeria Raimondi[1], Donald J. Viglione[2], Alessandro Zennaro[1], Luciano Giromini[1]

[1] Department of Psychology, University of Turin, Italy
[2] Alliant International University, San Diego, California, U.S.

francesca.ales@unito.it

ABSTRACT. Nel test di Rorschach, la Qualità formale (*Form Quality*, FQ) descrive il grado di somiglianza tra la risposta e la corrispondente localizzazione nella macchia, ed è derivata dalla frequenza con cui la risposta stessa è identificata e dal giudizio degli esaminatori (rater) riguardo l'aderenza della sua forma ai contorni della macchia. Un ampio numero di ricerche ha dimostrato che la FQ ha un'eccellente validità come misura dell'esame di realtà e di gravità della psicopatologia. Tuttavia, alcuni studi hanno riportato valori di *interrater reliability* (IRR) non ottimali. Nel presente articolo abbiamo esaminato 1588 risposte raccolte in 60 protocolli Rorschach d'archivio. Abbiamo esaminato la frequenza con cui FQ è stata ricavata dalle Tabelle (T), Estrapolata (E) o stabilita sulla base del Giudizio dell'esaminatore (*Judged,* J), e testato la forza dell'associazione tra il processo di siglatura della FQ e (a) i punteggi delle variabili FQ, e (b) la IRR. I risultati hanno mostrato che, quando confrontate alle risposte T, le risposte E e J erano caratterizzate da FQ progressivamente più scadente e IRR progressivamente meno ottimale. Nel complesso, questi risultati confermano che il processo di siglatura della FQ ha un notevole impatto sull'accuratezza della siglatura e sulla IRR della FQ. Al fine di ridurre le incoerenze riscontrate nella codifica della FQ, gli autori suggeriscono che gli sviluppi futuri dell'R-PAS possano provare a incorporare algoritmi computerizzati in grado di aiutare gli esaminatori nell'attribuzione della codifica FQ.

SUMMARY. Rorschach Form Quality (FQ) describes how well a response fits a given inkblot location and is derived from how frequently it is identified and whether raters judge it to be a good fit. A large body of research has established that FQ has excellent validity as a measure of reality testing and severity of psychological disturbance. However, some studies have reported sub-optimal interrater reliability (IRR). In this article we inspected 1588 responses from 60 archival Rorschach protocols. We examined the frequency with which FQ was Tabled (T), Extrapolated (E) or Judged (J), and tested the strength of the association of FQ determination path to FQ scores and IRR. Results showed that when compared to T responses, E and J responses were characterized by increasingly poorer FQ and less optimal IRR. Taken together, these results confirm that the determination path used to code FQ has a notable impact on the scoring accuracy and IRR of FQ. In order to reduce the FQ coding inconsistencies, the authors suggest that future R-PAS developments might try to incorporate computer algorithms to help with the attribution of FQ codes.

Keywords: *Rorschach, Form quality, Interrater*

INTRODUCTION

Rorschach Form Quality (FQ) measures how well a Rorschach response fits a particular inkblot location, and how frequently it is seen in that location (Meyer, Viglione, Mihura, Erard & Erdberg, 2011). To score FQ one determines whether the chosen inkblot area looks like the objects or object that the respondent sees. This is done by comparing the respondent's perceptions of the inkblot to other respondents' perception of the same inkblot. Thus, FQ is an essential measure of perceptual accuracy and reality testing and one of the key variables of the Rorschach test since its inception (Meyer et al., 2011; Mihura & Meyer, 2018).

Hermann Rorschach himself noted the relationship between the accuracy of response objects offered by the examinee in terms of whether their form matches the shape of the blots and the person's ability to perceive the world in a realistic way (Rorschach, 1921). Although Rorschach created a list of objects to help determine the quality of the forms perceived by the examinees, his premature death interrupted his work and his preliminary interpretations left much to debate (Exner, 1969). In the following years, several Rorschach systems were developed that varied in administration, coding, and interpretation (e.g., Beck, Beck, Levitt & Molish, 1961; Klopfer, Ainsworth, Klopfer & Holt, 1954). Nevertheless, every major Rorschach system included FQ coding, and research established it as a core variable when evaluating psychotic processes, regardless of the Rorschach system being used (e.g., Berkowitz & Levine, 1953; Dao, Prevatt & Home, 2008; Goldfried, 1962; Harder & Ritzler, 1979; Kimhy et al., 2007). These FQ scoring systems incorporated some version of fit and frequency ratings even if they were based on examiner judgment rather than by comparing the given response to tabled lists derived from previously collected quantitative data. In other Rorschach systems, the person scoring FQ looks up the verbalized response object(s) in a list organized by card number and location areas within each card (i.e., FQ tables).

The Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011) was introduced about ten years ago to overcome some of the known psychometric and validity limitations of the comprehensive system (CS; Exner, 2003; Mihura & Meyer, 2018). Like previous systems, it defines FQ as a function of two components of perceptual accuracy: (1) *fit*, i.e., whether the inkblot location looks like the object described, and (2) *frequency*, i.e., how common it is to see that object at that particular location. It improved on other systems by including much more fit and frequency data in its empirical foundation of its tables (Su et al., 2015). When participants respond to what the inkblot might be, the FQ of their visual percepts is categorized as either ordinary (FQo), unusual (FQu), minus (FQ–), or none (FQn). FQo responses are accurate, relatively common, and thus quickly and easily seen (e.g., "a butterfly" to the whole of Card I). FQu responses are less accurate and typically less common. However, they are not extremely inconsistent with stimuli contours (e.g., "bones" to the D7 of Card III). FQ– responses are inaccurate, infrequent, and difficult to see (e.g., "a face" to the D1 of Card X). Therefore, FQ–, FQu and FQo lie on a continuum of increasing accuracy and frequency (Meyer et al., 2011). Finally, FQn responses are typically impressions of the blot based on the color or shading of the ink without any reference to form or shape (e.g., "Blood, it's all red there, there's no particular shape" to the whole of Card II). Unlike the other FQ codes, FQn responses are not coded based on their degree of fit to the stimuli.

It is worth mentioning that these criteria are theoretically and empirically grounded in the Exner's notions of distal properties and critical bits of the blots. Distal properties are defined as true components of the inkblots, while critical bits are powerful visual features of the blots that contribute to the perceptual organization of many responses (Exner, 1996). As such, drawing from the distal properties of the stimuli and recognizing the critical bits in the inkblot can lead to conventional responses, which are currently scored ordinary. Similarly, those percepts that exceed the distal properties of a certain stimulus, may results in non-conventional responses, and, consequently, they are likely to be coded with poorer formal quality (i.e., FQu or FQ–).

Whereas a large body of research has established that FQ codes possess excellent validity as a measure of reality testing abilities and psychopathology (Meyer et al., 2011; Mihura, Meyer, Dumitrascu & Bombel, 2013; Su et al., 2015), some recent studies have reported sub-optimal results with regard to interrater reliability (IRR), as it had been the case with the CS (Viglione & Meyer, 2008). More specifically, four studies were designed to address IRR of Rorschach variables, including FQ: two were conducted at protocol-level, with IRR evaluated via intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979); the other two examined IRR at response-level via Cohen's *k* (Cohen, 1960).

In the first report of R-PAS IRR at protocol-level, Viglione and colleagues (Viglione, Blume-Marcovici, Miller, Giromini

& Meyer, 2012) found that FQo and FQ– were characterized by an excellent IRR, with ICC values of .84 and .81 respectively; these values were comparable to the average ICC of .88 across all variables. FQu, instead, was characterized by good IRR (ICC = .64) which was still satisfactory, but less optimal. More recently, Pignolo and colleagues (2017) provided the first account of R-PAS IRR at protocol-level in a non-American context, basing on raw data and complexity adjusted scores. Concerning raw data, the average IRR for all the 60 variables was excellent, with an ICC of .78; FQo reached an excellent IRR, with an ICC of .82, whereas less satisfactory findings emerged for FQ– and FQu, with fair ICC values of .53 and .59, respectively (the results did not change significantly with complexity adjusted scores).

The two recent studies assessing IRR at response-level yielded comparable results. Kivisalu and colleagues (Kivisalu, Lewey, Shaffer & Canfield, 2016) reported an average κ across 50 variables of .66, reflecting good IRR; while they found that FQo was characterized by excellent agreement (k = .77), FQ– and FQu showed barely good agreement (k = .62 and .59, respectively). The IRR was re-assessed on the same protocols by different raters in a subsequent study by Lewey and collaborators (Lewey, Kivisalu & Giromini, 2018); in this newer study the authors reported excellent agreement for FQo (k = .73), whereas FQu and FQ– were characterized by fair agreement (k = .53 and k = .52, respectively).

Taken together, the results of these four IRR studies indicate that when compared to other R-PAS variables, FQ codes (especially FQu and FQ–) yield relatively poorer IRR, both at the protocol- and at the response-level. From an applied, clinical perspective, only protocol-level IRR results are crucial to ensure that FQ–based clinical interpretations are made reliably. This is because ultimately clinicians only interpret scale level data and do not overly focus on item level results. However, we argue that response-level IRR data are important too, for at least two reasons. Firstly, consistent with our years-long teaching experience, empirical evidence (Viglione, Meyer, Resende & Pignolo, 2017) indicates that learning how to reliably code FQ at the response level is particularly challenging, which potentially contributes to discouraging new learners from wanting – or feeling confident enough – to adopt the Rorschach in their clinical practice. Secondly, response-level uncertainties and disagreements may give to both novel and more experienced Rorschach users an uneasy feeling that their coding may be inaccurate or arbitrary. As a result, they might take some extra-time to score FQ codes and ultimately their FQ based clinical interpretations may be under-weighted or considered with more skepticism that they probably should. By saying this, we do not intend to dramatize FQ as a critical code, but merely to acknowledge that all of these weigh on the cost side of the cost-benefit ratio and thus diminish test utility so that improving both protocol-level and response-level IRR of FQ codes would be beneficial.

In this article, we hypothesize that a possible explanation for the sub-optimal IRR of FQu and FQ– codes is that these codes are at times coded based on the examiner's subjective judgment of the degree of fit between the form of the response object and the contour of the blot where it was seen. The section in the R-PAS manual addressing these procedures (Meyer et al., 2011) is an extension of Exner's CS approach (1974, 2003) and largely derived from refinements to the procedure (Viglione, 2002, 2010). A few years after publishing the manual, the authors identified some limitations to the procedure in the R-PAS manual and uploaded a document (Viglione et al., 2016) on the R-PAS website (www.r-pas.org), which specifies three distinct FQ determination paths: Tabled, Extrapolated, and Judged. Tabled FQ determination occurs when the important response objects are found in the FQ tables. For example, in Card I, W, "The face of a witch" would be coded FQu and would consist of a Tabled determination because in the FQ tables, "Face, Witch" is listed as FQu. At times, however, the response object is not found in the FQ tables and an extrapolation process is required. Typically, Extrapolated FQ determination occurs when FQ is derived from similarly shaped tabled item, e.g., when extrapolating from a rat to a mouse or a hat to a bonnet. For example, in Card V, upside-down, W, one might say "A flower". In the FQ tables, W(v), no objects resemble a flower. However, in the standard position, flower is FQu. Thus, by extrapolation, "A flower" seen upside-down also is coded FQu. This would be called an obvious extrapolation. Extrapolation may also be less obvious and occur when, based on examination of multiple, tabled items, the preponderance of the evidence clearly favors one FQ score over another – or a more reasonable middle way. For example, in Card VIII, D3, "Skull of Bigfoot". By looking at the FQ tables, "Skull (Animal)" is coded FQo, while "Skull (Human)" is coded FQ–. Since Bigfoot has both some animal and human features, and given that there is equal evidence for FQo and FQ–, a reasonable coding would be FQu. It is important to specify that the R-PAS extrapolation procedure, similarly to the CS, is based on the fact that the degree of fit

is relative to the shape of the percept and not to the content per se. Lastly, Judged determination requires the examiner to look at the response in the location where the response was seen, so to establish FQ by answering the question: "Can I see that object in this location quickly and easily?". Coders may resort to Judgment in two situations. First, when FQ tables do not provide comparable responses for extrapolation; second, when FQ tables provides support for both FQ– and FQu (or for FQu and FQo), without a clear basis for preferring one over the other. For example, in Card IX, W, "The hand of a person, kinda like making the sign of peace… like with the two fingers up, you know what I mean?" would be coded using examiner judgment. In the FQ tables, "Hand" or "Fingers" are not listed with reference to W and looking for a rationale among similar or near-W locations also does not help. There is no location at the bottom half of the card to look for the palm of the hand; D3 would likely be the two fingers, but there is nothing similar in shape to fingers there. Thus, there are no guidance or comparable responses for extrapolation in the FQ tables.

Moreover, it is worth mentioning that multiple-object responses represent a tricky element that could generate inconsistencies in FQ coding. Basically, three cases should be taken into account here. Firstly, the FQ tables contain entries that refer to overarching percepts, such as landscape or anatomy. These are superordinate categories that could be used as tabled entries for multiple-objects responses composed by multiple objects or components (the FQ determination path would be Tabled). Secondly, the examiner should search for the most common multiple-object responses that are already listed in the FQ tables (also in this case, the FQ determination path would be Tabled). When the overarching category cannot be used, and the multiple-object response is not listed in the appropriate location area of the FQ tables, the guideline is to determine the FQ code for each important object following the procedure outlined above for single-object responses, and then to use the code down principle by choosing the least accurate (or lowest) FQ code and apply it to the overall response. Here, the attribution of the FQ determination paths follows the same rules described above (Viglione et al., 2016): if FQ is determined based on FQ, the path will be Tabled; if extrapolation is needed, the path will be Extrapolated and, finally, if the FQ is determined via judgment of fit, the path will be Judged. It should be noted that, when coding FQ (and its determination path) for multiple objects responses, it might be difficult to distinguish between important and unimportant objects.

AIM

Because no research has yet reported on the frequency with which FQ is coded based on Tabled (T) *versus* Extrapolated (E) *versus* Judged (J) determination paths, we inspected FQ codes from 60 archival Rorschach protocols and examined the percentage of cases in which FQ was determined based on each of those three paths. Next, as we anticipated that the more a response object is likely to be seen in a specific location of a given inkblot, the higher the likelihood that such a response object would appear also on the FQ tables, we tested the extent to which non-tabled, i.e., E and J paths, associated with poorer FQ outcomes. Lastly, and most importantly, we aimed at quantifying the extent to which the greater the use of some judgment (i.e., E and J paths) in the determination of FQ, the lower the IRR of the resultant FQ codes.

MATERIALS AND METHODS

Rorschach data

Rorschach protocols. For this study, we randomly selected 60 protocols from a broader data set we had access to, consisting of 96 Rorschachs from healthy, undergraduate volunteers with no previous neurological/psychiatric disorders. As further detailed in the journal article describing that data set (Burin et al., 2019), participants' recruitment was undertaken in Turin, in the north of Italy, either at the University of Turin or via snowball sampling, and Rorschach administrations were carried on using standard R-PAS guidelines. Most of the protocols analyzed for the current paper were from women (83.3%), and our sample mean age was 21.48 years (SD = 2.69). The total number of responses was 1588 with an average of 26.47 responses per protocol (SD = 2.77). Six out of the 1588 responses received the code FQn by rater 2, so the number of responses on which the analyses are based is 1582 (i.e., the total number of responses having a form demand).

Rorschach coders. Two of the authors of the current article (i.e., Ghirardello - DG - and Ales - FA) coded the great majority of the protocols originally analyzed in Burin et al. (2019) and all of the 60 protocols selected for the current study. Additionally, together with a third rater (Raimondi - VR), Ghirardello and Ales also independently

re-coded all responses of a selected number of protocols, so that the second coders were blind to any previous coding. Thus, all 60 protocols were eventually coded twice by two different and independent raters. To prevent the same protocol from being coded twice by the same rater, half of the protocols initially coded by Ales were randomly assigned to Ghirardello for the second coding; the other half were assigned to Raimondi. Similarly, half of the protocols originally coded by Ghirardello were randomly assigned for a second coding to Ales; the other half to Raimondi. All three raters were graduate students who had been trained by a member of the R-PAS Research and Development Group (last author).

Procedure

As noted above, the 60 protocols examined for the current study were coded twice, by two different and independent judges. More specifically, at t_1, coding was performed with the purpose of conducting Burin et al.'s (2019) study; at t_2, coding was performed to examine the frequency with which FQ was coded based on Tabled (T), Extrapolated (E), and Judged (J) determination paths, and to test the IRR of FQ codes. As such, in addition to coding FQ, t_1 raters also reported, for each response, what determination path was used to code FQ; t1 occurred in 2016, t_2 occurred in 2017. At both times, when coding FQ, all coders relied on both the coding guidelines reported on the R-PAS manual (Meyer et al., 2011) and the online document elaborated by Viglione et al. (2016) and uploaded in the R-PAS website (www.r-pas.org).

To test the IRR of the FQ determination path classifications, a subsample of 16 protocols from t_1 (8 protocols coded by Ghirardello and 8 coded by Ales) was randomly extracted, and the same raters who had coded FQ at t_1 were asked to re-examine the same responses a second time, to indicate what FQ determination path characterized the attribution of their FQ codes. For these 16 protocols comprising a total of 436 responses (27.5% of the total sample), the FQ determination paths were thus assigned twice (i.e., at t_1 and at t_2), by two independent judges (the 8 records coded by Ghirardello at t_1 were independently re-corded by Ales at t_2, and the 8 records coded by Ales at t_1 were independently re-corded by Ghirardello at t_2). Analyses of the IRR of the FQ determination path yielded a highly satisfactory Cohen's k of .79 (Cicchetti, 1994). Two out of the 436 responses received an FQn code, so the number of responses on which these analyses are based is 434 (i.e., the total number of responses having a form demand).

It should be noted that all judges were blind to the chief hypotheses of the study at t_1 and at t_2. Also, at t_2 each rater was blind to the other rater's codes provided at t_1.

Statistical analysis

Statistical analyses mainly focused on descriptive statistics and χ^2 analyses to determine the frequency with which Tabled, Extrapolated and Judged determination paths were used to code FQ across the ten inkblots. Next, FQ IRR was assessed both at response-level (using Cohen's k) and protocol-level (using ICC). IRR classification are based on Cicchetti (1994) and Shrout and Fleiss (1979): k or ICC values lower than .40 indicate poor IRR, between .40 and .59 fair IRR, between .60 and .74 good IRR, and values at or above .75 suggest excellent IRR. In many studies focusing on Rorschach variables, IRR evaluated via ICC was computed using the two-way random effect model (e.g., Acklin, McDowell, Verschell & Chan, 2000; Viglione et al., 2012), which assumes that the same pair of raters have rated each protocol. In our study, the pair of raters was not the same for all protocols, thus we used a one-way random effects model (for details, see Meyer et al., 2002; Shrout & Fleiss, 1979).

RESULTS

Tabled, Extrapolated, and Judgment determination paths and Form Quality

Table 1 shows the percentage of responses in which FQ was coded based on Tabled, Extrapolated, or Judged determination paths, divided by card. In total, about 60% of the responses were found in the R-PAS FQ tables (T), extrapolation (E) was required in about 30% of the cases, and judgment (J) was required in about 10%.

The distribution of T, E, and J, however, varied across all ten cards, $\chi^2(18) = 59.0$, $p<.001$. More specifically, when compared to all other cards, Card IV was characterized by a significantly higher proportion of E responses ($z = 2.1$), Card

Table 1 – Total and card by card FQ determination path

		T	E	J	Total
Card I	R	118	41	11	170
	% in Card	*69.4%*	*24.1%*	*6.5%*	
	Std. Residuals	1.7	−1.7	−1.2	
Card II	R	99	52	13	164
	% in Card	*60.4%*	*31.7%*	*7.9%*	
	Std. Residuals	.2	.1	−.6	
Card III	R	108	43	14	165
	% in Card	*65.4%*	*26.1%*	*8.5%*	
	Std. Residuals	1.0	−1.2	−.3	
Card IV	R	72	**59**	11	142
	% in Card	*50.7%*	***41.5%***	*7.7%*	
	Std. Residuals	1.3	**2.1**	−.6	
Card V	R	100	**29**	12	141
	% in Card	*70.9%*	***20.6%***	*8.5%*	
	Std. Residuals	1.8	**−2.3**	−.3	
Card VI	R	91	51	16	158
	% in Card	57.6%	*32.3%*	*10.1%*	
	Std. Residuals	−.3	.2	.3	
Card VII	R	107	39	12	158
	% in Card	*67.7%*	*24.7%*	*7.6%*	
	Std. Residuals	1.4	−1.5	−.7	
Card VIII	R	92	51	16	159
	% in Card	*57.9%*	*32.1%*	10.1%	
	Std. Residuals	−.2	.1	.3	
Card IX	R	**60**	**69**	**22**	151
	% in Card	***39.7%***	***45.7%***	***14.6%***	
	Std. Residuals	**−3.1**	**3.1**	**2.1**	
Card X	R	90	64	20	174
	% in Card	*51.7%*	36.8%	*11.5%*	
	Std. Residuals	−1.3	1.2	1.0	
Total	R	937	498	147	1582
	% in Card	*59.2%*	*31.5%*	*9.3%*	

Note. Bolded values represent standardized residuals greater than $|z| = 1.96$.

Legenda. T = Tabled; E = Extrapolated; J = Judged.

V was characterized by a lower proportion of E responses ($z = -2.3$), and Card IX was characterized by a lower number of T responses ($z = -3.1$), and by a higher number of E ($z = 3.1$) and J ($z = 2.1$) responses.

Also noteworthy, different FQ determination paths were associated with different FQ coding outcomes, $\chi^2_{(4)} = 391.1$, $p<.001$. Indeed, Table 2 shows that T responses were positively associated with FQo ($z = 8.1$), and negatively associated with FQu ($z = -7.7$) and FQ– ($z = -3.7$). Conversely, E responses associated positively with FQu ($z = 6.7$) and negatively with FQo ($z = -6.3$), and J responses associated positively with FQu ($z = 7.0$) and FQ– ($z = 6.4$) and negatively with FQo ($z = -8.8$). That is, in line with our hypotheses, compared to T determination path, E and J paths associated with increasingly poorer FQ.

Form Quality interrater reliability at response and protocol levels

The third step of our analyses entailed the evaluation of the impact of the determination path, i.e. Tabled (T), Extrapolated (E) and Judged (J), on FQ IRR at response and protocol level (see Table 3). Focusing on response-level IRR, when disregarding the type of determination path used to code FQ, a general good agreement was found, with $k = .68$. Comparing Cohen's k separately for T, E and J responses, however, revealed that IRR was excellent ($k = .77$) for T, but dropped to fair ($k = .48$) and to poor ($k = .37$) for E and J, respectively.

We next focused on protocol-level IRR (see Table 4). Overall, a good to excellent IRR was observed in all cases

Table 2 – Response-level percentage of Tabled (T), Extrapolated (E) and Judged (J) responses along with their FQ codes

FQ determination path	FQ–	FQu	FQo	Total
T	72	194	671	937
% in T	7.7%	20.7%	71.6%	
Std. Res.	**–3.7**	**–7.7**	**8.1**	
E	71	267	160	498
% in E	14.3%	53.6%	32.1%	
Std. Res.	1.6	**6.7**	**–6.3**	
J	44	103	0	147
% in J	29.9%	70.1%	0%	
Std. Res.	**6.4**	**7.0**	**–8.8**	
Total	187	564	831	1582
%	11.8%	35.7%	52.5%	

Note. Bolded values represent standardized residuals greater than $|z| = 1.96$.

Table 3 – Response-level IRR based on FQ determination path

FQ determination path	N	Cohen's *k*	Classification
Tabled	937	.77	Excellent
Extrapolated	498	.48	Fair
Judged	147	.37	Poor
Total	1582	.68	Good

Note. Cohen's *k* classification based on Cicchetti (1994) and Shrout & Fleiss (1979).

Table 4 – Protocol-level IRR

FQ determination path	FQ	ICC	Classification
All responses			
	FQo%	.77	Excellent
	FQu%	.66	Good
	FQ–%	.68	Good
T & E only			
	FQo%	.75	Excellent
	FQu%	.65	Good
	FQ–%	.64	Good
T only			
	FQo%	.79	Excellent
	FQu%	.75	Excellent
	FQ–%	.77	Excellent

Note. ICC classification based on Cicchetti (1994) and Shrout & Fleiss (1979).

(ICCs were comprised between .66 and .77). However, in line with our expectations, ICCs was notably higher when T determined responses only were examined (ICCs were comprised between .75 and .79).

Additional analyses

Sub-optimal agreement for tabled responses. It is surprising that FQ ratings were inconsistent between raters when the path for determining FQ, as reported by Rater 2 at t_2, was T (k = .77, see Table 3). Obviously, raters were using different approaches to derive their FQ, but what is the nature of these differences? To answer this question, we inspected the 16 protocols with 434 responses for which both independent raters identified the FQ determination paths, in addition to the FQ codes themselves (see Procedure). Confirming this hypothesis, we found 20 out of the 266 responses that had been classified as T by Rater 2, had been classified as E or J by Rater 1 (see Table 5), and that these inconsistencies typically resulted in FQ coding inconsistencies too.

To our surprise, inconsistencies on FQ coding occurred also for 29 of 246 (11.8%) responses classified as T by both raters. That is, it did happen – albeit relatively infrequently – that both raters considered the FQ determination to be Tabled, yet disagreed on FQ. We thought that raters were likely using different tabled entries to derive their FQ, so we examined the verbatim responses and location documentation to better understand this puzzling outcome.

This review revealed a number of sources for these Tabled FQ coding inconsistencies. The first involved multi-object responses and whether or not a given, tabled, response object should be considered to be an "important object". For example, the response "a flower and a bush"

Table 5 – Contingency table for the IRR of the FQ determination path

		Rater 1 FQ determination path			
		T	*E*	*J*	*Total*
Rater 2 FQ determination path	*T*	246	19	1	266
	E	13	114	8	135
	J	2	5	26	33
Total		261	138	35	434

could have one or two important objects depending on how they are elaborated. If both are tabled and have different FQ, disagreement on which objects are important would lead to different FQ. A second source of disagreement is whether or not a multi-object response would qualify as an overarching table entry. For example, the response "these look like lungs [tabled], these like bones [tabled]" could have a different FQ, if one looks up lungs and bones in the FQ Table but a second rater found "anatomy" tabled for the entire response location. Additional sources of inconsistencies included raters' misunderstandings related to the location of the response objects, particularly in the case of quasi-W or quasi-D responses and linguistic ambiguities in the description of a response and/or FQ tables entry (e.g., can "a cockroach" be automatically coded based on an FQ tables' entry such as "bug" or does one only use the "bug" entry as Tabled FQ determination when that exact word used by the examinee?).

A tentative approach to reduce judgment in FQ determination. As noted above, at the response-level, the characterization of Cohen's *k* was excellent for Tabled (T) responses, but fair and poor for Extrapolated (E) and Judged (J) responses respectively (see Table 3). At the protocol-level, when only T responses were analyzed, the characterization of ICC was excellent for all three FQ codes, whereas it decreased to good, for FQu% and FQ–%, when considering also the E and J responses (Table 4). Overall, Non-T responses (i.e., E and J responses) thus appeared to be characterized by relatively poorer IRR.

Given that, we wondered whether one could possibly predict the FQ data obtained when scoring the entire protocol basing on the FQ codes assigned in the T responses only. Ideally, such procedure could notably simplify the coding procedures of FQ, while increasing IRR. Indeed, as noted above, teaching how to code FQ is particularly challenging (Viglione et al., 2017) and FQ coding difficulties potentially discourage practitioners from using the Rorschach in their practice as they require a lot of time and effort. Consequently, difficulties in learning and uncertainties about the accuracy of the coding are time-consuming, impacting the cost-benefit ratio associated with using the Rorschach. We thus ran three hierarchical regression models to predict the three key protocol-level scores of FQ, i.e., FQo%, FQu% and FQ–%.

Because when compared to T responses, Non-T responses associated with poorer FQ, in each model we considered two predictors. One predictor (step 1) was the percentage of the target FQ code found in the T responses only. For example, the predictor of FQo% at the protocol-level was represented by the proportion of the FQo responses given to T responses divided by the total number of T responses in that protocol. The second predictor, entered at step 2, consisted of the number of responses whose FQ determination was not T, divided by the total number of responses in the protocol (i.e., the % of Non-T responses in the protocol).

The results of these three models are reported in Table 6. Their adjusted R^2 values were comprised between .50 (for FQ–%) and .70 (for FQo%), thus indicating that at least half of the variance of the overall score of each FQ variable could theoretically be estimated basing on two predictors only. Besides, ΔR^2 decreased from FQo to FQu and FQ–, which suggests that adding the % of Non-T responses to the models impacted more notably the prediction of FQo% than that of FQu% or FQ–%.

DISCUSSION

This study aimed at shedding some light on why IRR of FQ is sometimes less optimal than that of other R-PAS variables, despite its well-established validity. To this aim, we coded the percentage of different FQ coding paths, namely Tabled (T), Extrapolated (E) and Judged (J), and tested some hypotheses concerning FQ and its IRR across judges. In line with our hypotheses, we found that E and J responses were characterized by increasingly poorer FQ and less optimal IRR compared to T responses. Noteworthy, using the % of E and J responses (i.e., Non-T) and the FQ assigned to T responses, we were able to predict 50% to 70% of the variance of the FQ values found when coding FQ for the entire protocol. Taken together, these results confirm that the FQ determination path used to code FQ may have a notable impact on IRR.

An interesting result is that, as shown in Table 1, in approximately 60% of the cases, the percepts to be considered to code FQ were found in the FQ tables, without the necessity to make any extrapolations or judgments. This may be the reason why, even though subject to a certain degree of variability, the IRR of FQ is usually satisfactory across studies, albeit at times lower than optimal (Kivisalu et al., 2016; Lewey et al., 2018; Pignolo et al., 2017; Viglione et al., 2012). Moreover, extrapolation and judgment were required in about 30% and 10% of the cases, respectively.

Table 6 – Hierarchical regression models

Criterion/predictors	β_1	β_2	R	R^2	Adj. R^2	ΔR^2
FQo%						
(step 1) FQo% (T only)	.69**	.72**	.69	.48	.47	–
(step 2) % of Non-T	–	−.48**	.84	.71	.70	.23**
FQu%						
(step 1) FQu% (T only)	.71**	.69**	.71	.50	.50	–
(step 2) % of Non-T	–	.37**	.80	.64	.62	.13**
FQ−%						
(step 1) FQ−% (T only)	.65**	.72**	.65	.42	.41	
(step 2) % of Non-T	–	.33**	.72	.52	.50	.10**

* $p<.05$, ** $p<.01$

Since this is the first study to document the use of T, E, and J coding paths, we have no reference parameters to evaluate these frequencies in the context of a non-clinical sample. Nonetheless, these percentages represent an evidence of the unique contribution that each person can bring to the Rorschach task. When inspecting the percentages of T, E, and J responses across cards, however, we found that Card IX produced a notably greater number of Non-T responses. As such, it might be useful, for future R-PAS developments, to try to extend the FQ tables' list of percepts especially for that specific inkblot. It should be noted that Card IX could be considered one of the most difficult ones in the test, as it is typically characterized by fewer responses, and its Popular response is not so common or obvious (Berry & Meyer, 2019; Pianowski, Meyer & de Villemor-Amaral, 2016).

A second interesting result is the strong association between J responses and FQ– and, more generally, the decline in FQ when moving from T to E to J responses. This result was somehow expected based on technical and theoretical grounds; nonetheless, this is the first study to provide evidence on this matter. On the technical side, the criteria to code FQo when judgement of fit is required are quite strict, since the only case when FQo can be assigned is when the FQ tables provide conflicting support for both FQu and FQo, without clear guidance to help the decision (Viglione et al., 2016). To code FQo rather than FQu, the answer to the question "Can I see that object in this location quickly and easily?" is closer to "Yes. I can see that. It matches the blot

pretty well", whilst to code FQu rather than FQ– or FQo the answer is closer to "A little. If I work at it, I can sort of see that". When FQ tables do not provide comparable responses for extrapolation, the two possible codes are FQu or FQ–. On the theoretical side, the negative association between E (and J) responses and FQo has a basis on the critical bits concept (Exner, 1996), as implemented in the extrapolation for FQo decisions (Viglione et al., 2016), that is: to extrapolate FQo (vs FQu) it is required that the response includes critical bits matching those included in the FQ tables. By definition, when the rater has to resort to judgement, there could be no match between critical bits of the response objects and the critical bits of the tabled ones.

A third interesting finding obtained from this investigation is that response-level IRR tended to decrease when moving from T to E to J responses. While this result was largely expected, to date no study had yet empirically documented the existence of this phenomenon. In this regard, two main considerations may be drawn. First, Rorschach trainers should try to make some extra efforts when teaching trainees how to code FQ if the relevant percepts are not in the FQ tables, and therefore the examiner has to rely on E or J determination paths. Second, if possible, it would be useful to try to further extend the list of percepts included in the FQ tables, so to minimize the need to use E or J to code FQ. It should be noted, however, that when inspected at the protocol-level, the IRR values of FQ codes were always highly satisfactory, even when including Non-T responses. As such, these recommendations for future improvements may be considered to be 'desirable' but certainly not 'mandatory.'

Indeed, a possible source of interrater disagreement could be the weight of local coding conventions (see Meyer, Shaffer, Erdberg P. & Horn, 2015). The Rorschach coders in this study strictly followed the coding guidelines provided by the R-PAS manual (Meyer et al., 2011), along with the guidance provided by Viglione and colleagues (2016), and this should have avoided the IRR being affected by local coding conventions. Nonetheless, our results help to pinpoint two important aspects about FQ coding. First, a disagreement on the FQ determination path could end up in a disagreement concerning the FQ coding. Secondly, the fact that two raters found a response in the FQ tables does not guarantee agreement on the resulting FQ. Indeed, the FQ coding procedure is complex and it is often much more difficult than just "Look it up in the FQ table".

When closely examining possible sources of FQ coding inconsistencies, we found that differences in the determination of which objects are important in a multi-object response often leads to inconsistent FQ scoring. In addition, examiners sometimes disagree on whether or whether not to use overarching category entries such as anatomy or landscape. For instance, in a response such as "these are lungs and these are bones", to what degree one can safely code the FQ of the response by relying on an overarching category such as "anatomy"? For some locations, the FQ tables clarify whether the potentially overarching category "anatomy" may or may not be used to code a specific anatomic part of the body such as the lungs. For instance, on Card VIII, W, the FQ tables present different entries for "anatomy (unspecified)" versus "anatomy (specific)." However, this distinction is not made explicit for other locations (e.g., on D2 in Card I, or W in Card III, the FQ tables only report "anatomy," with no distinction between unspecified vs specific), so that different examiners could treat the same anatomy-related response differently, for those areas.

Some other sources of FQ coding inconsistency identified in our Additional analyses section involved possible uncertainties or misunderstandings about the location of the important response objects and the use of potentially ambiguous categories and synonyms in the description of the response in the FQ tables itself. These were all cases where, despite the existence of seemingly clear rules, a minimum degree of judgment was still somehow required. In fact, Pignolo et al. (2021) stated that FQ judgments made by individual examiners are not always reliable. Therefore, when scoring FQ, one should carefully scrutinize the empirically supported FQ tables and base the FQ score on these rather than personal judgments (Pignolo et al., 2021). We believe that future developments of the R-PAS should therefore make an effort to address each and every one of those issues, so to further improve interrater reliability. Indeed similar issues led to the publication of more thorough coding procedural instructions for the CS in 2002 (Viglione, 2002), many of which were adopted into R-PAS (Meyer et al., 2011).

From a broader perspective, we believe that many of the FQ coding inconsistencies result from failures to search the FQ tables thoroughly, forgetfulness about complex coding guidelines, and the need for subjective examiner judgment. To reduce the resulting, observed inconsistencies, one could add details, distinctions, and clarifications to the FQ guidelines and tables. However, doing so would make it more and more

difficult for the Rorschach examiner to remember all specific FQ coding procedures at the right times. To avoid this from happening, it would be best if FQ coding were delegated to computers as much as possible. We argue that advances in computer technology should be applied to increase reliability and decrease FQ coding time and effort and thus increase utility in terms of the cost-benefit analyses given the unique contributions to assessment offered by the Rorschach in general and FQ in particular (Meyer et al., 2011; Mihura et al., 2013). To be clear, we are not stating that all Rorschach problems could (nor should) be solved by exclusively relying on computer algorithms. Yet, if the administration and coding processes were more automated, the examiner could dedicate more attentional resources to other interpretively meaningful, subtle behavioral manifestations put in place by the examinee while taking the Rorschach. In this direction, it is perhaps noteworthy that the R-PAS team is trying to develop a new feature that will allow an advanced, speech-to-text function, which will likely simplify the examiner's task during the administration phase.

Because IRR was lower for Non-T than for T responses and learning how to code FQ based on E or J determination paths is challenging and intricate (Viglione et al., 2017), we investigated if one could avoid coding the Non-T FQ, by estimating the FQ scores at protocol level on the basis of two predictors: T FQ%, and % of Non-T responses. Results showed that the information generated by using these data alone was sufficient to estimate, with relative accuracy, what FQ values one would obtain if FQ was coded across the entire protocol. Given that (1) Non-T responses represented almost 40% of the total number of responses, and (2) extrapolating FQ for non-tabled objects has been rated by R-PAS new learners as challenging or difficult and time-consuming (Viglione et al., 2017), this approach could potentially notably simplify the learning and practical usage of the test while increasing IRR. This notwithstanding, presently this approach is going to lose some important clinical information, mainly because the accuracy of the estimation is less satisfactory particularly for FQ– %, a key variable for reality testing interpretation. Thus, future studies should replicate our findings by including some validity criteria, so to test the extent to which the supposedly increased IRR would have any influence on FQ validity.

On the basis of the points discussed above, for the time being we recommend using the online R-PAS document authored by Viglione et al. (2016) to solve any extrapolation and judgment issues/doubts. We also suggest that it might be useful, in the future, to code the path used to determine FQ, i.e., T, E or J, as it might add context to the interpretation. Given their higher IRR, T FQ scores will be the ones on which to ground the interpretation. In turn, E and J responses will be treated more tentatively because of their lower IRR, while at the same time potentially providing a more nuanced interpretation. In fact, J responses appear to document a stronger deviation from what is commonly seen in the card (as documented in the FQ tables), since they are generally characterized by a higher percentage of FQ– compared to E responses (30% vs 14%, respectively).

A few limitations of this study should be kept in mind, while reading this article. Firstly, the study was conducted on a non-clinical sample, comprising undergraduate volunteers. As such, the generalizability of our findings may be questioned. Thus, future studies should inspect both clinical samples and controls composed of subjects pertaining to other professional areas and with different ages. This is important because the prevalence of T over Non-T responses, and of FQ– and FQu over FQo, may significantly change in clinical samples. In fact, one would expect clinical protocols to include a higher number of percepts that are not listed in the FQ tables, thus impacting IRR. Moreover, FQ– is interpreted as a perceptual lapse or distortion, and high FQ– % is strongly associated with reality testing problems and psychopathology. Therefore, the conclusions we drew from our results might be questionable in a clinical sample with a higher proportion of FQ–. Somewhat related to this point, one cannot rule out that possible examiners' disagreements on coding FQ– could in fact associate with (and thereby possibly even indicate) the presence of severe problems in the examinee's psychological functioning. To investigate this possibility, one should test the association between the presence of psychopathology and the amount and possibly type of J responses (e.g., using external criteria such as psychiatric diagnosis). Secondly, the study was focused on IRR, so validity was not evaluated. Thus, criterion measures to assess validity should be included in future research. Despite these limitations, this study is the first to analyze FQ scores with respect to the FQ determination paths, contributing to a deeper understanding of both the FQ variability and the issues regarding the IRR of FQ codes.

Conflicts of interest. Donald Viglione (fourth author) owns a share in the corporate (LLC) that possesses rights to Rorschach Performance Assessment System.

References

ACKLIN, M.W., McDOWELL, C.J., VERSCHELL, M.S. & CHAN, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment*, *74*, 15-47. doi.org/10.1207/S15327752JPA740103

BECK, S.J., BECK, A.G., LEVITT, E.E. & MOLISH, H.B. (1961). *Rorschach's test: I. Basic processes (3rd ed.)*. Grune & Stratton.

BERKOWITZ, M. & LEVINE, J. (1953). Rorschach scoring categories as diagnostic "signs". *Journal of Consulting Psychology, 17*, 110-112. doi.org/10.1037/h0062113

BERRY, B.A. & MEYER, G.J. (2019). Contemporary data on the location of response objects in Rorschach's inkblots. *Journal of Personality Assessment*, *101* (4), 402-413. doi.org/10.1080/00223891.2017.1408016

BURIN, D., PIGNOLO, C., ALES, F., GIROMINI, L., PYASIK, M., GHIRARDELLO, D., ZENNARO, A., ANGILETTA, M., CASTELLINO, L. & PIA, L. (2019). Relationships between personality features and rubber hand illusion: An explorative study. *Frontiers in Psychology*. doi.org/10.3389/fpsyg.2019.02762

CICCHETTI, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. doi.org/10.1037/1040-3590.6.4.284

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, *XX* (1), 37-46. doi.org/10.1177%2F001316446002000104

DAO, T.K., PREVATT, F. & HOME, H.L. (2008). Differentiating psychotic patients from nonpsychotic patients with the MMPI-2 and Rorschach. *Journal of Personality Assessment*, *90*, 93-101. doi.org/10.1080/00223890701693819

EXNER, J.E. (1969). *The Rorschach Systems*. Grune & Stratton.

EXNER, J.E. (1974). *The Rorschach: A comprehensive system*. John Wiley & Sons.

EXNER, J.E. (1996). Critical bits and the Rorschach response process. *Journal of Personality Assessment*, *67* (3), 464-477. doi.org/10.1207/s15327752jpa6703_3

EXNER, J.E. (2003). The Rorschach: A comprehensive system, *Vol. 1: Basic foundations (4th ed.)*. Wiley.

GOLDFRIED, M.R. (1962). Rorschach developmental level and the MMPI as measures of severity of psychological disturbance. *Journal of Projective Techniques*, *26*, 187-192. doi.org/10.1080/08853126.1962.10381095

HARDER, D.W. & RITZLER, B.A. (1979). A comparison of Rorschach developmental level and 51 form-level systems as indicators of psychosis. *Journal of Projective Techniques*, *43*, 347-354. doi.org/10.1207/s15327752jpa4304_2

KIMHY, D., CORCORAN, C., HARKAVY-FRIEDMAN, J.M., RITZLER, B., JAVITT, D.C. & MALASPINA, D. (2007). Visual form perception: A comparison of individuals at high risk for psychosis, recent onset schizophrenia and chronic schizophrenia. *Schizophrenia Research*, *97*, 25-34. doi.org/10.1016/j.schres.2007.08.022

KIVISALU, T.M., LEWEY, J.H., SHAFFER, T.W. & CANFIELD, M.L. (2016). An investigation of interrater reliability for the Rorschach Performance Assessment System (R-PAS) in a nonpatient U.S. sample. *Journal of Personality Assessment*, *98* (4), 382-390. doi.org/10.1080/00223891.2015.1118380

KLOPFER, B., AINSWORTH, M.D., KLOPFER, W.G. & HOLT, R.R. (1954). *Developments in the Rorschach technique. Vol. 1. Technique and theory*. World Book Co.

LEWEY, J.H., KIVISALU, T.M. & GIROMINI, L. (2018). Coding with R-PAS: Does prior training with the exner comprehensive system impact interrater reliability compared to those examiners with only R-PAS-Based training? *Journal of Personality Assessment*, *101* (4), 393-401. doi.org/10.1080/00223891.2018.1476361

MEYER, G.J., HILSENROTH, M.J., BAXTER, D., EXNER JR, J.E., FOWLER, J.C., PIERS, C.C. & RESNICK, J. (2002). An examination of interrater reliability for scoring the Rorschach comprehensive system in eight data sets. *Journal of Personality Assessment*, *78* (2), 219-274. doi.org/10.1207/S15327752JPA7802_03

MEYER, G.J., SHAFFER, T.W., ERDBERG P. & HORN S.L. (2015). Addressing issues in the development and use of the composite international reference values as Rorschach norms for adults. *Journal of Personality Assessment*, *97* (4), 330-347. doi.org/10.1080/00223891.2014.961603

MEYER, G.J., VIGLIONE, D.J., MIHURA, J.L., ERARD, R.E. & ERDBERG, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation and technical manual*. Rorschach Performance Assessment System.

MIHURA, J.L. & MEYER, G.J. (Eds.). (2018). *Using the Rorschach Performance Assessment System (R-PAS)*. The Guilford Press.

MIHURA, J.L., MEYER, G.J., DUMITRASCU, N. & BOMBEL, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin, 139* (3), 548-605. doi.org/10.1037/a0029406

PIANOWSKI, G., MEYER, G.J. & DE VILLEMOR-AMARAL, A.E. (2016). The impact of R-Optimized administration modeling

procedures on Brazilian normative reference values for Rorschach scores. *Journal of Personality Assessment*, *98* (4), 408-418. doi.org/10.1080/00223891.2016.1148701

PIGNOLO, C., GIROMINI, L., ANDO', A., GHIRARDELLO, D., DI GIROLAMO, M., ALES, F. & ZENNARO, A. (2017). An interrater reliability study of Rorschach Performance Assessment System (R-PAS) raw and complexity-adjusted scores. *Journal of Personality Assessment*, *99* (6), 619-625. doi.org/10.1080/00223891.2017.1296844

PIGNOLO, C., VIGLIONE, D.J. & GIROMINI, L. (2021). How reliably can examiners make Form Quality (FQ) judgments in the absence of the Form Quality (FQ) tables? *Rorschachiana*, *42*, 21-34. doi.org/10.1027/1192-5604/a000135

RORSCHACH, H. (1921). *Psychodiagnostik*. Hans Huber.

SHROUT, P.E. & FLEISS, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420-428. doi.org/10.1037//0033-2909.86.2.420

SU, W.-S., VIGLIONE, D.J., GREEN, E.E., TAM, W.-C. C., SU, J.-A. & CHANG, Y.-T. (2015). Cultural and linguistic adaptability of the Rorschach Performance Assessment System as a measure of psychotic characteristics and severity of mental disturbance in Taiwan. *Psychological Assessment, 27* (4), 1273-1285. doi.org/10.1037/pas0000144

VIGLIONE, D.J. (2002). *Rorschach coding solutions: A reference guide for the comprehensive system*. Donald J. Viglione.

VIGLIONE, D.J. (2010). *Rorschach coding solutions: A reference guide for the comprehensive system (2nd ed.)*. www.rorschachcodingsolutions.com

VIGLIONE, D.J., BLUME-MARCOVICI, A.C., MILLER, H.L., GIROMINI, L. & MEYER, G. (2012). An interrater reliability study for the Rorschach Performance Assessment System. *Journal of Personality Assessment*, *94* (6), 607-612. doi.org/10.1080/00223891.2012.684118

VIGLIONE, D.J. & MEYER G.J. (2008). An overview of Rorschach psychometrics for forensic practice. In C.B. Gacono & F.B. Evans with N. Kaser-Boyd (Eds.), *Handbook of forensic Rorschach psychology*. Lawrence Erlbaum Associates.

VIGLIONE, D., MEYER, G., MIHURA, J., ERARD, B., ERDBERG, P. & GIROMINI, L. (2016). *Guidance for coding form quality requiring "Judgment of fit" and an important supplement on FQ coding for former CS user*s. www.r-pas.org.

VIGLIONE, D.J., MEYER, G.J., RESENDE A.C. & PIGNOLO, C. (2017). A survey of challenges experienced by new learners coding the Rorschach. *Journal of Personality Assessment*, *99* (3), 315-323. doi.org/10.1080/00223891.2016.1233559

DOI: 10.26387/bpa.289.3

An analysis of the literature about the application of Artificial Intelligence to the Recruitment and Personnel Selection

Andrea Rezzani, Andrea Caputo, Claudio G. Cortese

Department of Psychology, University of Turin

claudio.cortese@unito.it

ABSTRACT. L'intelligenza artificiale (IA) applicata ai processi di ricerca e selezione del personale (R&S) è la nuova frontiera della gestione delle Risorse Umane, che ha permesso da un lato di velocizzare alcune attività più meccaniche e dall'altro di introdurre modalità innovative come l'analisi di grandi quantità di dati e delle caratteristiche para-verbali dei candidati. Il contributo presenta una rassegna della letteratura sull'introduzione della IA nei processi R&S, considerando aspetti etici e pragmatici, potenzialità e limiti, oltre che la percezione dei candidati e gli impatti sull'immagine aziendale.

SUMMARY. Artificial intelligence, aiming to develop machines solving cognitive problems and thinking like humans, has become one of the most promising solutions to improve certain HR functions. Currently, it primarily affects Recruitment and Personnel Selection. Despite the wide interest of researchers and organizations in recent years there are still many questions to be analysed. The literature review provides an overview of the changes related to the use of AI in these HR processes, analyzing scholarly research on Human - AI tools Interaction, considering AI's pragmatic and ethical aspects as well as the wider HRM processes. We focus on sustainability for people and organizations. Results regard issues of potential AI activities in Recruitment and Selection, AI tools users' perception and acceptance, and ethical concerns

Keywords: *Artificial intelligence, Human resource management, Personnel selection, Review*

INTRODUCTION

The Fourth Industrial Revolution, or Industry 4.0, is altering the dynamics of jobs, workers, and organizations.

To remain competitive in this revolution, companies search for highly qualified and specialized employees. Consequently, their ability to attract new talents is a major function to take the possible advantages and opportunities from these changes (Ghislieri, Molino & Cortese, 2018). For this purpose, they are investing in new technologies to optimize and increase recruitment and personnel selection effectiveness and efficiency.

The organizational practice of recruiting and selection (R&S) is usually divided into the following phases: job analysis, candidate profile definition, date scheduling, interview, psychological test, individual or group trials, delineation of a shortlist of candidates, interview, signature of the contract and insertion in the work environment (Chamorro-Premuzic & Furnham, 2010; Cortese & Del Carlo, 2017).

In the last decade Artificial Intelligence, as a broad discipline that studies and realizes systems that simulate human behaviour and thinking (Russel & Norvig, 2012), has become one of the most promising solutions to improve R&S processes.

AI-enabled tools, due to the latest technological advances can perform tasks beyond human capability (Lucci & Kopec, 2016). In the R&S, these include Big Data Analytics, Intelligent Robots, Face Recognition, Voice Interaction (Jia, Guo, Li & Chen, 2018). Nowadays AI systems can attend, even if partially, to tasks that previously were considered only done by humans (Jarrahi, 2018). Activities such as data extraction from curricula, analysis of the professional profile, candidate engagement, job interview, contract proposal are theoretically to be performed by AI algorithms. Through these tools, some researchers have even supposed that it will be possible to automate the entire process and replace humans in decision- making.

High-performance computing is the ability to process data and perform complex calculations at high speeds. It allows AI tools to optimize the R&S process in time, cost saving, and quality (Geetha & Bhanu Sree Reddy, 2018).

Current investigations are often difficult to compare and put questions about reliability, validity, ethical concerns, personal data treatment (van den Broek, Sergeeva, & Huysman, 2020), in user's perception (van Esch & Black, 2019).

In this scenario, Psychological Sciences should define, in collaboration with other disciplines, the theoretical and methodological aspects related to the application of Artificial Intelligence in Human Resource Management. It can highlight opportunities and advantages, as well as risks and limits.

AIM

The literature provided a basic understanding of the changes related to the use of AI in the R&S processes but more generally in Human Resource Management (HRM). To survive this purpose a sustainability perspective focused on people and organization well-being was adopted. Publications in the field of the interaction between human and AI tools were considered with a focus on the practical and ethical aspects but not on technical ones.

METHOD

The present study is completely based on literature reviews. The library database used was Scopus (https://www.scopus.com/home.uri). The main keywords used to the research include Recruitment, Personnel Selection, Human Resources and Artificial Intelligence. The time period of the selected articles was from 2010 because the articles prior to this timeline were considered not representative of the current technological scenario. The total number of articles included in this literature was 262. The query is dated September 2020.

To identify and access the relevant publication, information related to the title and the abstract were analysed and papers not related to the objective were excluded. Furthermore, articles from ICT and Engineering fields, or in any case purely technical nature, were eliminated. The result is 19 papers overall.

Date of publication is from 2017 and the most recent is from 2020. In detail of the reviewing papers involved in the study are 2017 (2 papers), 2018 (2 papers), 2019 (13 papers) and 2020 (2 papers).

The selected journals come from India, China, Europe, Bahrain, New Zealand, USA and Russia. The publications, where indicated, were from the field of HR services and IT companies.

RESULTS

The research about the application of AI and selection processes within organizations involves a large number of disciplines with different theoretical and methodological perspectives. The result is the overlapping of different theories and hypotheses that are difficult to compare themselves. To provide an overview from a psychological view, the results are discussed on these issues: i) potential AI activities in the Recruitment and Selection phases; ii) AI tools users' perception and acceptance; iii) ethical concerns.

Potential AI activities in the Recruitment and Selection phases

Van Esch and Black (2019) asserted "three related drivers have moved AI-enabled recruiting from a peripheral curiosity to a critical capability" (p. 730). First, the increase of the applicants' time spent in digital spaces implies that companies have to recruit new talents in digital space with digital technologies and tools. Therefore, the number of applicants per position from 100 per job in 2013 to 250 in 2018 forced companies to adopt AI-enabled tools to screen ever-growing numbers of job applicants. Finally, AI-enabled tools have improved to the point where their superiority to humans in terms of both efficiency and effectiveness, especially in the early stages, of recruiting is beyond debate.

AI technologies can provide a large contribution to deal with different activities of the R&S: collect and order data to specific criteria, update, and maintain information on the database, interact with applicants simulating the human behaviours. The greatest advantage of their use is the time and cost-saving that contribute to improving the efficiency and effectiveness of the entire process. The high-performance computing of the AI-enabled tools allows them to reduce human efforts in some decision-making (Nawaz, 2019a, 2019b). They can analyse information about the experience and applicants' skills to select the right candidate for the commitment (Chakraborty, Giri, Aich & Biswas, 2020).

Natural Language Processing (NLP) techniques and computer vision can be used to evaluate candidates' glossary, tone of voice, way of speaking, and body language to analyse their integrity and personality traits (Gupta, Fernandes & Jain, 2018; van Esch & Black, 2019). It is also possible to automate the data collection, grow the number of applicants per position thanks to an easier and more appealing process; screen the candidates; answer most common issues and questions; provide feedback; schedule the interviews (Nawaz & Gomes, 2019).

Recruiters agree to consider the application of AI especially to the early stages of R&S process to analyse information and schedule calendars. The controversial issue of the AI role in HRM is still open for the stages where human-machine interaction is needed (Nawaz, 2019a, 2019b). Despite the ability of AI-enabled tools to make decisions, currently most of the applications involve a human verification of the output (Jia et al., 2018).

Users' perception and acceptance of AI-enabled tools

HR staff and candidates have a central role to drive the HR transformation process. Evaluating their perception and acceptance is a way to postulate what will be the impact of these transformations.

Deloitte (2018) highlighted that even if 72% of the managers agree to apply innovation tools, only 31% assert their companies are able to achieve potential benefits (Deloitte Insights, 2018). A LinkedIn research in 2019 shows that HR profession is one the five professions with the highest turn-over (Rąb-Kettler & Lehnervp, 2019). The reason may be that HR activities (e.g. screening hundreds of curricula, schedule and conduct interviews) are often repetitive, with high effort and few occasions of recognition and gratification. AI-enabled recruiting allows HR to focus on the monitoring and decision-making aspects, reducing their cognitive stress and boredom. As a result, it can reduce the level of turnover within the HR area (Bhardwaj, Singh & Kumar, 2020). Other studies instead alert of the possible negative effects for recruiters, who may feel easily replaced by current technologies (Simonova, Lyachenkov & Kravchenko, 2020). Anyway, to take advantage from the AI application it is necessary for the HR to know the tools. Tambe, Cappelli and Yakubovich (2019) highlighted that the 41% of the CEOs are not confident about their ability to use new tools and analyse data. Only the 4% said to be highly prepared.

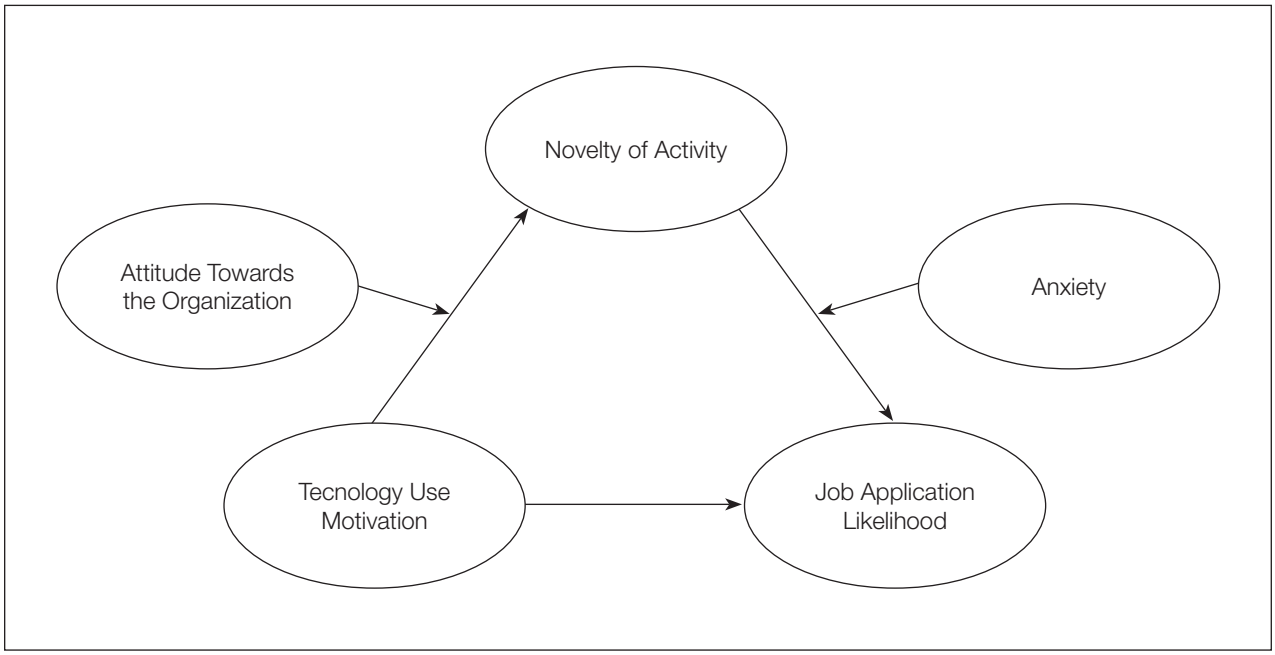
Van Esch and Black (2019) evaluated which factors influence new-generation candidates (i.e. social media users job seekers from 18 to 35 years old) to engage with and completely digital, AI-enabled recruiting. The 293 participants

were enlisted through a crowdsourcing platform; they were invited to read a job application scenario and to answer some questions. Authors found a positive relationship between the use of social media, intrinsic rewards, fair treatment, and perceived trendiness on the intention to engage with a completely digital, AI-enabled recruiting process.

Van Esch, Black and Ferolie (2019) created a theoretical framework to explain the job applicant likelihood in an AI-enabled process. The factors involved in the relationship between technology use motivation and job applicant likelihood, in order of relevance, were the novelty of the activity, attitude towards the organization, and anxiety (see Figure 1).

Specifically, there is a positive effect on the relationship between technology use, motivation, and job application likelihood (β = .38, p<.01). The novelty factor of using AI in the recruitment process mediates and further positively influences job application likelihood (p<.01, 95% CI = .13-.33). Novelty of activity is another measure of intrinsic motivation and like technology use, motivation is a measure of anticipated intrinsic benefits of using AI in the recruitment process. The investigation of conditional indirect effects further supports attitude towards the organization as a moderator and anxiety as a moderator of job application probability. Attitudes towards organizations that use AI and anxiety significantly influence the plausibility that potential

Figure 1 – Conceptual framework (van Esch et al., 2019)



candidates will complete the application process. Anyway, job applicant anxiety towards the use of AI recruitment is secondary to an applicants' attitude towards the hiring organization. Attitude towards the organization significantly predicted novelty of activity ($\beta = .22$, $t_{(528)} = 5.53$, $p = .01$), as did technology use motivation ($\beta = .50$, $t_{(528)} = 16.55$, $p = .01$). Anxiety significantly predicted job application likelihood ($\beta = -.16$, $t_{(528)} = -5.07$, $p = .01$), as did novelty of activity ($\beta = .80$, $t_{(528)} = 16.30$, $p = .01$).

Furthermore, as additional benefits, AI recruiting technologies allow candidates to set any time anywhere for the interview. That increases the potential positive effect on the candidate experience. An additional incentive would be represented by the perception of candidates to adhere to a fast selection process, with fewer biases (van Esch et al., 2019).

A better candidate experience during the application process would allow to further increase the talent pool of an organization.

Job Vite report (2017) shows that only the 8.52% of job applicants, when they visit a job posting site, completes all the steps necessary to conclude the application process and most of them, if they do not receive a feedback following application, will not apply for other positions in the same company.

Ethical concerns

The R&S transformation implies new constraints about personal data's protection and treatment. Especially, it emphasizes different ethical concerns for the HR area.

The General Data Protection Regulation 2016/679 (GDPR) is a regulation in EU law on data protection and privacy. While it provides general guidelines, recent updates have introduced applicable regulations on how data is processed using digital tools.

Articles 9 and 22 are often involved in the application of the AI in the R&S processes. The article 9 concerns the processing of special categories of personal data. It prohibits the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.

If, during the recruitment process, organizations gather that kind of additional characteristics, they could use that information to further categorize candidates and to discriminate, where possible, in terms of job selection. This raises several ethical and privacy concerns, not to mention the determination of both an organization and the values of candidates (van Esch et al., 2019).

Artificial Intelligence has been shown to not only be able to detect a person's data as gender and ethnicity but also more personal data: Wang and Kosinski (2017) showed that deep neural networks can detect sexual orientation from faces. In particular, given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style).

The main purpose of the aforementioned article was not to stereotype people with different sexual orientation but to highlight the link between facial features and psychological traits and the ability of the AI tools to detect them. As a consequence, it signals the need for guidelines for the AI tools in the human-computer interaction, especially when applied in any type of selection process.

The article 22 establishes automated individual decision-making, including profiling. It forbids the use of data subject to be subject to a decision based solely on automated processing, including profiling, without human intervention.

In any case, paragraph 1 does not apply if the decision is necessary for the conclusion or execution of a contract between the data subject and data controller or is based on the explicit consent of the data subject. Such consensus during a selection process, however, could be questioned. Furthermore, the digital environment is very difficult for the police to control. Data can be easily moved across borders, stolen, or recorded without users' consent (Wang & Kosinski, 2017). In conclusion, even in the presence of articles on the GDPR that protect the improper use of personal data, there remains the need to rethink this regulation on the basis of these new technologies.

Another controversial topic concerns the biases of AI tools. If machines think and act like humans, probably they have the same cognitive prejudices (Osoba & Welser, 2017). Biased decision-making is certainly not unique to AI, but

the growing reach of AI makes addressing it particularly important.

Those who object to the decision of Artificial Intelligence systems often argue that humans have consciousness and are potentially able to avoid being swayed by bias during the research and development process. The opacity and lack of transparency of the operating mechanisms of Artificial Intelligence systems make these tools perceived unreliable and dangerous. On the other hand, it is also true that this distinction is only valid when human beings are aware of their prejudices and this does not happen very often. A common bias in the selection process is the halo effect: a cognitive bias in which a single positive trait or characteristic of someone influences judgment on other factors that are not correlated. It can be based on characteristics such as appearance, communication skills and usually occurs on an unconscious level.

In the hiring process there are some procedures to avoid this type of bias such as blind recruitment, a process of removing all identifying details from the candidate's resume and application.

In this direction, Zou and Schiebinger (2018) in the article "AI can be sexist and racist: it's time to make it fair" have mapped several possible strategies to provide systematic solutions to the omnipresent nature of the problem. For example, they suggested developing algorithms to avoid the use of biased data, tagging the content of training datasets with standardized metadata, and accompanying the training data with information about how it was collected and annotated, incorporating constraints and essentially pushing the machine learning model to ensure it achieves fair performance between different subpopulations and between similar individuals, by modifying the learning algorithm to reduce its dependence on sensitive attributes, such as ethnicity, gender, and income.

However, the authors have concluded with some questions: "Should the data be representative of the world as it is, or of a world that many would aspire to? Likewise, should an AI tool use to assess potential candidates for a job, evaluate talent or the probability that the person will assimilate well into the work environment? Who should decide which notions of fairness to prioritize?" (Zou & Schiebinger, 2018, p. 326).

The concept of ethics can be associated with an approach aimed at avoiding discrimination and prejudice, increasing sustainability etc. However, analysing this concept, it emerges that it may not be so uniform and may have different facets depending on the point of view. Van den Broek and colleagues (2020) studied ethical issues within a multinational company after the implementation of AI tools in research and development processes. They found out the use of AI did not always improve or worsen the ethical values of hiring but rather they observed a several of mismatches between the notions of fairness.

In particular, prior to the AI application, the HR team considered it correct to set a cut-off threshold for the evaluation of candidates: "We need a very structured process, because we are dealing with so many candidates. And everyone is assessed in the same way. We need to be objective - Field notes weekly HR team meeting" (van den Broek et al., 2020, p. 6).

However, during their day-to-day work with the AI, the HR professionals experienced that the fixed threshold did not allow for differentiation between situated contexts of the programs, locations and temporary changes in supply, and demand.

Furthermore, candidates contested the notions of fairness. The HR team was confronted with several candidates who expressed during recruiting events, selection events, or via email that they did not feel they had a fair chance to prove their worth. In contrast, the HR team was also confronted with candidates who aimed to gain an unfair advantage over other candidates in the selection process by "gaming the system". HR professionals found out that several candidates bypassed the system by creating a new account with a different email address, in the hope to improve their AI scores. For example, a candidate expressed about a specific game in which he/she had to memorize changing figures: "You could actually cheat on those games. If you would do the game with two people, hold your phone in your hand, and both make a picture [of the figure you have to memorize], I am sure you would pass the game - Candidate 1" (p. 7).

Finally, the managers of the other areas also experienced feelings of frustration following the application of AI when it did not allow the hiring of their favourite candidates. Some managers have also reported that in their opinion the application of AI would result in a lower rate of diversity within the organization, as the algorithms would tend to search for attributes similar to those indicated as ideals in the candidates: "We will have less diversity because we will hire more of the same profile, right? - Field notes group panel" (p. 7).

DISCUSSION

Although the hypothesis relating to a possible replacement of man in the process of research and selection through AI tools appears to be suggestive, it is not supported by data in the literature.

Forecasts converge considering these tools, in the short term, valid support in the analytical phases of the selection process, e.g. publication of job advertisements, extraction, and categorization of data from curricula. So currently there is an agreement in considering the potential offered by AI tools to support and not replace human work.

The users' perception and acceptance have been investigated by a small number of studies and consequently, it is not possible to formulate hypotheses on the factors involved.

Rynes, Colbert and Brown (2002) showed how often the practices of professionals differ from the suggestions of researchers, especially with regard to the area of personnel selection. Specifically, HR professionals were quite sceptical about the use of intelligence or personality tests to evaluate employee performance even though these tools were widely supported by data in the literature.

Factors that influenced professionals' beliefs about research results were the seniority within organizations, the SPHR certification (Senior Professional in Human Resources), and the general knowledge of the academic literature. Probably, the acceptance of AI tools by HR professionals may depend on the same factors.

The positive perception of candidates that emerged from the results of van Esch and colleagues (2019) is in line with what emerged from the study by Sylva and Mol (2009) that suggested that candidates appear more satisfied with technologically advanced recruitment.

However, if the adoption of advanced technologies such as AI would end war for talent, van den Broek and colleagues (2020) highlighted the risk of gaming the systems, where candidates circumvent artificial scoring systems for their advantage. In this scenario, the HR would find themselves engaged in an attempt to detect these using time and energy and thus the advantage of adopting innovative tools.

Analysing the processing of personal data and the studies that have investigated ethical concerns, it emerges that with the practical adoption of artificial systems is necessary to create guidelines in order to understand the specificities related to AI tools and the risks associated with them.

The use of AI in recruitment and selection processes can bring a quantitative added value in an initial phase, e.g. simplifying and speeding up activities such as screening of CVs and analyzing job seekers' social interactions. However, the use of these tools, to date, cannot replace the qualitative value that a human relationship can allow, in particular the knowledge job seekers can find in a two-way contact, a human-level exchange, asking the recruiter (during the interview) for some information regarding organizational culture, values and climate. These dynamics are also important from an Employer Branding perspective (Ambler & Barrow, 1996), as the recruiter, managing the exchange with the candidate, is able to communicate the company image, focusing on the most important value propositions (Backhaus & Tikoo, 2004).

CONCLUSIONS

AI takes and will play an increasingly central role within organizations to cope with the changes imposed by digitalization to attract talent, reduce time and costs, and improve the matching between supply and demand.

Recent evidence shows that the hypotheses of a total human replacement are currently unfounded. HR should not be afraid of the automation of R&S. On the contrary, they should exploit the potential of the AI tools to encourage the development and growth of internal resources within the organization.

The use of AI for recruiters could make the job more meaningful and focused on the candidate and allows them to use their psychological and managerial knowledge. Human intuitive, understanding, and adaptation abilities are skills that not even the most sophisticated robot can simulate.

But, the lack of data regarding the feeling of perception and acceptance does not allow us to fully understand the point of view of the users involved in this transformation. A hypothetical aversion towards AI tools, that we cannot exclude *a priori*, could cause a failure to exploit the potential of the tools currently on the market. Indeed, the fact that these tools are objectively effective in terms of reliability and validity does not imply that they are perceived or experienced as such.

In order to avoid risks, a crucial aspect is the training and sharing of information material within the human resources

departments. If HR professionals will not be trained to grasp the potential offered by technological advancements, the management of processes that are currently overseeing the HR area will be supervised by other company areas (for example, ICT area).

The positive perception of candidates toward AI tools, emerged in the study by van Esch and colleagues (2019), would translate the application of innovative tools into an opportunity for companies to create value both within, from an employer branding perspective, and outside the organization, increasing the pool of talents.

However, there is a need for digital ethics and data management that can be processed automatically. Data masking is a partial solution and not always applicable. While some data represent a potential source of bias, they could be necessary also for an in-depth evaluation and for the establishment of decision-making strategies such as cognitive heuristics, shortcuts extrapolated from reality, which guarantee faster decision making and for this reason, sometimes, even more effective.

We believe it is essential that AI developers interact with social scientists and experts in the humanities to gain the completeness of the dynamics they will have to face for the elaboration, and development of valid, reliable and above all ethical software.

From this perspective, Psychological Sciences can and should follow technological progress hand in hand to understand how to support this transformation without risking forgetting the value of individual and organizational psychological well-being.

Among the limitations present in the literature to date, we can mention the small number of studies examining the effectiveness of AI in R&S processes performed with tools with robust statistical properties according to Psychological Sciences. Although there are early attempts at a more psychological approach to the study of AI, such as, for example, the acceptance of an entirely AI-managed R&S process (Wright & Atkinson, 2019), these studies lack in presenting more robust tools about, for example, their reliability.

Future studies could continue to investigate in parallel both the effectiveness of AI from an instrumental point of view (simplifying processes perceived as more mechanical) and from a psychological one (such as greater engagement of the job seeker). Taking into consideration the use of AI with respect to human capital, studies on the acceptance by job seekers, on the one hand, and on the respect of ethical principles, on the other, show how it is complicated, as well as expensive, to date, to create an AI capable to take into account all these dynamics. On the contrary, psychological literature shows that an expert (human) recruiter is able to understand and to manage these interactional aspects, thus becoming the vehicle spreading a positive employer image (Cortese & Del Carlo, 2017).

References

AMBLER, T. & BARROW, S. (1996). The employer brand. *Journal of brand management*, *4* (3), 185-206.

BACKHAUS, K. & TIKOO, S. (2004). Conceptualizing and researching employer branding. *Career development international*, *9*, 501-517.

BHARDWAJ, G., SINGH, S.V. & KUMAR, V. (2020). An empirical study of artificial intelligence and its impact on human resource functions. *Proceedings of International Conference on Computation, Automation and Knowledge Management, ICCAKM 2020*. https://doi.org/10.1109/ICCAKM46823.2020.9051544

CHAKRABORTY, S., GIRI, A., AICH, A. & BISWAS, S. (2020). Evaluating influence of artificial intelligence on human resource management using PLS SEM (Partial Least Squares Structural Equation Modeling). *International Journal of Scientific & Technology research*, *9* (03), 5876-5880. Retrieved from http://www.ijstr.org/

CHAMORRO-PREMUZIC, T. & FURNHAM, A. (2010). The psychology of personnel selection. In *The Psychology of Personnel Selection*. Cambridge University Press https://doi.org/10.1017/CBO9780511819308

CORTESE, C.G. & DEL CARLO, A. (2017). *La selezione del personale. Come scegliere il candidato migliore ai tempi del web*. Raffaello Cortina Editore.

GEETHA, R. & BHANU SREE REDDY, D. (2018). Recruitment through artificial intelligence: A conceptual study. *International Journal of Mechanical Engineering and Technology*, *9* (7), 63-70.

GHISLIERI, C., MOLINO, M. & CORTESE, C.G. (2018). Work and organizational psychology looks at the Fourth Industrial Revolution: How to support workers and organizations? *Frontiers in Psychology*, *9*, 2365. https://doi.org/10.3389/fpsyg.2018.02365

GUPTA, P., FERNANDES, S.F. & JAIN, M. (2018). Automation in recruitment: A new frontier. *Journal of Information Technology Teaching Cases*, *8* (2), 118-125. https://doi.org/10.1057/s41266-018-0042-x

JARRAHI, M.H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61* (4), 577-586. https://doi.org/10.1016/j.bushor.2018.03.007

JIA, Q., GUO, Y., LI, R., LI, Y. & CHEN, Y. (2018). A conceptual artificial intelligence application framework in human resource management. *Proceedings of the International Conference on Electronic Business (ICEB)*.

LUCCI, S. & KOPEC, D. (2016). *Artificial intelligence in the 21st century: A living introduction. mercury learning and information*. Duxbury.

NAWAZ, N. (2019a). Artificial intelligence interchange human intervention in the recruitment process in Indian Software Industry. *International Journal of Advanced Trends in Computer Science and Engineering*, *8* (4), 1433-1441. Retrieved from http://www.warse.org/IJATCSE/static/pdf/file/ijatcse62842019.pdf

NAWAZ, N. (2019b). Artificial intelligence is transforming recruitment effectiveness in CMMI level companies. *International Journal of Advanced Trends in Computer Science and Engineering*, *8* (6). https://doi.org/10.30534/ijatcse/2019/56862019

NAWAZ, N. & GOMES, A.M. (2019). Artificial intelligence chatbots are new recruiters. *International Journal of Advanced Computer Science and Applications*, *10* (9), 1-5. https://doi.org/10.14569/ijacsa.2019.0100901

OSOBA, O. & WELSER, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. In *An intelligence in our image: The risks of bias and errors in artificial intelligence*. https://doi.org/10.7249/rr1744

RĄB-KETTLER, K. & LEHNERVP, B. (2019). Recruitment in the times of machine learning. *Management Systems in Production Engineering*, *27* (2), 105-109. https://doi.org/10.1515/mspe-2019-0018

RUSSEL, S. & NORVIG, P. (2012). Artificial intelligence: A modern approach 3rd edition. In *The Knowledge Engineering Review*. https://doi.org/10.1017/S0269888900007724

RYNES, S.L., COLBERT, A.E. & BROWN, K.G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management*, *41* (2), 149-174. https://doi.org/10.1002/hrm.10029

SIMONOVA, M., LYACHENKOV, Y. & KRAVCHENKO, A. (2020). HR innovation risk assessment. *E3S Web of Conferences 157*.

SYLVA, H. & MOL, S.T. (2009). E-Recruitment: A study into applicant perceptions of an online application system. *International Journal of Selection and Assessment*, *17* (3), 311-323. https://doi.org/10.1111/j.1468-2389.2009.00473.x

TAMBE, P., CAPPELLI, P. & YAKUBOVICH, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61* (4), 15-42. https://doi.org/10.1177/0008125619867910

VAN DEN BROEK, E., SERGEEVA, A. & HUYSMAN, M. (2020). Hiring algorithms: An ethnography of fairness in practice. *40th International Conference on Information Systems, ICIS 2019*.

VAN ESCH, P. & BLACK, J.S. (2019). Factors that influence new generation candidates to engage with and complete digital, AI-enabled recruiting. *Business Horizons*, *62* (6), 729-739. https://doi.org/10.1016/j.bushor.2019.07.004

VAN ESCH, P., BLACK, J.S. & FEROLIE, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215-222. https://doi.org/10.1016/j.chb.2018.09.009

WANG, Y. & KOSINSKI, M. (2017). Deep neural networks can detect sexual orientation from faces. *Journal of Personality and Social Psychology*, 1-47.

WRIGHT, J. & ATKINSON, D. (2019). The impact of artificial intelligence within the recruitment industry: Defining a new way of recruiting. *Carmichael Fisher*, 1-39.

ZOU, J. & SCHIEBINGER, L. (2018). AI can be sexist and racist: It's time to make it fair. *Nature*. https://doi.org/10.1038/d41586-018-05707-8

DOI: 10.26387/bpa.289.4

Development and validation of the Post-Vacation Work Adjustment Scale (P-VWAS): Study of a Portuguese sample

Cátia Sousa [1,2], Gabriela Gonçalves [1,2]

[1] University of Algarve, Faro, Portugal
[2] Centre for Research in Psychology (CIP/UAL) & University of Algarve, Portugal

cavsousa@ualg.pt

ABSTRACT. L'obiettivo di questa ricerca è quello di sviluppare una scala di adeguamento al lavoro post-ferie e di testarne la struttura fattoriale e le proprietà psicometriche. Attraverso i risultati di due studi (*n* = 232 e *n* = 332), è possibile ottenere una scala composta da 19 item e due dimensioni (Adattamento organizzativo ed Equilibrio lavoro-vita). La scala ha mostrato dei buoni valori per la coerenza interna e valori accettabili per gli indici di adeguamento. La scala ha mostrato validità predittiva del livello di produttività e del grado di concentrazione durante il primo giorno di rientro al lavoro dopo le ferie. Studi aggiuntivi sono richiesti per rafforzare e adeguare la scala, che fornisce un contributo nella comprensione del processo di adeguamento al lavoro dopo le ferie. Il riconoscimento del grado di adeguamento del dipendente permetterà la definizione di una serie di misure e strategie per la sua ottimizzazione nel contesto lavorativo delle organizzazioni.

SUMMARY. The objective of this research is to develop a scale of post-vacation work adjustment and test its factorial structure and psychometric properties. By carrying out two studies (n = 232 and n = 332), the results allow to obtain a scale composed of 19 items and two dimensions (Organizational adjustment and Work-life balance). The scale showed good values of internal consistency and acceptable adjustment indexes. The scale showed predictive validity on the productivity level and concentration degree on the first day of return to work after vacations. The scale proved to be invariant between genders and in relation to the time of return from vacation. Additional studies are needed to reinforce and adjust the scale, which is a contribution to understanding the process of adjusting to work after vacations. The identification of the employee's adjustment degree will allow the definition of a set of measures and strategies for their optimization in the organizations' work contexts.

Keywords: *Work adjustment, Post-vacations, Scale, Validation, Factor analysis*

INTRODUCTION

Vacations, defined as a cessation of work, or a time when a person is not actively participating in his/her work (Lounsbury & Hoopes, 1986), are identified in the literature as an essential and significant period for the recovery of workers (Blomm et al., 2010; Fritz & Sonnentag, 2006).

Work, a significant sphere of life, requires individuals to use cognitive, physical, emotional and psychological resources on a daily basis; not only for the job performance, but also in the continuous and persistent confrontation with countless factors that enhance wear, which in extreme situations can lead to fatigue and exhaustion, with negative consequences for the health and performance of employees (Kinnunen & Feldt, 2013), making it essential to provide periods for their recovery.

Korpela and Kinnunen (2011) point to recovery as a necessary and determining process for individuals who, faced with the perception of fatigue, need to break with their daily work obligations, restoring their internal resources. Undertaking low effort activities outside working hours (e.g., watching television, reading a book), physical activity (where despite the effort spent, internal resources other than work are mobilized) or socializing with family and friends, promotes the recovery of resources and increases the perception of well-being (Blasche, Arlinghaus & Dorner, 2014; Tucker, Dahlgren, Akerstedt & Waterhouse, 2008; Zijlstra & Sonnentag, 2006). The weekend, the post-work periods and vacations are pointed out by the researchers as relevant for this purpose, since they allow individuals to disconnect or reduce the confrontation with the demands of work, greater relaxation and the performance of leisure activities (Blasche et al., 2014; Binnewies, Sonnentag & Mojza, 2009; Koerber, Rouse, Stanyar & Pelletier, 2018), promoting health and well-being benefits for employees (Bloom, Geurts & Kompier, 2012; Mitas & Kroesen, 2019).

Numerous studies based on the understanding of this issue, confirm the effectiveness of the vacation for workers in the recovery of physical and psychological resources (Bloom et al., 2011; Kühnel & Sonnentag, 2011; Sonnentag, 2018). These studies have shown that during and after vacations, workers demonstrate greater satisfaction with life (Kawakubo & Oguchi, 2019; Lounsbury & Hoopes, 1986; Mitas & Kroesen, 2019), better sleep quality (Strauss-Blasche et al., 2005) and humor (Nawijn, Marchand, Veenhoven & Vingerhoets, 2010; Strauss-Blasche, Ekmekcioglu & Marktl, 2000).

Its repercussions extend to work contexts, since, in general, after vacations, the workers present better performance, greater involvement in the work (Fritz & Sonnentang, 2006; Kühnel & Sonnentag, 2011) and reduced levels of stress and burnout (Etzion, 2003; Kuhnel & Sonnentag, 2011).

Thus, even though the vacations represent an effective cost for organizations (which are temporarily deprived of their human resources), the gains also become evident, since more satisfied employees and with better performance levels, contribute to improving organizational results.

However, the process of adapting to work after a vacation is still a little explored topic. The return to work will consequently imply a new readjustment, the return to daily routines, to the experience and the articulation that results from the inherent performance of different roles (work, family, social), where the allocation of individual resources is important, but also of organizational strategies that facilitate this process (Sousa & Gonçalves, 2019). In this regard, Sousa and Gonçalves (2019) grouped the difficulties associated with returning to work in 4 dimensions: work-related difficulties, difficulties at the social level, general difficulties related to the reconciliation of the professional and family spheres and a lack of identification with both their colleagues and organization. This is because, during the absence from the workplace, there was an interruption of the shared history and collective unconsciousness (Sousa & Gonçalves, 2019), which can lead to what Pryzbylski and colleagues (Pryzbylski, Murayama, Dehaan & Gladwell, 2013) called fear of missing out (FoMO), that is, fear of losing opportunities, experiences, building professional relationships, obtaining valuable information and contributing to the main organizational decisions and projects (Budnick, Rogers & Barber, 2020; Pryzbylski et al., 2013).

In summary, it is possible to observe that back to work after vacations is a process that implies initial difficulties, and an effort of readjustment and adaptation, which allows to return to the professional routine.

Inspired by the work of Sousa and Gonçalves (2019), who identified the main difficulties associated with this process, calling it a tune-up day, we tried to develop a scale that allows measuring the adjustment to work after vacations. Developing a measurement instrument that makes it possible to accurately assess the determinants of the work adjustment process and the degree of that adjustment, within an organization, based on the current social and organizational context, proves to be an issue of important relevance. In this sense, this study aims

to develop and validate the *Post-Vacation Work Adjustment Scale* (*P-VWAS*), as well as the analysis of its psychometric properties: exploratory factor analysis (EFA), confirmatory factor analysis (CFA), analysis of internal consistency and predictive validity on the productivity level, concentration degree and the difficulty in getting back to the pace of work on the first day after the vacation. It is also objective to observe the metric invariance of the scale with regard to gender and time of return from vacation. As a determining factor for the increase in productivity, the capacity to adjust to work by the human capital of organizations, this research aims to be a contribution to the understanding of this process, therefore constituting itself as a facilitating platform for the definition of a set of measures and strategies for its optimization in the organizations' work contexts, within the scope of good human resource management practices.

STUDY 1

Study 1 aims to construct and analyze the psychometric properties of P-VWAS through EFA, CFA, internal consistency and predictive validity.

Study 1: Method

Construction of the Post-Vacation Work Adjustment Scale. For the construction and validation of the *Post-Vacation Work Adjustment Scale*, we tried to be faithful to the recommendations proposed by Furr (2010). According to the author, there are four steps that must be respected when building a new scale: 1) articulation between the construct and the context; 2) choice of response format and construction of the set of initial items; 3) data collection; and 4) examination of the psychometric properties and quality of the scale.

Preliminary construction of the Post-Vacation Work Adjustment Scale. Since the literature on the topic is relatively scarce and recent, an attempt was made to articulate existing constructs, which can be adjusted to the theme in question. Thus, and considering that adjustment to work can be understood as a kind of socialization/integration in the company, this instrument was inspired by the contents and descriptions of problems reported in the study by Sousa and Gonçalves (2019) and in the *Newcomer Socialization Questionnaire* (*NSQ*) developed by Haueter and colleagues (Haueter, Macan & Winter, 2003). The NSQ is a questionnaire composed of 35 items distributed over 3 dimensions: organizational socialization, socialization with the group and socialization with tasks, assessed on a 7-point Likert scale (1 = I totally disagree to 7 = I totally agree). Of the 35 items on the original scale, 29 were used, which were adapted and modified for the present study, according to the evaluation carried out by the panel of experts. The remaining 6 items were excluded since they did not fit the objective of our study (example of excluded items: "I understand the expertise - e.g., skill, knowledge - each member brings to my particular work group" and "I know who my customers - internal and external - are").

Instrument pre-test. After the construction of the instrument, a group of 5 experts in the field of Organizational Psychology was asked to review the proposed items in order to increasing the content validity (DeVellis, 2016). They were given an assessment protocol, consisting of two parts: the first part was intended to request a global assessment of the general characteristics of the questionnaire; and the second part intended to evaluate the operationalization of the concept of adjustment to work after an interruption of work. Thus, at first they were asked to evaluate: 1) the presentation of the questionnaire and 2) response instructions (1 = Not suitable to 5 = Very suitable); 3) the degree of difficulty in answering the questionnaire (1 = Very difficult to answer to 5 = Very easy to answer); 4) the dimensions for knowing the difficulties of adjusting to work after vacations (1 = Nothing relevant to 5 = Very pertinent); 5) the order of the questions (1 = Not at all appropriate to 5 = Very adequate); 6) extension/amplitude of the instrument (1 = Very short to 5 = Very long). In the second part, regarding the specific aspects of the question groups, the group of experts was asked to evaluate (from 1 = Strongly disagree to 5 = Strongly agree) both instruments, regarding:1) relevance of the questions to the objective to be measured; 2) writing the questions and conditioning the answer; 3) clarity of the questions; 4) use of comprehensible terms for respondents; and 5) inclusion of all possible alternatives in the contemplated responses. Suggestions/ comments regarding the instrument were also requested. The evaluations obtained in both parts of the evaluation protocol were positive (M = 4.6). According to the group of experts' suggestions, the wording of some items of the questionnaire was revised, and items related to adapting

to schedules, and those in relation to the work-family interface/personal life were added, totaling 32 items.

Subsequently, a group of participants (n = 20) with heterogeneous demographic characteristics (i.e., with different educational qualifications, area of training and professional activity) were asked to answer the questionnaire, in order to identify possible semantic or comprehension difficulties. This pretest showed a Cronbach's alpha greater than .70. These participants were not included in the final sample.

Study 1: Sample

The application of the work instrument resulted in a sample of 232 respondents, of which 65.9% are female (n = 153) and 34.1% male (n = 79), aged between 20 and 73 years (M = 41.35; SD = 10.45). With regard to marital status, the majority of the sample, 56.5% (n = 131) is married or living in common law; 27.2% (n = 63) reported being single and 16.4% (n = 38) divorced/widowed. All respondents are Portuguese nationals, and the majority of the sample has higher education, 64.7% (n = 150); 25.9% (n = 60) secondary education and 9.5% (n = 22) completed basic education. The vast majority of participants work in full time, 95.3% (n = 221). Regarding professional activity, data analysis shows that there is no response from 31 of the sample elements (13.4%), as well as different areas of activity, with a greater distribution to the administrative area (19.4%, n = 45) and senior technicians (17.2%, n = 40). About 55% work in the public sector.

Regarding vacations and when respondents were asked to report to the last vacation period with 15 days or more of absence from work, it was found that for the vast majority of the sample, the extended vacation period had been taken 3 or more weeks ago (n = 185; 79.7%), 7.3% (n = 17) had returned to work just 2 weeks ago, and 12.5% (n = 29) had their vacation ended in the week before completing the questionnaire. Regarding the variable's concentration degree and productivity level on the first day after vacations, there is a greater representativeness of the sample in the third quartile and a distribution without very significant differences in the second and last quartiles (see Figure 1), which puts the most respondents in the upper half of the graph, with a medium to high concentration degree and the productivity level on the first day immediately after vacations.

Study 1: Instruments

Post-Vacation Work Adjustment Scale. After the evaluation carried out by the experts and the necessary changes and corrections were made, the work adjustment scale resulted in an initial instrument composed of 32 questions, assessed using a Likert scale from 1 = None difficulty to 7 = Very difficult. In the questionnaire instructions, respondents were asked to indicate the degree of difficulty in readjusting in relation to the need to adjust to work, routines and colleagues again [e.g., item 7: "... to the goals of my work team and their contribution to the goals of the organization"; item 14: "... to the way I operate the tools I use in my work (e.g., email, software, programs, machines, thermometer)"; item 16: "... how to execute forms / paperwork (e.g., timesheets, expense reports, reports) in the course of doing my job"].

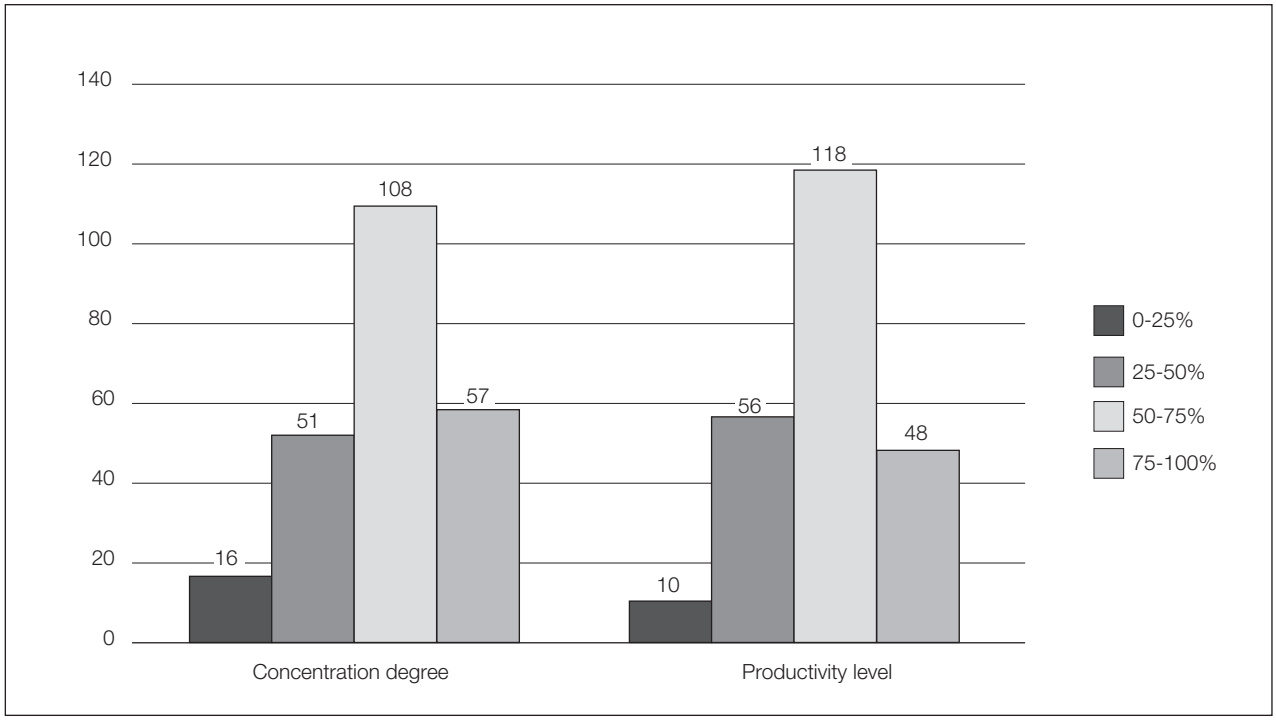
Another questions. In addition to the scale participants were asked about the return to work, in particular the concentration degree, productivity level and pace of work on the first working day after vacations, assessed on a 4-option response scale: a) 0-25%; b) 25-50%; c) 50-75%; and d) 75-100%. Participants were also asked about when they returned from their last vacation (1 week; 2 weeks; 3 or more weeks).

Sociodemographic data. In order to characterize the sample, questions about the participants' sociodemographic data, namely, gender, age, marital status, nationality, educational qualifications, and how long ago they returned from vacation were asked.

Study 1: Procedures

The questionnaire was applied both online and in person, in public places, universities, commercial facilities and companies. It was considered as an inclusion criterion to be professionally active (employed). The exclusion criteria were being under 18 years old and unemployed or retired. Approximately 15 minutes were estimated for filling. This study was approved by the Scientific Committee (protocol number UID/PSI/04345/2020). Participants were assured of the anonymity of their responses through fulfilment of ethical guidelines for administration questionnaires. Participation in the survey was voluntary, and participants did not receive any reward for their participation. The administration period was between August and September 2019.

Figure 1 – Number of respondents by concentration degree and productivity level on the first day of work after vacations



Study 1: Data analysis

Data analysis was performed using the SPSS (v.26) and SPSS AMOS (v.21) software. The psychometric properties of the work adjustment scale were assessed through exploratory factor analysis, confirmatory factor analysis and internal consistency.

In confirmatory factor analysis, the following criteria were considered (Byrne, 2001): χ^2, which represents a test of the significance of the minimized discrepancy function during model adjustment and the lower its value, the better the adjustment (Marôco, 2011); CMIN/*df*, corresponds to the probability of adjustment of the data to the theoretical model and its values should vary between 2 and 5; Comparative Fit Index (CFI) and Goodness of Fit Index (GFI) vary between 0 and 1, assuming .90 as a good adjustment value (Bentler & Bonett, 1980); Root Mean Square Error of Approximation (RMSEA) whose ideal value is between .05 and .08, accepting values up to .10. Internal consistency was assessed using Cronbach's alpha, which can vary on a scale from 0 to 1, with acceptable values starting from .70 (Nunnally, 1978).

Study 1: Results

Exploratory factorial analysis. In order to understand the structure of the P-VWAS, an exploratory analysis was carried out. The KMO index had a value of .912, and there was also a correlation between the items under study (Bartlett's sphericity test = 4478.889; $df = 496$; $p \leq .001$). The analysis of the main components, considering the criterion of variance extracted by factor and total extracted variance, using Promax rotation, allowed us to observe 4 factors, which

explain 71.40% of the variance of the results obtained. Items with a saturation value of less than .50 were then removed, as well as items that saturated in two or more factors, for a total of 12 items.

A new analysis was performed, which resulted in a two-dimensional structure. The KMO index showed a value of .930, with the existence of a correlation between the items under study (Bartlett's sphericity test = 3830.383; *df* = 171; $p \leq .001$). The analysis of the main components, considering the criterion of eigenvalues greater than 1 for the determination of the factors to be retained, allowed us to observe 2 factors (see Figure 2), which explain 65.97% of the variance of the results obtained and with factor weights ranging from .60 (item 9) to .96 (item 4) (see Table 1).

The means of the items ranged from 1.89 (item 2) to 3.07 (item 19). In terms of corrected item-total correlation, all items are above .30 (Nunnally & Bernstein, 1994), and are statistically acceptable. Asymmetry and kurtosis measurements show that the distribution of the 19 items is normal (symmetry values between .56 and 1.42 and kurtosis values between –.78 and 2.25), since the values are between 2 and 7, respectively (Bentler & Wu, 2002; Finney & DiStefano, 2006) (see Table 2).

Confirmatory Factorial Analysis. The 19 items of P-VWAS were subjected to a confirmatory factor analysis using the maximum likelihood estimator (ML). The adjustment values obtained were: $\chi^2_{(152)}$ = 885.002 which translates into a CMIN/*df* of 5.82, which is an acceptable value (Byrne, 2001).

Figure 2 – Screeplot of P-VWAS items

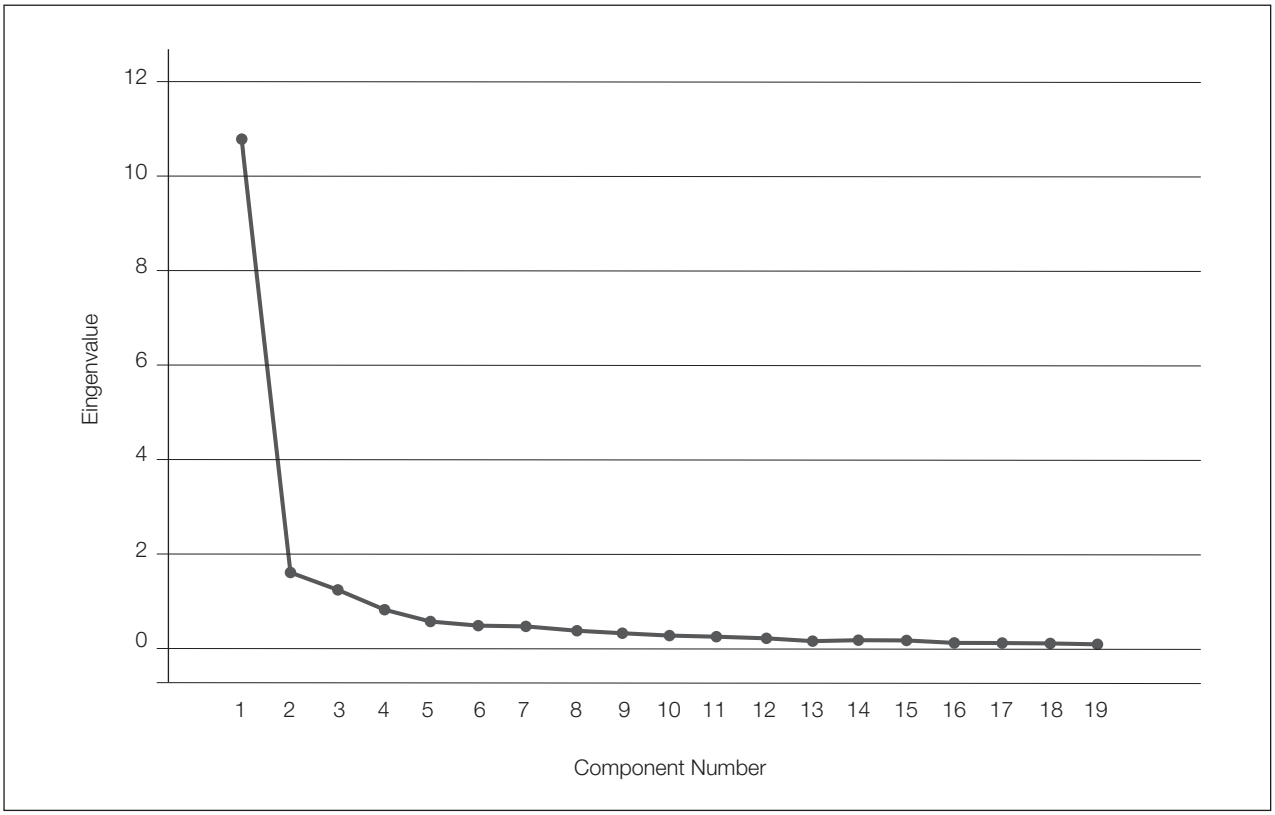


Table 1 – Components extracted from P-VWAS (factorial weights and communalities)

	Factor 1	Factor 2	Communalities
Item 1	.76		.48
Item 2	.92		.67
Item 3	.89		.68
Item 4	.96		.76
Item 5	.80		.66
Item 6	.75		.66
Item 7	.76		.71
Item 8	.67		.60
Item 9	.60		.65
Item 10	.65		.63
Item 11	.74		.73
Item 12	.78		.74
Item 13	.75		.57
Item 14	.72		.51
Item 15	.77		.58
Item 16	.66		.48
Item 17		.89	.72
Item 18		.93	.80
Item 19		.91	.81

Table 2 – Descriptive statistics of the items (n = 232)

Item	*M*	*SD*	Correlation corrected item-total	Cronbach's alpha (α) if item deleted	Asymmetry *SE* = .16	Kurtosis *SE* = .32
1	1.94	1.12	.61	.95	1.21	.95
2	1.89	1.06	.71	.95	1.37	1.42
3	1.94	1.13	.76	.95	1.35	1.65
4	1.97	1.11	.78	.95	1.42	2.25
5	2.31	1.41	.74	.95	1.15	.97
6	2.38	1.48	.79	.95	1.12	.69
7	2.12	1.28	.81	.95	1.20	.89
8	2.17	1.31	.73	.95	1.17	.98
9	2.27	1.38	.77	.95	1.06	.78
10	2.06	1.28	.75	.95	1.40	1.66
11	2.03	1.19	.82	.95	1.30	1.56
12	2.02	1.15	.83	.95	1.32	2.09
13	2.05	1.28	.69	.95	1.32	1.41
14	2.10	1.26	.66	.95	1.08	.47
15	1.99	1.17	.71	.95	1.22	1.19
16	2.16	1.35	.64	.95	1.26	1.23
17	2.61	1.76	.51	.95	.98	−.04
18	2.92	1.74	.56	.95	.59	−.68
19	3.07	1.79	.60	.95	.56	−.78

The values of CFI (.81), NFI (.78) and TLI (.76) are close to the value 1, which reveals a good adjustment (Marôco, 2011). The RMSEA (.10) is above the desirable value (Ullman, 2006).

Internal consistency. The scale presented a Cronbach's alpha of .95, and the two dimensions an alpha of .96 (*Organizational adjustment*) and .88 (*Work-life balance*) (see Table 3).

Descriptive statistics. Table 3 shows the means, standard deviations and Cronbach alphas of the scale, as well as the correlation values between their dimensions. It is possible to observe that the adjustment to work has a mean of 2.22 (*SD* = 1.00), with the dimension of *Work-life balance* being the one with the highest mean (*M* = 2.85; *SD* = 1.57) and the dimension of *Organizational adjustment* a lower mean (*M* = 2.12; *SD* = 1.01).

Predictive validity. In order to observe the predictive power of P-VWAS on issues related to return and job performance, regression analyzes were performed. P-VWAS showed a predictive power of about 4.9% on the productivity level on the first day of work after the vacation (β = −.222; p = .001) and 4.6% on the concentration degree in work activities, on the first day of work after the vacation (β = −.215; p = .001). The scale also explains 9% of the difficulty in getting back to work on the first day after vacation (β = .298; p = .001).

STUDY 2

Study 2 aims to assess the invariance of the scale with respect to gender and time of return from vacation.

Study 2: Sample

The sample consists of 332 participants, 220 of whom are female (66.3%) and 112 are male (33.7%) and aged between 19 and 73 years old (*M* = 38.86, *SD* = 11.39). Regarding marital status, 142 (42.8%) are married or living in common law, 118 are single (35.5%) and only 72 of the participants are divorced or widowed. Most participants have higher education (74.4%).

Table 3 – Means, standard deviations and Cronbach alphas - P-VWAS and correlation

	M	*SD*	α	1	1.1
1. Work adjustment	2.22	1.00	.95	–	
1.1. Organizational adjustment	2.12	1.01	.96	.982**	–
1.2. Work-life balance	2.85	1.57	.88	.668**	.514**

In relation to professional activity, this is spread over several areas, the most representative are: senior technicians (25.2%), health sector (17%), administrative (13.7%) and commerce sector (7.6%). About 48% of the sample works in the public sector.

Study 2: Instruments

The participants in this sample responded to the version of the P-VWAS scale obtained in Study 1, consisting of 19 items and 2 dimensions. The scale presented a Cronbach's alpha of .954, the dimension *Organizational adjustment* (16 items) an alpha of .951 and the dimension *Work-life balance* (3 items) obtained an internal consistency value of .900.

In addition to the P-VWAS scale, questions were also asked about the time the participants returned from vacation and sociodemographic questions to characterize the sample.

Study 2: Procedures

The procedures were the same as in Study 1. The questionnaire was applied both online and in person, in public places, universities, commercial facilities and companies. The same inclusion and exclusion criteria used in Study 1 were considered. Participants took about 15 minutes to complete a self-reported questionnaire. Freedom of participation and data confidentiality were previously guaranteed, in accordance with the ethical principles of the protocol mentioned in the previous study. The administration period was between November and December 2019.

Study 2: Data analysis

To analyse the measurement invariance across gender and period of return from vacations we used a multi-group confirmatory factor analysis adopting the maximum likelihood estimator (ML). As suggested by Chen (2007) the following criteria were used to determine acceptable model fit: $\Delta CFI \leq -.01$, $\Delta RMSEA \leq .015$, for tests of metric and scalar invariance. The period from return from vacations variable was operationalized in two groups: group 1 - individuals who returned from vacation 2 or less weeks ago ($n = 75$); group 2 - individuals who returned from vacation more than 3 weeks ago ($n = 257$).

Study 2: Measurement invariance across gender and across period of return from vacations

Analysis of measurement invariance of the P-VWAS scale across gender and period of return from vacations was conducted using multigroup confirmatory factorial analysis (MGCFA) with the 19 items two-factor model as the baseline model. As shown in Table 4, the configural invariance model across gender appeared to provide an acceptable fit to the data, although RMSEA is slightly above what is considered acceptable. Next, the comparison of the configural model with the metric model showed that ΔCFI and ΔRMSEA were all within the recommended ranges (e.g., Chen, 2007) and there was adequate statistical support for metric invariance across gender groups. After establishing metric invariance, the scalar invariance model was fitted to the data provided empirical support for scalar invariance across gender groups. Regarding the vacations return period, the configural invariance model provide an acceptable fit to the data. Similar to the indices previously obtained, the RMSEA value is considered high, compared to the values recommended as acceptable. The values obtained (ΔCFI and ΔRMSEA) allow to verify empirical support for scalar and metric invariance.

DISCUSSION

The main objective of our research was the development and initial validation of an adjustment scale to work after vacations in a Portuguese sample. Duo to the little existing literature on the subject, an attempt was made to articulate existing constructs, namely socialization/integration in organization. Thus, from the study by Sousa and Gonçalves (2019) and the adaptation and modification of the *Newcomer Socialization Questionnaire* of Haueter and colleagues (2003) and according to the evaluation carried out by the panel of experts, the results obtained through EFA and CFA allowed us to observe a two-dimensional structure of 19 items, which presented good values of internal consistency and reasonable adjustment indexes. The predictive validity of the scale was observed with regard to the productivity level and the concentration degree on the first day of work after vacations. The second study aimed to observe the extent to which the scale configuration and parameters are invariant (equivalent) for different groups. The MGCFA carried out confirmed the

Table 4 – Measurement invariance test across gender and across period of return from vacations

Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	RMSEA [90% CI]	AIC	ΔCFI	ΔRMSEA
Gender Invariance									
Configural	1293.07	302			.809	.10 [.094-.105]	1525.07		
Metric	1319.20	319	26.13	17	.807	.097 [.092-.103]	1517.20	−.002	−.003
Scalar	1334.94	338	15.74	19	.808	.095 [.089-.100]	1494.94	.001	−.002
Return from vacations									
Configural	1251.57	302			.817	.098 [.092-.103]	1483.57		
Metric	1322.62	319	71.05	17	.807	.098 [.092-.103]	1520.62	−.01	0
Scalar	1352.65	338	30.03	19	.805	.096 [.090-.101]	1512.65	−.002	−.002

Legenda. df = degree of freedom; CFI = *Comparative Fit Index*; RMSEA = *Root Mean Square Error of Approximation*; AIC = *Akaike Information Criterion*.

scale's invariance both between genders and in relation to the period of return from vacation. This result reinforces the possible generalization of the scale to different populations.

The final version of the scale (see Appendix) consists of the *Organizational adjustment* dimension (16 items) and the Work-life balance dimension (3 items). The first dimension is associated with work-related factors, namely adjustment to the processes and practices inherent to the function (e.g., culture, values, norms, team objectives, task execution, etc.). The *Work-life balance* dimension (3 items) refers to the adaptation to working hours, the management of the family-work interface and the management of personal commitments (e.g., leisure, hobbies, socializing with friends, etc.).

Despite the important contributions of this study, several limitations suggest avenues for future research. First, the adjustment indices obtained, namely the RMSEA, are not entirely satisfactory. The recommendations for the RMSEA cut-off points have been reduced considerably in recent years, since until the early 1990s, an RMSEA between .05 to .10 was considered an indication of fair adjustment and values above .10 indicated an inadequate adjustment (Hooper et al., 2008; MacCallum et al., 1996). Currently an RMSEA≤.08 is considered acceptable. However, according to Kenny and colleagues (Kenny, Kaniskan & McCoach, 2015) there is a greater sampling error for models with few degrees of freedom and small samples, which can lead to artificially large values of the RMSEA. Thus, further testing of the scale with a more representative sample should be considered in the future. Another of the limitations resulting from this study is related to the period of application of the scale to

participants. Participants were asked to report their last vacation with 15 days or more of absence from work. For the vast majority, this period had already occurred more than 3 weeks ago. This is a possible justification for the means of adjustment to work after vacations to present low values, that is, the participants in general, did not present a high degree of adjustment difficulty to work. To overcome this problem, it is suggested that in future investigations the instrument be applied immediately after returning to work. Further analysis must be carried out to test this initial validation of P-VWAS, for example, the convergent validity of the scale. This analysis can be performed with the cognitive and/or emotional demands of the job (e.g., Wach, Stephan, Weinberger & Wegge, 2020), considered as stressors challenges, since they are work demands that involve the possibility of future gains and personal growth (Crawford, Lepine & Rich, 2010). Other analysis for possible items reduction (e.g., Item Response Theory), test-retest and cross-cultural validation should also be considered. The application of the scale to other populations, such as teleworkers or businessmen/entrepreneurs, will also allow better testing of the instrument.

CONCLUSION

Adjusting to work after vacations is an extremely relevant topic for organizations, as it has implications for the productivity, performance and well-being of employees. This study focused on returning after vacations, but we believe that this scale can be adjusted to other situations that imply a prolonged absence (i.e., 10 to 15 days) from the workplace. For example, maternity leave, sick leave, or even returning home after an expatriation process. Considering the current global situation, a consequence of the COVID-19 pandemic, which forced many employees to be away from their work for at least 2 months, the application of this scale would be an asset for organizations to understand the main difficulties of adjusting to work by part of its employees. The identification of the degree of adjustment to work after a period of absence, will allow the outline of organizational strategies aimed at facilitating the return and respective adjustment to routines. Namely, intervention strategies that enhance a policy of reintegration and reduction of labor requirements after the return from vacation. For example, performing a return to work debriefing, with the objective of assessing the level of preparation for the return, defining an action plan for better adjusting the employee to work and updating the employees about the events that occurred during their absence (Sousa & Gonçalves, 2019). Or, adjust the workload, in the first two weeks, in order to facilitate the transition to resume the reconciliation of personal and professional life (Sousa & Gonçalves, 2019). It would be important to understand the adjustment strategies that people make, but also that individual variables (e.g., psychological capital, self-efficacy) or attitudes towards work are facilitators of the new readjustment. In short, an organization that adopts measures that facilitate the readaptation and readjustment of employees, will contribute to the creation of positive work environments, a greater commitment on the part of employees, a greater perception of organizational support, which will certainly translate into better performance and greater job satisfaction.

Funding: This work was funded by national funds through FCT - Fundação para a Ciência e a Tecnologia - as part the project CIP/UAL - Refª UID/PSI/04345/2020

Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest

References

BENTLER, P. & BONETT, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606. doi.org/10.1037/0033-2909.88.3.588

BENTLER, P. & WU, E. (2002). *EQS for windows user's guide*. Encino, CA: Multivariate Software, Inc.

BINNEWIES, C., SONNENTAG, S. & MOJZA, E. (2009). Daily performance at work: Feeling recovered in the morning as a predictor of day-level job performance. *Journal of Organizational Behavior*, *30* (1), 67-93. doi.org/10.1002/job.541

BLASCHE, G., ARLINGHAUS, A. & DORNER, T. (2014). Leisure opportunities and fatigue in employees: A large cross-sectional study. *Leisure Sciences*, *36* (3), 235-250. doi.org/10.1080/01490400.2014.886981

BLOOM, J., GEURTS, S. & KOMPIER, M. (2012). Effects of short vacations, vacation activities and experiences on employee health and well-being. *Journal of the International Society for the Investigation of Stress*, *28* (4), 305-318. doi.org/10.1002/smi.1434

BLOOM, J., GEURTS, S., SONNENTAG, S., TARIS, T., WEERTH, C. & KOMPIER, M. (2011). How does vacation from work affect employee health and well-being? *Psychology & Health*, *26* (12), 1606-1622. doi.org/10.1080/08870446.2010.546860

BLOOM, J., GEURTS, S., TARIS, T., SONNENTAG, S., WEERTH, C. & KOMPIER, M. (2010). Effects of vacation from work on health and well-being: Lots of fun, quickly gone. *Work & Stress*, *24* (2), 196-216. doi.org/10.1080/02678373.2010.493385

BUDNICK, C., ROGERS, A. & BARBER, L. (2020). The fear of missing out at work: Examining costs and benefits to employee health and motivation. *Computers in Human Behavior, 104*, 1-13. doi.org/10.1016/j.chb.2019.106161

BYRNE, B. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

CHEN, F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504. doi.org/10.1080/10705510701301834

CRAWFORD, E.R., LEPINE, J.A. & RICH, B.L. (2010). Linking job demands and resources to employee engagement and burnout: a theoretical extension and meta-analytic test. *Journal of Applied Psychology*, *95* (5), 834-848. doi.org/10.1037/a0019364

DeVELLIS, R. (2016). *Scale development: Theory and applications (4th ed.)*. Newbury Park: Sage Publications, Inc.

ETZION, D. (2003). Annual vacation: duration of relief from job stressors and burnout. *Journal of Anxiety, Stress, and Coping*, *16* (2), 213-226. doi.org/10.1080/1061580021000069425.

FINNEY, S. & DiSTEFANO, C. (2006). Non-normal and categorical data in structural equation modeling. In G. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course*. Greenwich, CT: Information Age Publishing.

FRITZ, C. & SONNENTAG, S. (2006). Recovery, well-being, and performance-related outcomes: the role of workload and vacation experiences. *Journal of Applied Psychology*, *91* (4), 936-945. doi.org/10.1037/0021-9010.91.4.936

FURR, R. (2010). *Scale construction and psychometrics for social and personality psychology*. USA: Sage Retrieved from https://pdfs.semanticscholar org/245e/18045e05363a12e163acf388e245e202ec40.pdf

HAUETER, J., MACAN, T. & WINTER, J. (2003). Measurement of newcomer socialization: Construct validation of a multidimensional scale. *Journal of Vocational Behavior*, *63* (1), 20-39. doi.org/10.1016/S0001-8791(02)00017-9

KAWAKUBO, A. & OGUCHI, T. (2019). Recovery experiences during vacations promote life satisfaction through creative behavior. *Tourism Management Perspectives*, *30*, 240-250. doi.org/10.1016/j.tmp.2019.02.017

KENNY, D., KANISKAN, B. & McCOACH, D. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods Research*, *44* (3), 486-507. doi.org/10.1177/0049124114543236

KINNUNEN, U. & FELDT, T. (2013). Job characteristics, recovery experiences and occupational well-being: Testing cross-lagged relationships across 1 year. *Stress and Health*, *29* (5), 369-382. doi.org/10.1002/smi.2483

KOERBER, R., ROUSE, M., STANYAR, K. & PELLETIER, M-H. (2018). Building resilience in the workforce. *Organizational Dynamics*, *47* (2), 124-134. doi.org/10.1016/j.orgdyn.2017.08.002

KORPELA, K. & KINNUNEN, U. (2011). How is leisure time interacting with nature related to the need for recovery from work demands? Testing multiple mediators. *Leisure Sciences*, *33* (1), 1-14. doi.org/10.1080/01490400.2011.533103

KÜHNEL, J. & SONNENTAG, S. (2011). How long do you benefit from vacation? A closer look at the fade-out of vacation effects. *Journal of Organizational Behavior*, *32* (1), 125-143. doi.org/10.1002/job.699

LOUNSBURY, J. & HOOPES, L. (1986). A vacation from work: Changes in work and nonwork outcomes. *Journal of Applied Psychology*, *71* (3), 392-401. dx.doi.org/10.1037/0021-9010.71.3.392

MARÔCO, J. (2011). *Análise Estatística com o SPSS statistics (5ª ed.)*.

Pero Pinheiro: ReportNumber.

MITAS, O. & KROESEN, M. (2019). Vacations over the years: A cross-lagged panel analysis of tourism experiences and subjective well-being in the Netherlands. *Journal of Happiness Studies*, 1-20. doi.org/10.1007/s10902-019-00200-z

NAWIJN, J., MARCHAND, M., VEENHOVEN, R. & VINGERHOETS, J. (2010). Vacationers happier, but most not happier after a holiday. *Journal of Applied Research Quality Life*, *5* (1), 35-47. doi.org/10.1007/s11482-009-9091

NUNNALLY, J. (1978). *Psychometric theory (2nd ed.)*. New York, NY: McGraw-Hill.

NUNNALLY, J. & BERNSTEIN, I. (1994). *Psychometric theory (3rd ed.)*. New York, NY: McGraw-Hill.

PRYZBYLSKI, A.K., MURAYAMA, K., DEHAAN, C.R. & GLADWELL, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*, *29*, 1841-1848. doi.org/10.1016/j.chb.2013.02.014

SONNENTAG, S. (2018). The recovery paradox: Portraying the complex interplay between job stressors, lack of recovery, and poor well-being. *Research in Organizational Behavior*, *38*, 169-185. doi.org/10.1016/j.riob.2018.11.002

SOUSA, C. & GONÇALVES, G. (2019). Back to work bang! Difficulties, emotions, and adjustment strategies in returning to work after vacation. *International Journal of Human Resource Management* (online). doi.org/10.1080/09585192.2019.1602784

STRAUSS-BLASCHE, G., EKMEKCIOGLU, C. & MARKTL, W. (2000). Does vacation enable recuperation? Changes in well-being associated with time away from Work. *Occupational Medicine*, *50* (3), 167-172. doi.org/10.1093/occmed/50.3.167

STRAUSS-BLASCHE, G., REITHOFER, B., SCHOBERSBERGER, W., EKMEKCIOGLU, C. & MARKTL, W. (2005). Effect of vacation on health: Moderating factors of vacation outcome. *Journal of Travel Medicine*, *12* (2), 94-101. doi.org/10.2310/7060.2005.12206

TUCKER, P., DAHLGREN, A., AKERSTEDT, T. & WATERHOUSE, J. (2008). The impact of free-time activities on sleep, recovery and well-being. *Applied Ergonomics*, *39* (5), 653-662. doi.org/10.1016/j.apergo.2007.12.002

ULLMAN, J. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87* (1), 35-50. doi.org/10.1207/s15327752jpa8701_03

WACH, D., STEPHAN, U., WEINBERGER, E. & WEGGE, J. (2020). Entrepreneurs' stressors and well-being: A recovery perspective and diary study. *Journal of Business Venturing*. doi.org/10.1016/j.jbusvent.2020.106016

ZIJLSTRA, F. & SONNENTAG, S. (2006). After work is done: Psychological on recovery from work. *European Journal of Work and Organizational Psychology*, *15* (2), 129-138. doi.org/10.1080/135943205500513855

APPENDIX

Final version of the Post-Vacations Work Adjustment Scale (P-VWAS)

Regarding the need to adjust to work, routines and colleagues (etc.), to what extent do you find it difficult to readjust to…:
1. ... the specific names of products and services produced or supplied by the organization.
2. ... the organization's culture (e.g., values, rituals).
3. ... the structure of the organization (e.g., organization chart, departments).
4. ... the organization's operations (e.g., who does what).
5. ... the organization's internal policy (e.g., chain of command, who is influential, what needs to be done to move forward).
6. ... the management style of the organization.
7. ... the goals of my work team and their contribution to the organization goals.
8. ... what the supervisor expects from the work team.
9. ... the management style of the team supervisors.
10. ... the performing tasks according to team standards.
11. ... the rules and procedures of my work team.
12. ... the team policy (e.g., who is influential, what needs to be done to move forward).
13. ... the responsibilities, tasks and projects for which I was hired.
14. ... the way of operating the tools I use in my work (e.g., email, software, programs, machines, thermometer).
15. ... the way and the people to whom I must go to acquire the necessary resources to perform my work (e.g., equipment, facilities).
16. ... how to execute forms/paperwork (e.g., timesheets, expense reports, etc.) in the course of doing my job.
17. ... the work schedules.
18. ... the family-work interface management.
19. ... the management of my personal commitments (e.g., leisure, hobbies, socializing with friends, etc.).

Note. Dimensions: *Organizational adjustment* (items 1 to 16); *Work-life balance* (items 17 to 19).

DOI: 10.26387/bpa.289.5

Advanced interpretation of WAIS-IV. The application of the CHC model to a WAIS-IV protocol

Lina Pezzuti[1], Clara Michelotti[2], Marco Lauriola[3], Margherita Lang[2]

[1] Department of Dynamic, Clinical and Health Psychology, Sapienza University of Rome
[2] Studio Associato A.R.P. - Milan
[3] Department of Psychology of Developmental and Socialisation Processes, Sapienza University of Rome

lina.pezzuti@uniroma1.it

ABSTRACT. Con la pubblicazione delle scale Wechsler di quarta generazione (WPPSI-IV, WISC-IV e WAIS-IV) avviene un cambiamento rilevante determinato dalle teorie differenti delle neuroscienze cognitive fondate sulla ricerca clinica e neuropsicologica. Dalle prime analisi fattoriali confermative condotte sui campioni di standardizzazione statunitense e italiano della WAIS-IV emerge la medesima struttura a quattro fattori. La WAIS-IV, in particolare, permette quindi il computo di quattro indici (o fattori): Comprensione verbale (ICV), Ragionamento visuo-percettivo (IRP), Memoria di lavoro (ML) e Velocità di elaborazione (IVE). Ciascuno degli indici concorre al computo del punteggio composito totale o Quoziente Intellettivo. Tuttavia, alla fine del secolo scorso sono comparsi numerosi modelli di intelligenza, alcuni dei quali hanno portano alla realizzazione di nuovi strumenti per la valutazione del costrutto o all'aggiornamento di quelli esistenti, e ricerche successive statunitensi e italiane hanno dimostrato che i dati della WAIS-IV possono anche essere letti alla luce della *Cattell*, *Horn*, *Carroll theory of intelligence* (o teoria CHC) distinguendo 5 fattori: Intelligenza cristallizzata (Gc); Elaborazione visiva (Gv); Intelligenza fluida (Gf); Memoria a breve termine (Gsm); Velocità di elaborazione (Gs). L'obiettivo del presente lavoro è quello di evidenziare attraverso un caso clinico l'utilità di avvalersi del modello a cinque fattori CHC invece di quello a quattro fattori, in particolare quando uno dei primi fattori risulta non interpretabile come abilità unitaria e coesa.

SUMMARY. A relevant change occurs with the publication of the fourth generation Wechsler Scales (WPPSI-IV, WISC-IV and WAIS-IV), determined by the different theories of cognitive neuroscience based on clinical and neuropsychological research. The first confirmatory factor analyses conducted on the US and Italian standardization samples of the WAIS-IV show the same four-factor structure. The WAIS-IV, in particular, allows the calculation of four indices (or factors): Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI) and Processing Speed Index (PSI). Each of the indices contributes to the total composite score or Intellectual Quotient. However, at the end of the last century, numerous models of intelligence appeared, some of which led to the creation of new tools for assessing the construct or updating existing ones, and subsequent U.S. and Italian research have shown that the WAIS-IV data can also be read in the light of the Cattell, Horn, Carroll theory of intelligence (or CHC theory) distinguishing 5 factors: Crystallized Intelligence (Gc); Visual Processing (Gv); Fluid Intelligence (Gf); Short-Term Memory (Gsm); Processing Speed (Gs). The objective of this paper is to highlight through a clinical case the usefulness of using the five-factor CHC model instead of the four-factor model, particularly when one of the first factors is not interpretable as a unitary and cohesive ability.

Keywords: *WAIS-IV, CHC model, Clinical case*

INTRODUCTION

Numerous models of intelligence appeared at the end of the last century, some of which led to the development of new instruments for assessing the construct or updating existing ones. Since the tests most frequently used to measure cognitive abilities are built on the psychometric model (Neisser et al., 1996), we will focus our attention on the latest generation of psychometric models that have guided the implementation of the instruments and the reading of the results.

In the late 1990s, McGrew (1997) proposed a model that integrates the one proposed by Carroll (1993) and those proposed by Horn and Cattell.

Carroll cognitive abilities are differentiated into three layers (Strata) or levels. The architecture of the model is hierarchical and can be represented as a pyramid, at the apex of which is Stratum III, which is the conceptual equivalent of Spearman's and Vernon's g-factor. Stratum II is composed of a relatively small number of broad cognitive abilities (Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Visual Perception, Auditory Perception, Retrieval Ability, Cognitive Speediness, and Reaction Time). Beneath these broad skills, there are countless narrow skills (about 69) or abilities that are part of Stratum I.

Horn and Cattell's Gf-Gc model is a "truncated" hierarchical model, as it does not include a g-factor at the apex or a two-stratum model, in which first-order factors form the upper stratum and second-order factors form the lower stratum. The upper stratum includes several broad cognitive abilities; the lower stratum includes Thurstone's primary abilities (Horn, 1985) and the Cattell Horn Carroll theory of intelligence (CHC), a multicomponential hierarchical model with an unprecedented empirical basis (Schneider & McGrew, 2018).

The CHC model includes operationalized broad and narrow abilities: broad abilities are the basic constitutional characteristics of people that endure and can govern or influence a wide range of behaviours in a specific area, narrow abilities represent specific (detailed) aspects of the broad ability to which they belong). Broad abilities are: Crystallized Intelligence (Gc), Visual Processing (Gv), Quantitative Knowledge (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Fluid Intelligence (Gf), Processing Speed (Gs), Long-Term Storage and Retrieval (Glr), Auditory Processing (Ga), and Decision-Making Speed/Reaction Time (Gt). The narrow abilities underlying each broad ability are multiple.

Based on research data, the model has undergone some updates. In 2012 and 2018, Schneider and McGrew proposed significant revisions with the addition of new skills, the elimination of others, and a focus on the relationship between skills and information processing.

With the publication of the fourth-generation Wechsler Scales (WPPSI-IV, WISC-IV and WAIS-IV), a major change occurs in the history of this family of instruments, a change brought about by "different theories of cognitive neuroscience grounded in clinical and neuropsychological research" (Weiss, Saklofske, Coalson & Raiford, 2010, p. 62). In summary: the WPPSI-IV is an instrument for assessing cognitive functioning of subjects from 2 years, 6 months, and 0 days to 7 years, 3 months, and 30 days; the WISC-IV is an instrument for assessing cognitive functioning of subjects from 6 years, 0 months, and 0 days to 16 years, 11 months, and 30 days; the WAIS-IV is an instrument for assessing cognitive functioning of subjects from 16 years, 0 months, and 0 days to 89 years, 11 months, and 30 days.

From confirmatory factor analyses conducted on the U.S. (Wechsler, 2008) and Italian (Orsini & Pezzuti, 2013, 2015) standardization sample of the WAIS-IV, an important finding emerges: the same four-factor structure both considering only the 10 core subtests and all 15 subtests including the supplementary ones (the same result was found for the WISC-IV; Wechsler, 2003). The subtests are then grouped into four factors that assess specific cognitive domains. The WAIS-IV allows the calculation of four indices (factors): Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI) and Processing Speed Index (PSI). Each of the indices contributes to the computation of the total composite score or Intellectual Quotient (IQ). To compute the four indexes, it is sufficient to administer the 10 core subtests.

The 5 supplementary subtests can be administered in two circumstances: 1) when the clinician needs to replace a core subtest with a subtest of the supplementary ones (for example, if a person has physical or sensory limitations, or if the score of a core subtest is invalidated because of errors in administration or because the person always answers "I don't know"); 2) there is a need for clinical investigation of a particular cognitive ability, and complete the diagnosis by analyzing discrepancies between different subtests.

The factorial structure of the WAIS-IV has been the

subject of several analyses from which alternative models have emerged - in addition to the one formed by four factors - that allow a better understanding of the patient's "functioning".

The first factor analyses on the WAIS-IV data were conducted by Benson, Hulac and Kranzler (2010). Subsequently, Weiss, Keith, Zhu and Chen (2013) compared, based on U.S. data, both the four-factor and five-factor structures that best met Cattell, Horn, and Carroll's model (CHC; McGrew, 1997). The results of their analyses showed that:

- the Crystallized Intelligence factor (Gc) was saturated by the Similarity, Vocabulary, Information, and Comprehension subtests;
- the Visual Processing factor (Gv) was represented by the subtests Block Design, Matrix Reasoning, Puzzles, Figure Weights, and Picture Completion;
- the Fluid Intelligence factor (Gf) was saturated by the Matrix Reasoning, Figure Weights, and Arithmetic Reasoning subtests. Analyses also revealed a narrow ability in Quantitative Reasoning (QR), saturated by Figure Weights and Arithmetic Reasoning;
- the Short-Term Memory factor (Gsm) was represented by Digit Span, Letter and Number Sequencing and Arithmetic Reasoning;
- the Processing Speed factor (Gs) was represented by Coding, Symbol Search, and Cancellation.

These findings have been confirmed in more recent work such as that of Ryan and colleagues (Ryan, Kreiner, Gontkovsky, Golden & Myers-Fabian, 2019) and that conducted on the Italian calibration data of the WAIS-IV (Pezzuti, Lang, Rossetti & Michelotti, 2018).

Thus, the factorial structure of the WAIS-IV (in the US and Italian editions) allows us to read the results according to both the four-factor model (Wechsler, 2008; Orsini & Pezzuti, 2013, 2015) and the five-factor model or CHC model (Pezzuti et al., 2018; Weiss et al., 2013) for all age groups (16-90 years).

Regardless of the model chosen, the clinician can compute a total composite score (IQ) and some partial composite scores related to specific cognitive domains, which are particularly useful for understanding the subject's cognitive functioning (Kaufman, Raiford & Coalson, 2016; Weiss, Saklofske, Holdnack & Prifitera, 2016). On the other hand, if the clinician uses the CHC model, they must administer 15 subtests, which can be particularly burdensome for the patient. Hence the search for an alternative, which constitutes an "acceptable compromise" for both the patient and the clinician.

We therefore considered the hypothesis already explored by Lichtenberger and Kaufman in a 2009 paper: keeping the CHC theory as the reference theory and reducing the number of subtests to be administered to two subtests for each of the five CHC factors:

- Crystallized Intelligence (Gc): Vocabulary and Information;
- Visual Processing (Gv): Block Design and Puzzles;
- Fluid Intelligence (Gf): Matrix Reasoning and Figure Weights (supplemental subtest);
- Short-Term Memory (Gsm): Digit Span and Letter and Number Sequencing (additional subtest);
- Processing Speed (Gs): Symbol Search and Coding.

The choice of the pairs of subtests to be administered for each factor was guided by the results of Keith's (2009) confirmatory factor analysis and by the effects that the single broad ability measured by the clusters has in the clinic and, consequently, in the person's functioning in daily life.

In light of the considerations of Lichtenberger and Kaufman (2009) and the work of Pezzuti and colleagues (2018) to assess the broad ability of Crystallized Intelligence (Gc), we believe that the most appropriate subtests are Vocabulary and Information (core subtests), which have high levels of saturation across all age groups. The two subtests are excellent measures of the background knowledge possessed by a person and are less influenced by fluid reasoning than the other two subtests (Similarities and Comprehension) that contribute to the Verbal Comprehension Index (VCI) computation (Lichtenberger & Kaufman, 2009).

Block Design and Puzzles (core subtests) are the subtests that best appear to measure the broad Visual Processing (Gv) ability, as high saturations on the factor emerge for both age groups. The Picture Completion subtest, although it measures the broad Visual Processing skill (Gv), also requires Crystallized Intelligence (Gc) and the narrow skills of Flexibility of Closure (CF) and General Information (K0) for a correct performance and is therefore not very relevant to the broad skill.

The broad ability of Fluid Intelligence (Gf) can be measured by Matrix Reasoning (fundamental) and Figure Weights (supplemental), which have high saturations on the factor and strong representation of the construct (Flanagan, Ortiz & Alfonso, 2013; Lichtenberger & Kaufman, 2009). Figure Weights, moreover, as demonstrated in work on data from the Italian calibration of the WAIS-IV by Pezzuti and Rossetti (2017a, 2017b), can also be administered to older subjects.

The broad Short-Term Memory (Gsm) ability is measured by the fundamental Digit Span subtest and the supplementary Letter and Number Sequencing subtest, which are the subtests that most saturate the factor and represent it for both age groups. According to the results of the work of Pezzuti and Rossetti (2017a, 2017b), the Letter and Number Sequencing subtest can also be administered to Italian subjects over 69 (saturations on the Gsm factor are almost the same for both age groups considered). It is important for the psychologist to pay particular attention to Arithmetic Reasoning (core subtest) because, although it can be considered a measure of short-term memory, it also measures other broad abilities, such as crystallized knowledge, fluid reasoning and quantitative reasoning, as well as some other variables such as distractibility and anxiety (Lichtenberger & Kaufman, 2009).

To assess the broad ability of Processing Speed (Gs), the core subtests of Symbol Search and Coding, which are the same subtests that contribute to the Index of Processing Speed (PSI) [4-factor model in the U.S. and Italian manuals] appear to be adequate.

5-FACTOR CHC MODEL AND WAIS-IV

Having another model available (in addition to the 4-factor model) to interpret the WAIS-IV data, without this implying an excessive workload for the patient and the clinician, makes it possible to reduce the risk that the psychologist finds himself in the condition of not being able to explain the data obtained according to the "traditional" method of interpretation (4 factors), since one or more of these composite scores may sometimes not be unitary, namely internally cohesive. In fact, the clinician must keep in mind that when reading all composite scores (including the IQ reading), the unitary nature of the score must be considered. A composite score is unitary if the difference between the highest and lowest scores of its component values is less than a "threshold value" (Pezzuti, 2016). The "threshold value" corresponds to the minimum difference required, for a score to occur in a very low percentage (6.7%) of the general population. "Threshold values" for the Italian population are available in Orsini, Pezzuti and Hulbert (2015), Pezzuti (2016), and Lang, Michelotti, Bardelli and Pezzuti (in press).

For example, suppose that a 32-year-old patient obtains a total IQ of 119, to decide whether this IQ is representative of a unitary and internally cohesive ability, we need to analyse the difference between the highest and the lowest score among the 4 indexes that compose the total IQ. The same patient scored an VCI = 131, PRI = 121, WMI = 109 and PSI = 89, so we calculate the difference between the highest IQ (131 of VCI) and the lowest IQ (89 of PSI) which is 42 and compare it to the cut-off value which is ≥38, since 42 IQ points is higher than the cut-off we can reasonably conclude that the total IQ of 119 is not unitary and cohesive within it.

The lack of unity of a score can be a real obstacle with respect to the purposes for which the test was administered, namely to have nomothetic data to confirm or disconfirm clinical hypotheses. Moreover, it can induce the clinician to privilege idiographic interpretations which have many limitations, because they are often based on a qualitative reading of the data that is affected by the subjectivity of the clinician and/or his model of psychopathology. It is also possible that the non-uniformity of a score induces the clinician to "fall back" on the results to the single subtests and/or on the ipsative analysis.

Some researchers are of the opinion that it is possible to use discrepancy scores between subtests that make up the same index to determine whether the score is interpretable. In their view, a high dispersion among the scores that make up the index makes it uninterpretable (Flanagan & Kaufman, 2009). Other researchers take a different view. For example, for Reynolds no level of dispersion among scores makes an Index uninterpretable (Reynolds & Kamphaus, 2015).

Thus, the clinician must ask another question: when an unusual level of variability is detected among the scores that make up an index, what interpretation may be appropriate? This question is consequential to the findings of the research. In fact, no data emerges from the research showing that an index score has less predictive efficacy because of the level of dispersion present among the scores that comprise it (Ryan, Kreiner & Burton, 2002). Reynolds and Kamphaus (2015) are of the opinion that the belief that high variability negatively affects the predictive validity of the index is fundamentally a myth. The clinician can make some assumptions about this finding.

– If the clinician finds an unusual level of dispersion among the scores that make up an index, he or she can ask himself or herself whether the index is a good summary statistical indicator for the variable in question. For example, if there

is a difference too large between the subtests that make up the PRI, is the index a good overall representation of Visual Perceptual Reasoning?

- Next, the clinician needs to shift the focus to the subject's functioning and formulate hypotheses congruent with the data available. For example, he might make a further interpretation and hypothesize that the cognitive skills measured come into play in the everyday life. In this case, it is possible to consider what the fallout of a low PSI might be with respect to both specific levels of school/academic performance and everyday life situations.
- The clinician may add in his textological report that the degree of variability among PRI abilities appears unusually high.
- It is possible that the clinician may interpret the level of dispersion as a stand-alone variable or refer, if other data are available, to more specific constructs/skills underlying the index itself.

In summary, if the clinician administers 10 subtests of which 8 are foundational and two are supplemental (Figure Weights and Letter and Number Sequencing), he or she can read the data by referring to 5 broad skills described by the CHC model.

If the clinician also wants to assess the 4 primary factors, i.e., the indices (VCI, PRI, WMI, and PSI), he or she will also need to administer the fundamental subtests Similarities and Arithmetic Reasoning: therefore, to make a double interpretation, 12 subtests must be administered.

CLINICAL CASE: WAIS-IV READING OF RESULTS ACCORDING TO TWO MODELS

We propose as an illustrative clinical case of the use of the CHC model for the reading of the results the case of a young man of 23 years (Giovanni), who requested a consultation, because at a time of difficulty in the continuation of university studies: "I cannot study ... I am more easily distracted than usual and I remember only some information ... if they ask me those, then I pass the exam, otherwise ... I cannot". This difficulty seems to reduce not only his decision-making power (he could do "something else"), but it also has repercussions on his interpersonal relationships and on the consideration he has of himself, for which feelings of inadequacy have appeared, which he cannot justify, in addition to the impossibility of seeing alternatives.

After attending a technical institute - with results "more or less average ... sometimes it was good, sometimes not so much" - he decided to enroll in the faculty of mechanical engineering. The results are discontinuous, but he completes the three-year course. The discontinuity of performance is not an object of concern. The real difficulties begin the first year of the master's degree: he is unable to pass his exams and complains of attentional problems and difficulty in concentrating. All of this translates into feelings of anxiety and depression combined with a feeling of inability to commit to learning.

After a collection of bio-psycho-social data, the following are administered: the WAIS-IV, the Rorschach and the *Dimensional Assessment of Personality Pathology – Basic Questionnaire (DAPP-BQ*; Livesley & Jackson, 2009, It. ed. 2014). Since the patient does not report a "frank" symptomatology, but rather complains of disorders that can be attributed to a multiplicity of causes, it was considered essential to investigate the cognitive-adaptive and personality areas. The assumption is that between cognitive functioning and personality there is a biunivocal relationship as claimed in the literature.

The WAIS-IV, as mentioned elsewhere, allows not only the assessment of operationalized cognitive abilities, but also the effects of emotional interference. Using an instrument that allows to consider specific cognitive functioning puts the clinician in the position to detect the presence/absence of a flexion in a cognitive ability; the failure of any compensatory modalities; the incidence of emotional variables on cognitive functioning.

The administration of the WAIS-IV took place in two successive moments in order to avoid excessive fatigue.

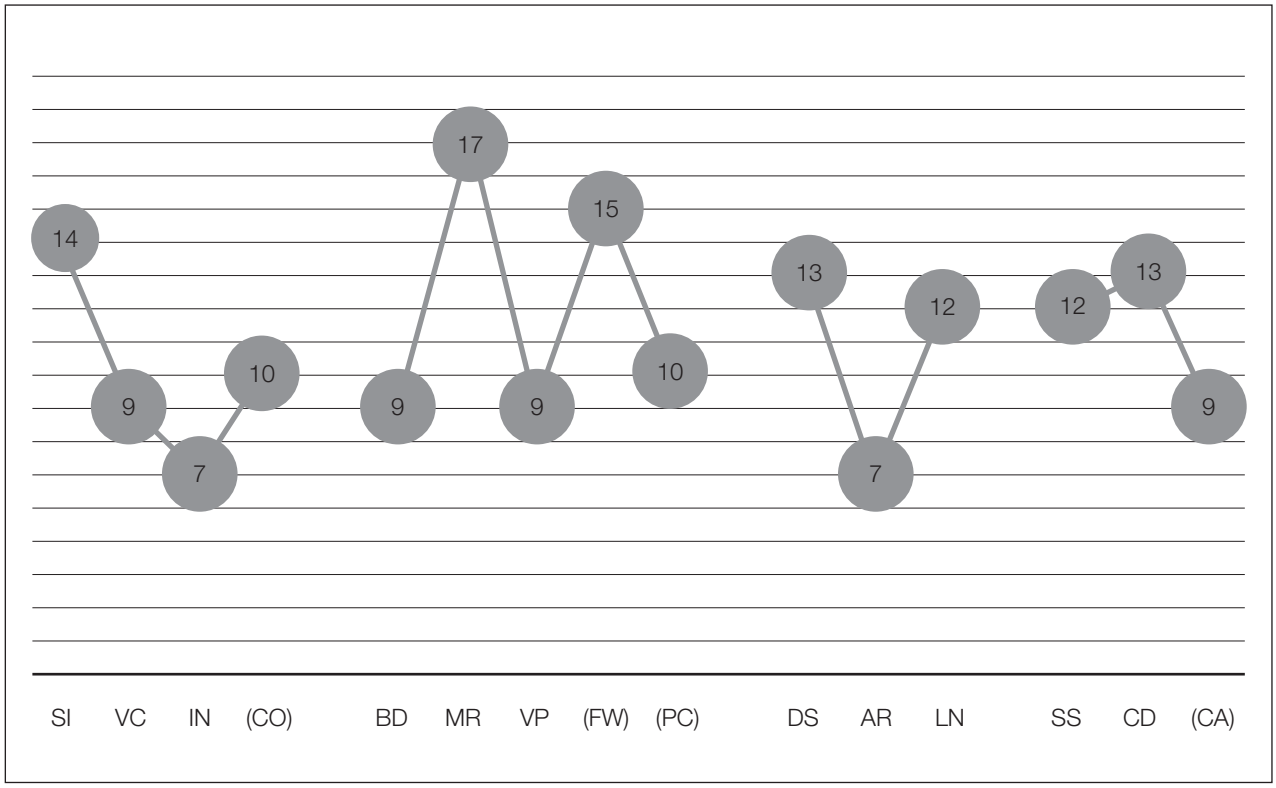
The scale was administered by a clinical psychologist with significant experience in the use of the instrument, and the evaluation of the protocol was supervised by the authors of the article.

Reading according to the 4-factor model

Figure 1 shows the weighted subtest scores that Giovanni obtained on the WAIS-IV.

The patient's performance is not homogeneous meaning

Figure 1 – Profile of Giovanni's scores at WAIS-IV



there are particularly high scores on some subtests, for example, Matrix Reasoning (MR) and Figure Weights (FW) and at the lower end of the mean on other subtests such as Information (IN) and Arithmetic Reasoning (AR).

The composite scores (Indices), which are the "preferred level for clinical interpretation" (Weiss, Chen, Harris, Holdnack & Saklofske, 2010, p. 61), were all found to be of medium/medium-high level.

Table 1 shows the composite scores with their respective confidence intervals (95%), percentile ranks, and qualitative descriptor.

If we consider a fundamental parameter for the interpretation of the Indexes, namely their unitarity, further information emerges. If a score is not unitary, it is improper to attribute to it the meaning it would have if it were unitary because it is not an adequate descriptor of the abilities that the index is intended to measure, and in Giovanni's case (see Table 2) the only unitary index is PSI (medium/medium-high level), so the data that emerged from the WAIS-IV read with the 4-factor model provide the clinician with two pieces of information: the patient has a medium/medium-high total composite score and a Processing Speed Index (PSI) that in turn is medium/medium-high. The non-uniformity of the other indices does not allow for reliable hypotheses regarding the interaction between the different abilities measured by the WAIS-IV.

Table 1 – Giovanni's composite scores

Index	Composite score (IQ)[a]	CI (95%)	Percentile rank	Level[b]
Verbal Comprehension Index (VCI)	100	93-107	50	average
Perceptual Reasoning Index (PRI)	110	103-116	75	average/high-average
Working Memory Index (WMI)	100	93-108	51	average
Processing Speed Index (PSI)	114	104-121	82	average/high-average
Intellectual Quotient (IQ)	108	103-113	69	average/high-average

Legenda. CI = Confidence Interval.
Note. [a] See tables in Orsini & Pezzuti, 2013b, for the conversion of the sum of the weighted scores of each scale in the corresponding Index.
[b] See Orsini & Pezzuti (2013, p. 20, table 2-5)

Table 2 – Assessment of the unitarity of Giovanni's indexes

Index	Max-Min	Max-Min difference	Unitary cut-off (16-69 years)[a]	Unitary Ability
Verbal Comprehension Index (VCI)	14 – 7	7	≥6	No
Perceptual Reasoning Index (PRI))	17 – 9	8	≥7	No
Working Memory Index (WMI)	13 – 7	6	≥5	No
Processing Speed Index (PSI)	12 – 13	1	≥5	Yes
Intellectual Quotient (IQ)	114 – 100	14	≥38	Yes

Note. [a] See Orsini, Pezzuti & Hulbert (2015), Pezzuti (2016) and Lang, Michelotti, Bardelli & Pezzuti (in press) for cut-off values.

Based on this finding, the clinician can formulate the following hypotheses: the subject has the prerequisites (i.e., the cognitive skills) to pass the master's degree examinations; his information processing speed - defined by the authors of the CHC model (Schneider & McGrew, 2012, 2018) as the average speed with which a subject completes a series of simple tasks in succession - is in fact medium/medium-high.

Processing speed is a construct that has been the subject of multiple discussions in the literature because there has been no agreement on its operationalization. For some authors it would be an index of complex attention, mental speed, reaction time, or inspection time, or even information processing time, etc. It is a construct, which is often confused with working memory and attention and consequently has been used interchangeably (Martin & Bush, 2008). We lean toward DeLuca's (2008, p. 266) definition that it is "the time required to perform a cognitive task or the amount of work that can be completed in a defined time frame".

What is of most interest in the clinic of this construct is some data that we list:

- in factor analyses for the study of cognitive abilities, mental speed of information processing has been identified as an important domain of cognitive functioning (Carroll, 1993; Horn & Noll, 1994, 1997; McGrew, 1997; Schneider & McGrew, 2012);
- there is evidence for connections between this construct and other cognitive constructs, such as working memory and fluid intelligence (Fry & Hale, 1996; Kyllonen & Christal, 1990), including the interaction between Baddeley's central executive and this construct (DeLuca, Barbieri-Berger et al., 1994);
- the fact that a slowdown in processing speed adversely affects verbal and visuospatial abilities (Sherman, Strauss et al., 1997), long-term episodic memory (DeLuca, Barbieri-Berger & Johnson, 1994; DeLuca, Gaudino, Diamond, Christodoulou & Engel, 1998; Gaudino, Chiaravalloti, DeLuca & Diamond, 2001), working memory, executive functions, problem-solving skills, and visuospatial skills and school skills such as reading and arithmetic (Chiaravalloti, Christodoulou, Demaree & DeLuca, 2003; Demaree, DeLuca, Gaudino & Diamond, 1999; Kennedy, Clement & Curtiss, 2003; Lengenfelder et al., 2006; Madigan, DeLuca, Diamond, Tramontano & Averill, 2000);
- mental speed correlates less with general intelligence than working memory and is the one that declines first with age as early as age 34 (Pezzuti, Lauriola, Borella, De Beni & Cornoldi, 2019).

The most recent research data only allow us to state that there is "some sort of global, biologically determined mechanism that limits the speed at which information is processed" (DeLuca, 2008, p. 272).

This information, although very important, in this context does not allow the clinician to formulate hypotheses because of the non-unitarity of the other indices.

The lack of unity of the other three indices forces the clinician to become aware of it and to "fall back" on a more idiographic reading. We use the term "fall back" because - as reported in literature - an idiographic reading has many limits.

Reading according to the 5-factor CHC model

If one can make use of the CHC model, the clinician can make a nomothetic evaluation of the data and formulate - based on the above literature data - some additional hypotheses regarding the patient's cognitive functioning; given the purpose of the article, we intentionally do not consider the links to emotional and personality variables.

As it is evident from the results reported in Table 3, because the criterion of unitarity of the broad CHC abilities is met, hypotheses can be made regarding Giovanni's cognitive functioning based on the broad CHC abilities.

The breadth and depth of the knowledge acquired by Giovanni with resp ect to his culture of belonging and the effective use of this knowledge, are partially adequate, given that the patient does not belong to a linguistic minority, has not had language problems in pre-school age and has a level of culturalization for which given the years of schooling should have acquired more knowledge. During the administration, moreover, the subject did not express any particular difficulty in dealing with the tasks proposed by the subtests, except for Vocabulary, where he stated that he had reduced lexical knowledge due to the fact of "being a bad reader" and to prefer video communication.

The score reported at Fluent Intelligence (Gf), even taking into account schooling is well 2 standard deviations above the mean. The authors of CHC and Lichtenberger and Kaufman (2009) operationalize Fluid Intelligence (Gf) as the ability

Table 3 – CHC model applied to the clinical case Giovanni

CHC broad ability	Subtest (WS[a])	Max-Min difference (ws[a])	Unitary cut-off (16-69 years)	Unitary Ability	Sum pp[a]	IQ	CI (95%)	PR[b]	Level[c]
Crystallized Intelligence (Gc)	Vocabulary (9) + Information (7)	2	≥5	Yes	16	89	83-96	27	average/low-average
Fluid Intelligence (Gf)	Matrix Reasoning (17) + *Figure Weights* (15)	2	≥5	Yes	32	135	125-140	98	high/very high
Visual Processing (Gv)	Block Design (9) + Visual Puzzle (9)	0	≥5	Yes	18	94	87-102	40	average/low-average
Short-Term Memory (Gsm)	Digit Span (13) + *Letter and Number Sequencing* (12)	1	≥5	Yes	25	114	105-121	84	average/high-average
Processing Speed (Gs)	Symbol Search (12) + Coding (13)	1	≥5	Yes	25	114	104-121	82	average/high-average

Legenda. CI = Confidence Interval.

Note. [a] Weighted Score; [b] Percentile Rank; [c] Orsini & Pezzuti (2013, p. 20, tables 2-5). The subtests in italics are the supplementary ones.

of inductive and deductive reasoning aimed at identifying common and different aspects, forming concepts, identifying general rules and applying rules to solve new problems. In other words, Giovanni is able to adequately and quickly solve new problems/situations, such as those posed to him by the two subtests, which propose tasks that cannot be performed automatically. Hence the need to be able to make inferences, identify the possible relationships that may exist between the different elements as well as formulate and verify the hypotheses formulated.

The question then arose as to what might be the possible causes of the current difficulties. There are two other clinically interesting pieces of data: Visual Processing (Gv), defined as the ability to create, store, retrieve and transform visual images (e.g., flipping or rotating shapes in space) shows a slight decline compared to other abilities and also considering his level of education. The level of performance in Short-Term Memory (Gsm), which detects the ability to grasp and maintain at a level of awareness information elements present in the current situation, is slightly above normal and 1 *SD* higher than the average performance of subjects of equal education. Giovanni is therefore able to activate cognitive resources to maintain information at a conscious level. This prevents the system, which has a limited capacity, from losing them quickly as they decay.

Another interesting fact is that having split the composite PRI index (which cannot be interpreted as a unitary ability) into two indexes according to the CHC model we also have an explanation for the non-unitarity of the PRI as it is due to a different performance of two distinct cognitive constructs (Gf and Gv), where performance is decidedly higher in Gf and poorer in Gv with a difference of about 41 IQ points.

CONCLUSIONS

The latest research regarding the assessment of scores on the WAIS-IV allows the clinician to make use of "new" scores that support him/her in the interpretation of the results achieved. However, we would like to focus attention on an aspect that we consider fundamental. In clinical practice, the interpretation of test results cannot be divorced from a context of assessment understood as the systematic process of forming and testing hypotheses to detect "the difficulties or failures [one encounters] in dealing with developmental problems and tasks" (Price & Zwolinski, 2010, p. 19). The purpose of an assessment process, therefore, is not to obtain a single score or even a series of test scores (testing), but to consider multiple pieces of information obtained altogether from testing and bio-psycho-social data collection and behavioral observations "in order to arrive at a coherent and comprehensive understanding of the person being assessed" (Bornstein, 2010, p. 147). The sole purpose of testing is "to provoke a phenomenon that is not seen so that it is revealed through its effects on behavior. The test must put the hypothetical construct into action in a way that causes observable outcomes" (Gottfredson & Saklofske, 2009, p. 187). Psychological testing, in fact, is "a process of data collection in which an individual's behaviors are taken as a sample and observed systematically in a standardized setting" (Zhu & Weiss, 2005, p. 310) and is only the beginning of psychological assessment.

The administration and reading of the results to the WAIS-IV occurs in the context of psychological testing, the goal of which is to obtain valid and reliable scores. The reading of test results is therefore one of the indispensable pre-requisites for the drafting of the psychodiagnostic report. Only afterwards the clinician integrates the interpretation of scores with other information coming from different sources (e.g. other instruments, clinical interviews, history, informants etc.) and are re-evaluated in order to understand the specificity of the single case. Only at this point can one speak of psychodiagnostic assessment and it is here that explanatory and intervention hypotheses are generated (Zhu & Weiss, 2005).

References

BENSON, N., HULAC, D.M. & KRANZLER, J.H. (2010). Independent examination of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22* (1), 121.

BENTLER, P.M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software (www.mvsoft.com).

BORNSTEIN, R.F. (2010). Psychoanalytic theory as a unifying framework for 21st century personality assessment. *Psychoanalytic Psychology*, *27* (2), 133-152.

CARROLL, J.B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York, NY: Cambridge University Press.

CATTELL, R.B. (1987). *Intelligence: Its structure, growth and action*. North-Holland, Oxford, England.

CATTELL, R.B. & HORN, J.L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement*, *15*, 139-164.

CHIARAVALLOTI, N.D., CHRISTODOULOU, C., DEMAREE, H.A. & DeLUCA, J. (2003). Differentiating simple versus complex processing speed: Influence on new learning and memory performance. *Journal of Clinical and Experimental Neuropsychology*, *25* (4), 489-501.

DeLUCA, J., BARBIERI-BERGER, S. & JOHNSON, S.K. (1994). The nature of memory impairments in multiple sclerosis: Acquisition versus retrieval. *Journal of Clinical and Experimental Neuropsychology*, *16* (2), 183-189.

DeLUCA, J., GAUDINO, E.A., DIAMOND, B.J., CHRISTODOULOU, C. & ENGEL, R.A. (1998). Acquisition and storage deficits in multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology*, *20* (3), 376-390.

DEMAREE, H.A., DeLUCA, J., GAUDINO, E.A. & DIAMOND, B.J. (1999). Speed of information processing as a key deficit in multiple sclerosis: implications for rehabilitation. *Journal of Neurology, Neurosurgery & Psychiatry*, *67* (5), 661-663.

FLANAGAN, D.P. & KAUFMAN, A.S. (2009). *Fondamenti per l'assessment con la WISC-IV.* It. ed. 2012, Firenze: Giunti OS Organizzazioni Speciali.

FLANAGAN, D.P., McGREW, K.S. & ORTIZ, S.O. (2000). T*he Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation.* New York, NY: Allyn & Bacon.

FLANAGAN, D.P., ORTIZ, S.O. & ALFONSO, V.C. (2013). *Essentials of Cross-Battery Assessment (3rd ed.).* Hoboken, NJ: John Wiley.

FRY, A.F. & HALE, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7* (4), 237-241.

GAUDINO, E.A., CHIARAVALLOTI, N.D., DeLUCA, J. & DIAMOND, B.J. (2001). A comparison of memory performance in relapsing–remitting, primary progressive and secondary progressive, multiple sclerosis. *Cognitive and Behavioral Neurology*, *14* (1), 32-44.

GOTTFREDSON, L. & SAKLOFSKE, D.H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/ Psychologie Canadienne*, *50* (3), 183.

HORN, J.L. (1985). Remodeling old models of intelligence. In B.B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications*. New York, NY: John Wiley.

HORN, J.L. (1991). Measurement of intellectual capabilities: A review of theory. In K.S. McGrew, J.K. Werder & R.W. Woodcock (Eds.), *Woodcock-Johnson Technical Manual.* Chicago, IL: Riverside Publishing.

HORN, J.L. & NOLL, J. (1994). A system for understanding cognitive capabilities: A theory and the evidence on which it is based. *Current Topics in Human Intelligence*, *4*, 151-203.

HORN, J.L. & NOLL, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York, NY: Guilford Press.

KAUFMAN, A.S., RAIFORD, S.E. & COALSON, D.L. (2016). *Intelligent testing with the WISC-V*. Hoboken, NJ.: John Wiley.

KENNEDY, J.E., CLEMENT, P.F. & CURTISS, G. (2003). WAIS-III processing speed index scores after TBI: The influence of working memory, psychomotor speed and perceptual processing. *The Clinical Neuropsychologist, 17* (3), 303-307.

KYLLONEN, P.C. & CHRISTAL, R.E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14* (4), 389-433.

LANG, M., MICHELOTTI, C., BARDELLI, E. & PEZZUTI, L. (in press). *WAIS-IV: Lettura dei risultati e interpretazione clinica (2a ed.)*. Milano: Raffaello Cortina Editore.

LENGENFELDER, J., BRYANT, D., DIAMOND, B.J., KALMAR, J.H., MOORE, N.B. & DeLUCA, J. (2006). Processing speed interacts with working memory efficiency in multiple sclerosis. *Archives of Clinical Neuropsychology*, *21* (3), 229-238.

LICHTENBERGER, E.O. & KAUFMAN, A.S. (2009). *Essentials of WAIS-IV assessment*. New York, NY: John Wiley.

MADIGAN, N.K., DeLUCA, J., DIAMOND, B.J., TRAMONTANO, G. & AVERILL, A. (2000). Speed of information processing in traumatic brain injury: Modality-specific factors. *The Journal of Head Trauma Rehabilitation*, *15* (3), 943-956.

MARTIN, T.A. & BUSH, S.S. (2008). Special Issue: Geriatric Neuropsychology. *NeuroRehabilitation*, *23* (5), 447-454.

McGREW, K.S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D.P. Flanagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York, NY: Guilford Press.

NEISSER, U., BOODOO, G., BOUCHARD, T.J., BOYKIN, A.W., BRODY, N., CECI, S.J., HALPERN, D.F., LOEHLIN, J.C., PERLOFF, R., STERNBERG, R.J. & URBINA, S. (1996). Intelligence: Knowns and unknowns. A*merican Psychologist*, *51* (2), 77-101.

ORSINI, A. & PEZZUTI, L. (2013). *WAIS-IV. Contributo alla taratura italiana (16-69)*. Firenze: Giunti OS Organizzazioni Speciali.

ORSINI, A. & PEZZUTI, L. (2015). *WAIS-IV. Contributo alla taratura italiana (70-90 anni*). Firenze: Giunti OS Organizzazioni Speciali.

ORSINI, A., PEZZUTI, L. & HULBERT, S. (2015). The unitary ability of IQ and indexes in WAIS-IV. *European Journal of Psychological Assessment*, 1-8. doi: 10.1027/1015- 5759/a000306

PEZZUTI, L. (2016). The GAI and CPI in the Italian standardization of the WAIS-IV and their clinical implications. *BPA-Applied Psychology Bulletin*, *276*, 19-37.

PEZZUTI, L., LANG, M., ROSSETTI, S. & MICHELOTTI, C. (2018). CHC model according to Weiss. *Journal of Individual Differences*, *39* (1), 53-59.

PEZZUTI, L., LAURIOLA, M., BORELLA, E., DE BENI, R. & CORNOLDI, C. (2019). Working Memory and Processing Speed mediate the effect of age on a General Ability Construct: Evidence from the Italian WAIS-IV standardization sample. *Personality and Individual Differences*, *138*, 298-304.

PEZZUTI, L. & ROSSETTI, S. (2017a). Letter-Number sequencing, Figure Weights and Cancellation subtests of WAIS-IV administered to elders. *Personality and Individual Differences*, *104*, 352-356.

PEZZUTI, L. & ROSSETTI, S. (2017b). Norms for Letter and Number sequencing, Figure Weights and Cancellation subtests for the elderly Italian population. *BPA-Applied Psychology Bulletin*, *279*, 15-21.

PRICE, J.M. & ZWOLINSKI, J. (2010). The nature of child and adolescent vulnerability: History and definitions. In R.E. Ingram & J.M. Price (Eds.), *Vulnerability to psychopathology: Risk across the lifespan (2nd ed.)*. New York, NY: The Guilford Press.

REYNOLDS, C.R. & KAMPHAUS, R.W. (2015). *The Reynolds Intellectual Assessment Scales, and the Reynolds Intellectual Screening Test*. Lutz, FL: Par.

RYAN, J.J., KREINER, D.S. & BURTON, D.B. (2002). Does high scatter affect the predictive validity of WAIS-III IQs? *Applied Neuropsychology*, *9*, 173-178.

RYAN, J.J., KREINER, D.S., GONTKOVSKY, S.T., GOLDEN, C.J. & MYERS-FABIAN, A. (2019). Frequency of occurrence of four and five-factor WAIS-IV profiles. *Applied Neuropsychology: Adult*, 1-11. Available at https://nsuworks.nova.edu/cps_facarticles/1627.

SCHNEIDER, W.J. & McGREW, K.S. (2012). The Cattell-Horn-Carroll model of intelligence. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues (3rd ed.)*. New York, NY: The Guilford Press.

SCHNEIDER, W.J. & McGREW, K.S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D.P. Flanagan & E.M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York, NY: The Guilford Press.

SHERMAN, E.M., STRAUSS, E. & SPELLACY, F. (1997). Validity of the Paced Auditory Serial Addition Test (PASAT) in adults referred for neuropsychological assessment after head injury. *The Clinical Neuropsychologist*, 11, 34-45.

WECHSLER, D. (2003). *WISC-IV. Wechsler Intelligence Scale for Children-Fourth Edition*. It. ed. A. Orsini & L. Pezzuti (Eds.), 2012. Firenze: Giunti OS Organizzazioni Speciali.

WECHSLER, D. (2008). *WAIS-IV. Wechsler Adult Intelligence Scale-Fourth Edition*. It. ed. A. Orsini & L. Pezzuti (Eds.), 2013. Firenze: Giunti OS Organizzazioni Speciali.

WEISS, L.G., CHEN, H., HARRIS, J.G., HOLDNACK, J.A. & SAKLOFSKE, D.H. (2010). WAIS-IV use in societal context. In *WAIS-IV clinical use and interpretation*. New York, NY: Academic Press.

WEISS, L.G., KEITH, T.Z., ZHU, J. & CHEN, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31* (2), 94-113.

WEISS, L.G., SAKLOFSKE, D.H., COALSON, D. & RAIFORD, S.E. (Eds.) (2010). *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives*. New York, NY: Academic Press.

WEISS, L.G., SAKLOFSKE, D.H., HOLDNACK, J.A. & PRIFITERA, A. (2016). *WISC-V: Advances in the assessment of intelligence*. San Diego, CA: Academic Press.

ZHU, J. & WEISS, L.G. (2005). The Wechsler Scales. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues*. New York, NY: The Guilford Press.