University of Milano Bicocca
Phd program in Statistics and Mathematical Finance

ASSESSING PSYCHOMETRIC SCALES THROUGH IRT-BASED MODELLING WITH APPLICATIONS TO COVID-19 DATA

Supervisor: Prof.ssa Francesca GRESELIN
Co-supervisor: Prof. Matteo BONZINI

Coordinator: Prof.ssa Emanuela ROSAZZA GIANIN

Author:
Anna COMOTTI
Registration number: 827395

Academic Year 2020 - 2021

A mio papà

Ringraziamenti

L'unica gioia al mondo è cominciare.
È bello vivere, perché vivere è cominciare, sempre, ad ogni instante.
(Cesare Pavese)

Non ho mai concluso una tesi con dei ringraziamenti, nonostante la gratitudine fosse sempre stata molta, un po' perché non sono brava ad esprimere a parole certi sentimenti e un po' perché ritengo che in un grazie detto a voce ci sia molto di più. Mi è però così evidente che non avrei mai ricominciato né concluso questo percorso senza l'aiuto di certe persone, che desidero accompagnare queste pagine con un ringraziamento rivolto a loro. Niente è andato come avevo immaginato, ma alla fine tutto è stato prezioso, in primis la loro presenza.

Ringrazio di cuore il prof. Matteo Bonzini, per avermi dato l'opportunità di (ri)cominciare questo lavoro con immeritate fiducia e stima e per la cura e la discrezione con cui mi ha guidato a concluderlo. Allo stesso modo ringrazio la prof.ssa Francesca Greselin, per averci messo l'entusiasmo che non avevo, per non essersi mai stancata di incoraggiarmi, per tutto il tempo che mi ha dedicato e per l'enorme aiuto che mi ha dato.

Sono riconoscente al Collegio dei Docenti di Dottorato e in particolare al coordinatore, prof.ssa Emanuela Rosazza Gianin, per il supporto e la comprensione mai venuti meno.

Grazie a tutti i miei amici, senza dei quali non sarei riuscita - nè riuscirei - a vedere il lato buono delle cose e il cui affetto è stato un grande conforto. Tra loro, in modo particolare, ringrazio i tanti che vivono il loro lavoro nel mondo della ricerca, pur in ambiti diversi, come curiosa apertura e mai paga conoscenza di se stessi e del mondo; è anche grazie alla loro sana passione che ho deciso di riscommettere su questa strada. Ad Andrea e Gregorio rivolgo un grazie speciale perché, anche se non lo sanno, in certi momenti il confronto con loro è stato per me davvero importante.

Ringrazio anche i miei genitori per il loro sostegno e dedico questa tesi a mio padre Fausto, che mi ha mostrato la bellezza di poter ricominciare.

Contents

Preface

This thesis stems from a wider project - *PostCovid: a systematic evaluation of health-care workers' psychological well-being before and after the COVID-19 pandemic* - carried out at Fondazione IRCCS Ca' Grande Ospedale Maggiore Policlinico (Milan) by the Dipartimento di Medicina del Lavoro, along with the Dipartimento di Psichiatria. Said program, which started in 2020, is still ongoing. PostCovid study design is observational and longitudinal, non-pharmacological and entails a multi-step evaluation of mental health in all workers of Ospedale Policlinico, in order to: (i) evaluate with standardized tests the psychological well-being of the participants with a structured medical-assisted interview in the context of occupational health surveillance; (ii) draw up further questionnaires to better assess possible psychological distress for those subjects who express signs of distress in the first interview; (iii) offer a specialist evaluation to whomever showed specific symptoms at the second-level questionnaire, followed, when needed, by an individual psychological support and/or psychiatric treatment; (iv) follow up workers over an extended period of time, with a re-evaluation after 6 months for workers with sub-optimal psychological well-being at first level, and after 12 months for others, so as to evaluate trends in psychological burden, recognize delayed onset of symptoms and evaluate the efficacy of specialist treatments.

The whole study design, the methodology and preliminary results on a subsample referred to subject enrolled during the first six months are published and available in the manuscript Fattori et al. (2021).

This thesis is the result of a series of clinical questions arising from a specific medical environment, the same medical environment where our work experience took place. In

order to find appropriate answers, it was necessary to rely on - and deepen - the methodological topics which would have enabled a better understanding of the outcomes.

The structure of our dissertation is as follows.

Chapter 1 introduces the main topics of the above-mentioned study, presents the sample of 990 subjects gathered throughout the enrollment period (July 2020 - July 2021), describes the involved subjects and all the observed variables, hints at the contents of the psychometric scales, and contains a statistical analysis of the risk factors for the psychological impairment. Said analysis - the first to be carried out on the gathered data - focused on one of the main issues tackled by occupational medicine, i.e. identifying the factors that might impair one's work ability, assessing them and implementing all the necessary improvement and prevention strategies. The first phase, where we pinpointed all the possible associations between factors and outcomes of the psychometric scales, was followed by an actual risk factor analysis through multiple logistic regression. The procedure was repeated three times, one for each of the questionnaires involved in the study and dedicated to psychological distress, post-traumatic symptoms and anxiety. Moreover, the same procedure was applied to the patients (363) taking part in the second-level screening in order to point out possible risk factors for depression, dissociation and psychiatric conditions.

In our study, an established set of questionnaires played a crucial role in the assessment of the interviewees' well-being. The core of the following dissertation is therefore an analysis of said scales as an essential measurement instrument, relying on psychometric models and expressing a relation between the answers provided and the construct to be measured. Our objective was, on the one hand, exploring the clinical implications of the scales' outcomes; and on the other hand, trying to fill the gaps that are still present in the literature. Therefore, Chapter 2, after a brief introduction about general properties and psychometric characteristics of questionnaires (focusing on validity and reliability), presents the psychometric scales chosen for the assessment of psychological well-being within the study. For each of them, the main characteristics are described, their origin and the major contributions present in the literature are summarized, with a particular

focus on dimensionality issue: each scale, in fact, has the purpose of measuring a particular characteristic and it is therefore assumed that the elements that compose it measure the same trait. This is not always true and by investigating the number of latent factors, the multidimensionality of the scale can instead emerge.

Classical Test Theory (CTT) and Item Response Theory (IRT) represent two different measurement frameworks in questionnaires assessment. The unit of analysis in CTT is usually the total score, often a sum score, of the person's responses to the set of items. IRT is considered one of the several modern test theory methods, and, as the name implies, the unit of analysis is the individual item response. While CTT methods are easy to use and computationally efficient, CTT has several limitations including its test-level approach and scores dependent upon the test items and sample. IRT may provide a more informative and thorough evaluation of the items and the person's latent trait. IRT alleviates several limitations of CTT. IRT theory has been extensively analyzed and studied from a methodological point of view, and its application in psychometry has recently increased. Nonetheless, the number of contributions where said theory was applied to the above-mentioned psychometric scales are limited.

In Chapter 3 we present the most used models in the IRT approach, both for binary and polytomous items, illustrates the item parameters and presents the general idea behind IRT. The recognition of the complexity of underlying traits in psychological testing and the limitations of a single trait, or unidimensional modelling, have led to extensions to account for the multidimensionality of response data. In Chapter 3 the multidimensional version of IRT is reported too, named Multidimensional Item Response Theory (MIRT). MIRT theory has great potential for solving many problems in psychological assessment. Although many researchers believe that psychological tests measure multiple constructs, MIRT modeling is still not prevalent. MIRT analysis can help to provide that test scores are being properly used and interpreted. If the results of a test are reported as a single score, it is implicitly assumed that all the items are measuring the same skill or same composite of skills. Dimensionality analyses can help establish the degree to which this is true.

Chapter 4 is dedicated to one of above mentioned scales - GHQ-12 (General Health

Questionnaire-12) - and to its assessment with the IRT. There are several reasons why this scale has been chosen over the others: firstly, it's one of the most widespread tools used in the psychological screening of patients and the related literature is particularly wide; secondly, the number of items it consists of (i.e., 12 items) made it more suitable for subsequent assessment; and lastly, following the risk factor analysis described in Chapter 1 this scale was the best to screen the psychological status of the interviewees.

The IRT enabled us to delve in three areas of interest of GHQ-12. The first issue was the dimensionality of this scale. Literature offers a wide variety of interpretations, which reflect the complexity of the tool. The vast majority of them resorts to factor analysis, though, both exploratory and confirmatory, while the application of IRT on GHQ-12 in order to assess its dimensionality is scarce. In this study we focused on the outcomes of both approaches, and we outlined the main similarities and differences.

The second issue was the *predictive* value of this scale on the second level assessment, an aspect that falls slightly more in the medical field. In order to investigate this topic, we used a version of IRT models with discrete distribution (latent classes), which allowed us to classify patients according to how they answered the twelve questions and to analyze the second-level assessment. Said analysis, with its predictive value, shows how the multi-step evaluation of the patients - lengthy and quite articulated - could be simplified by enhancing the role of GHQ-12.

The third and final issue was the difference between the outcomes of the healthcare workers who were directly and marginally involved in COVID-19 units. As a matter of fact, it needs to be borne in mind that our study population was made up of professionals dealing on different levels with a pandemic. It was therefore sensible to analyze more thoroughly the divergence in their psychological status. One of the several ways this could be done is to resort to IRT in order to detect the so-called DIF (differential item functioning), a tool which allows to identify the items with a differential behavior among groups. Different answers to the same items exemplify further the multi-dimensionality of GHQ-12 and give another approach to the issue of dimensionality.

In Chapter 5 we introduce the first steps of our new research within the robust statistic framework and implement a forward search algorithm for identifying atypical subjects/ob-

servations in IRT models for binary data (Rasch models).

This manuscript ends with some conclusions and future research directions.

Chapter 1

PostCovid study

Aim of this chapter is to present the starting point of this thesis. Our scientific journey was generated, owes its origin, as well as its development, toward the the present complete form to the project conducted in Ospedale Policlinico. The COVID-19 pandemic has been (and continues to be) a severe challenge for health-care workers, with a considerable impact on their mental health. Therefore, in 2020, Ospedale Policlinico started an observational study (for brevity named *PostCovid*) with the purpose of evaluating the psychological well being of the health care workers, to plan and propose a support when needed.

Section 1.1 briefly describes the study. Section 1.2.1 illustrates the whole protocol and instruments used for data collection, with a preliminary description of the psychometric tests used for the screening.

Section 1.2.2 presents sample characteristics and results of the evaluation of factors associated with mental distress, mainly performed through multiple logistic regressions and commented in Section 1.2.3.

Section 1.2.3 introduces the general ideas and open questions which have generated the main presented in the next chapters of this thesis.

1.1 Introduction

Italy was the first western country to be affected by the COVID-19 pandemic since February 2020, when the exponential rise of cases required a national lockdown and imposed an extraordinary amount of work on the healthcare system dramatically increasing in terms of critical care and reorganization.

Under such circumstances, health care workers (HCW) experienced heavy workload, physical exhaustion, frustration and helplessness and also fear of infecting themselves and their relatives (Boccia et al., 2020). Thus, besides physical safety, HCW's mental health was a major concern for authorities and occupational physician, because studies among HCW during previous epidemics (SARS, MERS, Ebola Maunder, 2004; de Pablo et al., 2020) and primary studies conducted in China at the very beginning of the COVID-19 pandemic, showed high prevalence of post-traumatic stress disorder (PTSD), depression and anxiety disorders (Liu et al., 2020a, Dai et al., 2020) in HCW. From the very beginning of the COVID-19 pandemic several studies have been conducted on HCW mental health and confirmed that a considerable proportion of workers developed adverse psychological outcomes during the COVID-19 pandemic (Pappa et al., 2020, d'Ettorre et al., 2021). Furthermore, these studies found that being frontline workers, female sex, younger age, lower job seniority and nursing profession predicted worsened mental health (Hao et al., 2021, De Kock et al., 2021).

Most of these studies focused on critical care workers and collected data through web-based questionnaires, being able to collect only a proportion of workers in several cases. Thus, results could be partially affected by self-selection of respondents and the comparison of mental health outcomes between more exposed workers and other colleagues are limited. Another common limitation is the lack of information on non-occupational important risk factors (such as COVID-19 infection in the family). Beside the fact that HCW of intensive care units (ICUs) faced a large number of COVID-19 deaths and substantial work-related stress, all health care professionals were also exposed to personal grief and family concerns (Rabow et al., 2021). Occupational exposure was often collected within questionnaires possibly influenced by workers' perception and mental well-being. Finally, since most of the published studies were conducted during the first phases of the

pandemic, results are focused on the early onset symptoms with little evidence on the persistence of symptoms and delayed-onset PTSD, which typically occurs a few months after exposure.

This is why, even in the current pandemic scenario, it is crucial to evaluate and monitor the mental health of HCW during different phases and waves of the COVID-19 pandemic. The primary purpose are: to prevent possible mental disorders, discover work-related and individual risk factors that can exacerbate psychological distress and target rehabilitation strategies on more vulnerable people. For these reasons Occupational Medicine Department designed a prospective study that systematically evaluate mental well-being of all workers employed in Ospedale Policlinico. HCW were followed by the occupational physician health surveillance, using a multistep approach to assess psychological workload and symptoms with validated scales. The study period covered almost one year. It has been characterized by two waves of epidemic and by a massive and rapid campaign of HW vaccination that occurred in January-February 2021 in our region (Lombardy).

1.2 The study

1.2.1 Protocol and instruments

An extensive description of the methodology adopted for this study was illustrated in a previous report (Fattori et al., 2021). A multi-step process has been designed to evaluate the mental health of all workers employed in our hospital through three different levels:

1. first level: to assess psychological well-being with standardized scales during a structured medical-assisted interview in the context of occupational health surveillance;

2. second level: when first-level scales show psychological impairment, workers are invited to undergo a second-level questionnaire to assess possible psychological distress better;

3. third level: to offer a specialist evaluation, psychological support and/or psychiatric treatment to workers showing specific symptoms at the second-level questionnaire.

A follow-up re-evaluation within 12 months since enrollment to all participant is also planned, in order to evaluate trends in psychological burden, recognize possibly delayed onset of symptoms, and assess the efficacy of specialist treatments.

The study is conducted jointly by the Occupational Medicine and Psychiatry Units. From July 2020 on wards, all workers have been invited to participate, independently from age, sex, department and job title. The only two exclusion criteria were: being employed after the beginning of the study and refusing to sign the informed consent. There were no exclusion criteria on pre-existing pathologies, to include the overall and most general population. The workers sign on an extended informed-consent form before the first-level evaluation. Formal ethical approval has been obtained by the Hospital ethical committee in July 2020.

The first level evaluation began with by an occupational physician interview collecting socio-demographic questions (age, sex), occupational data (occupational role, working experience in Covid-19 areas, with specific details on intensity and time spent) and clinical questions regarding chronic conditions and habitual medications, with a distinction for those taken after the onset of the pandemic. After the interview, a psychometric questionnaire, collected directly on digital support, has been administered. It consisted of:

1. The General Health Questionnaire (GHQ-12) (Goldberg, 1986) in the validated Italian version for assessing psychological distress and short-term changes in mental health. We adopted the dichotomous scoring method (0-0-1-1) and a score above or equal to 4 as the cut-off point;

2. Impact of Event Scale-Revised (IES-r) for assessing post-traumatic stress symptoms. A brief description guides subjects to answer the following questions by assessing their subjective responses related to COVID-19 emergency in the previous 7 days with 22 questions exploring intrusion, avoidance, and hyperarousal symptoms.

3. Generalized Anxiety Disorders (GAD-7) to screen anxiety symptoms. With robust psychometric properties and strong validity, a score of 10 or greater represents a reasonable cut-off point to identify cases of GAD; increasing scores on the GAD-7

are also strongly associated with multiple domains of functional impairment and disability.

4. A section collecting individual COVID-19 exposure and COVID-related health concerns/believes: having been infected to COVID-19 and duration of the condition, having been in quarantine and duration, having family members infected/hospitalized/deceased for COVID-19, personal concern for infecting family members, having experienced social discrimination outside the hospital, changes in family's habits, thoughts about changing job, fear for their own safety, experience of moral injury at work.

The second-level questionnaire contains specific scales to further investigate psychopathological symptoms and disorders

1. Symptom Checklist-90-Revised (SCL-90-R) is a self-administered scale for the evaluation of psychiatric symptomatology;

2. The Dissociative Experiences Scale II (DES II). Dissociative symptoms are frequently found in the aftermath of trauma. They occurs to some degree in individuals without mental disorders and are thought to be more prevalent in persons with major mental illnesses. The DES II has been developed to offer a means of reliably measuring dissociation in normal and clinical populations;

3. Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 is aimed at assessing depression disorder.

A specialist psychiatric feedback of second-level evaluation results is sent to the occupational physician who, if tests indicate an impairment in psychological functioning, proposes to the worker a specialist consultation in person. The third-level evaluation comprised by the specialist consultation within one week from the second-level evaluation- It is followed, according to every single case, by an eventual psychiatric follow-up or psychotherapy.

All subjects repeat the tests after no more than 12 months to individuate late signs and to assess individual changes in psychological distress.

1.2.2 Statistical analysis

Data are collected through an automatic database generated by REDCap platform. An independent coded dataset accessible only to the PI guarantees data protection linking individual information (i.e. name and surname) with an alphanumeric code.

Statistical analysis is aimed to individuate risk factors for sub-optimal psychological wellbeing and/or impaired psychological function. First, the relationship between each potential risk factor and the outcomes, treated as continuous variables, was preliminarily investigated in terms of mean differences across subgroups through independent samples t-test and one-way analysis of variance (ANOVA). Comparison in the percentage of subjects with a total score higher than the cutoff for each scale was evaluated through Chi-square test. Secondly, each potential risk factor is included in multiple logistic regression models to explore the relative contributions (in terms of Odds Ratios-OR) of the various risk factors to the dependent variables, including potential covariates and confounders. The overall significance of each variable was tested through Likelihood-Ratio (LR) test. The relationship between personal concerns and feelings about COVID-19, collected through six questions with multiple answers (not at all, little, enough, very), and first level outcome variables was graphically explored and the differences in the distribution was investigated through the Kolmogorov-Smirnov test for discrete variables. To study their effect on first-level scores in terms of risk factors, they have been converted into dichotomous variables (Yes = not at all, little, No = enough, very) and put one by one in the multiple logistic regression model. The effect of vaccination on psychological scales has been investigated by exploring differences between workers enrolled before and after the COVID-19 vaccination campaign, which started in January 2020. To study how the effect of risk factors, in particular of the variables related to COVID-19 exposure, varied after the vaccination, we performed logistic regressions on first-level screening dividing the dataset into two sub-samples ($N = 584$ and $N = 406$, before and after the vaccination campaign, respectively). The significance of the relationship between these variables and vaccination was evaluated by including an interaction term in the multiple logistic regression model on the whole dataset, using a binary variable indicating enrollment before or after the vaccination campaign.

The Occupational Medicine unit, where workers underwent the periodical health surveillance already prescribed by the current Italian legislation, proposed the study protocol to all workers since July 2020. Up to July 2021, we enrolled a total of 990 subjects. Participation rate was 86%. In detail, 83 (8%) workers did not answer our calls or were unavailable, 70 (6%) refused to participate. Table 1.1 summarizes main characteristics of enrolled subjects and results of the first-level questionnaires. The percentage of subjects scoring above the cutoff of the first-level scales widely differed by sex, age, occupational role, COVID-19 exposure at work and in their own family. No significant differences were found when considering the two groups of subject with or without a previous COVID-19 infection (stated by a positive swab). Similar results were found considering average values in each psychometric scale instead of cut-offs.

Table 1.3 presents logistic regression analysis for first-level screening scales. Adjusted OR showed that sex, occupational role, working experience with COVID-19 patients and having a family member with previous COVID-19 infection were risk factors for psychological impairment. Women had increased risk of developing anxiety symptoms by around 70% (see GAD-7 scale). Being a nurse almost triplicated the risk for developing symptoms of post-traumatic distress (see IES-R scale), it almost doubled the risk of anxiety (GAD-7) and increased by 41% the risk of general discomfort (GHQ-12). Direct experience with COVID-19 patients was associated with an increased risk of psychological impairment in all the three scales. In detail, risk to score above the cut-off (for all measured scales) increased with increasing time spent in the COVID-19 area, with a higher level of clinical intensity, or dividing subjects with none, former, or current involvement in COVID-19 units. Subjects with a family member previously infected by COVID-19 showed an OR for GHQ-12 score above the cut-off equal to 1.48; age was not found as a significant risk factor for psychological impairment.

Table 1.2 shows the analysis for the second-level scales, collected among 316 subjects. Similar to firs-level screening, sex and occupational role resulted as statistically significant factors associated with psychological distress: means and percentage of scoring above the cutoff were higher for females, nurses and health assistants (although the latter are composed by few cases). However, contrary to first-level outcomes, working exposure to COVID-19 and having a family member with previous COVID-19 infection were not

		GHQ-12		IES-R		GAD-7	
	N (%)	**Mean (sd)**	**N(%) >cutoff**	**Mean (sd)**	**N(%)>cutoff**	**Mean (sd)**	**N(%)>cutoff**
Sex							
Male	297 (30%)	2.79 (3.07)	96 (32%)	16.2 (15.3)	46 (16%)	4.58 (4.43)	44 (15%)
Female	693 (70%)	3.27 (3.32)	270 (39%)	20.5 (17.0)	146 (21%)	6.38 (5.30)	161 (23%)
p value		0.03*	0.06***	<0.001*	0.05***	<0.001*	0.003***
Age group							
20-30	137 (14%)	3.73 (3.54)	62 (45%)	20.6 (16.5)	30 (22%)	6.55 (4.93)	33 (24%)
30-40	276 (28%)	3.21 (3.17)	110 (40%)	19.3 (15.5)	55 (20%)	5.92 (4.84)	56 (20%)
40-50	245 (24.5%)	3.27 (3.43)	90 (37%)	19.9 (18.6)	53 (22%)	6.13 (5.60)	60 (25%)
>50	332 (33.5%)	2.72 (3.02)	104 (31%)	17.9 (16.0)	54 (16%)	5.27 (5.02)	56 (17%)
p value		0.01**	0.02***	0.35**	0.32***	0.06**	0.17***
Occupational role							
Administrative staff	119 (12%)	2.44 (2.83)	34 (29%)	16.8 (14.3)	14 (12%)	5.32 (4.92)	20 (17%)
Health assistant	63 (6.5%)	2.67 (3.45)	17 (27%)	23.1 (18.2)	15 (24%)	5.98 (5.23)	17 (27%)
Nursing staff	416 (42%)	3.79 (3.52)	188 (45%)	23.0 (18.4)	115 (28%)	6.71 (5.52)	111 (27%)
Physician	233 (23.5%)	2.81 (2.89)	80 (34%)	15.0 (13.6)	27 (12%)	4.96 (4.49)	34 (15%)
Others	159 (16%)	2.55 (2.97)	47 (29%)	15.6 (14.0)	21 (13%)	5.20 (4.68)	23 (14%)
p value		<0.001**	<0.001***	<0.001**	<0.001***	<0.001**	<0.001***
COVID-19 area working experience							
Never	544 (55%)	2.54 (2.92)	160 (29%)	16.7 (14.3)	72 (13%)	5.27 (4.79)	90 (17%)
Yes							
previously	202 (20%)	3.63 (3.47)	86 (43%)	21.5 (17.9)	48 (24%)	6.04 (5.25)	46 (23%)
currently	244 (25%)	4.01 (3.52)	120 (49%)	23.9 (18.6)	72 (30%)	7.04 (5.49)	69 (28%)
p value		<0.001**	<0.001***	<0.001**	<0.001***	<0.001**	<0.001***
<130 days	227 (23%)	3.93 (3.54)	107 (47%)	22.7 (18.6)	58 (26%)	6.38 (5.44)	54 (24%)
>130 days	219 (22%)	3.74 (3.45)	99 (45%)	23.1 (18.1)	62 (28%)	6.81 (5.38)	61 (28%)
p value		<0.001**	<0.001***	<0.001**	<0.001***	<0.001**	<0.001***
high-intensity area	345 (35%)	4.01 (3.51)	169 (49%)	24.1 (19.0)	101 (29%)	6.85 (5.52)	94 (27%)
low-intensity area	101 (10%)	3.26 (3.41)	37 (37%)	18.6 (15.2)	19 (19%)	5.70 (4.94)	21 (21%)
p value		<0.001**	<0.001***	<0.001**	<0.001***	<0.001**	<0.001***
Positive nasoph. swab							
Yes	153 (15%)	3.15 (3.40)	55 (36%)	18.9 (16.2)	31 (20%)	5.89 (4.84)	28 (18%)
No	837 (85%)	3.13 (3.23)	311 (37%)	19.3 (16.7)	161 (19%)	5.83 (5.17)	177 (21%)
p value		0.93*	0.83***	0.83*	0.85***	0.87*	0.48***
Family member positive to Covid-19							
Yes	209 (21%)	3.43 (3.15)	89 (43%)	19.1 (15.6)	44 (21%)	6.04 (4.72)	45 (22%)
No	781 (79%)	3.16 (3.29)	277 (36%)	19.3 (16.9)	148 (19%)	5.79 (5.22)	160 (21%)
p value		0.30*	0.07***	0.86*	0.56***	0.55*	0.9***

Table 1.1: First level screening scales across subgroups: Number of enrolled subjects, means, standard deviations and frequencies of scorings above the cutoff at the different first level psychometric scales. (*t-test, **One-way ANOVA, ***Chi-square test)

associated with higher psychological scales scoring.

Table 1.4 presents logistic regression analysis for psychological distress (second-level questionnaire results). Nurses and health assistants had sensibly higher adjusted OR

		PHQ-9		DES		SCL-90	
	N (%)	Mean (sd)	N(%) >cutoff	Mean (sd)	N(%) >cutoff	Mean (sd)	N(%) >cutoff
Sex							
Male	81 (26%)	8.63 (4.79)	22 (27%)	9.94 (10.4)	12 (15%)	0.66 (0.48)	17 (21%)
Female	235 (74%)	9.54 (5.44)	88 (37%)	13.2 (13.7)	50 (21%)	0.84 (0.64)	73 (31%)
p value		0.16*	0.12***	0.03*	0.27***	0.01*	0.12***
Age group							
20-30	57 (18%)	9.11 (5.15)	15 (26%)	11.3 (9.29)	9 (16%)	0.73 (0.57)	13 (23%)
30-40	91 (29%)	8.95 (5.29)	27 (30%)	14.4 (14.2)	27 (30%)	0.78 (0.61)	29 (32%)
40-50	81 (25.5%)	9.98 (5.49)	34 (42%)	11.3 (13.1)	13 (16%)	0.84 (0.60)	25 (31%)
>50	87 (27.5%)	9.18 (5.21)	34 (39%)	11.9 (13.8)	13 (15%)	0.81 (0.63)	23 (26%)
p value		0.59**	0.14***	0.37**	0.04***	0.72**	0.62***
Occupational role							
Administrative staff	27 (8%)	8.44 (5.01)	9 (33%)	14.4 (17.5)	6 (22%)	0.86 (0.71)	10 (38%)
Health assistant	16 (5%)	12.2 (4.62)	11 (69%)	21.3 (19.1)	8 (50%)	1.27 (0.83)	9 (56%)
Nursing staff	173 (55%)	10.3 (5.43)	64 (37%)	14.1 (13.3)	43 (25%)	0.86 (0.60)	55 (32%)
Physician	62 (20%)	7.34 (4.58)	13 (21%)	6.48 (6.29)	2 (3%)	0.54 (0.32)	6 (10%)
Others	38 (12%)	7.18 (4.46)	13 (34%)	8.92 (9.17)	3 (8%)	0.67 (0.61)	10 (26%)
p value		<0.001**	0.008***	<0.001**	<0.001***	<0.001**	<0.001***
COVID-19 area working experience							
Never	138 (44%)	8.65 (4.92)	47 (34%)	12.2 (12.8)	3 (17%)	0.80 (0.62)	41 (30%)
Yes							
currently	114 (36%)	9.64 (5.56)	37 (32%)	12.8 (13.9)	23 (20%)	0.80 (0.59)	34 (30%)
previously	64 (20%)	10.1 (5.47)	26 (41%)	12.0 (11.9)	16 (25%)	0.77 (0.60)	15 (24%)
p value		0.13**	0.53***	0.89**	0.37***	0.95**	0.63***
>130 days	96 (30%)	9.65 (5.55)	21 (33%)	11.6 (12.5)	17 (18%)	0.81 (0.58)	30 (31%)
<130 days	82 (26%)	10.0 (5.50)	31 (38%)	13.5 (14.1)	22 (27%)	0.77 (0.60)	19 (23%)
p value		0.14**	0.79***	0.61**	0.15***	0.93**	0.47***
high-intensity area	148 (47%)	9.94 (5.54)	53 (36%)	12.4 (13.1)	31 (21%)	0.79 (0.57)	40 (27%)
low-intensity area	30 (9%)	9.20 (5.46)	10 (33%)	12.9 (14.2)	8 (27%)	0.80 (0.69)	9 (30%)
p value		0.12**	0.93***	0.95**	0.39***	0.97**	0.86***
Positive nasopharyngeal swab							
Yes	51 (16%)	9.69 (5.08)	18 (35%)	13.4 (15.7)	8 (16%)	0.76 (0.56)	12 (24%)
No	265 (84%)	9.23 (5.33)	92 (35%)	12.1 (12.5)	54 (20%)	0.80 (0.61)	78 (29%)
p value		0.56*	0.9***	0.58*	0.56***	0.63*	0.47***
Family member positive to COVID-19							
Yes	76 (24%)	9.14 (4.72)	23 (30%)	10.9 (9.91)	11 (15%)	0.74 (0.51)	16 (21%)
No	240 (76%)	9.36 (5.46)	87 (36%)	12.8 (13.9)	51 (22%)	0.81 (0.63)	74 (31%)
p value		0.74*	0.41***	0.19*	0.25***	0.32*	0.12***

Table 1.2: Second level screening scales (N=316): means, standard deviations and frequencies of scorings above the cutoff across subgroups. (*t-test, **One-way ANOVA, ***Chi-square test)

for developing depression or other psychological symptoms than physicians. Namely, ORs were greater in women considering all the three scales (even if not statistically significant). Similar to previous analysis, the occupational exposure with COVID-19 seemed not to be

	N (%)	GHQ-12 AdjOR (95% CI)	IES-R AdjOR (95% CI)	GAD-7 AdjOR (95% CI)
Sex				
Male	297 (30%)	1.00	1.00	1.00
Female	693 (70%)	1.37 (1.01, 1.85)	1.44 (0.99, 2.13)	1.72 (1.19, 2.54)
p value		0.04	0.06	0.003
Age				
>50	137 (14%)	1.00	1.00	1.00
20-30	276 (28%)	1.12 (0.72, 1.76)	0.69 (0.39, 1.20)	1.02 (0.59, 1.72)
30-40	244 (24.5%)	1.05 (0.73, 1.51)	0.79 (0.50, 1.24)	0.96 (0.61, 1.49)
40-50	233 (33.5%)	1.05 (0.73, 1.46)	1.06 (0.68, 1.66)	1.35 (0.88, 2.07)
p value		0.03	0.31	0.17
Occupational role				
Physician	233 (23.5%)	1.00	1.00	1.00
Administrative staff	119 (12%)	1.07 (0.63, 1.80)	1.58 (0.74, 3.27)	1.44 (0.75, 2.75)
Health assistant	63 (6.5%)	0.66 (0.34, 1.22)	2.27 (1.09, 4.61)	2.07 (1.04, 4.05)
Nursing staff	416 (42%)	1.41 (1.00, 2.01)	2.90 (1.82, 4.73)	1.95 (1.26, 3.06)
Others	159 (16%)	0.99 (0.62, 1.56)	1.60 (0.84, 3.05)	1.14 (0.75, 2.75)
p value		0.003	<0.001	0.007
COVID-19 area working experience				
Never	544 (55%)	1.00	1.00	1.00
Yes				
previously	202 (20%)	2.27 (1.59, 3.25)	2.80 (1.82, 4.34)	1.96 (1.29, 2.96)
currently	244 (25%)	1.75 (1.20, 2.52)	2.08 (1.31, 3.29)	1.43 (0.91, 2.22)
p value		<0.001	<0.001	0.007
<130 days	227 (23%)	1.95 (1.35, 2.82)	2.66 (1.71, 4.15)	1.93 (1.26, 2.96)
>130 days	219 (22%)	2.07 (1.44, 2.97)	2.26 (1.45, 3.54)	1.49 (0.97, 2.29)
p value		<0.001	<0.001	0.009
high-intensity area	345 (35%)	2.22 (1.61, 3.09)	2.69 (1.81, 4.05)	1.80 (1.23, 2.66)
low-intensity area	101 (10%)	1.41 (0.87, 2.28)	1.67 (0.90, 3.03)	1.35 (0.75, 2.37)
p value		<0.001	<0.001	0.009
Positive nasoph. swab				
Yes	153 (15%)	1.00	1.00	1.00
No	837 (85%)	0.78 (0.53, 1.15)	0.94 (0.58, 1.48)	0.73 (0.45, 1.16)
p value		0.55	0.98	0.21
Family member positive to Covid-19				
Yes	209 (21%)	1.00	1.00	1.00
No	781 (79%)	1.48 (1.05, 2.08)	1.17 (0.77, 1.76)	1.11 (0.74, 1.65)
p value		0.02	0.64	0.61

Table 1.3: Multiple logistic regression for f first level screening scales: Adjusted OR for scoring above the cut-offs with associated 95% confidence intervals and corresponding LR test p-values.

an independent risk factor for psychological distress.

Figure 1.1 illustrates the distribution of health beliefs and COVID-19 concerns for each answer, which significantly differed according to first level screening result (Kolmogorov-Smirnov test). Worries, discomfort and fear were expressed more frequently by subjects who scored above the cut-off in at least one scale compared to colleagues with no evidence

	N (%)	PHQ-9 AdjOR (95% CI)	DES AdjOR (95% CI)	SCL-90 AdjOR (95% CI)
Sex				
Male	81 (26%)	1.00	1.00	1.00
Female	235 (74%)	1.40 (0.77, 2.60)	1.68 (0.80, 3.79)	1.48 (0.78, 2.94)
p value		0.14	0.10	0.11
Age				
>50	87 (27.5%)	1.00	1.00	1.00
20-30	57 (18%)	0.39 (0.17, 0.88)	0.78 (0.27, 2.21)	0.68 (0.328 1.64)
30-40	91 (29%)	0.44 (0.21, 0.91)	1.71 (0.72, 4.17)	1.05 (0.50, 2.24)
40-50	81 (25.5%)	0.93 (0.48, 1.79)	0.88 (0.35, 2.20)	1.19 (0.58, 2.48)
p value		0.14	0.05	0.61
Occupational role				
Physician	62 (20%)	1.00	1.00	1.00
Administrative staff	27 (8%)	2.12 (0.70, 6.40)	8.23 (1.61, 62.65)	5.41 (1.62, 19.6)
Health assistant	16 (5%)	9.45 (2.79, 36.3)	26.7 (5.48, 202.3)	11.9 (3.29, 47.5)
Nursing staff	173 (55%)	2.79 (1.34, 6.10)	8.53 (2.39, 54.6)	4.81 (1.99, 13.6)
Others	38 (12%)	2.35 (0.89, 6.30)	2.53 (0.39, 20.5)	3.52 (1.13, 11.8)
p value		0.004	<0.001	<0.001
COVID-19 area working experience				
Never	138 (44%)	1.00	1.00	1.00
Yes				
currently	114 (36%)	1.32 (0.70, 2.50)	1.19 (0.54, 2.62)	1.20 (0.62, 2.33)
previously	64 (20%)	1.59 (0.79, 3.20)	1.41 (0.60, 3.29)	0.71 (0.32, 1.51)
p value		0.34	0.65	0.38
>4 months	96 (30%)	1.35 (0.70, 2.60)	1.05 (0.46, 2.39)	1.27 (0.65, 2.49)
<4 months	82 (26%)	1.51 (0.78, 2.96)	1.55 (0.69, 3.48)	0.71 (0.34, 1.46)
p value		0.38	0.44	0.27
high-intensity area	148 (47%)	1.47 (0.82, 2.67)	1.20 (0.58, 2.52)	0.97 (0.52, 1.81)
low-intensity area	30 (9%)	1.19 (0.44, 3.09)	1.80 (0.57, 5.55)	1.03 (0.37, 2.75)
p value		0.37	0.55	0.97
Positive nasopharyngeal swab				
No	51 (16%)	1.00	1.00	1.00
Yes	265 (84%)	0.92 (0.44, 1.88)	0.85 (0.32, 2.02)	0.80 (0.36, 1.71)
p value		0.60	0.52	0.31
Family member positive to COVID-19				
No	76 (24%)	1.00	1.00	1.00
Yes	240 (76%)	0.77 (0.41, 1.42)	0.66 (0.29, 1.41)	0.61 (0.30, 1.17)
p value		0.40	0.29	0.13

Table 1.4: Multiple logistic regression for second level scales: Adjusted OR of scoring above the cut-offs with associated 95% confidence intervals (CI) and corresponding LR test p-values.

of psychological impairment. Adjusted OR of having a first-level scale above the cut-off by dividing subjects according to their personal concerns and believes about COVID-19 are presented in Table 1.5.

Each variable resulted in a statistically significant risk factor with a high odds-ratio, indicating a strong relationship with psychological distress. The highest risk, increased by more than six times, was associated with thoughts about changing job and fear for

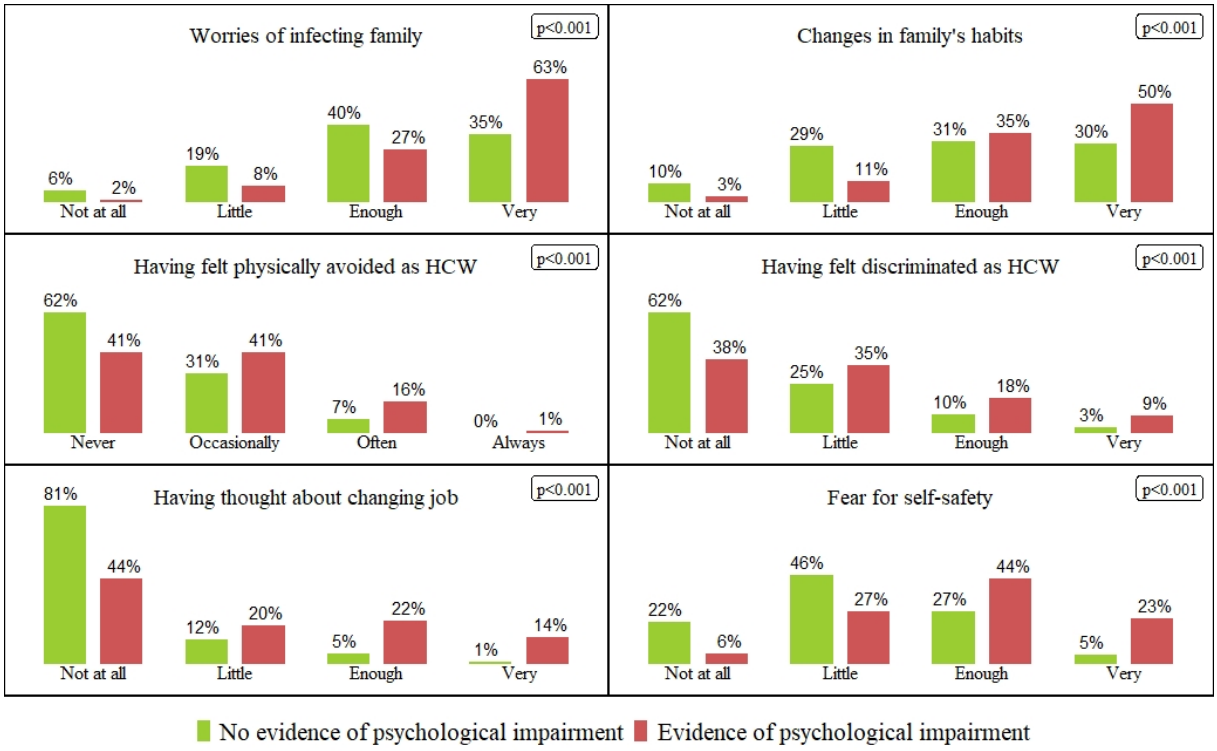

Figure 1.1: Health beliefs and COVID-19 related concerns: percentage of each answer dividing subjects with evidence of psychological impairment (red columns) and without psychological impairment (green columns)

		GHQ-12	**IES-R**	**GAD-7**
	N of positive (%)	AdjOR* (95% CI)	AdjOR* (95% CI)	AdjOR* (95% CI)
Worries of infecting family	792 (80%)	2.43 (1.60, 3.47)	4.13 (2.30, 8.11)	2.15 (1.34, 3.59)
Changes in familyâs habits	695 (70%)	3.22 (2.31, 4.54)	4.89 (3.04, 8.25)	4.34 (2.78, 7.04)
Having felt physically avoided as HCW	111 (11%)	1.72 (1.13, 2.61)	3.50 (2.25, 5.43)	2.54 (1.63, 3.91)
Having felt discriminated as HCWs	179 (18%)	2.07 (1.44, 2.86)	3.46 (2.37, 5.03)	2.16 (1.48, 3.13)
Having thought about changing job	175 (18%)	6.71 (4.58, 10.0)	6.17 (4.21, 9.08)	6.38 (4.36, 9.37)
Fear for self-safety	445 (45%)	3.59 (2.72, 4.77)	5.65 (3.89, 8.35)	3.92 (2.79, 5.56)

Table 1.5: Personal concerns about COVID-19 and risk to score above the cut-off at the first levels scales (reference subject answering No). ORs are adjusted by sex, age group, occupational role, Covid-19 area, personal infection and family member infection.

self-safety.

Figure 1.2 shows the time trends in the percentage of subjects scoring above the cut-

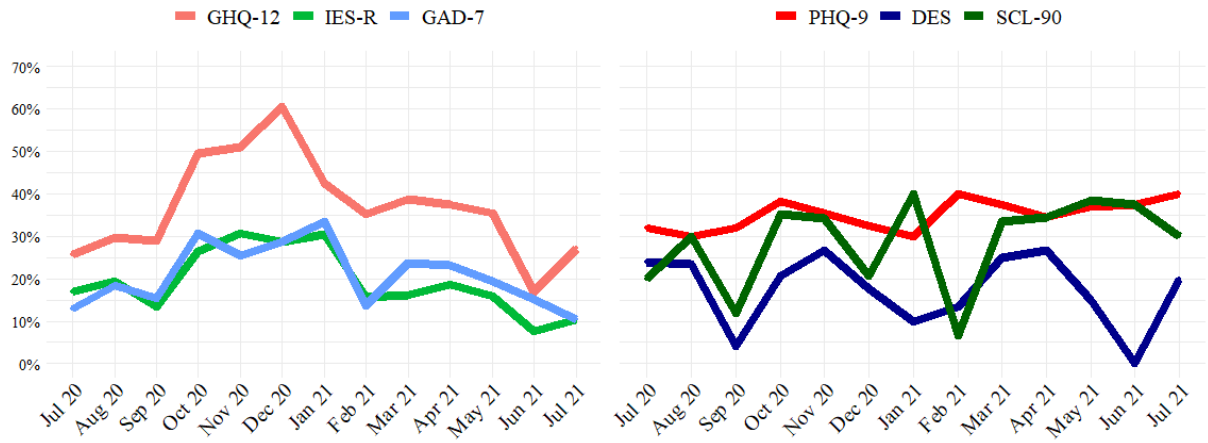

Figure 1.2: Time trend of first level screening (left) and second level evaluation (right). Percentage of subjects scoring above scales cut-off over time.

off in first and second-level scales.At first-level screening, the highest levels were reached between October and December 2020, during the second pandemic wave in Italy. In particular, the percentage above cut-off of GHQ-12 scale increased from September to December, reaching a peak of around 60%. A rapid increase in September-October was also present for GAD-7 and IES-R scales. From January 2021, the percentages of subjects with psychological impairment started to decrease, returning to baseline values in a few months. Time trends of second-level questionnaires were more irregular and different from each other: the percentage of overpass PHQ-9 cut-off was constant around 30-40%, and for DES and SCL-90 no clear trend during the study period was found.

In January-February 2021 more than 90% of HCW received anti COVID-19 vaccination. We explored the effect of vaccination on psychological well-being, comparing results in subjects evaluated before and after the vaccination campaign started. Values of OR for psychological impairment related to exposure to COVID-19 working area did not vary with vaccination: although statistical significance was lost in the post-vaccine subsample, results showed a stable increased risk among subjects working in the COVID-19 area. Similarly, a personal COVID-19 infection was not a risk factor before or after vaccination. Having a family member previously infected was a risk factor for psychological impairment only for workers enrolled before the vaccination campaign (ORs are equal to 2.25 for

GHQ-12, 1.46. for IES-R, and 1.71 for GAD-7) but not for vaccinated workers (ORs equal to 1.18, 1.10, 0.86 respectively). Detailed data for GHQ-12, IES-R and GAD-7 scales are illustrated in Table 1.6.

	N (%)		**GHQ-12**			**IES-R**			**GAD-7**		
	PRE	POST	AdjOR (95% CI)		p	AdjOR (95% CI)		p	AdjOR (95% CI)		p
			PRE	POST		PRE	POST		PRE	POST	
COVID-19 area working experience	N = 584	N = 406									
Never	249 (43%)	295 (73%)	1.00	1.00		1.00	1.00		1.00	1.00	
Yes											
currently	202 (34%)	42 (10%)	1.99 (1.25, 3.18)	2.64 (1.23, 5.70)	0.71	2.25 (1.29, 4.01)	2.55 (0.94, 6.63)	0.73	1.86 (1.07, 3.28)	1.66 (0.66, 4.01)	0.64
previously	133 (23%)	69 (17%)	1.72 (1.05, 2.83)	1.54 (0.82, 2.87)	0.75	1.97 (1.08, 3.64)	1.60 (0.72, 3.49)	0.67	1.44 (0.79, 2.63)	1.15 (0.54, 2.38)	0.71
high-intensity area	270 (46%)	75 (18%)	2.17 (1.40, 3.39)	1.80 (0.99, 3.27)	0.57	2.47 (1.45, 4.31)	1.74 (0.80, 3.69)	0.68	1.78 (1.05, 3.07)	1.43 (0.70, 2.85)	0.58
low-intensity area	65 (11%)	36 (9%)	1.11 (0.59, 2.06)	2.04 (0.87, 4.76)	0.30	1.19 (0.52, 2.59)	2.30 (0.77, 6.61)	0.56	1.31 (0.61, 2.72)	0.99 (0.33, 2.72)	0.71
Positive nasoph. swab											
No	515 (88%)	322 (79%)	1.00	1.00	0.28	1.00	1.00	0.84	1.00	1.00	0.58
Yes	69 (12%)	84 (21%)	0.58 (0.32, 1.03)	1.03 (0.60, 1.77)		1.00 (0.51, 1.88)	1.00 (0.47, 2.01)		0.52 (0.24, 1.04)	0.98 (0.49, 1.85)	
Family member infected											
No	500 (86%)	281 (69%)	1.00	1.00	0.06	1.00	1.00	0.28	1.00	1.00	0.11
Yes	84 (14%)	125 (31%)	2.25 (1.34, 3.83)	1.18 (0.73, 1.91)		1.46 (0.81, 2.58)	1.10 (0.57, 2.05)		1.71 (0.95, 3.03)	0.86 (0.47, 1.54)	

Table 1.6: ORs (adjusted for sex, age, occupational role) of scoring above cut-off of first level screening scales before and after vaccination campaign. P-values are referred to the significance of the interaction term

1.2.3 Some comments on results of descriptive analysis

We conducted a 12 months-long systematic evaluation of mental health in all workers who underwent occupational surveillance (n = 990) in a tertiary hospital in Milan, identified as one of the COVID-19 hub centers in the Lombardia Region (Italy). Our study investigated psychological well-being (by GAD-7, IES-R, GHQ-12) and specific psychiatric symptoms (by PHQ-9, DES, SCL-90) focusing on risk factors associated with mental health issues.

As consistently stated by previous investigations, psychological impairment was more frequent among nurses and female workers (Hao et al., 2021, De Kock et al., 2021). By comparing psychological scales in workers with or without direct involvement with COVID-19 patients, we observed an increased risk for impairments (in all considered scales) in the exposed workers. Such a result was confirmed when considering the duration of employment in COVID-19 wards ($>$ 6 months, $<$ 6 months, none) and the level of intensity of care (high, low, none). This assessment is consistent with research on pre-

vious coronavirus outbreaks, showing the exposure level as a major risk factor for mental health problems (Carmassi et al., 2020). On the other hand, we observed a not negligible proportion of workers with psychological impairment even in healthcare workers without direct experience with COVID-19 patients, and among administrative staff. These results are compatible with a background proportion of mental health issues in the working population, and with the effect of pandemic-related changes and concerns that involved the entire working population. The COVID-19 pandemic represents a psychological challenge and a trigger of psychological distress for all people. Our data confirmed that personal concerns and health believes related to COVID-19 (e.g. worries about infection or about infecting family members) strongly impact the risk of psychological impairments.

In this regard, the observed increased psychological distress among workers having a family member with a previous COVID-19 infection confirmed the multidimensional (occupational and non-occupational) impact of the pandemic in workers' mental health (Muller et al., 2020). Three hundred and sixteen workers (32%) presented sign of psychological impairment at the first screening level (i.e. with scores above the cut-off in at least one scale among GAD-7, IES-r and GHQ-12); among these, only a proportion of subjects presented clinically relevant symptoms (second-level screening) on PHQ-9 (35%), DES (20%), SCL-90 (28%). The observed relative frequency of psychological impairment was strongly associated with the pandemic trends in the region (with a rapid increase in the last trimester of year 2020) and a sensible decrease after January 2021, when almost all workers received the vaccination. Differently, specific psychiatric symptoms showed a different pattern of association with potential risk factors and different time trends compared to psychological impairment. In fact, scores of second-level scales were not associated to direct working experience with COVID-19 patients nor with COVID-19 experience in the family, and seem not to be influenced by pandemic waves or workers vaccination. Instead, pre-existing and more stable conditions (specifically sex and occupational level) resulted associated with sensibly higher ORs.

These results are not completely surprising. Psychiatric symptoms are more stable over time than impairments, and we cannot exclude that HCW involved in high-intensity wards have been previously self-selected in term of psychological well-being and resilience. However, a key challenge in terms of occupational medicine is precisely to detect suscepti-

ble people that may develop psychiatric problems in a context of generalized and persistent stress, as it was the pandemic experience. For example, the higher proportion of mental health issues observed among nurses and health assistants (compared with doctors) is a matter of concern. It suggests targeting specific efforts and care to preserve psychological well-being in such working groups.

Our results must be considered in light of several limitations. First of all, we have no data collected before COVID-19. Thus we cannot attribute to the pandemic all the observed psychological distress. We are aware that psychological symptoms are present in all working populations and that HCW, in particular, experienced high level of job stress and even burnout for shift-work, long working hours and several other job-related psychological risk factors. But the increasing trend in psychological impairments associated with longer direct working involvements within the COVID-19 area, suggested that COVID-19 patients' care had a specific and independent effect in determining psychological burden even if (or maybe because) HCW constitutes a population previously exposed to high levels of job strain.

We collected both exposure and effect with questionnaires; thus our study is prone to potential biases as self-selection of respondents and to common methods bias. We managed to minimize those risks grounding our investigation on the occupational physician health surveillance (obtaining a very high participation rate and minimizing the risk of untrue or uncompleted answer in describing job task) and by assessing individual COVID-19 exposure by objective data (hospital wards, duration of employments, swab results, vaccination, and other specific events).

Our results about the effect of the vaccination campaign among HCW are interesting and currently they represent one of the first evidences collected in Europe. Unfortunately, we were not able to evaluate each worker before and after vaccination, therefore we can only compare mental well-being in the same population, for the two parts of the population we interviewed before and after the vaccination campaign. Thus we cannot exclude that the better psychological scores observed after vaccination were a consequence of another unmeasured time-dependent factor, first of all the general improvement of the pandemic situation in Italy. In this respect, we must say that in Italy the vaccination among HCW was performed sensibly before (2-4 months as average) the general population. We

experienced, within the study period (March-July 2021), a marked increase of cases and hospital admissions (Covid-19 third wave) without observing an evident effect on workers' psychological burden after their vaccination.

Our study plans to follow all enrolled workers for another year, to properly assess late onset of symptoms, to analyses risk factors for symptoms persistence, and to overcome some of the above-mentioned limitations. The next results may provide further insights on preventive and beneficial interventions to support HCW mental health during and after a pandemic.

1.3 Open issues and next developments

The presented risk factor analysis was the first purpose of the *Post Covid* project. As we have already anticipated, this study plans a further investigation: all participants will undergo the same multi-step procedure after twelve months to re-evaluate their psychological status. In light of the results we got after one year of enrollment, we are planning a way to properly assess the mental HCW in the next steps.

One of the primary evidence of our findings is that the GHQ-12 scale is a powerful instrument. Such scale (alone) was able to determine whether a subject required further evaluation through the second-level questionnaire. According to physicians, this result deserves a deeper investigation and understanding. For this reason, we started a statistical study focused on this questionnaire. Firstly, we review the broad literature on this topic. Afterward, we applied various statistical methods, more advanced than the ones commonly used in clinical research, to examine properties and characteristics of the GHQ-12. Through model-based inferential methods, we gained valuable interpretations of the results. For example, we were asked to determine which traits are captured by the questionnaire and describe the patterns in people's answers. Specialists who offered psychological support to individuals with specific symptoms reported particular traits of suffering, sometimes even new ones, in some categories of frontline COVID-19 workers. Therefore, another goal was to reach further information about significant differences between subjects directly involved in the COVID-19 area and subjects without such experience.

Motivated by the suggestions and the clinical open questions mentioned above, we developed a further analysis based on the Item Response Theory (IRT). IRT is the specific statistical model for evaluating questionnaires, also in the field of psychological assessment. It is a more suitable tool, in comparison to the usual methodologies based on Classical Test Theory, whose use is still prevalent in the field. Within this theoretical framework, we have found many ways to contribute to clinical research. Indeed, to our knowledge, few authors proposed analyzing the GHQ-12 scale via an IRT approach, and nobody has already applied such methods in the context of the test factor structure (multidimensional IRT). About the latter, we have found in literature several contributions on the assessment of the dimensionality of GHQ-12. Therefore, we reviewed the topic and compared the new results with the findings based on factor analysis we previously got to give a deeper insight and interpretation of the collected data.

We are aware that our results cannot be generalized, neither are they comparable with results obtained in different scenarios. Our population was directly affected by the pandemic consequences and this exceptional situation such be carefully taken into account.

Chapter 2

Psychometric scales

The purpose of the present chapter is to give this study a clear position within the field of psychometry and to analyze the applied psychometric scales on a deeper level. We aimed at providing an overview of the topic and focusing on the meaning of the validation of a scale and the properties of questionnaires based on validated tests. Subsequently, we offered a preliminary and simple analysis of the scales of the *PostCovid* study - presented in Chapter 1. One area of interest was emphasizing the role of a psychometric scale as a measurement instrument and pinpointing its methodological and statistical bases. For instance, we expressed all the study-s psychometric indexes, resorting to the underlying formulae that are generally implied. As a matter of fact, we found out a clear foundation of the mathematical elements used to determine the properties of a scale.

Section 2.1 focuses on the quality of measurement instruments, defining the properties a good test needs to have. We will adopt the definitions and concepts reported in rigorous reviews on the psychometric characteristics of health measurement scales (Terwee et al., 2007, Souza et al., 2017, Keszei et al., 2010). Such studies deeply explore the concepts of reliability and validity. Validity is the extent to which a trait is measured in a quantitative study, while reliability refers to the accuracy of the used instrument.

Section 2.2 presents the same concepts of quality, under the Classical Test Theory (CTT) approach, introduced by Spearman (1904), and developed in (Thurstone 1925a, Kuder and Richardson 1937, Guttman 1945, and Cronbach 1951). Some limitations of this methodology, thus anticipating the Item Response Theory (IRT) approach, are also

discussed.

In Section 2.3 the six psychometric scales used in the *PostCovid* study are explored and analyzed. For each of them, we presented the internal structure and a brief review of the corresponding literature, focusing on the dimensionality issue, which is still a matter of debate, and a crucial topic for clinical assessment. We then applied the classical psychometric techniques to the six scales of the *PostCovid* study, assessing their quality.

2.1 Properties of questionnaires: an overview

Measurement instruments play an essential role in research, clinical practice, and health assessment. Our aim here is to introduce and understand the instruments that have been used for the *PostCovid* project, to be able to extract all the information they could convey about the mental health of the interviewed population.

Psychometrics - coined from the Greek words for mental and measurement - refers to the field in psychology devoted to testing, measurement, assessment and related activities. The psychometrics field looks at the theory and techniques of psychological measurement, which quantifies knowledge, abilities, attitudes and personality traits.

Therefore, we can look at the following psychometric tests like *scales*, which are typically used to uncover a personal trait that cannot be observed directly via one or more variables or items. In other words, a set of items is the instrument for measuring an underlying latent variable, continuous and unidimensional, which is not directly observable. In addition, items are usually categorical (dichotomous or polytomous). Using multiple items to measure an underlying latent construct can account for item-specific measurement error, leading to more accurate research findings. Thousands of scales have been developed to assess a range of social, psychological, and health behaviors and experiences.

Reliability and *validity* are considered the main measurement properties of such instruments and indicators of their quality. Briefly, reliability is defined as the ability to replicate a result consistently in time and space; validity refers to how well a test is actually measuring what it is intended to measure.

2.1.1 Reliability

Reliability refers mainly to the concepts of *stability*, *internal consistency* and *equivalence* of a measure. Reliability is not a fixed property of a questionnaire, as it relies on the function of the instrument, on the population in which it is used, on the circumstances, on the context; that is, the same instrument may not be considered reliable under different conditions (Souza et al., 2017).

Stability estimates how the measurement repetition is consistent, i.e. it measures how similar the results are when measured at two different times.

One way to assess stability can be the so called *test-retest method*. The procedure consists of administering the same test to the same sample at two different times, requiring that the factor to be measured remains the same in both tests moments and any change in score can be attributed to random errors. The amount of time allowed between the two measures is critical and it will influence the interpretation of reliability in the test-retest. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. Therefore, a time span from 10 to 14 days is generally considered adequate. Regarding sample size, at least 50 subjects should be considered. For the results interpretation, correlations greater than 0.70 are taken into account (Terwee et al., 2007).

The *intra-class correlation coefficient* (ICC) is one of the most used tests to estimate continuous variables stability, because it takes into account the measurement errors. Other correlation coefficients, such as Pearson or Spearman, are not suitable for this type of reliability test, because they do not consider such errors.

Internal consistency, or homogeneity, indicates if all sub-parts of an instrument measure the same characteristic. Conversely, low internal consistency may indicate that the items measure different constructs or that the answers to the questions of the instrument are inconsistent (Keszei et al., 2010).

One of the most used tools to assess internal consistency is *Cronbach's α coefficient*. Cronbach's α coefficient computes the covariance level between the items of a scale. Thus, the lower the sum of items variance is, the more consistent the instrument will be. Although Cronbach's α coefficient is the most used in assessing internal consistency, there is

no consensus on its interpretation. Even though some studies establish that values higher than 0.7 are ideal (Nunnally, 1994; Terwee et al., 2007), some researches consider values under 0.70 (but close to 0.60) as satisfactory (Streiner, 2003). Values of Cronbach's α coefficient are highly influenced by the number of items of the test. A small number of items per domain in an instrument may reduce α values, affecting the internal consistency. Concerning the average correlation between the items, if it is low, the value of Cronbach's α coefficient will be low, too. When the α coefficient increases, the average correlation also increases. Therefore, if the correlations are high, there is evidence that the items measure the same construct, fulfilling the reliability assessment.

Equivalence is the degree of concordance of two or more observers regarding an instrument's scores. The most common way of assessing equivalence is the *inter-rater or inter-observer reliability*, which involves the independent participation of two or more raters who filled the instrument. High concordance between the raters indicates that measurement errors were minimized. The inter-observer reliability depends mainly on an adequate training process of the raters and on standard practices for the test application (Rousson et al., 2002). The *Kappa coefficient* is a measure used to assess inter-observer agreement, applied to categorical items. It is a concordance measure between the raters and has a value from -1 to 1. The higher the Kappa value is, the higher the concordance between the raters.

2.1.2 Validity

Scale validity is the extent to which "an instrument indeed measures the latent dimension or construct it was developed to evaluate" (Raykov and Marcoulides, 2011). The validity of an instrument can be examined in numerous ways; the most common tests of validity are *content validity*, which can be done before the instrument is administered to the target population, *criterion validity* (predictive and concurrent) and *construct validity* (convergent, discriminant, differentiation by known groups, correlations), which occurs after survey administration.

Content validity refers to the "adequacy with which a measure assesses the domain

of interest" (Hinkin, 1995). The need for content adequacy is vital if the items are to measure what they are presumed to measure (DeVellis, 2016). Additionally, content validity specifies content relevance and content representations, i.e., that the items capture the relevant experience of the target population being examined (Church and Waclawski, 2007).

Content validity entails the process of ensuring that only the phenomenon spelled out in the conceptual definition, but not other aspects that "might be related but are outside the investigator's intent for that particular construct are added" (DeVellis, 2016).

Criterion validity is the "degree to which there is a relationship between a given test score and performance on another measure of particular relevance, typically referred to as criterion" (DeVellis, 2016, Raykov and Marcoulides, 2011). There are two forms of criterion validity: *predictive* (criterion) validity and *concurrent* (criterion) validity. Predictive validity is "the extent to which a measure predicts the answers to some other question or a result to which it ought to be related with" (Fowler Jr and Fowler, 1995). Thus, the scale should be able to predict a behavior in the future. Concurrent criterion validity is the extent to which test scores have a stronger relationship with criterion measurement made at the time of test administration or shortly afterward (Raykov and Marcoulides, 2011). This can be estimated using Pearson product-moment correlation or latent variable modeling.

Construct validity is the "extent to which an instrument assesses a construct of concern and is associated with evidence that measures other constructs in that domain and measures specific real-world criteria" (Raykov and Marcoulides, 2011). Four indicators of construct validity are relevant to scale development: convergent validity, discriminant validity, differentiation by known groups, and correlation analysis.

- *Convergent validity* is the extent to which a construct measured in different ways yields similar results. Specifically, it is the "degree to which scores on a studied instrument are related to measures of other constructs that can be expected on theoretical grounds to be close to the one tapped into by this instrument". Evidence of convergent validity of a construct can be provided by the extent to which the

newly developed scale correlates highly with other variables designed to measure the same construct (Raykov and Marcoulides, 2011, Churchill Jr, 1979). It can be invalidated by too low or weak correlations with other tests which are intended to measure the same construct.

- *Discriminant validity* is the extent to which a measure is novel and not simply a reflection of some other construct (Churchill Jr, 1979). Specifically, it is the "degree to which scores on a studied instrument are differentiated from behavioral manifestations of other constructs, which on theoretical grounds can be expected not to be related to the construct underlying the instrument under investigation" (Raykov and Marcoulides, 2011). Discriminant validity is indicated by predictably low or weak correlations between the measure of interest and other measures that are supposedly not measuring the same variable or concept (Churchill Jr, 1979). The newly developed construct can be invalidated by too high correlations with other tests which are intended to differ in their measurements.

- *Differentiation* or *comparison* between known groups examines the distribution of a newly developed scale score over known binary items (Churchill Jr, 1979). This is premised on previous theoretical and empirical knowledge of the performance of the binary groups.

- Although *correlational analysis* is frequently used, bivariate regression analysis is preferred to correlational analysis for quantifying validity (Bland and Altman, 1990, Hébert and Miller, 1991). Regression analysis between scale scores and an indicator of the domain examined has a number of important advantages over correlational analysis. First, regression analysis quantifies the association in meaningful units, facilitating judgment of validity. Second, regression analysis avoids confounding validity with the underlying variation in the sample and therefore the results from one sample are more applicable to other samples in which the underlying variation may differ. Third, regression analysis is preferred because the regression model can be used to examine discriminant validity by adding potential alternative measures (Hébert and Miller, 1991).

Taken together, these methods enable the assessment of the validity of an adapted or a newly developed scale. In addition to predictive validity, existing studies in fields such as health, social, and behavioral sciences have shown that scale validity is supported if at least two of the different forms of construct validity discussed in this section have been examined.

A primer which well describes the process for scale development is illustrated in Boateng et al. (2018), according to which there are three phases to creating a rigorous scale: *item development*, *scale development* and *scale evaluation*. These can be further split into nine steps. Item development (i.e. coming up with the initial set of questions for an eventual scale) is composed of identifying the domain(s) and item generation, and consideration of content validity. The second phase, scale development (i.e., turning individual items into a harmonious and measuring construct), consists of pre-testing questions, sampling and survey administration, item reduction, and extraction of latent factors. The last phase, scale evaluation, requires tests of dimensionality, tests of reliability, and tests of validity.

2.2 Classical Test Theory approach

Classical test theory (CTT) is a traditional quantitative approach for testing the reliability and validity of a scale based on its items. Within CTT, also known as *true score theory*, it is assumed that each person i in a population of size n has a true score, T_i, that would be obtained if there were no errors in measurement. A person's true score is defined as the expected score over an infinite number of independent administrations of the scale. Scale users never observe a person's true score but only an observed score, X_i. It is assumed that

$$Y_i = X_i + \epsilon_i \qquad i = 1, ..., n.$$

True scores quantify values on an attribute of interest, defined as the underlying concept, construct, trait, or ability of interest. As values of the true score increase, responses to items representing the same concept should also increase (i.e., there should

be a monotonically increasing relationship between true scores and item scores), assuming that item responses are coded so that higher responses reflect more of the concept.

The measurement errors ϵ_i are usually assumed to have a continuous distribution, typically normally distributed with mean 0, and to be uncorrelated with true score X_i, with no systematic relationship between a person's true score and whether that person has positive or negative errors:

$$\mathbb{E}(Y_i) = \mathbb{E}(X_i), \qquad \text{cor}(X_i, \epsilon_i) = 0$$

and, as a consequence,

$$\text{var}(Y_i) = \text{var}(X_i) + \text{var}(\epsilon_i) \qquad i = 1, ..., n.$$

A fundamental premise of CTT is that items can be summed (without weighting or standardization) to produce a total score. Summing item scores is considered legitimate when

- the items are approximately parallel (i.e., they measure at the same point);
- they contribute similarly to the variation of the total score (i.e., they have similar variances, otherwise item scores should be standardized);
- they measure a common underlying construct, and
- they contain a similar proportion of information concerning the construct being measured.

As already mentioned in Section 2.1, reliability is the measure indicating the overall quality of a test, in terms of the degree of consistency exhibited when a measurement is repeated under identical conditions.

We can give two different definitions of reliability, according to Bartolucci et al. (2019).

The first index, based on proportion of variance, is the ratio between the true and the observed score variance

$$\rho^2_{xy} = \frac{\text{var}(X_i)}{\text{var}(Y_i)} = 1 - \frac{\text{var}(\epsilon_i)}{\text{var}(Y_i)}. \tag{2.1}$$

The second definition is given in terms of the squared correlation between the observed and the true scores. Therefore, reliability equal to 1 indicates that differences

between respondents' observed test scores are perfectly consistent with respondents' true test scores:

$$\rho^2_{xy} = \text{cor}(Y_i, X_i)^2 = 1 - \text{cor}(Y_i, \epsilon_i)^2.$$

Concerning reliability estimation, there are two main families of tools

- methods based on repeated measures, such as the *alternate form method* and the *test-retest method*, and
- methods based on a single measure, such as the *split-half method* and the *internal consistency reliability method*.

Some of them were already introduced in Section 2.1 and are here formally presented.

The *alternate form method* requires administering two similar forms of a test, with the same content, to the same group of respondents within a short period of time (Crocker and Algina, 1986). Half of the participants receive the forms in a given order and the other half in reverse sequence. The correlation between the observed scores on the two forms, called the *coefficient of equivalence*, is taken as an estimate of test reliability

$$\hat{\rho}^2_{xy} = cor^{(T_1,T_2)}$$

where the correlation between the two tests T_1 and T_2 is defined as

$$\text{cor}^{(T_1,T_2)} = \frac{\sum_{i=1}^{n}(y_i^{T_1} - \bar{y}^{T_1})(y_i^{T_2} - \bar{y}^{T_2})}{\sqrt{\sum_{i=1}^{n}(y_i^{T_1} - \bar{y}^{T_1})^2(y_i^{T_2} - \bar{y}^{T_2})^2}}$$

being $\bar{y}$ the mean individual score.

A high coefficient of equivalence indicates that scores from the different forms are interchangeable. A value equal or higher than 0.70 is acceptable.

The *split-half method* consists of dividing the items of a test into two subsets of the same dimension. Different methods can be used for dividing a test into halves (Crocker and Algina, 1986), for example assigning even-numbered items to form 1 and odd-numbered items to form 2 (or vice versa); ranking the items on the basis of their difficulty levels and

then assigning items with odd-numbered ranks to form 1 and items with even-numbered ranks to form 2 (or vice versa); assigning items to the two subsets randomly. The most common measure used to estimate the reliability of the full test (i.e., the test composed by the two halves) within the split-half procedure is the *Spearman-Brown formula*

$$\text{cor}^{(T_1,T_2)} = \frac{2\text{cor}^{(T_1,T_2)}}{1 + \text{cor}^{(T_1,T_2)}}.$$

The split-half method to estimate reliability requires the assumption that the two halves are parallel (e.g., the subsets should have the same means and variances of observed scores). Furthermore, there are many possible ways of dividing a test into halves. Thus, the method does not yield to a unique reliability estimate.
An alternative for avoinding such limitations of the split-half method is the *internal consistency reliability method*. This method is defined as an item-level approach, because it considers each item of a test as a separate test. The complete test is administered once to a sample of examinees. Subsequently, covariances are calculated between all the pairs of items. The index for measuring internal consistency is *Cronbach's* α, defined as

$$\alpha = \frac{J}{J-1} \frac{\sum_{j=1}^{J} \sum_{j'=1, j' \neq j}^{J} \text{cov}(Y_{ij}, Y_{ij'})}{\text{var}(Y_i)}$$

where $\text{cov}(Y_{ij}, Y_{ij'})$ is the pairwise covariance between items j, j' with $j \neq j'$.

Values in range $[0.70; 0.90]$ indicate high reliability; values < 0.60 weak reliability (e.g. item can be ambiguously defined or items are very different from each other); for values > 0.90 reliability is suspicious (e.g. redundant items).

Regarding *validity* within the CTT approach, the internal structure of a test is usually assessed through *factor analysis*, which allows to identify the number of factors corresponding to different subsets of correlated items. Thus, when a test measures only one factor, all items within the test are correlated with each other, reflecting only one psychological attribute or construct (unidimensional test). On the other hand, whenever factor analysis gives evidence of multidimensionality, it also allows *i)* to determine the number of factors, *ii)* to identify which items are linked to which factors, and *iii)* to evaluate the strength of associations between the dimensions.

Derived from CTT, factor analysis includes a variety of statistical procedures for exploring the relationships among a set of observed variables with the intent of identifying a smaller number of factors (EFA), the unobserved latent variables, thought to be responsible for these relationships among the observed variables. CFA is used primarily as a way of testing hypotheses about the latent structure underlying a set of observed data.

Classical test Theory is known to have some shortcomings, which are summarized below.

1. The assumption of a linear relationship between the latent and observed scores is restrictive and does not represent psychological constructs. Moreover, assuming such a linear function with different item locations implies that for certain values of the latent trait, no score is defined unless the item is metric and ranges from minus infinity to plus infinity. This is undesirable because it restricts the span of the latent variable if categorical variables are used. If such a linear relationship is assumed, it is not congruent with the idea of the different locations of items.

2. The true score cannot be estimated directly, but only with additional assumptions regarding the item-specific true scores. A scoring rule (e.g., simple or weighted sum) is implicitly assumed to be correct, but its adequateness cannot be tested. Furthermore, the simple sums of the item scores are often taken as an estimate of the person's latent trait value or the item's location, equating the expected true values with the sum of the observed scores. However, it is possible that a person with a lower score in a test has a higher position on the latent trait. This could be the case if this person fakes an answer while the person with the higher location answers truthfully. Therefore, the sum of observed scores is not appropriate for measuring such empirical situations.

3. Parameters such as reliability, discrimination, location, and factor loadings are *sample dependent*, which implies different reliabilities as well as different factor loadings of an item set for both homogeneous and heterogeneous samples. Hence, it often happens that different numbers of factors for different samples emerge. They apply only to the sample at hand and are unbiased for the population of interest only if the sample is a true random sample and is representative of the population of interest.

If we want to estimate a person's location on a latent trait, that value depends on the sample of items used for measurement and on the other people who are being assessed. Depending on the reference population, a person will also have a different position on the latent trait, even if the random sample is representative.

2.3 Psychometric scales used in *PostCovid* study

2.3.1 General Health Questionnaire (GHQ-12)

The General Health Questionnaire (Spearman, 1904) aims to provide information about an individual's mental well-being through the identification of distressing symptoms (Tait et al., 2002). The original version of this measure was developed to evaluate the mental health status of patients of general practitioners in the United Kingdom and it originally contained 60 items. Shorter versions have been developed from the original one, e.g., the GHQ-30, GHQ-28 and GHQ-12. All such scales have been used in clinical practice, epidemiological research, and psychological research (Hankins, 2008).

2.3.1.1 Scale structure and characteristics

Since its development and introduction, the shorter version of the GHQ (12-item) has probably become the most widely used scale for assessing psychological distress and short-term changes in mental health, and its popularity can be mostly attributable to its brevity, easy administration, and availability of normative data (Fernandes and Vasconcelos-Raposo, 2013).
The twelve questions regard general psycho-physical conditions during the last two weeks:

1. Have been able to concentrate on what you are doing?
2. Have you lost much sleep over worry?
3. Felt you have playing a useful part in things?
4. Felt capable making decisions about things?
5. Felt constantly under strain?

6. Felt you couldn't overcome your difficulties?
7. Been able to enjoy your day-to-day-activities?
8. Been able to face up to your problems?
9. Been feeling unhappy and depressed?
10. Been losing confidence in yourself?
11. Been thinking of yourself as a worthless person?
12. Been feeling reasonably happy, all things considered?

Four response categories are given for each item: *better than usual, same as usual, less than usual, much less than usual.*

Two scoring ways can be adopted, the conventional one (0-0-1-1) and the *Likert scoring* (0-1-2-3). According to Goldberg et al. (1997), the Likert method is worse than the other one: the dichotomous scoring (0-0-1-1) eliminates the problem of *middle and end users* and that of the *conceptual distance* between positions on the response scale (Piccinelli et al., 1993). Campbell and Knowles (2007) suggest that Likert scoring of the GHQ-12 allows better discrimination between competing models in confirmatory factor analyses. Six of the items are positively worded (1, 3, 4, 7, 11, 12) and six (2, 5, 6, 8, 9, 10) are negatively worded. The positive items were coded from 0 (better than usual) to 3 (much less than usual) and the negative ones from 3 (better than usual) to 0 (much less than usual), according to Likert scale.

There exists also a modified dichotomous system (0-1-1-1) of the conventional score, proposed in Goodchild and Duncan-Jones (1985); authors suggested that the answer 'no more than usual' to an item describing pathology should be treated as an indicator of chronic illness rather than of good health.

Several studies in the literature aim to find the cut-off that gives the best sensitivity and specificity. *Sensitivity* refers to the proportion of people who have a psychiatric disorder and who score above a cut-off on a measure of psychological symptoms. *Specificity* refers to the proportion of people without a psychiatric disorder who score below a cut-off on the same instrument. Choosing measures with high sensitivity as well as high

specificity can help physicians identify probable psychiatric disorders while limiting the over-diagnosis of patients who are not likely to have a disorder (i.e., false-positives) (Cano et al., 2001).

It has been discussed that socio-demographic characteristics of respondents may affect the screening properties of the questionnaire: for example, de Jesus Mari and Williams (1986) reported that general practice attendees aged less than 40 years had a higher risk of being classified as false negatives (Politi et al., 1994).
The most used cut-off values are equal to three or four (out of 12, the total score according to dichotomous scoring) and equal to twelve (using Likert scale).

The Italian version of the GHQ-12 is a reliable instrument, used first in a study of 18-years old male population, with a Cronbach's α of 0.81 (Politi et al., 1994). Later it was subjected to direct validation in general practice with sensitivity ranging between 0.71 and 0.75, specificity between 0.73 and 0.76, according to different scoring methods (Politi et al., 1994). In Italy, in a later study, Cronbach's α for the GHQ-12 total scale was 0.85 and for each dimension/factor ranged from 0.73 to 0.82, suggesting good reliability of the instrument.

2.3.1.2 Dimensionality

Although the GHQ-12 was originally intended as a unidimensional instrument, the structure and dimensionality of the GHQ-12 are still a matter of discussion. Many studies derived a unidimensional interpretation; nonetheless, several analyses also assessed multidimensional GHQ-12 structures. Some two-factor solutions have been proposed and validated using Confirmatory Factor Analysis (CFA) techniques. The two factors commonly uncovered are a *Depression/Anxiety* construct and a *Social Dysfunction* construct (Andrich and Van Schoubroeck, 1989), given by GHQ-12 items 2, 5, 6, 9, 10 and 11, and 1, 3, 4, 7, 8 and 12 respectively. *Depression/Anxiety* relates to the emotional component of psychological distress, whereas *Social Dysfunction* relates to the social functioning component of the individual experiencing the distress.

Politi et al. (1994) used a principal components analysis to explore the dimensionality of the GHQ-12 and identified a two-factor structure: a seven-item factor consisting of the anxiety and depression items (*General Dysphoria*), and a six-item factor, consisting

of the items relating to daily activities and ability to cope (*Social Dysfunction*). One item (item 12, 'Not feeling happy') loaded weakly onto both factors. Evidence of a two-factors structure, even if differently labelled, were found in other later studies: in Schrnitz et al. (1999) an alternative two-factor model has been proposed, consisting of a six-item 'Anxiety/Depression' factor and a five-item 'Daily Activities and Social Performance' factor. Other contributes from UK (Smith et al. 2010), New Zealand (Kalliath et al. 2004), Brazil (Gouveia et al. 2010) gave proof of the bi-dimensional structure.

According to Graetz (1991), the most accepted model is the three-dimensional, comprising anxiety (4-item), social dysfunction (6-item), and loss of confidence (2-item) dimensions. Available empirical evidence has suggested that there is little usefulness and clinical meaningfulness for GHQ-12 in estimating the above-mentioned dimensions. Hence, it is suggested that if clinically useful, reliable, and consistent domains/components of a disorder (e.g., anxiety, depression) are to be evaluated, clinicians and researchers should administer specific syndrome-oriented measures (e.g., Beck Anxiety Inventory, Beck Depression Inventory, Hospital Anxiety and Depression Scale) alongside a structured clinical interview, after the 'case-ness' identification defined by the GHQ (Fernandes and Vasconcelos-Raposo, 2013).

Systematic reviews and meta-analyses also consistently identify these two factors most commonly in both two-and-three factor solutions (Gnambs and Staufenbiel, 2018, Picardi et al., 2001, Werneke et al., 2000). It is interesting to note that these groupings also align with the positive or negative phrasing of the constituent items. *Social Dysfunction* items are all positively worded, and *Depression/Anxiety* items are all negatively worded, and as such, there has been debate as to whether this structure simply reflects differences in phrasing (Hankins, 2008).

A report in Smith et al. (2013) is entirely dedicated to the factor structure of GHQ-12. The paper explores the fact that item phrasing, item variance and levels of respondents' distress affect the factor structure observed for the GHQ-12 and may perhaps explain why different factor structures of the instrument have been found in different populations.

A recent work, due to Hystad and Johnsen (2020) summarizes this debate and provides a comparison between different methods, testing five alternative factor structures: a bi-factor structure with one general (principal) factor and two specific factors proved to be

the best representation of the data from a statistical perspective.

2.3.2 Impact of Event Scale - Revised (IES-R)

The Impact of Event Scale - Revised (IES-R) is a 22-item self-report measure that assesses subjective distress caused by traumatic events (Weiss, 2007).

The IES-R scale is an appropriate instrument to measure the subjective response to a specific traumatic event, especially in the response sets of *intrusion* (intrusive thoughts, nightmares, intrusive feelings and imagery, dissociative-like re-experiencing), *avoidance* (numbing of responsiveness, avoidance of feelings, situations, and ideas), and *hyperarousal* (anger, irritability, hypervigilance, difficulty concentrating, heightened startle). It is used for recent and traumatic events: according to the first version in fact, until traumatic experiences are psychologically assimilated, the individual will experience intrusive thoughts and feelings in one moment and avoidance strategies in the next. Following this model, the IES was constructed with two subscales, one detecting intrusions (e.g., repeated thoughts about the trauma) and the other for avoidance (e.g., avoidance of situations that serve as reminders of the trauma). The IES-R has 22 questions and is the revised version of an older version, the 15-item IES (Horowitz et al., 1979), and it was developed in order to include a scale measuring hyperarousal. As the scale was originally constructed to measure "the current degree of subjective impact experienced as a result of a specific event" (Horowitz et al., 1979) and in the first version the intrusion and avoidance subscales were not designed to represent independent constructs that would be distinct and unique from general distress. However, it is unclear to what extent these subscales assess trauma-specific phenomena. The lack of assessment of hyperarousal within the IES contributes further to uncertainty about how the measure corresponds to the tripartite model of PTSD. The authors of the IES-R intended for the scale to be comparable with the original scale and so, made only minor changes to the intrusion and avoidance subscales. The aim of the revisised version was to improve the utility of the IES and its applicability symptomatology for PTSD (Weiss, 2007).

Respondents are asked to identify a specific stressful life event and then indicate how much they were distressed or bothered during the past seven days by each *difficulty* listed.

The IES-R yields a total score ranging from 0 to 88 and scores can also be calculated

for the Intrusion, Avoidance, and Hyperarousal subscales.

The participant is asked to report the degree of distress experienced for each item in the past seven days. The five points on the scale are: 0 (not at all), 1 (a little bit), 2 (moderately), 3 (quite a bit), 4 (extremely). The 22 items are the following:

1. Any reminder of it brought back feelings about it
2. I had trouble staying asleep
3. Other things kept making me think about it
4. I felt irritable and angry
5. I avoided letting myself get upset when I thought about it or was reminded of it
6. I thought about it when I didn't mean to
7. I felt as if it hadn't happened or wasn't real
8. I stayed away from reminders of it
9. Pictures about it popped into my mind
10. I was jumpy and easily startled
11. I tried not to think about it
12. I was aware that I still had a lot of feelings about it, but I didn't deal with them.
13. My feelings about it were kind of numb
14. I found myself acting or feeling like I was back at that time
15. I had trouble falling asleep
16. I had waves of strong feelings about it
17. I tried to remove it from my memory
18. I had trouble concentrating

19. Reminders of it caused me to have physical reactions, such as sweating, trouble breathing, nausea, or a pounding heart
20. I had dreams about it
21. I felt watchful and on-guard
22. I tried not to talk about it.

Sub-scales are composed by the following items

- *intrusion sub-scale*: items 1, 2, 3, 6, 9, 14, 16, 20
- *avoidance sub-scale*: items 5, 7, 8, 11, 12, 13, 17, 22
- *hyper-arousal sub-scale*: items 4, 10, 15, 18, 19, 21

Both versions (IES and IES-R) have good psychometric properties (Weiss, 2007). Acceptable Cronbach's α for each sub-scale indicate good internal consistency: 0.87-0.94 for intrusion, 0.84-0.97 for avoidance and 0.7-0.91 for hyper-arousal. High correlations have been found between the IES-R and the original IES sub-scales: equal to 0.86 and 0.66 for intrusion and avoidance, respectively.

The validation of the Italian version found that the dimensions showed a significant level of correlation with each other, indicating that the questionnaire sub-scales measured several approaches of the impact of the event that are relatively distinct from one another, suggesting an acceptable level of score independence. All sub-scales α coefficients can be considered as good (hyper-arousal, $\alpha = 0.83$; avoidance, $\alpha = 0.72$; intrusion, $\alpha = 0.78$), this self-report instrument for assessing the dimensions of trauma has good psychometric properties and can be adopted usefully, both for research and in practice in Italy.

There is also an Italian version of the IES-6 (Thoresen et al., 2010), and an even shorter version of IES, tested in the bank robberies field (Giorgi et al., 2015). IES-6 seems particularly useful since it can be quickly administered for screening individuals with important traumatic stress symptoms as the first step of a further assessment.

There are no specific cut-off scores for the IES-R, although higher scores represent greater distress. Increased scores on all subscales may indicate the need for further evaluation. However, Creamer et al. (2003) reports that a total score of 33 on the IES-R yielded

a diagnostic sensitivity of 0.91 and specificity of 0.82. For this reason, in our study, we have adopted such value for the cut-off.

Despite its good psychometric properties, the factorial structure of the IES-R is debated. For example, Shevlin et al. (2000) reports three confirmatory factor analyses on the IES scale, finding that a two-factor model with additional cross-factor loadings for Items 2 and 12. Another contribution to the multidimensionality is in Andrews et al. (2004), which tests seven alternative factor models of the IES and concludes that the optimal structure of the IES consisted of four first-order factors (intrusion, avoidance, numbing, and sleep), with one second-order factor which assessed general distress. A four-factor structure representing intrusion, avoidance-numbing, hyperarousal, and sleep emerged as the preferred model in King et al. (2009).

In Craparo et al. (2013) the Italian version of the IES-R showed a clear factor structure with three independent dimensions: intrusion, avoidance and hyper-arousal.

2.3.3 Generalized Anxiety Disorders Scale (GAD-7)

The 7-item Generalized Anxiety Disorders Scale (GAD-7) is a valid and efficient tool for assessing general anxiety symptoms across various settings and populations. It was originally developed as a screener for generalized anxiety disorder (GAD) in primary care settings (Spitzer et al., 2006).

GAD-7 was born starting from the 13 original items GAD proposed in the Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition (DSM-IV). Items were correlated with the total score and the seven items with the highest correlation with the total 13-item scale were selected (Spitzer et al., 2006).

Participant is asked to answer how often he/she has been bothered by the following problems:

1. Feeling nervous, anxious, or on edge
2. Not being able to stop or control worrying
3. Worrying too much about different things

4. Trouble relaxing
5. Being so restless that it's hard to sit still
6. Becoming easily annoyed or irritable
7. Feeling afraid as if something awful might .

Each item has four possible responses, scored on Likert scale (0-1-2-3) with total scores ranging from 0 to 21, where higher scores reflect greater anxiety severity. The GAD-7 has several advantages: it is easy to use, has clear psychometric properties, and consists of only seven items. Thus, it can be safely adopted by clinicians. The scale is a useful tool with a strong criterion validity for identifying probable cases of anxiety. It is also an excellent severity measure, as demonstrated by the fact that increasing scores on the GAD-7 are strongly associated with multiple domains of functional impairment and disability days. Good properties of GAD-7, in terms of consistency, reliability and validity, were found measuring GAD symptoms in the psychiatric population (Kertz et al., 2013, Rutter and Brown, 2017), in general population (Löwe et al., 2008, Hinz et al., 2017) and in heterogeneous psychiatric samples (Beard and Björgvinsson, 2014, Johnson et al., 2019). A score greater or equal to 10 (Spitzer et al., 2006) on the GAD-7 represents a reasonable cut point for identifying cases of GAD. Cut points of 5, 10, and 15 might be interpreted as representing mild, moderate, and severe levels of anxiety on the GAD-7.
The factor structure has been investigated for GAD-7 too. A one-dimensional factor structure in a German representative study was reported in Löwe et al. (2008), while a confirmatory factor analysis in an acute psychiatric sample in Kertz et al. (2013) found three items measuring bodily symptoms (items 4-6). Beard and Björgvinsson (2014) proposed a two-factor structure for the GAD-7 using exploratory factor analysis, one factor reflecting bodily symptoms (items 4-6) and the other factor assessing the cognitive and emotional experience of anxiety (items 1-3 and 7). However, a recent study (Rutter and Brown, 2017) used confirmatory factor analysis and found evidence for a unidimensional factor, after accounting for the additional covariance found between items 4, 5, and 6.

2.3.4 Application of CTT, and first results on psychometric tests for the *PostCovid* study

This section collects preliminary results from a deeper analysis of scale scoring and items answering.

In terms of reliability, we checked whether internal consistency is reached. Table 2.1 reports Cronbach's α values which indicate very good internal consistency for scales and subscales.

	Cronbach's α (95%CI)
GHQ-12	0.87 (0.86, 0.88)
IES-R	0.93 (0.92, 0.94)
intrusion	0.88 (0.86, 0.89)
avoidance	0.81 (0.79, 0.83)
hyper-arousal	0.84 (0.82, 0.86)
GAD-7	0.90 (0.89, 0.91)

Table 2.1: Cronbach's α values

Table 2.2 shows the scales distribution: 37% of participants expressed signs of psychological distress (GHQ-12), 19% post-traumatic symptoms, and for 19% of workers reported signs of anxiety over the past two weeks. Out of 363 (36%) workers who underwent the second level screening, only 32 persons did not score above four in GHQ-12, meaning that the latter scale was the most 'sensitive' and basically caused participation to the following screening phase. At least one scoring above cutoff determines the access to fill second-level questionnaires.

	mean	**sd**	**range**	**cutoff**	**% > cutoff**
GHQ-12	3.13	3.25	$0-12$	4	37%
IES-R	19.2	16.6	$0-88$	33	19%
GAD-7	5.84	5.12	$0-21$	10	19%

Table 2.2: First level scales summary statistics

Items response distributions are shown, respectively, in Table 2.3 for GHQ-12, in Table 2.4 for IES-R, and in Table 2.5 for GAD-7.

Regarding GHQ-12, for which the dichotomous scoring was chosen, Item 5 and Item 7 reported the highest percentage of responses equal to 1. In particular, Table 2.3 show that half of the participants constantly felt under strain, while 45% could not enjoy their day-to-day activities. A huge percentage (92%) scored equal to zero in Item 11 (about thinking of themselves as worthless people). These results sound reasonable, and they explain the considerable percentage of subjects scoring above cut-off; we can assess that the 'reason' for overcoming cut-off is mainly due to stress and general discomfort.

	response=0	**response=1**
Item 1	75%	25%
Item 2	64%	36%
Item 3	86%	14%
Item 4	85%	15%
Item 5	49%	51%
Item 6	78%	22%
Item 7	55%	45%
Item 8	79%	21%
Item 9	68%	32%
Item 10	83%	17%
Item 11	92%	8%
Item 12	73%	27%

Table 2.3: GHQ-12: item response distribution

The IES-R was adapted to the pandemic context. Participants were asked to express how much they have been distressed or bothered by the difficulties listed above during the past seven days regarding COVID-19 emergency. The item with the highest score (see Table 2.4) is item 21 (*I felt watchful and on-guard*). This fact is easily understood when thinking of item 21 as a widespread consequence of the pandemic situation. On the contrary, Item 7 (*I felt as if it hadn't happened or wasn't real*), Item 19 (*Reminders of it caused me to have physical reactions, such as sweating, trouble breathing, nausea, or a pounding heart*) and Item 20 (*I had dreams about it*) obtained very few high responses. Item 1 (feeling nervous, anxious, or on edge) and Item 4 (trouble relaxing) showed the

highest percentage of high scores, while Item 5 (being so restless that it's hard to sit still) and Item 7 (feeling afraid as if something awful might happen) reported the lowest scoring.

	subscale	**0**	**1**	**2**	**3**	**4**
Item 1	intrusion	25%	34%	19%	18%	5%
Item 2	intrusion	48%	22%	11%	13%	5%
Item 3	intrusion	29%	36%	15%	17%	4%
Item 4	hyper-arousal	36%	30%	15%	13%	6%
Item 5	avoidance	33%	28%	15%	20%	4%
Item 6	intrusion	38%	32%	13%	14%	2%
Item 7	avoidance	84%	9%	4%	3%	1%
Item 8	avoidance	57%	25%	6%	8%	3%
Item 9	intrusion	54%	26%	8%	10%	2%
Item 10	hyper-arousal	50%	26%	10%	10%	4%
Item 11	avoidance	41%	29%	11%	16%	3%
Item 12	avoidance	61%	20%	6%	9%	3%
Item 13	avoidance	70%	16%	5%	7%	2%
Item 14	intrusion	58%	25%	8%	7%	2%
Item 15	hyper-arousal	50%	25%	11%	8%	6%
Item 16	intrusion	46%	29%	10%	11%	4%
Item 17	avoidance	68%	18%	5%	7%	2%
Item 18	hyper-arousal	54%	26%	7%	10%	3%
Item 19	hyper-arousal	79%	12%	4%	4%	1%
Item 20	intrusion	77%	12%	4%	4%	2%
Item 21	hyper-arousal	26%	33%	15%	16%	9%
Item 22	avoidance	66%	16%	6%	10%	2%

Table 2.4: IES-R: item response distribution

	1	**2**	**3**	**4**
Item 1	22%	52%	12%	14%
Item 2	48%	35%	6%	11%
Item 3	35%	43%	8%	14%
Item 4	31%	44%	10%	15%
Item 5	73%	17%	4%	7%
Item 6	36%	43%	9%	12%
Item 7	61%	26%	5%	7%

Table 2.5: GAD-7: item response distribution

2.3.5 Psycho-pathological scales

In this last section, we briefly present characteristics of three psycho-pathological scales employed in *PostCovid* study. Since items of these scales are not explored (especially for SCL-90, since they are a huge number), the following analysis is limited to the literature review regarding origins and dimensionality issues. Summary statistics of scores are reported, together with results about subscales.

2.3.5.1 Patient Health Questionnaire (PHQ-9)

The Patient Health Questionnaire (PHQ) is a self-administered diagnostic instrument for common mental disorders. The PHQ-9 is the module dedicated to depression, which scores each of the Diagnostic and Statistic Manual (DSM)-IV criteria (anhedonia, depressive feelings, dyssomnia, anergia, appetite problems, negative self-evaluation, concentration problems, suicidality) (Kroenke et al., 2001). The patients are asked to rate how much of the time symptoms persisted over the last two weeks. There are four response categories: not at all (0), several days (1), more than half the days (2) or nearly every day (3).
More specifically, the nine items are

1. Did you feel little interest or pleasure in doing things?

2. Feeling down, depressed, or hopeless?

3. Trouble falling or staying asleep, or sleeping too much?
4. Feeling tired or having little energy?
5. Poor appetite or overeating?
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down?
7. Trouble concentrating on things, such as reading the newspaper or watching television?
8. Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual?
9. Thoughts that you would be better off dead, or of hurting yourself in some way?

It was originally constructed for screening of major depression (Spitzer et al., 1999) but it can also be used for the measurement of depression severity, for the evaluation of depression over time (Kroenke et al., 2001, Löwe et al., 2004). These characteristics, associated with its brevity, make the PHQ-9 a highly useful screening tool, but it is not intended to be a stand-alone diagnostic test.
As a severity measure, the PHQ-9 score can range from 0 to 27; scores of 5, 10, 15, and 20 represent valid thresholds indicating limits of mild, moderate, moderately severe, and severe depression. In particular, scores less than 10 seldom occur in individuals with major depression while scores of 15 or greater usually signify the presence of major depression. The PHQ-9 is validated by many studies, in the medical and general population, with excellent psychometric values (Martin et al., 2006).
Regarding the dimensionality of PHQ-9, some recent studies reported uni-dimensional (González-Blanch et al., 2018, Keum et al., 2018) as well as two-dimensional structures, consisting of somatic and affective factors (Elhai et al., 2012, Guo et al., 2017, Chilcot et al., 2013, Krause et al., 2010, Richardson and Richards, 2008). Some authors combine PHQ-9 depression scale and GAD-7 scale as a composite measure of depression and anxiety (Kroenke et al., 2016), so giving a sort of new scale, whose properties are investigated. Factorial structure, validity, and association of PHQ-9/GAD-7 instruments are examined,

for instance, in Teymoori et al. (2020), based on longitudinal observational data. Similarly, Stochl et al. (2020) evaluate such properties for common measures of depression (PHQ-9) and anxiety (GAD-7) in a clinical sample undergoing psychotherapy. Results show that while both PHQ-9 and GAD-7 are multidimensional instruments with highly correlated factors, there is justification for sum scores as measures of severity. A comparison with another similar scale, the Hospital Anxiety and Depression Scale (HADS-D), is the content of Kendel et al. (2010), which analyzes the dimensionality and the item fit of both scales individually and across the scales. Therefore, the short form of the PHQ-9 seems to be an economical and valid instrument for screening depression, indicating the same latent construct captured by six items of the HADS-D.

2.3.5.2 Dissociative Experiences Scale (DES)

The Dissociative Experiences Scale (DES) is a simple questionnaire widely used to screen for dissociative symptoms. Dissociation is characterized by a disruption of and discontinuity in the normal integration of consciousness, memory, identity, emotion, perception, body representation, motor control, and behaviour. The dissociative disorders are frequently found in the aftermath of trauma, and many of the symptoms, including embarrassment and confusion about the symptoms or a desire to hide them, are influenced by the proximity to trauma. Dissociation occurs to some degree in normal individuals and but is more prevalent in people with major mental illnesses. DES scale has been developed to offer a means of reliably measuring dissociation in normal and clinical populations: it consists of 28 items that describe common dissociative experiences. Each individual item asks about the percentage of time (from 0% to 100%) that a particular dissociative symptom is experienced. The overall DES score is the average of all the individual scores. Scores of 20 or more are consistent with various kinds of post-traumatic or dissociative disorders. The original DES-I asks respondents to make slash marks on a 100-mm line estimating the percentage of time they have the specific dissociative experience (the score for each item is the nearest 5-mm point, e.g., 0, 5, 10, etc.). The DES-II asks respondents to circle a percentage number (e.g., 0%, 10%, 20%,. . .,100%) indicating the frequency they experienced dissociative symptoms. Anyhow, DES-II has the same reliability and validity as the original DES. The psychometric properties of the DES/DES-II are good, with

excellent internal consistency, good test-retest reliability, and good convergent validity.

The Italian translation of the DES-II showed high internal consistency, adequate item-to-scale homogeneity, and good split-half reliability (Schimmenti, 2016). The Italian version intended as a three-factor model, is the most frequently used one. According to this model, DES is composed of the following sub-subordinates: *dissociative amnesia*, which concerns actions of which the subject does not remember (item: 3, 4, 5, 8, 11), *absorption and imaginative involvement*, about being immersed in a certain activity to the point of becoming completely unaware of the surrounding environment (item: 2, 14, 15, 17, 18, 20, 24) and *depersonalisation-derealization*, i.e., perception of the Self and the environment, such as the feeling of being disconnected from one's body, one's thoughts, one's feelings (items: 7, 12, 13, 21, 22, 23, 27, 28).

2.3.5.3 Self-Report Symptom Inventory Revised (SCL-90-R)

Self-Report Symptom Inventory Revised (SCL-90-R) is a self-administered scale for the evaluation of psychiatric symptomatology (Derogatis, 1992).
The checklist consists of nine primary symptom dimensions, including: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. For each of the dimensions, the relative score is calculated as an average of the answered questions, and -in general- are considered of interest average scores equal to or greater than 100.

It also includes three global indices of psychological distress:

- Global Severity Index (number of symptoms endorsed and intensity of distress)
- Positive Symptom Distress Index (average level of distress for those items that were endorsed; exaggerating or attenuating response style)
- Positive Symptoms Total (total symptoms endorsed/breadth of distress).

The questionnaire contains 90 items, rated on a five-point scale from 0 (Not at All) to 4 (Extremely) specifying how much each item has bothered the interviewee during the past 7 days. It is designed to be appropriate for use with general individuals, as well as individuals with either medical or psychiatric conditions. This scale can capture three

profiles: global distress scores, elevation in particular symptom dimensions, and at the level of individual symptoms. As a self-report measure, the SCL-90-R provides a subjective report of current distress, and can be administered repeatedly to track changes in symptoms over time.
In Italy, the scale was validated in 2005 and all dimensions showed a high internal consistency measured by a Cronbach's α higher than 0.70. The results of the study seem to highlight some merits but also different psychometric limits of the SCL-90-R and suggest that the instrument would tend to provide an overall measure of the psychological discomfort of the subject, rather than measuring real distinct psychopathological dimensions. A number of studies have highlighted the psychometric weaknesses of the SCL-90-R; in a review of factor analytic studies of the SCL-90-R up to 1985, Cyr et al. (1985) reported relatively weak factorial invariance of the scale across diagnoses, social variables and gender. This lack of factorial invariance seriously questions the stability, factorial content and generalizability of this measure. More recent factor-analytic studies have provided further support for the conclusions drawn by Cyr et al. (1985) in their review paper. In a study of a large sample, Vassend and Skrondal (1999) concluded that different factorial strategies can produce very different results in terms of model selection and pointed out that the complex logical-semantic structure of SCL-90-R items can be the reason for the structural indeterminacy problem.

2.3.5.4 Summary statistics of scores and results about subscales, for the Psyco-pathological scales employed in *PostCovid* study

	mean	**sd**	**range**	**cutoff**	**% > cutoff**	**% > cutoff (on tot)**
PHQ-9	9.35	5.39	$0-27$	11	36%	11%
DES	12.19	12.77	$0-100$	20	22%	6%
SCL-90	0.78	0.60	$0-4$	1	29%	9%

Table 2.6: Second level scales summary statistics

Table 2.6 shows that 36% of participants who filled the second level questionnaire (11% of the total sample) reported depression symptoms, and 29% scored above SCL-90

DES subscales	% > **cutoff**
dissociative amnesia	10%
absorption and imaginative involvement	41%
depersonalisation-derealization	19%
dissociazione patologica	10%

Table 2.7: DES subscales

SCL-90 subscales	% > **cutoff**
Somatization	35%
Obsessive-compulsive	44%
Interpersonal sensitivity	25%
Depression	46%
Anxiety	30%
Hostility	32%
Phobic anxiety	12%
Paranoid ideation	30%
Psychoticism	15%

Table 2.8: SCL-90 subscales

cutoff. Only 22% overpassed the cutoff of DES (see Table 2.6), but almost double of them showed signs of absorption and imaginative involvement (see Table 2.8). Nearly half of the subjects expressed depressive and obsessive-compulsive symptoms, according to subscales of SCL-90, while very few presented phobic anxiety or psychoticism, as shown in Table 2.8.

Chapter 3

An introduction to Item Response Theory

This chapter introduces the fundamental notions concerning Item Response Theory (IRT), analyzing its difference with CTT and illustrates its origins and assumption. Our purpose is to present IRT models for binary and ordinal data and their respective extension to multidimensional case. There are many estimation methods used in this context, which are not reported here. We referred to Reckase (2009) and Bartolucci et al. (2019) for a complete report of such methods. Moreover, we compare multidimensional IRT (MIRT) with factor analysis framework.

IRT was developed in the context of the measurement of latent theoretical constructs. A latent construct is not observable by definition, and it can only be determined indirectly through the use of other manifest variables. Considering educational and psychological fields, where IRT is extensively used, the aim of IRT is to measure the abilities and attitudes of individuals through the responses on several test items.

The power of IRT is that it estimates item characteristics through the modeled item parameters, which permit the calculation of the expected score at the item level (e.g., probability of a yes, or correct answer if responses are binary or dichotomous), and at the test level. In addition, the person's *latent trait* is also estimated, taking into account specific item characteristics and how the person respond to each item.

IRT comprises a set of models whose basic premise is that a person's interactions with test items can be adequately represented according to probabilistic relations, containing a single parameter to describe the person's characteristics. The most straightforward representation we can give of the model is in equation 3.1

$$P(Y = y|\theta) = f(\theta, \boldsymbol{\eta}, y), \tag{3.1}$$

where θ is a parameter which describes the characteristic of the person, $\boldsymbol{\eta}$ represents a vector of parameters that describes the characteristic of the test item, and Y is the score on the test item with possible value y.

3.1 The origins of IRT

IRT finds its origins in the work of Thurstone (1925b), which introduced the fundamentals of IRT in the educational field to measure students' abilities (in the literature, in fact, the latent trait is commonly called *ability*), but its use in the measurement theory field is more recent: the work of Lord and Novick (1968) represents the first formalization of the theory on the basis of ideas and principles that were raised in the thirties and forties. Improvements of IRT were due to the necessity to overtake the shortcomings of the CTT, which does not explicitly model the way respondents with different latent trait levels perform on questionnaire items (Hambleton and Swaminathan, 1985, Hambleton et al., 1991, Van Der Linden and Hambleton, 1997). On the other hand, IRT focuses on items rather than on individual scores, while in the CTT the evaluation of test properties and item characteristics are not included. Moreover, IRT permits evaluating individual ability and describing the performances of the items on the test simultaneously. For these reasons, IRT seemed to be an alternative and a promising method to substitute CTT in psychometric field, showing a wide and effective framework.

Among the first and most relevant contributions to IRT models, we find the works of Richardson (1936), and Ferguson (1942), for the specification of the normal ogive model to describe the relationship between the responses and the person's latent trait, and the work of Lord (1952), who stated the substantial difference between observed test scores and latent traits. Another essential contribution is represented by the work of Rasch

(1960) who developed a family of IRT models (for the dichotomous case), which were later extended by scholars like Andrich (1978), Masters (1982) and Samejima (1969) for the treatment of polytomously scored items. Afterward, the work by Lord and Novick (1968) provided a first unified treatment of CTT and IRT; an important contribution by Birnbaum (1968) defined the two-parameter logistic model for dichotomously scored items.

Some distinctions should be addressed about IRT models. The first distinction is between parametric and nonparametric IRT models. Parametric models are far more popular in practice than nonparametric models, and, consequently, a wide variety of models is in that category. Additionally, different models can be specified depending on:

- the structure of the data: binary or polytomous (nominal or ordinal) responses;
- the number of latent dimensions: unidimensional or multidimensional models;
- the distribution functions used to link responses and ability;
- the number of item parameters introduced in the model.

Concerning the first point, IRT permits us to specify different models depending on the kind of items we are dealing with, i.e. items with two response categories or items with more than two response categories. The second point is a crucial choice in the model specification procedure: when only one ability affects the responses, we are assuming unidimensionality, while when we need two or more latent traits to describe the correlation among the responses, we are assuming multidimensionality. Moreover, the model depends on the probability distribution used to describe the relationship between the response and the respondent's latent trait and the number of parameters describing the item characteristics. The most common probability models used are the normal distribution function (normal ogive models) and the logistic distribution function (logit models). Finally, a distinction can be made with reference to the number of item parameters, one, two or three, introduced in the model.

3.2 Key assumptions

Several key assumptions underlie the original IRT framework, including

- unidimensionality of the measured trait,
- local independence,
- monotonicity.

Assumption of *unidimensionality* states that only one latent trait affects item responses. Evaluation of unidimensionality may be done in different ways. Some researchers suggest using factor analysis to test the one-factor solution. If multiple factors emerge, evidence of a 'dominant' factor (i.e., demonstrating that the first factor accounts for at least 20% of the variance) is needed. Others recommend conducting tests of model fit to determine unidimensionality; if misfit is detected in any item, it may indicate that the item is not closely related to the overall latent trait or that there is a lack of clarity in the item, causing respondents to interpret it differently.

Along with the hypothesis of a single person parameter θ, IRT models require *monotonicity*, which assumes that the probability of endorsing an item increases as individual's trait level θ increases. In other words, monotonicity implies that the conditional probability of responding correctly to item j, $P(Y_{ij} = 1|\theta_i)$ (binary case) or of responding category y or higher $P(Y_{ij} \geq y|\theta_i)$ (ordinal case) is a monotonic non-decreasing function of θ_i.

The third assumption is the so-called *local independence* which assumes independent responses to all test items by all respondents. The term 'local' indicates that responses are assumed independent at the level of an individual person with the same latent trait level θ, i.e., items within a measure should not be related except for the fact that they measure the same underlying trait. More formally, local independence guarantees that responses to a pair of items are statistically independent given the latent ability. Local independence holds when the assumption of unidimensionality is true.

As a consequence of the local independence assumption, denoting with $\mathbf{Y_i} = (Y_{i1}, ..., Y_{iJ})$ the random vector of the responses to J items for individual i, the probability of observing the responses $\mathbf{y_i}$ is the product of the probabilities of the responses y_{ij} to each item

$$P(\mathbf{Y_j} = \mathbf{y_j}|\theta_i) = \prod_{j=1}^{J} P(y_{ij}|\theta_i) = P(Y_{i1} = y_{i1}|\theta_i)P(Y_{i2} = y_{i2}|\theta_i) \cdots P(Y_{iJ} = y_{iJ}|\theta_i). \quad (3.2)$$

The three assumptions described above define a general class of IRT models.

3.3 IRT models for dichotomous items

The three most popular IRT models for dichotomously scored item responses are named according to the number of parameters that model the characteristics of an item and the form of the function. If a single characteristic of the item is modeled (e.g., item difficulty), the IRT model is called the *one-parameter model* (1-PL). If two characteristics of the item are modeled (e.g., item difficulty and item discrimination), the IRT model is called the *two-parameter model* (2-PL). When three characteristics of the item are modeled (e.g., item difficulty, item discrimination, and item pseudo-guessing, which is common for a multiple-choice item test), the IRT model is called *three-parameter model* (3-PL). The mathematical function that relates a person's latent trait or ability score and the expected item score is called the item response function (IRF). Popular choices of IRFs are the logistic distribution form and the normal ogive form.

According to 1-PL model, the probability of answering an item correctly depends on the respondent's ability level and the item difficulty. The difficulty level is thus the only parameter describing the item. In contrast, the discrimination power is assumed to be constant across items (and, more specifically, to be equal to any constant value in the one-parameter logistic model and to 1 in the Rasch model). On the other hand, in the two-parameter logistic (2-PL) model, two parameters are used to describe each item, corresponding to the difficulty level and the discrimination power. Finally, the three-parameter logistic (3-PL) model adds a pseudo-guessing parameter for each item, which defines the lowest horizontal asymptote for the probability of endorsing an item, even in the case of respondents with a very low ability level, as a result of guessing.

The 1-PL model (Rasch model)

The simplest case is a model with one parameter describing person characteristics and one parameter describing item characteristics. Called y_{ij} the score given by person i on item j (in this case 0 or 1), starting from (3.1) the model can be represented by

$$P(Y_{ij} = y_{ij}|\theta_i) = f(\theta_i, b_j, y_{ij}). \tag{3.3}$$

Going through its origin, for dichotomously scored test items, Rasch (1960) proposed a very simple function that relates the parameters to the probability of answering correctly ensuring the monotonicity property:

$$P(Y_{ij} = 1|A_i, B_j) = \frac{A_i B_j}{1 + A_i B_j} \tag{3.4}$$

where A_i and B_j are single person and item parameter respectively (now labeled θ_i and b_j). The model (3.4) in still frequently seen in psychometric literature even if a model based on a logarithmic transformation of the scales of the parameters (Fischer, 1995) is used:

$$P(Y_{ij} = 1|\theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}} = \Psi(\theta_i - b_j) \tag{3.5}$$

where Ψ is the cumulative logistic density function. Model (3.5) followed from (3.4) by logarithmic transformation of the parameters $\theta_i = \ln(A_i)$ and $b_j = -\ln(B_j)$. The scale of $th-$person parameter in (3.5) ranges from $-\infty$ to $+\infty$ rather than $0 - \infty$ for model (3.4). Item parameter has the same scale, with reverse direction: large values of B_j parameter indicate easy items and small values of b_j indicate easy items. For this reason, b_j parameter is called *difficulty parameter* while B_j is called *easiness parameter.*

Because it has only one item parameter and due to the use of the logistic density function, this model is called *one-parameter logistic IRT model.* Alternatively, it is noted as *Rasch model*,to acknowledge its original author.

Both θ_i and b_j are measured on the same scale and lie on $\mathbb{R}$. Thus, their values can be compared: if $\theta_i = b_j$ then $P(Y_{ij} = 1) = 0.5$, if $\theta_i > b_j$ then $P(Y_{ij} = 1) > 0.5$ and if $\theta_i < b_j$ then $P(Y_{ij} = 1) < 0.5$. The item difficulty represents the level of latent trait for which one has the 50% probability of responding correctly to that item. The item difficulty has a different meaning in the IRT framework and in the CTT setting. In the first case, it is a location parameter, as it identifies the point on the latent continuum at which the individual's latent trait is located. On the contrary, in CTT, the item difficulty

corresponds to the relative frequency of individuals that endorsed a certain binary item.

The 2-PL model

In the 2-PL model, a parameter is added to the Rasch model, the *discriminating parameter* for item j, to estimate the capacity of the item to distinguish between subjects with different latent trait levels. The equation to estimate the probability of person i providing a correct response given his ability θ_i, the item difficulty b_j, and item discrimination λ_j is

$$P(Y_{ij} = 1|\theta_i) = \frac{e^{\lambda_j(\theta_i - b_j)}}{1 + e^{\lambda_j(\theta_i - b_j)}}. \tag{3.6}$$

The Rasch model can be seen as a 2-PL where all items have the same discriminating parameter equal to 1. In other words, an important characteristic of the Rasch model is that all items are assumed to have the same discriminating capacity. From a practical point of view, this means that the ranking of item difficulty does not depend on the value of θ_i.

The 3-PL model

The 3-PL model includes the 2-PL model with an additional guessing parameter. The probability that person i, having an ability of θ_i, will score 1 (rather than 0) on item j, given the item's difficulty b_j, discrimination λ_j, and guessing parameter γ_j, is

$$P(Y_{ij} = 1|\theta_i) = \gamma_j + (1 - \gamma_j)\frac{e^{\lambda_j(\theta_i - b_j)}}{1 + e^{\lambda_j(\theta_i - b_j)}} \tag{3.7}$$

The guessing parameter γ_j describes the probability of answering correctly to item j when the individual guesses (i.e., when θ_i is very small). The 3-PL model is mainly used in the educational field.

3.4 IRT models for polytomous items

The majority of applications of IRT is in the analysis of data from test item dichotomously scored. Nevertheless, there has been a lot of development work on items with more than two score categories. Most contributions have focused on three models which

are illustrated here: the *partial credit model* (Masters, 1982), the *generalized partial credit model* (Muraki, 1982) and the *graded response model* (Samejima, 1969).

We assume that item j has l_j categories, from 0 to l_{j-1} and we denote for simplicity $p_{jy}(\theta_i)$ the probability that subject i with latent trait θ_i answers by category y to item j.

Denoting by $g_y[\cdot]$ the link function specific of category y we can give this general model formulation for IRT models for polytomous responses:

$$g_y[p_j(\theta_i)] = \lambda_j(\theta_i - b_{jy})$$

where b_{jy} is named *treshold difficulty parameter*.

Some example of link function are *global logits*

$$g_y[p_j(\theta_i)] = \log\frac{P(Y_{ij} \geq y|\theta_i)}{P(Y_{ij} \geq y-1|\theta_i)} \qquad y = 0, ..., l_j - 1 \tag{3.8}$$

or *local logits*

$$g_y[P_j(\theta_i)] = \log\frac{P(Y_{ij} = y|\theta_i)}{P(Y_{ij} = y-1|\theta_i)} \qquad y = 0, ..., l_j - 1. \tag{3.9}$$

The *Partial Credit Model* (PCM) is a generalization of the Rasch model. It is a two-level ordinal logistic model with local logit as link function. The model formulation is

$$\log\frac{p(Y_{ij} \geq y|\theta_i)}{p(Y_{ij} < y|\theta_i)} = \theta_i - b_{jy}, \quad j = 1, ..., J, \;\; y = 1, ..., l_j - 1. \tag{3.10}$$

The probability of scoring y on item j ($y = 0, 1, ..., l_j - 1$) is given by

$$p_{jy}(\theta_i) = \frac{e^{\sum_{h=1}^{y}(\theta_i - b_{jh})}}{\sum_{m=0}^{l_j-1} e^{\sum_{h=1}^{m}(\theta_i - b_{jh})}} = \frac{e^{y\theta_i - \sum_{h=1}^{y} b_{jh}}}{1 + e^{m\theta_i - \sum_{h=1}^{m} b_{jh}}}. \tag{3.11}$$

The difficulty parameter b_{jy} here indicates the difficulty of choosing response category y with respect to choosing category $y-1$, i.e. if $\theta_i = b_{jy}$ then $p_{jy}(\theta_i) = p_{j,y-1}(\theta_i)$, if $\theta_i > b_{jy}$ then $p_{jy}(\theta_i) > p_{j,y-1}(\theta_i)$ and if $\theta_i < b_{jy}$ then $p_{jy}(\theta_i) < p_{j,y-1}(\theta_i)$.

The *Graded Response Model* (GRM) is considered the generalization of 2-PL model; it is a two-level ordinal logistic model with global logit as link function. For each item we have two parameters: a difficulty parameter b_{jy} for each threshold of each item and a discrimination parameter λ_j. The model formulation is the following

$$\log\frac{P(Y_{ij} \geq y|\theta_i)}{P(Y_{ij} < y|\theta_i)} = \lambda_j(\theta_i - b_{jy}), \quad j = 1, ..., J, \;\; y = 1, ..., l_j - 1 \tag{3.12}$$

The probability of scoring y on item j is therefore

$$p_{jy}(\theta_i) = p^*_{jy}(\theta_i) - p^*_{j,y+1}(\theta_i) \tag{3.13}$$

where

$$p^*_{jy}(\theta_i) = \frac{e^{\lambda_j(\theta_i - b_{jy})}}{1 + e^{\lambda_j(\theta_i - b_{jy})}}. \tag{3.14}$$

with $p^*_{jl_j}(\theta_i) = 0,\ p^*_{j0}(\theta_i) = 1$.

The difficulty parameter b_{jy} represents the difficulty of choosing response category y or higher with respect to choosing category $y-1$ or smaller. Difficulty parameters are always ordered $b_{j1} < ... < b_{j,l_j-1}$ and their interpretation is the following: if $\theta_i = b_{jy}$ then $p^*_{jy}(\theta_i) = 1 - p^*_{j,y-1}(\theta_i)$; if $\theta_i > b_{jy}$ then $p^*_{jy}(\theta_i) > 1 - p^*_{j,y-1}(\theta_i)$ and if $\theta_i < b_{jy}$ then $p^*_{jy}(\theta_i) < 1 - p^*_{j,y-1}(\theta_i)$.

3.5 Graphical representation

In IRT frameworks, graphical representation is a very common approach and it is very useful for exploring item characteristics. The most typical graphs included in a IRT-based analysis are the following:

Item Characteristic Curve

The Item Characteristic Curve (ICC) depicts the probability of a correct response as a function of θ. The ICC is a probability curve that is monotonic, or continuously increasing, in nature. As an individual's trait level increases, the probability of endorsing an item also increases. The variable used to express the individual's underlying trait level θ

is measured along the x-axis. Higher values of θ are associated with greater levels of the underlying trait. The y-axis indicates the probability of endorsing an item and is scaled from 0 to 1. When an item has polytomous response options, the interpretation of ICCs is slightly different in that the ICC plots the expected item score over the range of the trait. To depict the probability of endorsing each response category for a polytomous item, Categorical Response Curves (CRCs) can be plotted, one curve for each response category.

Information function

The concept of *information* is used in IRT to reflect how precisely an item or scale can measure the underlying trait. Greater information is associated with higher measurement accuracy. Information is inversely related to the standard error of the estimate, so, at any level of θ, greater information will result in a smaller standard error associated with the estimated θ score. In IRT, the interest is typically in estimating the value of the ability parameter θ_i for an individual. If the amount of information is small, it means that the ability cannot be estimated with precision, and the estimates will be widely scattered around the true ability. Since, in IRT, each item of a test contributes to measuring the underlying latent trait, items also contribute to defining the precision of the measurement for each person. In particular, the *item information curve* (or *item information function*) expresses the information arising from a single item against the ability θ_i and is denoted as $I_j(\theta_i)$

$$I_j(\theta_i) = \sum_{y=0}^{l_j-1} I_{jy}(\theta_i) p_{jy}(\theta_i) \tag{3.15}$$

with the information associated with a specific response category y of an item j given by

$$I_{ij}(\theta_i) = - \qquad \frac{\partial^2 \log p_{jy}(\theta_i)}{\partial \theta_i^2}. \tag{3.16}$$

Increasing the number of response categories will result in an increase of the item information; thus, a polytomous IRT model will estimate more precisely the ability levels than a dichotomous IRT model.

The information produced by a single item is useful to evaluate the quality of this item. In general, we are more interested in evaluating the test as a whole: due to the local independence assumption, the overall information yielded by the test at any ability

level is simply the sum of the item information functions at that level (Birnbaum, 1968; Timminga and Adema, 1995). Thus, the *test information curve* (or *test information function*) is defined as

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta) \tag{3.17}$$

Since the test information function is calculated by summing up all the single item information functions, then the order of the items does not influence the total information. In addition, it is obvious that longer tests will usually measure an individual's ability with greater precision than shorter tests.

3.6 Using IRT for detecting DIF

An item is said to exhibit differential item functioning (DIF) when examinees from different groups who have the same ability have different response probabilities for that item. DIF items can lead to biased measurement of ability because the measurement is affected by so-called nuisance factors. The existence for DIF for item k can be investigated by the following model

$$\eta_{ij} = \text{logit}[P(Y_{ij} = 1)|\theta_i, \lambda_j, \beta_j)] = \lambda_j(\theta_i - \beta_{ij}) \tag{3.18}$$

where $\beta_{ij} = \beta_j + \delta_k(I_{j=k} \times x_i)$, θ_i is the ability for person i, λ_j is the item discrimination parameter and β_j the item difficulty parameter for item j, δ_k is the DIF parameter for item k, $I_{j=k}$ is an indicator variable for item k which takes value 1 when $j = k$ and x_i is a dummy variable for the group classes. Under this model δ_k is the item difficulty difference for item k between groups.

The most commonly used methods to assess DIF assume the data in analysis are unidimensional, and they can be classified into two general approaches (Magis et al., 2011). One approach, called *the IRT approach*, superimposes an IRT model on the provided data. The Wald test (Wald, 1943), the likelihood-ratio test (Thissen and Steinberg, 1988), and the latent class logistic regression (Zumbo et al., 2015) follow this approach. The other approach, called *the observed-score approach*, does not require an IRT model. The Mantel-Haenszel procedure (Mantel and Haenszel, 1959), the logistic regression (Swaminathan

and Rogers, 1990), and SIBTEST (Shealy and Stout, 1993) fall into the second category. For the IRT approach, there are two subcategories of methods; one focuses on identifying differences in estimated item parameters between groups, whereas the other on identifying differences in the estimated areas between item response function curves from different groups (Kim et al., 1995). A thorough review of these methods can be found in Lee et al. (2015).

It is worth noting that these methods are either originally designed to compare two groups (e.g., Swaminathan and Rogers 1990), or have not yet established the evidence that they apply to more than two groups (e.g., De Boeck (2008)). Five recent DIF methods have been demonstrated with acceptable performance for analyzing DIF across multiple groups in the unidimensional framework. They are the generalized Mantel-Haenszel method (Penfield, 2001), the generalized Lord's χ^2 test or Wald test (Kim et al., 1995) with the Wald-1 linking algorithm (Penfield, 2011), the generalized logistic regression procedure (Kim et al., 1995), the multiple-indicator multiple-cause modeling (MIMIC; Muthén 1985), and the multiple-group confirmatory factor analysis (MGCFA; Asparouhov and Muthén 2014).

3.7 IRT and the Latent Class approach

The IRT models are compatible with the assumption that the population under study is composed of homogeneous classes (or sub-populations) of individuals who have very similar unobservable characteristics. This is the formulation adopted in the Latent Class (LC) model (Bartolucci et al., 2019).

In some situations, this assumption is particularly convenient, as it allows to cluster individuals. Moreover, this assumption implies several advantages for the estimation process. Some example of models with dichothomous data are in Rost (1990), Lindsay et al. (1991), Formann (1995), Rost and von Davier (1995), Bartolucci (2007), while Langeheine (1988), and Heinen (1996) present a critical discussion about discretized variants of IRT models.

IRT models may be expressed as latent class models under the discreteness assumption for Θ_i, i.e. every random variable θ_i, $i = 1, ..., n$, is assumed to have a discrete distri-

bution with support points $\xi_1, ..., \xi_k$ and corresponding weights $\pi_1, ..., \pi_k$. Each weight $\pi_v,\ v = 1, ..., k$ represents the probability that a subject belongs to class v

$$\pi_v = p(\Theta_i = \xi_v), \quad \text{where} \sum_v \pi_v = 1;\ \pi_v \geq 0 \tag{3.19}$$

The number of latent classes can be assumed a priori, on the basis of theoretical knowledges or substantial reasons, or selected by comparing the fit of the model under different values of k. Individuals do not differ within latent classes, as the same latent trait level ξ_v is assumed for all individuals in class v. Item parameters β_j, λ_j, γ_j may be assumed to be constant or variable across classes.

3.8 Multidimensional IRT

Multimensional Item Response Theory (MIRT) is an extension of the unidimensional IRT models that seeks to explain an item response according to an individual's standing across multiple latent dimensions (Reckase, 2009). As already mentioned in Chapter 3, a key limitation of unidimensional models is that they may not be appropriate to commonly used multidimensional instruments. A questionnaire is often composed by some subsets of items measuring different but potentially related constructs. In such a case, the traditional IRT assumption of only one underlying latent variable may be too restrictive. In fact, the unidimensional approach has the clear disadvantage of ignoring the differential information about individual ability levels. Consequently, developments in MIRT provide an opportunity for examining the psychometric functioning of some scales. This is particularly relevant given the complexity of the constructs considered in psychology that help explain how an individual responds to an item.

In MIRT approach each individual is characterized by a vector $\boldsymbol{\theta_i} = (\theta_{i1}, ..., \theta_{im})$ of latent abilities, where m is the number of latent dimensions measured by a generic item, and this is what is new in contrast to the unidimensional case. MIRT models permit, in fact, to separate inferences concerning each distinct latent dimension of a respondent, by introducing a personality trait and item discrimination parameters for each ability measured by a test item.

Several authors have dealt with multidimensional extensions of traditional IRT models. Examples of such generalizations are the log-linear multidimensional IRT model, developed by Duncan and Stenbeck (1987), and Kelderman (1997); the multidimensional normal ogive graded response model (GRM) introduced by Muraki and Carlson (1995), the multidimensional partial credit model by Kelderman (1996) and Yao and Schwarz (2006). Multidimensional models based on a discrete latent class analysis are in Bartolucci (2007) and Bacci et al. (2014).

According to the famous work of Reckase (2009), that is probably the main reference for MIRT, there are two broad categories of MIRT models: *compensatory* and *non-compensatory*. Compensatory models allow examinees' increased standing on one latent trait to overcome a low position on another dimension in the estimation of a probability of a correct item response. In other words, a lack in one trait naturally compensates for the other. Non-compensatory (or, partially compensatory) models restrict examinees' standings across the multidimensional space so as not to influence the probability of an item response. Within the literature, compensatory-based MIRT models are more the commonly used.

3.9 MIRT models formulation

MIRT represents a broad class of probabilistic models designed to characterize an individual's likelihood of an item response based on item parameters and multiple latent traits. The basic form of the models, using the same notation of its unidimensional form in Chapter 3, is

$$P(Y_j = y|\boldsymbol{\theta}) = f(y, \boldsymbol{\theta}, \boldsymbol{\eta_j}), \tag{3.20}$$

where Y_j is the score on the item j with the corresponding assigned value y, $\boldsymbol{\theta}$ is the vector of parameters describing the location of the person in the multidimensional space and $\boldsymbol{\eta}$ is the vector of item characteristics.

In particular, MIRT situates an individual's standing on the latent traits in a multidimensional space of the dimensions hypothesized to be associated with an item response.

In the next sections we extend the unidimensional models shown in Chapter 3 in the multidimensional framework, for both dichotomous and polytomous scored items.

3.9.1 MIRT models for dichotomously scored items

Multidimensional extension of the 2-PL model

$$P(Y_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{a_j}, b_j) = \frac{e^{\boldsymbol{a_j}\boldsymbol{\theta_i'}+b_j}}{1 + e^{\boldsymbol{a_j}\boldsymbol{\theta_i'}+b_j}} \tag{3.21}$$

where

$$\boldsymbol{a_j}\boldsymbol{\theta_i'} + b_j = a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + \cdots + a_{jm}\theta_{im} + b_j = \sum_{l=1}^{m} a_{jl}\theta_{il} + b_j.$$

The $\boldsymbol{a}$ parameter is usually called the *slope* or *discriminatory parameter,* as it indicates the orientation of the equiprobable contours and the rate that the probability of correct answer changes from point to point in the $\theta-$space. This can be seen taking the first derivative of the expression (3.21) with respect to dimension θ_l :

$$\frac{\partial P}{\partial \theta_l} = a_l P(1 - P),$$

with $P = P(Y_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{a_j}, b_j)$.

The b parameter is called the *intercept parameter*, and it is not a difficulty parameter in the usual sense of a unidimensional IRT model because it does not give a unique indicator of the difficulty of the item. Instead, the quantity $-\frac{b_i}{a_l}$ (the point where the line intersects the θ_l if all of the elements of $\boldsymbol{\theta}$ are equal to 0 except θ_l) gives the relative difficulty of the item related to the corresponding coordinate dimension l.

Multidimensional extension of the 3-PL model

A straightforward extension of the multidimensional 2-PL model allows a non-zero lower asymptotic to the model giving the multidimensional 3-PL model:

$$P(Y_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{a_j}, b_j, \gamma_j) = \gamma_j + (1 - \gamma_j)\frac{e^{\boldsymbol{a_j}\boldsymbol{\theta_i'}+b_j}}{1 + e^{\boldsymbol{a_j}\boldsymbol{\theta_i'}+b_j}}. \tag{3.22}$$

3.9.2 MIRT models for polytomously scored items

Multidimensional GPCM

The multidimensional extension of the Generalized Partial Credit (MGPM) model describes the interaction of persons with items that are scored with more than two categories represented by $k = 0, 1, ..., K_j$. Called β_{jy} the threshold parameter for score category y,

$$P(Y_{ij} = k|\boldsymbol{\theta_i}) = \frac{e^{k\boldsymbol{a_j}\boldsymbol{\theta_i'} - \sum_{y=0}^{k}\beta_{jy}}}{\sum_{v=0}^{K_j} e^{v\boldsymbol{a_j}\boldsymbol{\theta_i'} - \sum_{y=0}^{v}\beta_{jy}}}.$$

Multidimensional PCM

The general form of the multidimensional extension of the Rasch model to the polytomous test item case is presented in Kelderman and Rijkes (1994) specifying the matrix of weights W_{jlk}

$$P(y_{ij} = k|\boldsymbol{\theta_j}) = \frac{e^{\sum_{l=1}^{m}(\theta_{il} - b_{jlk})W_{jlk}}}{\sum_{r=0}^{K-i} e^{\sum_{l=1}^{m}(\theta_{il} - b_{jlr})W_{jlr}}}, \tag{3.23}$$

where b_{jlk} is the difficulty parameter and W_{jlk} is the weight for item j on dimension l for score category k.

3.10 MIRT and Factor Analysis

Although many researchers believe that psychological tests measure multiple constructs, MIRT modeling is still not the prevalent methodology used in this field. Therefore, this theory has great potential in psychological assessment.

MIRT analysis can help ensure that test scores are properly used and interpreted. If the test results are reported as a single score, then it is implicitly assumed that all the items are measuring the same trait or same composite of traits. Thus, dimensionality analyses can help establish the degree to which this is true.

According to Reckase (2009), MIRT is an outgrowth of the two methodologies, unidimensional IRT and factor analysis, as shown in Figure 3.1. Then, one alternative approach for extending the unidimensional response model to the multidimensional case is to use

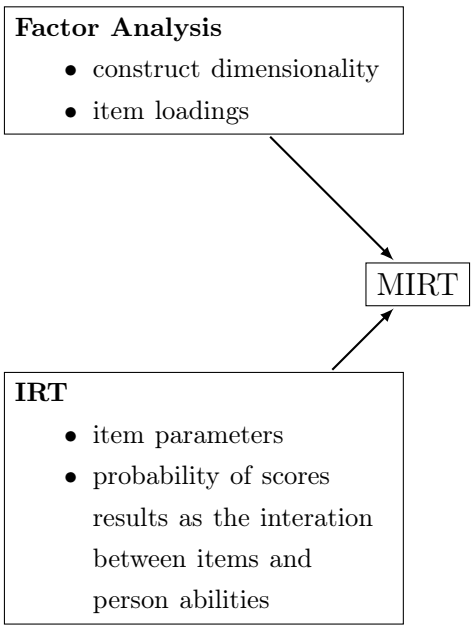

Figure 3.1: MIRT as a combination of IRT and FA

the factor analytic approach. In this section, the relationship between these two theories is illustrated, highlighting their similarities and differences and formulating the equivalence in terms of models.

3.10.1 Differences

Factor analysis and MIRT have virtually very similar statistical formulations when they are applied to item responses and item parameters can be expressed through factor loadings. Even if the statistical procedures are the same, the two methodologies have some important differences, that are here illustrated.

The first difference is that analyzing a correlation matrix in factor analysis is basically the same as analyzing standardized variables with zero mean and standard deviation equal to 1. In the MIRT approach, differences in item scores provide relevant information about test items, so variables are not standardized. MIRT is intended as a special form of non-standardized factor analysis for the item scores, conceived as type of discrete variables. This is detailed in McDonald (2000).

Second, MIRT can overcome the issues that arise when applying the factor analy-

sis method on discrete data. In fact, MIRT was designed for discrete transformations of continuous variables (yielding dichotomous or polytomous variables); in this field, the traditional factor analysis can be problematic, especially when the usual Pearson correlation is used.

The third difference is that exploratory factor analysis aims to identify the minimum number of factors to describe the relationship between data. Instead, the goal of MIRT is to accurately represent the relationship between the probability of response and the location in the multidimensional space, without emphasizing to minimize the number of dimensions.

The relationship between these two methodologies was analyzed by various works, like McDonald (1982), McDonald (2000), Wherry and Gaylord (1944) and summarized by Reckase (2009).

According to Kamata and Bauer (2008), MIRT is a fusion of factor analysis and IRT (Figure 3.1). Although factor analysis and IRT share common ground in that some parameterizations of model parameters in factor analysis can be transformed into parameters in item response theory, the underlying philosophy of IRT is vastly different from that of factor analysis, which belongs to CTT. For example, the psychometric attributes of an instrument yielded from factor analysis attach to the entire scale. In contrast, each item developed by IRT has its own characteristics, described by the difficulty parameter, discrimination parameter, guessing parameter, and item information function). In addition, when responses are dichotomous instead of ordinal (e.g., Likert scale ratings), conventional factor analysis utilizing Pearson's correlation matrix is invalid.

The difference in terms of fit is analyzed in Reckase (2009). Unlike traditional factor analysis which identifies factors as eigenvectors of the correlation matrix, IRT introduces a statistical model which is then fit to the observations. The model is used to calculate the likelihood function L, and Maximum Likelihood estimation techniques are often used for parameter estimation. A number of general model fit statistics have been developed from this, such AIC, BIC, Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI).

IRT is often thought of as a type of CFA. CFA assumes that the items are continuous and multivariate normal; Item response theory is simply a non-linear CFA with appropri-

ate binary or ordinal response functions. Some authors suggest a complementary use of both approaches: it's the case of the recent work in Bean and Bowen (2021), which gives an example of merging CFA and IRT techniques in the educational field. The mentioned article demonstrates the advantages of using both CFA and IRT in the development and evaluation of scales with dichotomous or polytomous items. According to the authors, most researchers commonly employed CFA and internal consistency reliability tests to validate scales, while IRT has been underused. They report findings from CFA and IRT analyses to demonstrate that scale development and validation can benefit from the complementary use of the two methods, each one contributing valuable information.

3.10.2 IRT interpretation of FA

There are formal similarities between IRT (both unidimensional and multidimensional) and Factor Analysis. They were investigated and shown by different contributions, with the aim of establishing a common statistical framework to unify these two methodologies. In particular, Takane and De Leeuw (1987), and Knol and Berger (1991) illustrated the relationship between the normal-ogive IRT model and the factor analysis model in the dichotomous case, showing that multidimensional normal-ogive model can also be derived from modeling assumptions born in factor analysis. Following this formulation, Glockner-Rist and Hoijtink (2003) underlined how both models could be combined to analyze the structure of item responses. Although this affinity was proved in several ways, the diffusion of MIRT application in the psychological framework is still limited, while traditional exploratory and confirmatory factor analysis are increasingly preferred. Kamata and Bauer (2008) discuss the relationship among several alternative parametrizations of the binary factor analysis model and the 2-PL model. Morevover, Muraki and Carlson (1995) developed a MIRT model for polytomously scored items, on the basis of graded response model in the factor analysis context.

In this section, we formulate a sort of algebraic equivalence of IRT models to factor analysis, and we convert FA model parameters to IRT parameters under possible parametrizations. Like the factor model, a MIRT model can provide the basis for either exploratory or confirmatory investigations of a test's latent dimensional structure. More importantly, because MIRT assumes categorically scored variables, and is nonlinear in

form, it often provides a better approximation to the interaction between items and examinees in comparison to the common factor model, which assumes continuous variables and is linear. That is why MIRT is often presented as an outgrowth of the two lines of research.

We start from the formulation of the traditional factor analytic model. Factor Analysis, given a vector of continuous variables $\mathbf{Y} = (Y_1, ..., Y_k)$, assumes that the covariance structure of the Y_j is explained by the common factor model. For each item $j = 1, ..., k$

$$Y_j = \mu + \lambda_{j1}\theta_1 + \cdots + \lambda_{jm}\theta_m + \epsilon_j \tag{3.24}$$

where ϵ_j is a residual term, assumed to be normally distributed $\mathcal{N}(0, \psi_j)$ and $\boldsymbol{\lambda}_j = (\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jm})$ is the factor loadings vector.

It follows that

$$Y_j \sim \mathcal{N}(\mu, \Sigma),$$

with

$$\Sigma = \Lambda\Lambda' + \Psi.$$

Λ is the matrix of factor loadings λ_{jm} and Ψ is the diagonal matrix with residual variances.

In matrix form, assuming for simplicity that each Y_j has mean equal to 0, (3.24) becomes

$$\mathbf{Y} = \Lambda\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\theta}$ and $\boldsymbol{\epsilon}$ independent of each other.

An identification constraint is necessary here, according to what suggested in Takane and De Leeuw (1987) and Glockner-Rist and Hoijtink (2003).

Since $\boldsymbol{Y} \sim \mathcal{N}(0, \Lambda\Lambda + \Psi)$ and $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, I)$, the conditional distribution $\mathbf{Y}|\boldsymbol{\theta}$ is multivariate normal with vector mean $\boldsymbol{\mu}^*$ and covariance matrix Σ^* given by

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y + \text{cov}(\mathbf{Y}, \boldsymbol{\theta})I(\theta - \boldsymbol{\mu}_Y) = \Lambda\theta$$

and

$$\Sigma^* = \Lambda'\Lambda + \Psi - \Lambda'\Lambda = \Psi$$

where Ψ is a diagonal matrix whose diagonal elements are the ψ_j.

In the case of dichotomous item responses, Y_j is considered as an unobserved continuous variable; we assume that each item score is determined by the location of Y_j relative to a fixed item threshold parameter τ_j :

$$X_j = \begin{cases} 1, & \text{if } Y_j > \tau_j. \\ 0, & \text{otherwise.} \end{cases} \tag{3.25}$$

It follows that, called Z the standard normal distribution

$$\begin{aligned} P(X_j = 1|\boldsymbol{\theta}) &= P(Y_j > \tau_j) \\ &= \int_{\tau_j}^{\infty} f(Y_j|\boldsymbol{\theta}) dY_j \\ &= \frac{1}{\sqrt{2\pi}\sigma_j^*} \int_{\tau_j}^{\infty} e^{-\frac{1}{2}\left(\frac{Y_j - \mu_j^*}{\sigma_j^*}\right)^2} dY_j \\ &= P\left(Z_j > \frac{\tau_j - \mu_j^*}{\sigma_j^*}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\frac{\tau_j - \mu_j^*}{\sigma_j^*}}^{\infty} e^{-\frac{Z_j^2}{2}} dZ_j \\ &= \Phi\left(\frac{\mu_j^* - \tau_j}{\sigma_j^*}\right), \end{aligned}$$

where

$$\mu_j^* = \boldsymbol{\lambda}_j\theta, \quad \sigma_j^* = \sqrt{\psi_j} = \sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j}.$$

Transforming

$$\delta_j = \frac{-\tau_j}{\sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j}}, \quad \beta_{jm} = \frac{\boldsymbol{\lambda}_j\theta}{\sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j}} = \frac{\lambda_{jm}}{\sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j}},$$

we obtain the multidimensional two-parameter normal-ogive model

$$P(X_j = 1|\theta) = \Phi(\beta_{jm}(\theta - \delta_j)). \tag{3.26}$$

.

The relationship between the logistic distribution function L and the cumulative standard normal distribution F has been proven by Lord and Novick (1968) and it can be expressed as

$$|F(z) - L(1.7z)| < 0.01 \text{ for all } z$$

Therefore, the item parameters in the normal-ogive IRT model can be transformed to the corresponding parameters in the logistic IRT model by multiplying them by a scaling factor of 1.7. This is why the logistic IRT model is often expressed as

$$P(X_{ij} = 1|\theta) = \frac{e^{1.7\boldsymbol{\lambda_j}(\boldsymbol{\theta_j} - b_j)}}{1 + e^{1.7\boldsymbol{\lambda_j}(\boldsymbol{\theta_j} - b_j)}}. \tag{3.27}$$

Through relation (3.27), it is possible to easily express the equivalence between factor analysis and MIRT in (3.26) for logistic 1-PL, 2-PL, and 3-PL IRT models.

Parameters $\boldsymbol{\lambda_j}$ and b_j have the same role in the logistic models as they do in the normal-ogive models.

Chapter 4

GHQ-12 assessment using IRT

The frequent use of GHQ-12 in different cultures and the different interpretations of what the scale measures, motivated us for a psychometric analysis to clarify the quality of the instrument, within the purposes of *PostCovid* study, and also to give our contribution to some open questions in the literature (e.g., the discussed dimensionality of GHQ-12). We have seen in the previous chapters that there are many different methods that can be used to assess the psychometric properties of a survey; for example, reliability coefficients illustrated in Chapter 2 can be used to describe how well each item relates to the total score. In addition, methods such as exploratory or confirmatory factor analysis can be used to test hypotheses about the dimensionality of the survey. Although these relatively straightforward techniques are helpful in assessing the composition of a new survey, it is clear from Chapter 3 that the use of IRT-based methods can provide more details about individual survey items. Nevertheless, a recent study about the assessment of health surveys reviewed four prominent journals within the field of Health Psychology founding that IRT-based models were used in less than 10% of the studies examining scale development or assessment, see Depaoli et al. (2018).

Therefore, the aim of this chapter is to perform an IRT-based analysis on GHQ-12, investigating potential benefits of such approach. Our analysis is limited to the dichotomously scored GHQ-12 (i.e. items have four possible answers but the first two response categories and the last two response categories are collapsed respectively in 0 and 1) because that was the choice of clinicians who composed the protocol. Answers are sum-

marized in two possible responses: *less than usual* and *more than usual*. Six items are positive phrased (i.e. answer equal 1 corresponds to *more than usual* category) and six are negative phrased (the reverse).

Section 4.1 presents the results of an IRT analysis based on 2-PL model and reports findings, through the latent class approach, on the possible use of GHQ-12 alone (i.e. without IES-R and GAD-7 in the first-level screening) to determine final outcomes of the evaluation.

In Section 4.2 we perform DIF analysis dividing the population in subgroups, according to involvement in COVID-19 units (Yes/No) and according to enrollment time (before or after vaccination campaign) to investigate differences among such categories.

In Section 4.3 we assess GHQ-12 dimensionality performing both traditional techniques and IRT methods, underling differences and similitarities.

4.1 First application: IRT model and LC on GHQ-12

In this Section we performed an IRT analysis on the GHQ-12 dichotomously scored scale, for investigating the relationship between an individuals' response to test item and their *performance* on the overall measure of the trait that item was intended to measure.

4.1.1 Model choice

The three fundamental logistic models in the IRT tradition for scoring dichotomous data presented in Chapter 3 are the one-parameter logistic (1-PL), the two-parameter logistic (2-PL), and the three-parameter logistic (3-PL) models. The simplest IRT model is the Rasch model or 1-PL model: it estimates only one parameter (difficulty) for each item. Within the Rasch model, however, we can hold the discriminatory ability of each item constant at 1, or we can estimate such parameter (different from 1, and constant across all items), while in the 2-PL model the discrimination parameter is estimated. The 1-PL model is a sub-model of 2-PL. A well-known statistical test procedure for the comparison of the two nested models is the likelihood ratio test. For this, one needs to fit a reduced model and a full model. The reduced model is the one that can be found by constraining some parameters of the full model. When testing one against the other, the

null hypothesis is that the reduced model (e.g., in the present case, the 1-PL) underlies the data while the alternative hypothesis is that the full model, the 2-PL, offers a better fitting model. The likelihood ratio test is conducted using the *anova* function. Results are in Table 4.1: the p-value of the likelihood ratio test is less than 0.001; thus, we reject the null hypothesis of the 1-PL in favor of the 2-PL. The output also shows two information criteria, AIC and BIC. The values of the information criteria for the 2-PL are smaller than those for the 1-PL, again supporting the 2-PL for this data set.

	AIC	BIC	log.Lik	LRT	df	p-value
1-PL (with discrim=1)	10241.47	10300.15	-5108.74			
1-PL	9681.75	9745.32	-4827.88	561.72	1	<0.001
2-PL	9554.13	9671.48	-4753.07	149.62	11	<0.001

Table 4.1: Comparison between 1-PL and 2-PL models

Item parameters estimation (difficulty and discriminatory) of such 2-PL model is collected in Table 4.2.

	β_j	λ_j
Item 1 (able to concentrate)	0.95	1.90
Item 2 (lose sleep)	0.50	2.42
Item 3 (play useful part)	1.94	1.18
Item 4 (capable of making decisions)	1.58	1.57
Item 5 (constantly under strain)	0.00	2.25
Item 6 (overcome difficulties)	0.96	2.68
Item 7 (enjoy day-to-day activities)	0.21	1.79
Item 8 (face up problems)	1.01	2.70
Item 9 (unhappy and depressed)	0.57	3.61
Item 10 (losing confidence)	1.12	3.00
Item 11 (worthless person)	1.73	2.42
Item 12 (reasonably happy)	0.71	4.16

Table 4.2: Estimated Item parameters (2-PL model)

As we could expect from descriptive analysis, we observe that Item 5 (followed by Item 7) is the *easiest* item, while Item 3 (feeling useless) and Item 11 (thinking of yourself as a worthless person) are the most *difficult* ones. We are using here parameter names as they come from the IRT literature, but we specify that in our framework the meaning

of *difficult* is not related to a *correct* response to an item. It is related to the severity of the psychological distress that is related to a positive answer to an item. For example, Item 5 has difficulty parameter equal to 0, meaning that a person with a latent trait θ at level 0 has 50% of probability of answering *more than usual* to Item 5. On the contrary, a level of latent trait θ equal to 1.73 (much higher) is needed for having the same probability to answer *less* or *more than usual* to Item 11. That is why the difficulty parameter is called also *location parameter* or *threshold* parameter. Item 5 and Item 7 were respectively about feeling constantly under strain and being able to enjoy day-to-day activities; it is not a surprising finding that it is not *difficult* for HCW in pandemic period answering *more than usual*. Concerning the discrimination parameter, Item 12 has the highest value, much greater than others; Item 12 is in fact a sort of *summarizing* final question (feeling reasonably happy, all things considered). On the other hand, Item 3 has the lowest discrimination parameter. When discrimination is high, then the item provides ample information about differences across individuals. If discrimination is low, then the item is not providing much information about differences of θ across individuals.

4.1.2 Graphical representation

In this section we have employed the typical IRT plots to visualize the results of the analysis.

Item Characteristic Curves are shown in Figure 4.1 and provide us with more information about the underlying construct. Each of such curves shows the probability of answering 1 to an item at varying levels of the latent trait, specifying how well an item discriminates between respondents at various levels of the latent trait. The easier item functions are on the left side of the plot, in the lower regions of the latent trait scale, while the more difficult item functions are on the right (in our case, they are the items that underlie more severe impairment in mental health). The discrimination parameter represents the slope, which refers to how well the item response options discriminate (or differentiate) between those subjects with high and low levels of latent trait. Item 5 and Item 7 are in fact the leftmost lines represented in the plot, while Item 3, which is also the less steep, is plotted on the right. Curve of Item 12 is indeed the steepest.

Item Information Curves show how well and precisely each item measures the latent

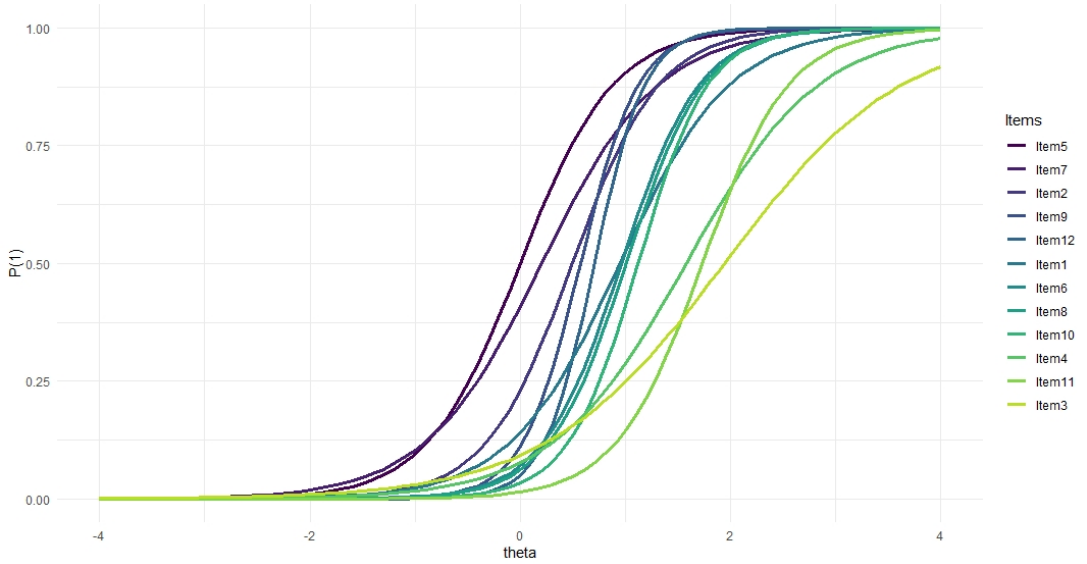

Figure 4.1: Item Characteristic Curves for GHQ-12 under 2-PL model

trait at various levels of the attribute. Item Information measures the strength of the relationship between an item and the latent trait. Some items may provide more information at low levels of the attribute, while others may provide more information at higher levels of the attribute. Moreover, items exhibiting relatively lower discrimination provide information over a wider range of the latent trait; in turn, items with larger discrimination parameters may provide useful information over a relatively narrower range of the latent trait. In Figure 4.2, the curve of Item 12 gives much information around a value of θ between 0 and 1, while Item 3 (the item with the lowest discrimination parameter) gives more or less the same (low) information over a wider range of values of θ.

The Test Information Function aggregates the Item Information Curves across all the items. It is a function of the unknown ability or true score θ, measuring the amount of information provided by the item responses on a test about θ. It tells us the Fisher information for θ contained in the item response vector, how well the test (on the whole) measures the latent trait at various attribute levels. Ideally, this line should have a peak at about the mean of the sample, because that is where the highest number of individuals is situated. It is represented in Figure 4.3, together with the standard error, serving as an estimate of the precision of the maximum likelihood estimator (MLE) $\hat{\theta}$ of the ability θ.

Equivalently, the inverse of the Test Information Function is an estimate of the variance of the MLE.

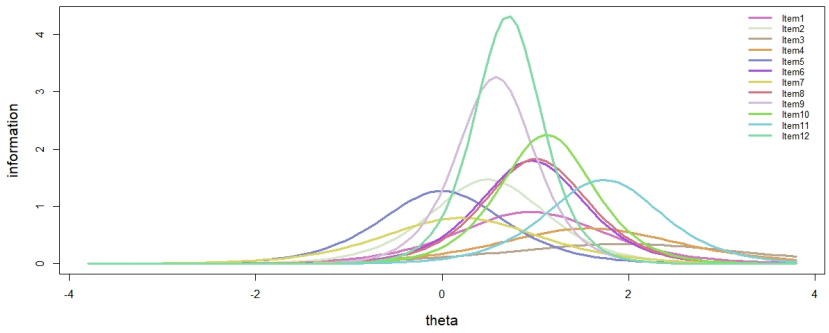

Figure 4.2: Item Information Curves for GHQ-12 under 2-PL model

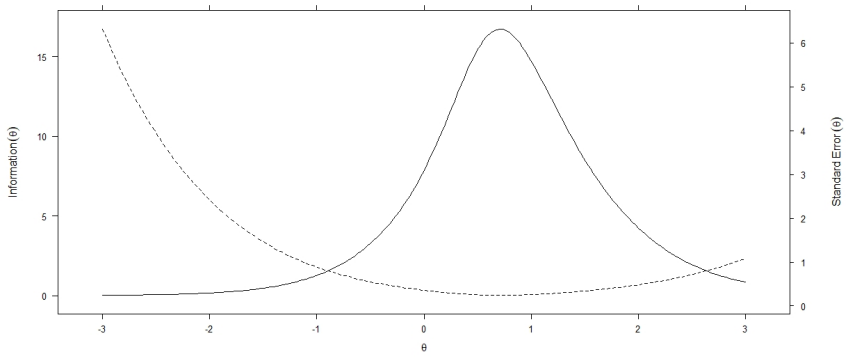

Figure 4.3: Test Information Function for GHQ-12 under 2-PL model (solid line) and standard error (dotted line)

Lastly, the plot in Figure 4.4 shows the relationship between estimated $\hat{\theta}$ scores and estimated true scores. These estimated true scores are model-based, meaning that they are generated using item parameters and are expressed in the original scale metric. We can observe that the curve (i.e., the observed total score) increases as θ increases, which sounds very reasonable. For example, a value of a latent trait around 3 corresponds to a subject giving an answer equal to 1 to all the twelve items, while respondents with two answers equal to 1 (i.e. total score equal to 2) are located around the mean of the latent

trait (zero). These subjects probably answered 1 only to Item 5 and Item 7.

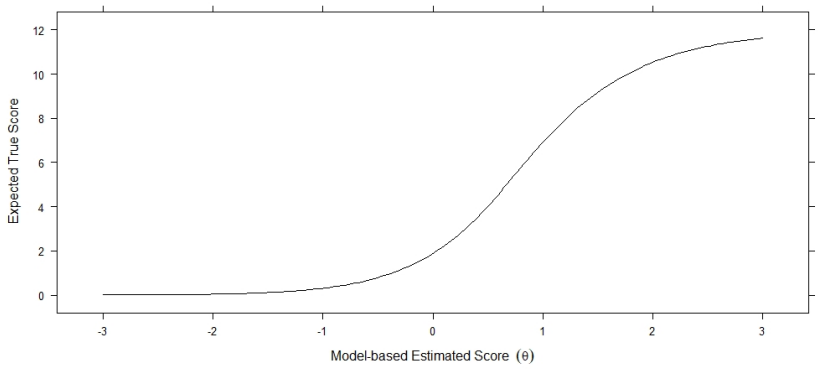

Figure 4.4: Estimated and expected latent trait scores under 2PL model

4.1.3 IRT and Latent Class Analysis

In order to fit a basic Latent class model (LC), as described in Chapter 3, we first have to search the optimal number of latent classes in the population, i.e., the support points of the latent distribution, which we suppose to be discrete. The basic hypothesis of such model is, in fact, that a discrete latent variable explains the association between responses and is used to cluster sample units on the basis of these responses (Bartolucci, 2007). The number of latent classes, called k, can be assumed a priori, on the basis of theoretical knowledge or reasonable assumptions, or selected by comparing the fit of the model under different values of k. We initially perform the procedure of comparing model fit with 1, 2 and 3 classes, as in Table 4.3.

	log-lik.	np	BIC
k=1	-6288.957	12	12660.589
k=2	-6288.957	12	12660.589
k=3	-4771.404	27	9728.827

Table 4.3: Fit comparison with $k = 1, 2, 3$, np is the number of estimated parameters

We choose the number of latent classes k equal to 3, at which we get the minimum

value of BIC. Thus, this model produced three latent classes with weights and levels of the latent trait for each dimension and latent class, i.e. it assigns to each individual some weights $\pi_v,\ v = 1, 2, 3$, as reported in Table 4.4

	class 1	class 2	class 3
θ	-3.3	-0.8	1.15
%	47%	38%	15%

Table 4.4: Latent classes values and corresponding weights

The model estimates three values for the latent trait θ, finding support points equal to -3.3 (low level of latent trait, labeled as *good mental well-being with no signs of distress*) with weight equal to 0.47, a *medium* level of latent trait (that can be interpreted as a level of *some distress but without severity*) for 38% of the population and a higher level of distress (1.15) with weight 0.15.

Computing posterior probabilities (i.e. percentage of participants assigned to each class) it seems that the 3-classes model works well. Percentages in Table 4.5 in fact are very similar to class weights.

class 1	class 2	class 3
49%	36%	15%

Table 4.5: Posterior probabilities

Moreover, the choice of the number of latent classes was discussed with the physicians who administered the questionnaires and they considered three classes as a suitable choice, since the final outcome of the multi-steps evaluation basically divided the population into three classes: a group with no evidence of psychological distress after first level screening, workers who expressed some discomfort, but without any severity according to specialists' evaluation, and finally the group of subjects who, having expressed signs of impairment, were offered psychological support. For this reason we match the three classes with the results on the whole evaluation, shown in Table 4.6. Moreover, in Table 4.7 we report how participants belonging to different classes answer to each item: once each class is identified to a certain level of θ, we want to explore the pattern of responses in each group.

	Stop after I	Stop after II	Psych. support
class 1	97%	1%	2%
class 2	44%	20%	36%
class 3	1%	14%	85%

Table 4.6: Latent classes vs results on second level screening

		Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**	**Item 9**	**Item 10**	**Item 11**	**Item 12**
class 1	0	97%	96%	96%	98%	83%	99%	84%	99%	99%	100%	100%	100%
	1	3%	4%	4%	2%	17%	1%	16%	1%	1%	0%	0%	0%
class 2	0	69%	48%	83%	83%	24%	76%	36%	79%	55%	85%	94%	67%
	1	31%	52%	17%	17%	76%	24%	64%	21%	45%	15%	6%	33%
class 3	0	24%	8%	62%	50%	4%	19%	10%	19%	5%	24%	59%	3%
	1	76%	92%	38%	50%	96%	81%	90%	81%	95%	76%	41%	97%

Table 4.7: Item answers distribution by latent classes

Participants assigned to the first class give answer equal to 0 to almost all the items; the highest percentage of the answer 1 is referred to Item 5 and Item 7 (but much lower than in the general distribution). On the contrary, considering such *easy* items (Item 5 and Item 7) almost everyone who belongs to the third class gives answer equal to 0. In addition, for the group with more severe signs of psychological distress, percentages of answers 1 are much higher (with respect to the general distribution) up to the items found to be the *most difficult* (Item 3 and Item 11). For the second class the distribution is more balanced and more than half of the participants answered 1 only to Item 2, Item 5 and Item 7.

We observe that almost all (97%) of subjects in class 1 did not undergo to the second level screening, i.e. they did not express any signs of discomfort through GHQ-12, IES-R and GAD-7. On the contrary, the majority (85%) of the subjects assigned to class 3 needed a psychological support, while only one third of participants in the second class required therapy.

Ou findings show that it is possible to classify the degree of severity of any impairment of psychological well-being by administering only GHQ-12 questionnaire and according to response patterns, focusing on the answers to each question, more than the scored obtained as a sum of responses equal to 1. Physicians think that this can potentially

simplify the patient's screening. One step of the evaluation (i.e. second-level) can be skipped. According to the severity, physicians plan to give immediate access to specialist evaluation for those who express symptoms of most serious psychological impairment (without testing them through second-level scales) and monitoring subjects with less severity (class 2), with a check evaluation after a certain period of time.

4.2 Second application: analysis of DIF on GHQ-12

An item is said to be affected by DIF when it performs differently for one subgroup of a population compared to another subgroup. In this sense, DIF exists when i) the probabilities of endorsement are uniformly unequal for the subgroups, i.e. the difficulty parameters are different (*uniform DIF*), or ii) when the conditional dependence (due to subgroups) of the endorsement of an item moves and changes direction at different direction at different locations of θ parameter space. This happens, for instance, when an item gives an advantage to a reference group at one end of the θ continuum while favors the other groups at the other hand (*non-uniform DIF*).

DIF can be detected using approaches drawing from both CTT and IRT. An IRT analysis also permits graphical representations of DIF. These plots provide valuable diagnostic insight for evaluating the potential effect of DIF both at the item level and at the level of the entire test. Through IRT techniques, DIF occurs when the trace lines in ICC for the same item differ between groups of respondents. Within the IRT framework, there are a number of ways to construct statistical tests of DIF. Bartolucci (2007) summarized many possible methods for DIF detection in both CTT and IRT framework. One of the most used approaches performs model-based likelihood ratio tests to evaluate the significance of observed differences in parameter estimates between groups, see Thissen et al. (1993) and Thissen (1988). This approach is closely integrated with conventional ML parameter estimation. Several methodological experts agree that DIF analyses using model-based likelihood ratio tests are more powerful and should be emphasized over other existing DIF detection approaches, see Teresi et al. (2000) and Wainer (1995).

In this section we have applied IRT methods to assess the presence of DIF and to evaluate it once it has been detected; we apply the IRT likelihood ratio (IRT-LR) test.

4.2.1 Analysis

The general process of IRT likelihood ratio (IRT-LR) testing involves first identifying a set of items that can serve as *anchors* to define the IRT ability scale, then testing the DIF of the remaining test items. It is important to perform DIF after controlling for the overall differences between subgroups on the construct being measured. If anchor items contain DIF, the ability scale used to evaluate the other items' DIF will be contaminated, which in turn will affect how accurately we detect DIF in the test items. Once we have a set of bias-free anchor items, we can test for uniform DIF, non-uniform DIF, and both uniform and non-uniform DIF jointly.

Analysis was conducted by steps, described here:

1. Ideally, with IRT-LR DIF testing, we already have a subset of items that we can use as anchor items. These items would have undergone previous DIF testing and content-expert fairness review evaluations and would have been evaluated as being free of item bias, (Thissen et al., 1993). With large-scale tests and surveys, organizations tend to have large item banks that contain well-established bias-free items in addition to newly developed items that are candidates for DIF testing. In testing the DIF of new items, these well-established items define the latent variable scale that is used to detect DIF in the new items being tested.

2. When we do not have well-established items that are known to be free of bias, we need to find anchor items. One approach to identifying anchors involves (a) estimating each item's DIF while holding all other items as anchors, (b) ranking the items based on their likelihood-ratio test statistics so we can can select a subset of items to serve as anchors, and (c) estimating the remaining items' DIF while using the designated anchors. If the sample size is large enough, this anchor item subset can consist of a single item, (Woods, 2009). Thus, we preliminary evaluated each item's uniform and non-uniform DIF against all others, which served as anchors. Ranking the items by their likelihood-ratio test statistic, we identify anchor items.

3. We estimated multiple-group 2-PL model specifying anchor items and we performed IRT-LR DIF testing on the items that are not anchors, asking for only those items with statistically significant DIF.

4. We separately examined uniform and non-uniform DIF, specifying parameter which will be inspected for DIF (difficulty or discrimination).

We performed the above mentioned analysis twice; for the choice of subgroups we followed two interests of the *PostCovid* study, i.e. the differences among workers with direct experience with COVID-19 patients with respect to those not directly engaged and differences between subjects enrolled before and after vaccination campaign. Less than half of the subjects, 446 (45%), were directly involved in COVID-19 units, while 544 (55%) were not. Proportion of workers enrolled before (584, 59%) and after (406, 41%) the vaccination campaign is similar.

4.2.2 Results

The first step was the search of anchor items, assuming that there is not a priori set of such items with that characteristic. We performed Step 2 of the analysis founding that all items had no DIF when all of others are used as the anchor items. Item 11 resulted the item with the lowest likelihood ratio test chi-square statistic and significant p-value, so it was chosen as anchor item. It is possible to estimate the multiple-group 2-PL model with Item 11 as the anchor. We see that the anchor item has the same difficulty and discrimination parameter estimates for the two groups but that the other items' parameters are allowed to differ. ICCs based on this multi-group model are shown in Figure 4.5. Because Item 11 is constrained to be equal in the two groups, its item characteristic curve is the same of the entire data set and is therefore a single line. The rest has two lines, and for many of the items, these two lines are very similar in difficulty and discrimination, so we might expect non-significant likelihood ratio tests. Items that seem to have discrepancies are Item 3 and Item 8. We can test uniform DIF and non-uniform DIF separately. Of these items, none is detected as having uniform DIF while Item 3 was found to have statistically significant non-uniform DIF.

Table 4.9 presents parameter estimation in the two groups and we found exactly what was tested as significant from DIF analysis. In particular, discrimination parameter for Item 3 for the sub-sample of exposed subjects is double (1.60) with respect to the other subsample (0.82), for whom, in addition, we observed a very high (2.71) level of θ for

having equal probability to respond 0 or 1.

A similar analysis was performed considering as sub-groups the two samples of enrolled subjects before and after vaccination campaign which started in January 2021 and almost all workers participated. We point out that, similarly to what already said in Chapter 1 about a preliminary analysis on differences when considering vaccination, the aim is not to evaluate a direct effect of vaccination on mental health (since we are not considering the same subject before and after vaccine) but just investigate in which the two sub-samples differ most. A set of four anchor items was chosen after Step 1: Items 3, 4, 10, 11. We observe that such items are the ones found to be the most *difficult* in the previous section. Probability to give answer equal to one to these items was basically similar before and after vaccination. Among the other items, Item 12 was found having both uniform and non-uniform DIF, while Item 7 (with the highest value of chi-square statistics and the lowest p-value) showed uniform DIF. Item 5 also had almost statistically significant uniform DIF (p-value=0.06).

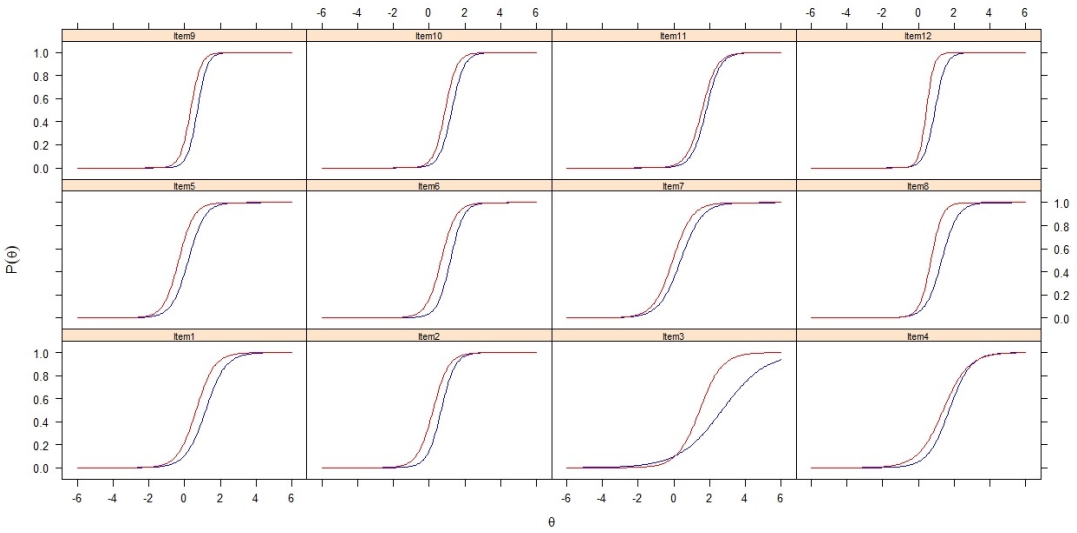

Figure 4.5: DIF detection by COVID-19 area engagement: Yes (red) - No (blue)

The DIF analysis was an extremely useful instrument to analyze how participants belonging to different groups responded to a particular item, in this case Item 3 (*Did you feel you were playing a crucial role in what you were doing? - Ha avuto la sensazione di giocare un ruolo utile in ció che stava facendo?*). The non-uniform DIF of Item 3 provided

a simple, yet interesting result. All participants were health workers that operated in the hospital during the COVID-19 pandemic, and some of them were handling COVID patients on a daily basis. It's easy to imagine how said professionals might feel differently in terms of their practical role, but our study showed this difference through the analysis of discrimination parameter of Item 3. In other words, such analysis specified how the above-mentioned question differentiates subjects on the latent trait continuum in the two groups. Regarding the graphical results, in fact, ICC for Item 3 for the subgroup of COVID-19 workers is in fact much steeper, in particular around a high level of θ. What emerged from the data was then confirmed by the practical experience of specialists, psychologists and therapists who gave support to workers whose reaction to the circumstances was drenched in sorrow and guilt.

Furthermore, when it came to analyzing the results for the health workers who were enrolled after the vaccination campaign, not surprisingly they gave different answers to Item 5 and Item 7 with respect to people interviewed before it started. As a matter of fact, they didn't always feel under strain, and they seemed to find it easier to enjoy their day-to-day activities (see parameters in Table 4.9).

	Area Covid = Yes		**Area Covid = No**	
	Discr	Diff	Discr	Diff
Item 1 (able to concentrate)	1.95	0.65	1.79	1.19
Item 2 (lose sleep)	2.28	0.22	2.59	0.68
Item 3 (play useful part)	1.60	1.45	0.82	2.71
Item 4 (capable of making decisions)	1.41	1.38	1.69	1.72
Item 5 (constantly under strain)	2.37	-0.31	2.19	0.22
Item 6 (overcome difficulties)	2.43	0.67	2.75	1.19
Item 7 (enjoy day-to-day activities)	1.86	-0.05	1.68	0.39
Item 8 (face up problems)	3.16	0.70	2.25	1.30
Item 9 (unhappy and depressed)	3.50	0.34	3.63	0.72
Item 10 (losing confidence)	3.10	0.90	2.92	1.27
Item 11 (worthless person)	2.39	1.57	2.51	1.80
Item 12 (reasonably happy)	4.79	0.44	3.31	0.92

Table 4.8: Item parameters for multi-group 2-PL model (COVID-19 area Yes/No)

	Pre-vaccine		**Post-vaccine**	
	Discr	Diff	Discr	Diff
Item 1 (able to concentrate)	2.07	0.80	1.68	1.13
Item 2 (lose sleep)	2.51	0.36	2.45	0.64
Item 3 (play useful part)	1.08	1.93	1.42	1.84
Item 4 (capable of making decisions)	1.47	1.52	1.87	1.54
Item 5 (constantly under strain)	2.52	-0.18	2.07	0.23
Item 6 (overcome difficulties)	2.76	0.80	2.36	1.19
Item 7 (enjoy day-to-day activities)	1.79	-0.06	1.88	0.57
Item 8 (face up problems)	2.98	0.84	2.28	1.25
Item 9 (unhappy and depressed)	4.10	0.44	2.80	0.71
Item 10 (losing confidence)	3.29	0.99	2.54	1.28
Item 11 (worthless person)	2.72	1.60	2.13	1.86
Item 12 (reasonably happy)	5.13	0.56	2.78	0.89

Table 4.9: Item parameters for multi-group 2-PL model (pre/post vaccination campaign)

4.3 Third application: dimensionality assessment on GHQ-12

In Chapter 2 we presented and discussed the dimensionality issue of psychometric scales, focusing in particular on that about GHQ-12. We have shown that, despite being widely used, the factor structure of such instrument is still controversial: some authors supported unidimensionality of the scale but the majority of studies that have explored the dimensionality of GHQ-12 gave evidence of its multidimensional structure. Some studies suggested unidimensional latent construct with or without adjusting the effects of negative and positive phrased items. Some other studies have shown evidence of the scale containing two, three, or four latent factors as multidimensional mental health constructs.

4.3.1 Factor Analysis approach

Aim of this section is to perform Factor Analysis, exploring results from different methods of scale scoring, factors extraction and factor rotations, looking for the best model in terms of fit and good interpretation. A core aspect of EFA is the determination of which observed indicator variables are associated with which latent factors through the use of factor loadings. Loadings are initially extracted using an algorithm and then

transformed (via the factor rotations) to make them more interpretable.

item	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.62	0.48	0.52	0.59	0.57	0.51	0.65	0.60	0.63	0.58	0.64
2		1	0.29	0.40	0.76	0.65	0.60	0.59	0.74	0.65	0.59	0.73
3			1	0.54	0.28	0.46	0.45	0.44	0.38	0.45	0.59	0.47
4				1	0.50	0.59	0.47	0.60	0.51	0.60	0.49	0.55
5					1	0.67	0.62	0.62	0.69	0.61	0.49	0.72
6						1	0.54	0.72	0.74	0.74	0.64	0.71
7							1	0.69	0.60	0.59	0.47	0.66
8								1	0.70	0.65	0.60	0.80
9									1	0.80	0.68	0.84
10										1	0.81	0.71
11											1	0.68
12												1

Table 4.10: Tetrachoric correlation between items dichotomously scored

When dealing with ordinal variables in factor analysis, the problem is having a good measure indicating the strength of the association, as the Pearson correlation does for continuous variables, since a *correlation measurement* is required to extract uncorrelated latent factors. The assumption is that the ordinal responses correspond to a discretization of a continuously distributed trait, using some thresholds for categorical decision. A measure of association between such continuous traits can therefore be obtained and it is referred to as the *polychoric correlation* (*tetrachoric* correlation in the binary case). Then, classical factor analysis could be performed on the polychoric correlation. Table 4.10 contains tetrachoric correlations between items: the highest correlation occurs between Item 10 and Item 11, while the lowest between Item 3 and Item 5. Moreover, Item 3, except with Item 11 and Item 4, has discrete correlation with the other items (very low with Item 5). Concerning the *extraction method*, EFA was performed using the method of *principal axes*. Common other methods include unweighted least squares, generalized least squares, maximum likelihood, principal components. The method selected should be based on the nature of the underlying distribution of the data. For example, maximum likelihood is recommended when data are multivariate normally distributed, while principal axis factoring makes no distributional assumptions. The Principal Axis factor analysis does an eigenvalue decomposition of the correlation matrix with the diagonal replaced by

the values estimated by the factors of the previous iteration. Regarding the number of *dimensions* employed to represent a correlation matrix, there are many solutions to this problem and none of them is uniformly the best. Techniques most commonly used include a) extracting factors until the chi square of the residual matrix is not significant; b) extracting factors until the change in chi square from n factors to $n+1$ factor is not significant; c) extracting factors until the eigenvalues of the real data are less than the corresponding eigenvalues of a random data set of the same size (parallel analysis), (Horn, 1965); d) plotting the magnitude of the successive eigenvalues and applying the scree test, (Cattell, 1966); e) extracting factors as long as they are interpretable; f) using the Very Structure Criterion (VSS), (Revelle and Rocklin, 1979); g) using Wayne Velicer's Minimum Average Partial (MAP) criterion, Velicer (1976); h) extracting principal components until getting an eigenvalue less than one. Figure 4.6 reports the obtained screeplot, which suggests to extract two factors and the parallel analysis plot which suggests to add a third factor.

Assuming a scale is multidimensional, factors *rotation* will be necessary to aid the interpretation of the model. There are two main classes for rotations, orthogonal and oblique. Orthogonal rotation seeks to find a solution that minimises the relationship between factors. This method has been criticised as most factors that make up a latent variable are expected to share some degree of relationship among them. Moreover, an oblique rotation could be used to estimate an orthogonal model, but not vice versa. Therefore, oblique rotation, which allows relationships between factors, should be preferred in most situations, unless a strong argument can be made as to why the factors should not be correlated. In this scenario, we allow multiple factors to be related, so an oblique transformation was performed and used to identify and characterize factors. We chose the oblique transformation *oblimin*, as it leads to a better interpretation. To assign an item to a particular factor, we selected factor loadings equal to or greater than 0.4.

4.3.1.1 Results

From the screeplot two factors seem sufficient. Since parallel analysis suggests three, we examined both models, looking for the best in terms of model fit and good interpretation. In the Table 4.11 we compare the percentage of variance explained, together with

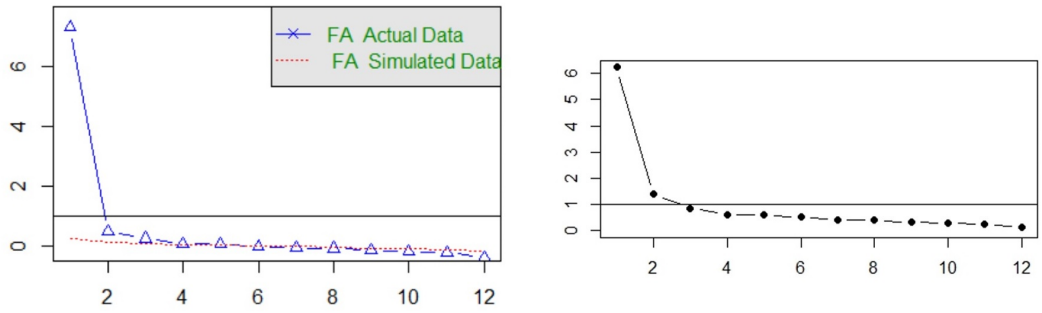

Figure 4.6: Parallel Analysis (left) and Screeplot (right)

	% var	RMSEA	BIC
one-factor	61%	0.18	1528.9
two-factors	67%	0.17	1109.4
three-factors	71%	0.17	742.5

Table 4.11: Fit comparison among one, two and three factors models

	Factor 1	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 3**
Item 1 (able to concentrate)	0.61		0.55		
Item 2 (lose sleep)	0.94		0.79		
Item 3 (play useful part)		0.74			0.57
Item 4 (capable of making decisions)	0.40	0.43	0.43		0.40
Item 5 (constantly under strain)	0.92		0.94		
Item 6 (overcome difficulties)	0.73		0.61		
Item 7 (enjoy day-to-day activities)	0.66		0.77		
Item 8 (face up problems)	0.72		0.79		
Item 9 (unhappy and depressed)	0.89		0.65		
Item 10 (losing confidence)	0.70			0.62	
Item 11 (worthless person)	0.53	0.41		0.89	
Item 12 (reasonably happy)	0.86		0.77		

Table 4.12: Factor loadings of the two-factors model and three-factors model

some goodness-of-fit indices. The two-factor model already achieves a good proportion of explained variance, while the three-factor model has a considerably lower BIC index. In either case, high values of RMSEA indexes do not indicate good fit.

In Table 4.12 all variables, except Item 3 and Item 4, load heavily (factor loading

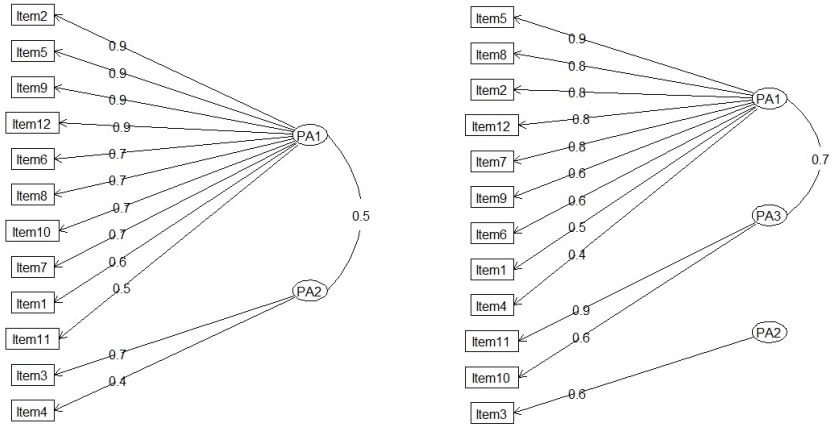

Figure 4.7: Factor Analysis diagram. Two-factors (left) and three-factor (right)

>0.5) on Factor 1. Item 3 is strongly associated with Factor 2, which has factor loadings equal to 0.4 for Item 4 and Item 11. These results agree with observations we previously derived via correlation analysis, where Item 3 was found having low correlation with other items. Item 3, 4, 10, 11 were found most *difficult* and related to most 'sensitive' traits of mental wellbeing. It seems that Factor 2 accounts for those items, while Factor 1, strongly associated with Item 5 (feeling under strain) and Item 2 (difficulty to sleep), is more related to a status of general stress and concerns.

Looking at the three-factor model (from Figure 4.7, which graphically represents factor loadings), factor analysis extraction groups Item 10 and Item 11 in a separate factor, also correlated with Factor 1 (which is the same as in the two-factor case, plus Item 4 with low factor loading), and identifies Item 3 as something much different from other items. This result is in line with DIF analysis performed in the previous section, in particular it supports the different behaviour that such question (regarding how much an individual can play an useful role) has shown in HCWs, confirming how it was considerably crucial during COVID-19 pandemic.

4.3.2 MIRT analysis

According to Hattie (1985), linear factor analysis with categorical items can produce distorted factor loading estimates for very difficult and very easy items. Therefore, in

this section we apply MIRT techniques to investigate under an alternative approach the GHQ-12 structure. Since unidimensionality seems not to hold from previous analyses, we use MIRT to support this hypothesis (i.e. confirmatory MIRT). We have already underlined the main difference between the two approaches: the aim of factor analysis (both exploratory and confirmatory) is to explain the covariance between items based on a certain number of latent factors. On the contrary, IRT tries to determine how respondents answer to each individual item.

Our analysis is based on the model in Equation (3.21), in which the vector of item discrimination parameters (or slope parameters) $\boldsymbol{\alpha_j}$ indicates the probability of correct response associated with changes in a respondent's standing along the m dimensions. Each α_{jm} is the slope, or discrimination parameter for item j on the m-th latent dimension θ_m.

Reckase (2009) proposed a single index of item difficulty and a single index of item discrimination in multidimensional modeling.

The single index of the multidimensional item discrimination (MIDISC_j) has form

$$\text{MIDISC}_j = \sqrt{\sum_{m=1}^{M} a_{jm}^2}. \tag{4.1}$$

The MIDISC_j represents the j-th item's overall multidimensional discrimination. In multidimensional modeling, it is defined as the value related to the slope of the multidimensional item response surface in the direction where the item response surface takes the steepest slope. The direction from the origin of the m-dimensional latent trait space in the steepest direction is quantified by

$$\alpha_{jm} = \cos^{-1}\left(\frac{a_{jm}}{\sqrt{\sum_{t=1}^{M} a_{jt}^2}}\right)$$

It is the value of the arccosine, the angle of the i-*th* item from the reference dimension, t, where $t = 1, 2, ..., M$.

The single index of the multidimensional item difficulty is defined as

$$\text{MDIFF}_j = -\frac{d_j}{\text{MDISC}_j}$$

It is obtained by the negative of the item intercept divided by the discimination index. This is similar to the definition of the unidimensional difficulty, which is the negative of the unidimensional intercept divided by the discrimination parameter. The interpretation of MDIFF_j is the same as the unidimensional β_j at the direction specified by α_{jt}.

We performed MIRT analysis within a confirmatory approach, as we have already obtained indications about the failure of unidimensionality. MIRT method allows us to estimate item difficulty and discrimination simultaneously with the factors extraction. For a confirmatory purpose, which items are loaded on to which factors should be identified before fitting the model. The relationship between factors investigated in the previous analysis has to be used here as a hypothesis to be tested through MIRT. Both EFA and DIF detection suggested us that Item 3 is probably associated with a separate factor. Also Item 4, Item 10, Item 11 may be considered in the same cluster of items. Among many possible models describing GHQ-12 factor structure, we found interesting the *bifactor model*, a specific design where all items load on a general or primary factor, and subsets of items each load on their own specific factor. In the psychological literature, the bifactor structure has gained increasing attention, see Reise (2012).

The general or primary factor and the specific factors are assumed to be uncorrelated. The specific factors are assumed to be uncorrelated to each other since, after accounting for their common association with the general or primary factor, the association with the specific factor is unrelated across the specific factors. This independence of the factors is part of the model identification constraints in the bifactor modeling. Physicians found that the bifactor modeling suits our situation in which we model the general or primary factor as a single construct while modeling each sub-domain's specific factor; that is why we decided to test it.

We initially focus on two factors (which allowed good interpretation through EFA analysis) and compare sequentially three models

1. two-factor model with uncorrelated factors
2. two-factor model with correlated factors
3. bifactor model with two sub-factors

Concerning the identification of the best model in terms of fit, Kline (2005) recommended interpreting and reporting four indices: the model chisquare (sample-based), the Steiger-Land root mean square error of approximation (RMSEA; population-based), the Bentler comparative fit index (CFI; population-based) and the standardized root mean square residual (SRMR; sample-based). Acceptable model fit is characterized by RMSEA < 0.05 and CFI > 0.96 or TLI> 0.96. In addition to these fit indices, we examined the Akaike information criteria (AIC; sample-based) and the Bayesian Information Criterion (BIC); sample-based). The optimal model minimizes both quantities. According to Jackson et al. (2009) a review of CFA journal articles published over the past decade identified these six fit indices as the most commonly reported. Also Hu and Bentler (1999) suggest using multiple indices to evaluate model fit. Table 4.13 compares the three models according to the above mentioned indices.

The two uncorrelated factors model gives the worst fit, while indices reveal good fit for both correlated and bifactor models (slightly better for the bifactor structure, except for BIC). The chi-square statistic tests the null hypothesis that the model has perfect fit in the population. Given its sensitivity to sample size, the chi-square test is often statistically significant. The bifactorial model seems plausible as it should reasonable to identify a dominant factor linked to all the items (a general state of stress) together with the presence of sub-factors, one related to stress, concerns and difficulties and one capturing those feelings of being useful (Item 3) and being able to make decisions (Item 4) which were considerably affected in HCW population who face COVID-19 pandemic.

Table 4.14 reports the factor loadings and the communalities of the bifactor model. Factor loadings can be interpreted as the strength of the relationship between an item and the latent variable. The communalities are squared factor loadings and are interpreted as the variance accounted for in an item by the latent trait. All of the items had a substantive relationship (loadings > 0.50) with the principal factor. What seems 'strange' is that the loading of Item 10 and Item 11 are very low (and negative in one case), indicating a poor relationship with the first specific dimension. For this reason, we compare this model with a bifactor model with three sub-domains, adding one related to Item 10 and Item 11 (and as previously suggested by the three-factor model in EFA). They are instead strongly related to the principal factor. Item 3 reports the lowest communality value, meaning

that under this model, only a small part of its variability is explained. We compare this model with a bifactor model with three sub-factors: one including Items 1, 2, 5, 6, 7, 8, 9, 10, 11, 12, the second one including Item 3 and Item 4. Factor loadings and item parameters are shown in Table 4.14 and 4.15 respectively. The mean of the slopes in the primary dimension is 2.6. The means of the slopes for the first, the second, and the third specific dimensions are 1.15, 1.08, and 1.59, respectively. Overall, the primary dimension is the strongest to affect item responses across all items. Of the three specific dimensions, the third one is the most salient. The means of the ability distribution for each dimension are fixed at zeros, and the covariance matrix of the four latent traits is the identity matrix, i.e., variances equal to 1 and all covariances equal to 0. These fixed values are the bifactor model identification constraints. Table 4.17 contains MDIFF and MIDISC of the bifactor model indices as computed in (4.3.2) and (4.1), which perfectly agree with the unidimensional analysis.

The bifactor model with three sub-factors gives very good fit (RMSEA 0.04; RMSR 0.03; TLI=0.99; CFI=0.99). All items have high loadings on the general factor. Item 2 (followed by Item 5) in strongly associated with the sub-domain F1, allowing to think that it represents insomnia, big concerns and strains; other items are poorly associated with F1. Sub-factor F2 is associated with Item 3 and Item 4 and it represents the sensation of feeling useless and unable to make decisions, while F3 accounts more on Item 11, indicating the sentiment of loss of trust and valueless.

In conclusion, our initial purpose was to investigate the factorial structure of GHQ-12. From our analysis what emerges, more than a clear and defined solution in terms of exact dimensionality assessment, is the capacity of this scale to describe and model different and complex level of psychological well-being of HCWs. We further specified the fact that subjects who completed this questionnaire were healthcare workers involved in a pandemic, which suddenly found them very unprepared and shocked. In such circumstances, questions about utility, capacity to make decisions, loss of trust and confidence showed peculiar responses, affecting in different ways the psychological status. From the IRT analysis we found that Item 3 (feeling useful) has the highest threshold (difficulty) parameter (so a high level of distress an individual needs to have in order to have 0.5 prob-

ability of giving response less than usual) but lowest discimination, not providing much information about differences across individuals. A significant difference in discrimination emerged dividing between individuals directly or indirectly involved with COVID-19 patients. According to de Gruijter and van der Kamp (2008), in fact, an item affected by DIF can be one sign of multidimensionality.

EFA assigned Item 3, together with Item 4 (capacity to making decision), to a distinct factor. Tetrachoric correlation matrix reported high correlations between almost all items and Item 3 having the lowest correlations. Confirmatory MIRT analysis, performed fitting a multidimensional 2-PL model, supported the failure of the unidimesnional model and showcased the bifactor structure with a general (or principal) factor related to all items and three specific sub-domains as the model with best fit. Results coming from factor analysis and IRT basically agreed. Both methods were informative for establishing factor structure and they were able to detect the multidimensional structure along with clusters of associated items. We found particularly useful the complementary use of the two methodologies. As a preliminary step, it is important to investigate the factorial structure from a correlation point of view using classical factor analysis. The IRT method, in addition to overcoming some methodological limitations that we have already illustrated (e.g. it is preferable in presence of discrete data), gives the best fit and allows the identification of properties and characteristics of each item, specifying better the relationships between items and latent factors. The use of both is recommended in literature, well reported and performed, for example, by Osteen (2010), and in more recent studies by Immekus et al. (2019) and Bean and Bowen (2021).

	M2 (χ^2)	RMSEA (95% CI)	SRMSR	TLI	CFI	AIC	BIC
Uncorrelated two-factors model	421***	0.08 (0.07, 0.09)	0.15	0.94	0.95	9757	9874
Correlated two-factors model	181***	0.05 (0.04,0.06)	0.04	0.98	0.98	9542	9664
Bifactor (two sub-factors)	130***	0.04 (0.34,0.05)	0.03	0.98	0.99	9506	9682

Table 4.13: Fit statistics

	G	F1	F2	Communalities
Item 1 (able to concentrate)	0.71	0.20		0.55
Item 2 (lose sleep)	0.72	0.45		0.73
Item 3 (play useful part)	0.58		0.34	0.46
Item 4 (capable of making decisions)	0.67		0.50	0.70
Item 5 (constantly under strain)	0.68	0.53		0.76
Item 6 (overcome difficulties)	0.81	0.19		0.70
Item 7 (enjoy day-to-day activities)	0.64	0.34		0.53
Item 8 (face up problems)	0.80	0.26		0.71
Item 9 (unhappy and depressed)	0.86	0.25		0.81
Item 10 (losing confidence)	0.93	-0.02		0.87
Item 11 (worthless person)	0.88	-0.20		0.82
Item 12 (reasonably happy)	0.85	0.33		0.84

Table 4.14: Factor loadings and communalities of the bifactor model (with two sub-factors)

	G	F1	F2 (Items 3-4)	F3 (Item 10-11)	Communalities
Item 1 (able to concentrate)	0.73	0.13			0.55
Item 2 (lose sleep)	0.72	0.56			0.84
Item 3 (play useful part)	0.59		0.44		0.55
Item 4 (capable of making decisions)	0.70		0.37		0.63
Item 5 (constantly under strain)	0.73	0.42			0.73
Item 6 (overcome difficulties)	0.84	0.09			0.72
Item 7 (enjoy day-to-day activities)	0.70	0.19			0.53
Item 8 (face up problems)	0.87	0.02			0.76
Item 9 (unhappy and depressed)	0.87	0.19			0.80
Item 10 (losing confidence)	0.86			0.31	0.84
Item 11 (worthless person)	0.78			0.49	0.85
Item 12 (reasonably happy)	0.89	0.18			0.84

Table 4.15: Factor loadings and communalities of the bifactor model

	slope G	slope F1	slope F2	slope F3	intercept	diff G	diff F1	diff F2	diff F3
Item 1 (able to concentrate)	1.83	0.58			-1.76	0.96	3.05		
Item 2 (lose sleep)	2.45	2.11			-1.41	0.58	0.67		
Item 3 (play useful part)	1.44		0.90		-2.61	1.81		2.91	
Item 4 (capable of making decisions)	2.10		1.27		-3.12	1.49		2.45	
Item 5 (constantly under strain)	2.21	1.90			0.07	-0.03	-0.04		
Item 6 (overcome difficulties)	2.65	0.60			-2.53	0.95	4.18		
Item 7 (enjoy day-to-day activities)	1.63	0.78			-0.32	0.20	0.41		
Item 8 (face up problems)	2.69	0.71			-2.71	1.01	3.84		
Item 9 (unhappy and depressed)	3.33	1.14			-1.91	0.57	1.67		
Item 10 (losing confidence)	4.28			1.44	-4.58	1.07			3.17
Item 11 (worthless person)	3.27			1.74	-5.66	1.73			3.25
Item 12 (reasonably happy)	3.77	1.44			-2.73	0.72	1.90		

Table 4.16: Item parameters for three-factor MIRT

	MDISC	MDIFF
Item 1 (able to concentrate)	1.92	0.92
Item 2 (lose sleep)	3.23	0.44
Item 3 (play useful part)	1.70	1.54
Item 4 (capable of making decisions)	2.45	1.27
Item 5 (constantly under strain)	2.91	-0.02
Item 6 (overcome difficulties)	2.72	0.93
Item 7 (enjoy day-to-day activities)	1.80	0.18
Item 8 (face up problems)	2.78	0.97
Item 9 (unhappy and depressed)	3.53	0.54
Item 10 (losing confidence)	4.52	1.01
Item 11 (worthless person)	3.71	1.53
Item 12 (reasonably happy)	4.04	0.68

Table 4.17: MDISC and MDIFF of the bifactor model (with three sub-factors)

Chapter 5

Forward Search for IRT robust estimation and for the detection of atypical response pattern

In this chapter we implement a forward search algorithm for identifying atypical subjects/observations in Item Response Theory models for binary data (Rasch models). Our proposal introduces diagnostic tools, based on robust high-breakdown methodologies, to avoid distortion in the estimation of the model, and to single out atypical response patterns. Forward plots of goodness of-fit statistics, residuals, and parameter estimates help us identify aberrant observations and detect deviations from the hypothesized model. Methods to initialize, progress, and monitor the search are explored. Simulation envelopes are constructed to investigate whether changes in the statistics being monitored are solely due to random variation. One real and one simulated datasets are used to illustrate the performance of the suggested algorithm. The simulated dataset explore the effectiveness of the method in the presence of a single outlier and a cluster of outliers.

5.1 Introduction and motivation

As we discuss till now, dichotomous data frequently arise in the social sciences, and Factor Analysis, Item Response Theory, and its multivariate extension MIRT, are the

most frequently employed models for assessing unobserved constructs, such as abilities, attitudes, exposure to an illness, and socioeconomic status.

Data are usually collected through surveys and measured over a scale. Atypical response patterns could arise due to cheating or guessing - when measuring abilities - and careless or systematic responding, and also erroneous interpretation of an item or exaggerated answer to, in all cases. Besides, inattention to easy questions and other surrounding factors may yield response patterns that do not reflect the person's score on the latent constructs. The presence of aberrant response patterns could severely bias parameter estimation, and robust methodologies have been introduced in the literature to overcome this issue.

While the most widely used statistical methods are based on the assumption of normality, real data are seldom normally distributed. Awareness of such issue has a long history, from Geary (1947), who stated: "Normality is a myth; there never was, and never will be a normal distribution" (p. 241). A few years later, Tukey gave his most decisive contribution to the Princeton Robustness Study: a clear conceptual recognition of the main underlying problem, in other words, the extreme sensitivity of some conventional statistical procedures to seemingly minor deviations from the assumptions. A pletora of authors empirically supported such thesis, starting from Micceri (1989) who compared the normal curve to the mythical unicorn. Micceri tested the distributional characteristics of 440 large-sample achievements and psychometric measures and found all to be significantly non-normal, at the $\alpha = 0.01$ significance level. Among many others in the recent years, see Cain et al. (2017) and references wherein, for a view on the influence of data departures from normality on the estimation.

In this chapter we will deal, in particular, with Maximum Likelihood (ML) estimation for IRT models in the case of dichotomous responses, but now in the framework of robust statistics. ML is the best known principled method for inference, and it owes its popularity also to its nice optimal properties. It is well known that ML estimation is particularly affected by the presence of outlying data. Typically, the nature and extent to which disturbances are present in data is unknown. Despite that, some researchers propose to consider more general models, based for instance on the skew normal of t distribution, to account for deviations from normality that could be captured by (excess

of) skewness and kurtosis. The latter are the most important indicators of the extent to which non-normality affects the usual inferences made in the analysis of variance. A broader, more complicated model could represent a viable solution. In many cases, however, data contamination follows unknown random patterns. Therefore, accounting for it by altering the model is not possible, as it would lead to increased model complexity and excessive sample size requirements. Such considerations motivated a fresh rethinking of statistical methods, strongly rooted within the Princeton Robustness Study. Many interesting proposals move toward introducing robust high-breakdown methodologies to avoid misleading results in inferential methods.

We argue here that with models designed to estimate severity of mental illness, or a temporaneous psychiatric impairments, like in the *PostCovid* study, our primary interest is in estimating correctly such severity we will be able to i) identify atypical response patterns, and ii) obtain unbiased inferential results. Atypical response pattern are of interest to the specialists in occupational medicine, and usually deserve further investigation. Within IRT, Panchapakesan (1969) suggested omitting low ability subjects entirely when estimating the item parameters, afterwards Wainer and Wright (1980) introduced a Jacknife estimator for person abilities, more resistant to occasional aberrant responses. Two years later, Mislevy and Bock (1982) presented an alternative robust estimator, based on the principle of Tukey's biweight. The biweight ignores changes that lie outside a substantial range from the bulk of the data, yet responds sensitively to changes in the middle portion of this range. Iterations toward ML estimates can easily be modified to iterate toward the biweight estimates, just by considering a modifies likelihood equation, where contribution of subject i is weighted by the Tuckey biweight function. The obtained reduction of bias gives it a smaller mean squared error than the ML estimator even when measurement disturbances are mild.

Recently, Yuan and Gomer (2021) made a careful review on the topic, in the broad framework of social and behavioural sciences, mainly focusing on methodologies based on M-estimators. Hong et al. (2020) introduce different data cleansing procedures to detect different manifestations of Insufficient Effort Responding (IER). After a first 'non robust' estimation, they removed any participant flagged by any (or a best subset) of the four IER detection methods, trying to find a tread-off between specificity and sensitivity. Hong and

Cheng (2018) proposed to leverage person-fit statistics to detect careless respondents and down-weight them in the estimation process. After a first unweighted full-data estimation procedure, they employ weights based on p-values of the same person-fit statistics, to perform a second, more robust estimation. Liu et al. (2020b) introduce the idea of using a mixture model, for response accuracy and response time to model different non-effortful and effortful individuals, and to improve item parameter estimation based on the effortful group. As non-effortfull responses distort the estimation of item parameters in IRT models, Liu and Liu (2021) proposed an iterative purification process based on a response time residual method with fixed item parameter estimates to detect such responses. The model should be re-estimated, each time that a non-effortful response is detected and discarded.

5.1.1 A brief review of research on robust IRT models

Several contributions about robust analysis in the IRT framework appeared in the literature. This section collects the most significant works, which are briefly summarized.

One of the first studies on the binary IRT model was in Waller (1974) with the purpose of removing the effects on random guessing in latent ability estimates. The idea is to base the estimate of any person's ability only upon items for which there is a reasonable chance that the person achieved the correct response through the interaction of his ability and the item characteristics. That is, an item which is very difficult for a particular individual is an item which invites guessing, and therefore it should be eliminated from consideration in estimating the person's ability (and the entire person's response is removed from the sample used to calibrate such an item). Whether or not the person guesses on such an item has no substantial effect on the estimate of his ability, because these item-person interactions are removed from the estimation procedure. For the ability estimation the authors propose a conditional ML Estimation method: the procedure simply omits from estimation any interaction for which this estimated probability is lower than some cutoff point, P_c.

A few years later, Wainer and Wright (1980) focuses on 1-PL model and use the

standard maximum likelihood method for estimating Rasch abilities, given a vector of item difficulties. It basically relies on the Rasch model property that raw score is a sufficient statistic for ability. The jackknife is shown to be useful for hypothesis testing as well. The way that it works in that application is to construct a matrix of abilities W where the first column contains the abilities associated with raw score r, calculated through the Rasch model. The second column collects the abilities based upon a test with the first item omitted; each succeeding column represents abilities estimated through previous scheme but with that item omitted. Thus the k-th column is a test containing all items except item $k-1$. The authors found that gains in recovering abilities in the presence of guessing, and untoward responses of other kinds, can be obtained through the use of a robust jackknife. The jackknife weighs these two extremes and places their estimate between i) treating the persons's getting the item correct as a wild guess and changing it to incorrect and ii) treating the response as one we fully believe.

Mislevy and Bock (1982) proposed to base the robust estimate of ability in 1-PL or 2-PL on a modified likelihood equation $\sum W_{ij}(X_{ij} - P_{ij})/S_i = 0$ where S_i is the item dispersion for item i (i.e. reciprocal of the item slope). They defined the residual as $r_j = \lambda_j(\theta - b_j)$. Thus, the larger the weighted differences between difficulty and ability, the heavier the down-weighting of the observation will be. As an alternative, they also propose an estimator weighted by Tukey's biweight function. This estimation approach has been shown to effectively reduce bias of ability estimates in the presence of response disturbances. However, some cons were pointed out by Schuster and Yuan (2011), who considered the latter method prone to produce infinite ability estimates for unexpected response patterns, whenever correct answers are sparse. To overcome such issue, an alternative robust estimator of ability which does not produce infinite estimates, the Huber-type estimator (essentially equivalent to MSE) have been employed.

A more recent paper by Patton et al. (2019) is based on one of the most popular approach to identify atypical responses under the IRT framework, the person-fit statistics l_z (Conijn et al. 2014, Karabatsos 2003). The authors propose to iteratively detect careless responders and cleanse the data by removing their responses. First, the careless respondents are detected using the person-fit statistics in its standardized log-likelihood

version, (Drasgow et al., 1985), then the corresponding responses are removed from the dataset, and this process continue by iteratively updating the item parameter estimates.

An analogous method, based again on the person-fit statistics, was used by Hong and Cheng (2019) and applied to the graded response models (polytomous scored items). The authors introduce robust estimation for the latent trait estimator when conducting outlier detection with person-fit, more precisely with the correction to the original l_z statistic, l^*_{z-d} for dichotomous items, due to Snijders's (2001). First, the item parameters are estimated from the full data, with equal weights for all participants. Next, the person-fit statistic, l^*_z, is evaluated for each participant, given the current estimates of the item parameters (obtained from the first step). Finally, normalized l^*_z's p-values are used as the new weights to re-estimate the model. Smaller p-values would indicate a greater severity of response aberrance, and the corresponding response pattern would be assigned a smaller weight. Such a weighting mechanism would allow to form a gradient differentiating between partial and complete carelessness. The item parameter estimates should be less biased, which would in turn improve the detection of careless responses. The two steps could also be iteratively performed until convergence.

A person fit test based on the Lagrange Multiplier (LM) test is presented for three item response theory models for polytomous items in Glas and Dagohoy (2007): the generalized partial credit model, the sequential model, and the graded response model. The LM test can also be used in the framework of multidimensional ability parameters. The log-likelihood is split into two parts, one pertaining to the marginal likelihood of m respondents (denoted by L_m) and one pertaining to the respondent that is the current focus of attention (say, observation $m+1$ whose likelihood is denoted by L_p). So, the method splits $\log L = \log L_m + \log L_p$, and solves the simultaneous system $\frac{\partial(\log L_m + \log L_p)}{\partial \xi} = 0$ and $\frac{\partial \log L_p}{\partial \theta} = 0$. An LM test accounting for the effects of estimation of the person parameter θ is derived.

Among some very recent contributions, Hong et al. (2020) investigated the efficacy of IER (Insufficient Effort Responding) detection methods as a data cleansing method. They evaluated six different IER detection method and demonstrated which IER detection

methods were best for flagging different types of IER. The considered IER statistics are the Mean Absolute Difference, the intra-individual response variability, the long string, the psychometric antonyms, the Mahalanobis distance and the standardized log-likelihood person-fit index.

A further iterative method was introduced by Liu and Liu (2021). The authors proposed to apply an iterative purification process based on a response time residual method with fixed item parameter estimates to detect non-effortful responses. The proposed method is compared with a) the traditional residual method and b) noniterative method with fixed item parameters in two simulation studies in terms of noneffort detection accuracy and parameter recovery.

Felt et al. (2017) propose an IRT approach for detecting person-level outliers which rely on likelihood-based statistics to determine the most typical response patterns given a specific model. The reduced $M2$ indicates an adequately fitting model when $p > 0.05$, and it is expressed as $M_2^* = N(p-\hat{\pi})'C(p-\hat{\pi})$ where N is the sample size, p is a vector of observed response probabilities, $\hat{\pi}$ is a vector of model implied response probabilities, and C is a weight matrix of response patterns. The reduced M_2 statistics is asymptotically χ^2 distributed with degrees of freedom related to the total number of first moments (Maydeu-Olivares and Joe, 2005). Therefore, the model is firstly estimated over the whole data and then the weights (derived from a potentially distorted estimation), are employed to correct it.

5.1.2 Why introducing the Forward Search for IRT?

All methods cited till now start upon a first inference performed on the entire dataset. Therefore, they are potentially unaffordable and conclusions derived from them could be misleading. In particular, they will not be preserved from the issue of *masking* and *swamping*. The latter are the two kinds of errors that may occur in the process of detecting outliers. The masking effect occurs when an outlier is undetected because of the presence of a cluster of outliers and the swamping effect occurs when a *good* observation is incorrectly identified as an outlier. Under the masking effect, the importance of the

observations is not evident unless several observations are deleted at once (Atkinson et al., 2004). Furthermore, a cluster of outliers will shift the mean from its true value and there is the possibility that some *good* observations may be classified as outliers (swamping effect).

Therefore, we opted for choosing the Forward Search (FS) to base our proposal for a robust estimation method in IRT models. The FS algorithm was initially introduced as an outlier detection tool for the estimation of covariance matrices, (Hadi, 1992), and regression models, (Atkinson et al., 2000). It was subsequently extended to standard multivariate methods, (Atkinson et al., 2004), and factor analysis, (Mavridis and Moustaki, 2009), and was recently applied in meta-regression, (Petropoulou et al., 2021).

The FS starts by fitting the hypothesized data generating model to a subset of the data which is gradually incremented by adding the remaining observed units according to their closeness to the postulated data-generating model. In each step of the FS algorithm, parameter estimates, measures of fit, and goodness-of-fit test statistics are monitored and plotted. The outlying behavior of the observations entering the initial subset is apparent through sharp changes in FS plots. The size of the initial subset of the data is obtained as a trade-off between a good initial parameter estimation (which would require a large subset of the data), and the need of having such subset *outlier-free* (therefore, the smallest, the safest). The procedure presented here represents an improvement over previous ideas in two important ways. First, the information contributed by a *safe* subset of subjects is used to explore contribution of all the other subjects; it is used only when one may be reasonably sure it is valid information, and the Forward Search plot gives a clear indication about such validity. In this way, distorted information does not bias first conclusions. Second, the procedure yields a criterion for identifying random guessing, or disturbances in data, and to single out deviations from the *regular* data. A special consideration of such responses, particularly when using survey data to assess abilities, severity of illness, or even social impairment is of interest.

The remaining part of the present chapter is organized as follows: Section 5.2 outlines the methodological extension of the FS algorithm to the IRT model; Section 5.3 presents

an application of the proposed methodology in published IRTs and simulated datasets; Section 5.4 discusses the main findings and provides directions for using the proposed diagnostic methodology for IRT.

5.2 The Forward Search for IRT models

We propose a procedure for the detection of multiple outliers in IRT data based on the Forward Search (FS). The Forward Search is a general method that uses graphs to understand the relationship between a model and the data to which the model is fitted. Along model estimation, in a classical setting, we can lose information about the effect of individual observations on inferences about the form and parameters of the model. FS methods, instead, reveals how the fitted model depends on individual observations and on groups of observations. Robust procedures can sometimes reveal this structure, by down-weighting or discarding some observations. The novelty in FS is to combine robustness and a *forward search* through the data with model diagnostics and computer graphics. We introduce here easily understood plots for IRT models that use information from the whole sample to display the effect of each observation on a wide variety of aspects of the fitted model. The examples on simulated data show the amount of information the plots generate and the insights they provide.

Let Y be an $n \times k$ data matrix representing n observations on k items. We first introduce a plausible method to select a *basic* subset which contains p *good* observations and a *non-basic* subset which contains the remaining $n - p$ observations. Second, we compute the relative distance from each point in the *non-basic* subset, relative to the model estimated on the basic subset. Third, we rearrange such $n - p$ observations in the *non-basic* subset in ascending order accordingly, then increase the *basic* subset by adding to it the closest observation and decrease the *non-basic* subset to $n - p - 1$ observations. This process is repeated until an appropriately chosen stopping criterion is met (usually derived from the FS plot itself). Below we give details for each of the above mentioned steps:

1. Starting from the n initial observations, we have to extract a sub-sample of size p, called the *basic set*. The best basic set is chosen according to a test statistic which

indicates how well the data are fitted by the estimated model. We have chosen to adopt the *limited information test statistic* M_2^* to evaluate such fit, due to its well known asymptotic properties. M_2^* needs to be introduced, therefore, and we will provide below its general definition.

Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_{2^k})$ and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, ..., \hat{\pi}_{2^k})$ be the vectors containing respectively the observed and the expected (under the model estimated on the data) proportions for all the possible distinct 2^k response patterns, being k the number of the items. The quantity denoted by

$$M_2^* = n(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})' C (\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}) \tag{5.1}$$

is the *limited information test statistic* for bivariate margins with positive responses to pair of items and it widely used to assess goodness-of-fit in the IRT field. C is a weight matrix of response patterns. The M_2^* statistics is asymptotically χ^2 distributed with degrees of freedom related to the total number of first moments and indicates an adequately fitting model when $p > 0.05$. The advantage of using a statistics such as M_2^*, evaluated on the bivariate margins, instead of the classical Pizzetti-Pearson Chi-square, resides on the fact that the former converges to its asymptotic distribution also when there is sparsity in the data.

Now, we are ready to describe the role played by M_2^* in the selection of the basic subset. Let $(S_1, ..., S_H)$ be a subset of size H extracted from all possible subsamples of size p and let $(M_2^{*(S_1)}, ..., M_2^{*(S_H)})$ be the respective limited information test statistics. We select the initial subset S_* having the minimum value of $M_2^{*(S_h, \hat{\boldsymbol{\beta}}_{\boldsymbol{h}})}$ where we specify that $\hat{\boldsymbol{\beta}}_{\boldsymbol{h}}$ is the estimated parameter vector based (only) on the p observation from S_h (and not from the whole dataset). As pointed out in (Mavridis and Moustaki, 2009) the asymptotic distribution of the limited information criterion does not hold for subsets of the data, but it is still a good measure of model fit as it evaluates the proximity between observed and estimated bivariate proportions under the fitted model. Moreover, (Mavridis and Moustaki, 2009) suggest to select initial subsets of $p = 50 - 200$ observations, depending on the overall sample size.

2. For progressing in the forward search we increase the *basic* set adding the response pattern with the highest likelihood contribution. The likelihood contribution of the

response pattern i, denoted by $\mathbf{Y_i} = (Y_{i1}, ..., Y_{ik})$, is given by

$$l_i = \prod_{j=1}^{k} p_{ij}^{Y_{ik}} (1 - p_{ij})^{1-Y_{ik}} \tag{5.2}$$

where, according to 1-PL model with $\lambda_j = 1 \ \forall j$, $p_{ik} = \dfrac{e^{\theta_i - \beta_k}}{1 + e^{\theta_i - \beta_k}}$. Likelihood contribution l_i is then weighted by the frequency of response pattern equal to i already present in the basic set, in order to avoid that those patterns observed only once are most likely to enter in the last steps of the search.

3. Once the basic set is increased by one response pattern, we monitor the effect of such addition of an observation has on parameter estimates, goodness-of-fit tests and residuals.

The procedure is simple, computationally inexpensive, suitable for automation, computable with widely available software packages, effective in dealing with masking and swamping problems and, most importantly, successful in identifying outliers, which in this process may enter in the basic in the last steps of FS.

5.3 Applications to simulated data

We apply our method to simulated data. We started with the simplest case and we generate synthetic data from a 1-PL model, with $k = 4$ items, with difficulty parameter estimation equal to $\beta_1 = -1, \beta_2 = -0.5, \beta_3 = 0.5, \beta_4 = 1$. In 1-PL model, the discrimination parameter is constrained to be the same for all items and we set it equal to 1. We obtained a set of $n = 500$ response patterns whose frequencies are reported in Table 5.1.

We determine the *basic* set choosing the *best* among $H = 10$ sub-samples randomly extracted from the initial sample. We choose this subset according to best fit given by minimum M_2^*. Sample size of subsets were set to $p = 100$. At each step of the FS, we select the best candidate from among the response patterns in the non-basic set (which progressively decreases) to enter in the basic set. Plots in Figure 5.1 shows the limited information goodness-of-fit statistic and the corresponding asymptotic p-values at each of the $n - p = 400$ steps of the forward search. It is clear that even when the search

1110	81
1111	62
1100	60
1000	55
0000	49
0100	41
1101	37
1010	36
1001	18
0010	15
1011	13
0101	11
0110	10
0001	7
0111	3
0011	2

Table 5.1: Response pattern frequencies in the simulated data

is initialized from a subset that yields large values for the M_2 fit statistic, the *basic* set is incremented so that the fit is improved and stabilized for the larger part of the search; it substantially deteriorates in the last steps. The same, of course, holds for the corresponding p-values: it reached values close to 1 after 100 steps and started to rapidly decrease in the last steps. This may also be indicative of the fact that a small subset of the data that is not well fitted by the hypothesized model does not necessarily contain outliers. Figure 5.3 represents the forward plot for the adjusted residuals for each response pattern r given by

$$r_{adj} = \frac{\pi_r - \hat{\pi}_r}{\sigma_r} \tag{5.3}$$

where σ_r is the estimated standard deviation for $\pi_r - \hat{\pi}_r$.

In order to test the proper functioning of the algorithm, we changed the last 25 response patterns of the initial dataset, forcing them to be equal to the most atypical response pattern (i.e., 0-0-1-1). Such a response pattern is the least plausible (having a score of 0 on the easiest items and a score of 1 on the most difficult ones); it appears only twice in the initial synthetic dataset. The forward plots for the bivariate limited information statistics obtained from 25 forward searches is given in Figure 5.2, showing the efficacy of

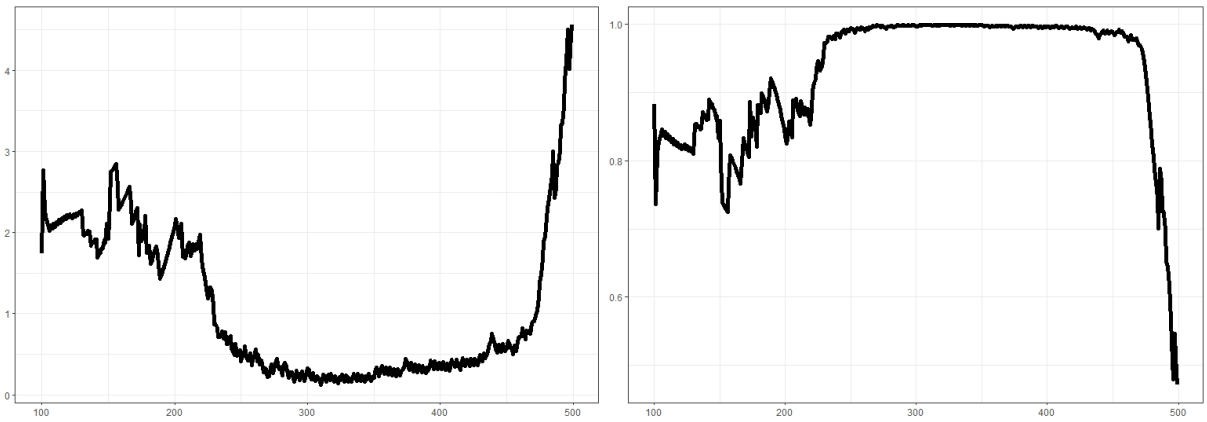

Figure 5.1: Limited information statistics (left) and corresponding p-values (right). In the X axis the size of the basic set is indicated.

such a method. In all the cases, the 25 atypical responses 0-0-1-1 enter in the last 30 steps of the search. In the adjusted residual plot, such a pattern reports a very high residual even in the last FS steps. Moreover, also the response pattern 0-0-0-1 (it appears only seven times in the initial synthetic dataset) has residuals far from 0 for all the FS steps as shown in Figure 5.3.

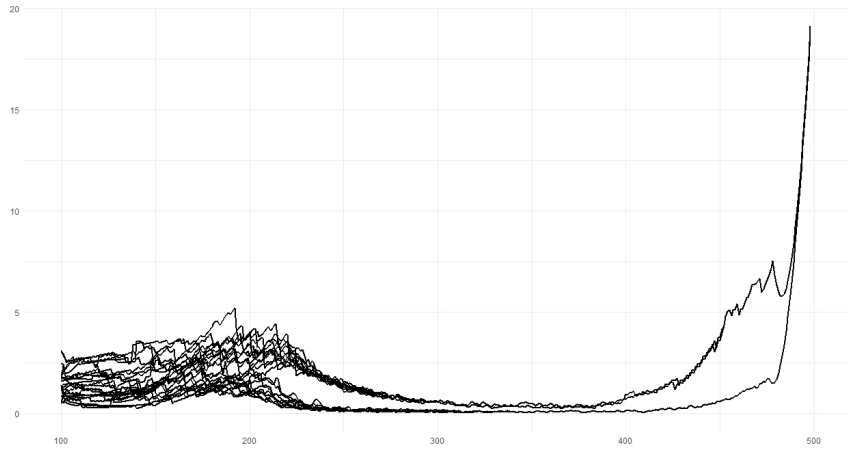

Figure 5.2: Limited information statistic from 25 forward searches. In the X axis the size of the basic set is indicated.

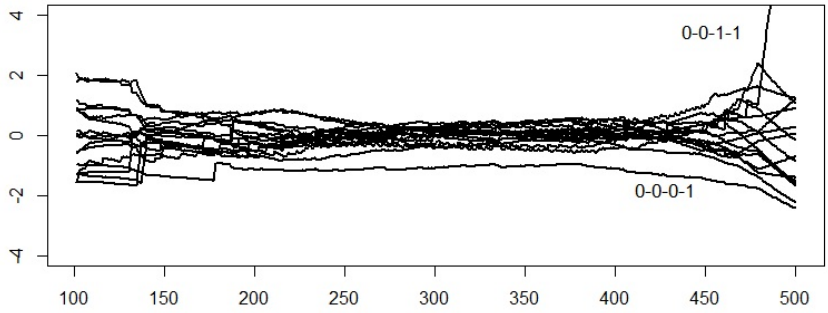

Figure 5.3: Adjusted residuals for each response pattern, along the FS. The more extreme residuals are found for 0-0-1-1 and 0-0-0-1. In the X axis the size of the basic set is indicated.

5.4 Conclusion and future directions

In this Chapter we have started to explore the topic of atypical response pattern detection using the Forward Search in the IRT modelling approach. Such method was applied to the simplest IRT model, the Rasch model with discrimination parameter constrained across items and equal to 1. Even if this work is still in progress, the good results obtained on simulated data generated from a 4-items Rasch model encourage further developments, which will be performed on 2-PL model and polytomous models (GRM, PCM,..). They also established the first promising application of FS algorithm to IRT model. The algorithm based on the Rasch model was applied and tested on GHQ-12 data, choosing a basic set of 200 response patterns and increasing it at each FS step according to the likelihood contributions. We are awared that GHQ-12 data are not well fitted by 1-PL model and this explains why the results in Figure 5.4 are not much informative. After 100 steps of the FS, the goodness-of-fit statistic starts to rapidly increase (looking at the entry order, when pattern 0000100000000 entered after the more "typical" pattern of all zeros), returning low and stable for the central 50 steps (when different patterns are gradually "allowed" to enter) and then it starts again to increase.

Within this application we found computationally intensive the likelihood contribution calculation for each response pattern, caused by the increasing number of elements in the basic set, to be executed at each step of the FS. Further research is needed to find a way

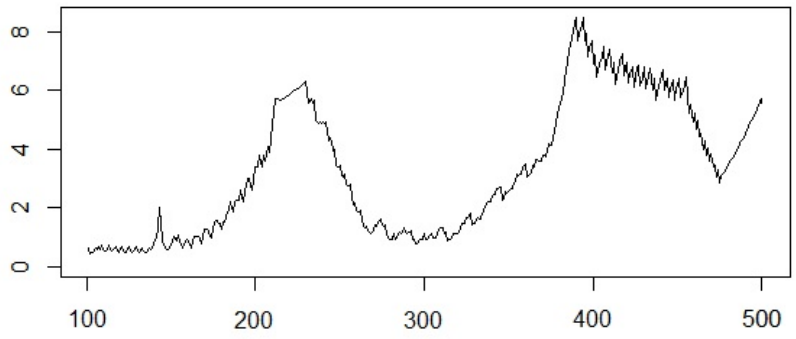

Figure 5.4: Limited information statistic on GHQ-12 binary data

to reduce such computational burden. A second option that deserves to be considered, is to use the person-fit statistic, instead of the likelihood, for selecting which observed unit is to be added to the basic set. As we recalled in the extensive literature review opening the present Chapter, the person-fit statistics has been widely employed as a criterion to enter in the basic set. Response pattern with anomalous l_z statistics should enter in the last steps of the search. The use of such statistic in the IRT framework is often applied with polytomous models, and therefore it could open the path for robust IRT analysis in such more general case.

The bad fit of GHQ-12 data with 1-PL model confirms the importance of taking into account different discrimination among items, as previous illustrated in several ways.

Lastly, our interest in approaching robust models for psychometric scales gave us the chance to start discussions with clinicians (i.e., psychologists and occupational physicians) about the meaning of an atypical response in their assessments. Clinicians confirmed that there is motivation to proceed along such line of research: when anomalous response patterns appeared in data, such respondents should be carefully examined and they perhaps needs a more specific treatment.

Conclusions

This thesis, developed within a research project based in the Dipartimento di Medicina del Lavoro, has motivated and gathered some contributions in the fields of Occupational Medicine, psychometrics, and robust statistics. All the analyses performed here were based on data collected among health care workers in Ospedale Policlinico di Milano, during the COVID-19 pandemics. Such data mainly contains information about mental health status of such workers, measured through self-reported questionnaires.

The first results concerned a risk factor analysis of psychological suffering and was carried out in strict collaboration with Occupational Physicians. Using multiple logistic regression, we found that women, nurses, and workers with direct experience in COVID-19 units showed a significantly higher risk for psychological distress (first-level screening). In contrast, specific psychiatric symptoms (second-level screening) showed a different pattern of association with potential risk factors and different time trends, compared to psychological impairment. The content of this work was already submitted (*One year facing Covid. Systematic evaluation of risk factors associated with mental distress among hospital workers in Italy.* Bonzini, M; Comotti, A; Fattori, A; Tombola, V; Colombo, E; Nava, C; Bordini, L; Riboldi, L; Brambilla, P.)

The second contribution is the statistical analysis of the most crucial questionnaire used in the study project, the General Health Questionnaire-12, an instrument widely used to screen mental health status of individuals. It is commonly analysed and interpreted according to Classical Test Theory (CTT) rules. We provided a basic overview of CTT, firstly, and of IRT in a second step. Starting from some questions of clinical interest

derived from the medical environment, we decided to employ the classical approach and to complement it by performing an analysis based on Item Response Theory (IRT). As outlined in our work, drawing on the strengths of IRT as an alternative to, or ideally in conjunction with, CTT analyses support researchers' development of rigorous measures, and valuable results interpretation. Through the analysis of the data collected via the GHQ-12 questionnaire, we gave a demonstration of the utility of IRT as compared with CTT-based methods.

The psychometric properties and the latent factor structure of the scale were evaluated using IRT (with its multidimensional version MIRT) and Factor Analysis techniques together. To the best of our knowledge, in the literature this is the first application of MIRT to the GHQ-12 questionnaire. Although designed as a unidimensional instrument, our model fit analysis using FA and MIRT did not support this solution. In particular, beyond general distress and concerns, the feeling of uselessness and inability to make decisions that health professionals have strongly experienced when assisting COVID-19 patients emerged as an essential dimension of psychological well-being measured by GHQ-12. It was perceived differently and much strongly by HCWs who directly worked in COVID-19 units.

Moreover, the IRT analysis allowed to simplify the multi-step protocol for the evaluation of mental health. In the light of our results, we proposed to use GHQ-12 as the unique measurement tool. Through an item-based analysis, we were able to determine the outcome of the screening without considering the other questionnaires, previously part of the first-level evaluation. Instead of considering the single overall score, in which each item accounts equally, IRT was found as a more suitable tool for scale assessment.

This work met the predominantly clinical objectives set at the very beginning of the *PostCovid* study, leading to satisfying results. It also opened further investigations. The diversified application of IRT to the collected data enriched the physicians involved in the study - who use GHQ-12 in their profession - and made them more aware of how versatile this tool can be. The idea of focusing more on the analysis of single items rather than on

the scale as a whole was met with interest, seen -as it is- a widely unexplored approach. Moreover, it set foundations for further research on outlier detection in IRT framework. Regarding the atypical response patterns detection, the ongoing study using the Forward Search algorithm looks promising. A small simulation study on Rasch models tested the efficacy of this method and gave the first example of the application of such robust analysis techniques in IRT framework. The diagnostic plots offered by the Forward Search allow to single out anomalous data and, therefore, obtain unbiased inferential results. In addition to detecting outliers, it is of clinical importance to determine the reason for the outlier status or anomalous response. Such response patterns should be carefully analyzed. They may shed more light on the individual's mental status to devise a specific follow-up.

Concerning future developments, the application of IRT analysis to the other psychometric tests used in the project is forthcoming. The factor structure assessment is still an open issue for each scale. We think that MIRT approach can be extended to IES-R and GAD-7 questionnaires. To the best of our knowledge, the application of multidimensional IRT techniques to the above-mentioned questionnaires has never been performed. In addition, it has been shown how IRT can improve the knowledge and the use of such tests.

Moreover, the validation of the Italian version of the so called *Psychosocial Safety Climate-4* (PSC-4) scale using data from *PostCovid* project is in progress in our research group; we then proposed to use IRT method for this purpose.

In conclusion, all workers enrolled within the study have recently started to be under their follow-up: we are collecting information about their mental health status within one year from the first evaluation, using the same questionnaires and the multi-step procedure. We will perform longitudinal analysis to evaluate differences over time. An IRT modelling approach, where a latent variable measuring the severity of the impairment is estimated for each time point, should be engaging in this framework too.

Bibliography

R. J. Adams, M. Wilson, and W.-c. Wang. The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, 21(1):1–23, 1997.

L. Andrews, M. Shevlin, N. Troop, and S. Joseph. Multidimensionality of intrusion and avoidance: Alternative factor models of the impact of event scale. *Personality and Individual Differences*, 36(2):431–446, 2004.

D. Andrich. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, 2(4):581–594, 1978.

D. Andrich and L. Van Schoubroeck. The general health questionnaire: a psychometric analysis using latent trait theory. *Psychological medicine*, 19(2):469–485, 1989.

W. Arrindell, D. P. Barelds, I. C. Janssen, F. M. Buwalda, and J. van der Ende. Invariance of scl-90-r dimensions of symptom distress in patients with peri partum pelvic pain (pppp) syndrome. *British Journal of Clinical Psychology*, 45(3):377–391, 2006.

T. Asparouhov and B. Muthén. Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):495–508, 2014. doi: 10.1080/10705511.2014.919210. URL https://doi.org/10.1080/10705511.2014.919210.

A. C. Atkinson, M. Riani, and M. Riani. *Robust diagnostic regression analysis*, volume 2. Springer, 2000.

A. C. Atkinson, M. Riani, A. Cerioli, et al. *Exploring multivariate data with the forward search*, volume 1. Springer, 2004.

S. Bacci, F. Bartolucci, and M. Gnaldi. A class of multidimensional latent class irt models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods*, 43(4):787–800, 2014.

F. Bartolucci. A class of multidimensional irt models for testing unidimensionality and clustering items. *Psychometrika*, 72(2):141, 2007.

F. Bartolucci, S. Bacci, and M. Gnaldi. *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Chapman and Hall/CRC, 2019.

G. J. Bean and N. K. Bowen. Item response theory and confirmatory factor analysis: Complementary approaches for scale development. *Journal of Evidence-Based Social Work*, pages 1–22, 2021.

C. Beard and T. Björgvinsson. Beyond generalized anxiety disorder: psychometric properties of the gad-7 in a heterogeneous psychiatric sample. *Journal of anxiety disorders*, 28(6):547–552, 2014.

A. L. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

J. Bland and D. Altman. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in biology and medicine*, 20(5):337–340, 1990.

G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149, 2018.

S. Boccia, W. Ricciardi, and J. P. Ioannidis. What other countries can learn from italy during the covid-19 pandemic. *JAMA internal medicine*, 180(7):927–928, 2020.

L. Cai. A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4):581–612, 2010.

M. K. Cain, Z. Zhang, and K.-H. Yuan. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, 49(5):1716–1735, 2017.

A. Campbell and S. Knowles. A confirmatory factor analysis of the ghq12 using a large australian sample. *European Journal of Psychological Assessment*, 23(1):2–8, 2007.

A. Cano, R. P. Sprafkin, D. J. Scaturo, L. J. Lantinga, B. H. Fiese, and F. Brand. Mental health screening in primary care: a comparison of 3 brief measures of psychological distress. *Primary care companion to the Journal of clinical psychiatry*, 3(5):206, 2001.

J. C. Cappelleri, J. J. Lundy, and R. D. Hays. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics*, 36(5):648–662, 2014.

C. Carmassi, C. Foghi, V. Dell'Oste, A. Cordone, C. A. Bertelloni, E. Bui, and L. Dell'Osso. Ptsd symptoms in healthcare workers facing the three coronavirus outbreaks: What can we expect after the covid-19 pandemic. *Psychiatry research*, page 113312, 2020.

K. M. Carpenter and J. B. Hittner. Dimensional characteristics of the scl-90-r: Evaluation of gender differences in dually diagnosed inpatients. *Journal of Clinical Psychology*, 51 (3):383–390, 1995.

R. B. Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

J. Chilcot, L. Rayner, W. Lee, A. Price, L. Goodwin, B. Monroe, N. Sykes, P. Hansford, and M. Hotopf. The factor structure of the phq-9 in palliative care. *Journal of psychosomatic research*, 75(1):60–64, 2013.

A. H. Church and J. Waclawski. *Alternative validation strategies: Developing new and leveraging existing validity evidence*, volume 19. John Wiley & Sons, 2007.

G. A. Churchill Jr. A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, 16(1):64–73, 1979.

J. M. Conijn, W. H. Emons, and K. Sijtsma. Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement*, 38(2):122–136, 2014.

G. Craparo, P. Faraci, G. Rotondo, and A. Gori. The impact of event scale–revised: psychometric properties of the italian version in a sample of flood victims. *Neuropsychiatric Disease and Treatment*, 9:1427, 2013.

M. Creamer, R. Bell, and S. Failla. Psychometric properties of the impact of event scale-revised. *Behaviour research and therapy*, 41(12):1489–1496, 2003.

L. Crocker and J. Algina. *Introduction to classical and modern test theory*. ERIC, 1986.

L. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16: 297–334, 1951.

E. E. Cureton. Corrected item-test correlations. *Psychometrika*, 31(1):93–96, 1966.

J. Cyr, J. McKenna-Foley, and E. Peacock. Factor structure of the scl-90-r: is there one? *Journal of personality assessment*, 49(6):571–578, 1985.

L.-L. Dai, X. Wang, T.-C. Jiang, P.-F. Li, Y. Wang, S.-J. Wu, L.-Q. Jia, M. Liu, L. An, and Z. Cheng. Anxiety and depressive symptoms among covid-19 patients in jianghan fangcang shelter hospital in wuhan, china. *Plos one*, 15(8):e0238416, 2020.

P. De Boeck. Random item irt models. *Psychometrika*, 73(4):533–559, 2008.

J. de Jesus Mari and P. Williams. Misclassification by psychiatric screening questionnaires. *Journal of Chronic Diseases*, 39(5):371–378, 1986.

J. H. De Kock, H. A. Latham, S. J. Leslie, M. Grindle, S.-A. Munoz, L. Ellis, R. Polson, and C. M. O'Malley. A rapid review of the impact of covid-19 on the mental health of healthcare workers: implications for supporting psychological well-being. *BMC Public Health*, 21(1):1–18, 2021.

G. S. de Pablo, J. Vaquerizo-Serrano, A. Catalan, C. Arango, C. Moreno, F. Ferre, J. I. Shin, S. Sullivan, N. Brondino, M. Solmi, et al. Impact of coronavirus syndromes on

physical and mental health of health care workers: Systematic review and meta-analysis. *Journal of affective disorders*, 275:48–57, 2020.

S. Depaoli, J. Tiemensma, and J. M. Felt. Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, health & medicine*, 23(sup1): 1299–1317, 2018.

L. R. Derogatis. Scl-90-r: Administration, scoring & procedures manual-ii for the (revised) version and other instruments of the psychopathology rating scale series. *Clinical Psychometric Research.*, pages 1–16, 1992.

G. d'Ettorre, G. Ceccarelli, L. Santinelli, P. Vassalini, G. P. Innocenti, F. Alessandri, A. E. Koukopoulos, A. Russo, G. d'Ettorre, and L. Tarsitani. Post-traumatic stress symptoms in healthcare workers dealing with the covid-19 pandemic: a systematic review. *International Journal of Environmental Research and Public Health*, 18(2):601, 2021.

R. F. DeVellis. *Scale development: Theory and applications*, volume 26. Sage publications, 2016.

F. Drasgow, M. V. Levine, and E. A. Williams. Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1):67–86, 1985.

O. D. Duncan and M. Stenbeck. Are likert scales unidimensional? *Social Science Research*, 16(3):245–259, 1987.

J. D. Elhai, A. A. Contractor, M. Tamburrino, T. H. Fine, M. R. Prescott, E. Shirley, P. K. Chan, R. Slembarski, I. Liberzon, S. Galea, et al. The factor structure of major depression symptoms: a test of four competing models using the patient health questionnaire-9. *Psychiatry research*, 199(3):169–173, 2012.

A. Fattori, F. Cantù, A. Comotti, V. Tombola, E. Colombo, C. Nava, L. Bordini, L. Riboldi, M. Bonzini, and P. Brambilla. Hospital workers mental health during the covid-19 pandemic: methods of data collection and characteristics of study sample in a university hospital in milan (italy). *BMC Medical Research Methodology*, 21(1):1–12, 2021.

J. M. Felt, R. Castaneda, J. Tiemensma, and S. Depaoli. Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, 8(MAY):1–9, 2017. ISSN 16641078. doi: 10.3389/fpsyg.2017.00863.

G. A. Ferguson. Item selection by the constant process. *Psychometrika*, 7(1):19–29, 1942.

H. M. Fernandes and J. Vasconcelos-Raposo. Factorial validity and invariance of the ghq-12 among clinical and nonclinical samples. *Assessment*, 20(2):219–229, 2013.

A. K. Formann. Linear logistic latent class analysis and the rasch model. In *Rasch Models*, pages 239–255. Springer, 1995.

F. J. Fowler Jr and F. J. Fowler. *Improving survey questions: Design and evaluation*. Sage, 1995.

J.-P. Fox. *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media, 2010.

R. C. Geary. Testing for Normality. *Biometrika*, 34(3/4):209, dec 1947. ISSN 00063444. doi: 10.2307/2332434. URL https://www.jstor.org/stable/2332434https://www.jstor.org/stable/2332434?origin=crossref.

G. Giorgi, J. Leon Perez, D. Castiello, A. Antonio, F. Fiz Perez, G. Arcangeli, et al. The general health questionaire (ghq-12) in a sample of italian workers: mental health at individual and organizational level. *World Journal of Medical Sciences*, 11(1):47–56, 2014.

G. Giorgi, F. S. F. Perez, A. C. D'Antonio, N. Mucci, C. Ferrero, V. Cupelli, and G. Arcangeli. Psychometric properties of the impact of event scale-6 in a sample of victims of bank robbery. *Psychology research and behavior management*, 8:99, 2015.

C. Glas and A. V. T. Dagohoy. A person fit test for irt models for polytomous items. *Psychometrika*, 72(2):159–180, 2007.

A. Glockner-Rist and H. Hoijtink. The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10(4):544–565, 2003.

T. Gnambs and T. Staufenbiel. The structure of the general health questionnaire (ghq-12): two meta-analytic factor analyses. *Health psychology review*, 12(2):179–194, 2018.

D. Goldberg. Use of the general health questionnaire in clinical work. *British medical journal (Clinical research ed.)*, 293(6556):1188, 1986.

D. P. Goldberg and V. F. Hillier. A scaled version of the general health questionnaire. *Psychological medicine*, 9(1):139–145, 1979.

D. P. Goldberg, R. Gater, N. Sartorius, T. B. Ustun, M. Piccinelli, O. Gureje, and C. Rutter. The validity of two versions of the ghq in the who study of mental illness in general health care. *Psychological medicine*, 27(1):191–197, 1997.

C. González-Blanch, L. A. Medrano, R. Muñoz-Navarro, P. Ruíz-Rodríguez, J. A. Moriana, J. T. Limonero, F. Schmitz, A. Cano-Vindel, and P. R. Group. Factor structure and measurement invariance across various demographic groups and over time for the phq-9 in primary care patients in spain. *PloS one*, 13(2):e0193356, 2018.

M. Goodchild and P. Duncan-Jones. Chronicity and the general health questionnaire. *The British Journal of Psychiatry*, 146(1):55–61, 1985.

V. V. Gouveia, G. A. Barbosa, E. d. O. Andrade, and M. B. Carneiro. Factorial validity and reliability of the general health questionnaire (ghq-12) in the brazilian physician population. *Cadernos de Saúde Pública*, 26:1439–1445, 2010.

B. Graetz. Multidimensional properties of the general health questionnaire. *Social psychiatry and psychiatric epidemiology*, 26(3):132–138, 1991.

B. Guo, C. Kaylor-Hughes, A. Garland, N. Nixon, T. Sweeney, S. Simpson, T. Dalgleish, R. Ramana, M. Yang, and R. Morriss. Factor structure and longitudinal measurement invariance of phq-9 for specialist mental health care patients with persistent major depressive disorder: Exploratory structural equation modelling. *Journal of affective disorders*, 219:1–8, 2017.

L. Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.

A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):761–771, 1992.

R. K. Hambleton and H. Swaminathan. A look at psychometrics in the netherlands. 1985.

R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of item response theory*, volume 2. Sage, 1991.

M. Hankins. The factor structure of the twelve item general health questionnaire (ghq-12): the result of negative phrasing? *Clinical Practice and Epidemiology in Mental Health*, 4(1):1–8, 2008.

Q. Hao, D. Wang, M. Xie, Y. Tang, Y. Dou, L. Zhu, Y. Wu, M. Dai, H. Wu, and Q. Wang. Prevalence and risk factors of mental health problems among healthcare workers during the covid-19 pandemic: a systematic review and meta-analysis. *Frontiers in Psychiatry*, 12, 2021.

J. Hattie. Methodology review: assessing unidimensionality of tests and ltenls. *Applied psychological measurement*, 9(2):139–164, 1985.

J. R. Hébert and D. R. Miller. The inappropriateness of conventional use of the correlation coefficient in assessing validity and reliability of dietary assessment methods. *European journal of epidemiology*, 7(4):339–343, 1991.

T. Heinen. *Latent class and discrete latent trait models: Similarities and differences.* Sage Publications, Inc, 1996.

T. R. Hinkin. A review of scale development practices in the study of organizations. *Journal of management*, 21(5):967–988, 1995.

A. Hinz, A. M. Klein, E. Brähler, H. Glaesmer, T. Luck, S. G. Riedel-Heller, K. Wirkner, and A. Hilbert. Psychometric evaluation of the generalized anxiety disorder screener gad-7, based on a large german general population sample. *Journal of affective disorders*, 210:338–344, 2017.

M. Hong and A. Cheng. Robust maximum marginal likelihood (rmml) estimation for item response theory models. 2018.

M. Hong, J. T. Steedle, and Y. Cheng. Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and psychological measurement*, 80(2):312–345, 2020.

M. R. Hong and Y. Cheng. Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior research methods*, 51(2):573–588, apr 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-1150-4. URL http://link.springer.com/10.3758/s13428-018-1150-4http://www.ncbi.nlm.nih.gov/pubmed/30350024.

J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

M. Horowitz, N. Wilner, and W. Alvarez. Impact of event scale: A measure of subjective stress. *Psychosomatic medicine*, 41(3):209–218, 1979.

L.-t. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

S. W. Hystad and B. H. Johnsen. The dimensionality of the 12-item general health questionnaire (ghq-12): Comparisons of factor structures and invariance across samples and time. *Frontiers in Psychology*, 11:1300, 2020.

J. C. Immekus, K. E. Snyder, and P. A. Ralston. Multidimensional item response theory for factor structure assessment in educational psychology research. In *Frontiers in Education*, volume 4, page 45. Frontiers, 2019.

D. L. Jackson, J. A. Gillaspy Jr, and R. Purc-Stephenson. Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological methods*, 14(1):6, 2009.

S. U. Johnson, P. G. Ulvenes, T. Øktedalen, and A. Hoffart. Psychometric properties of the general anxiety disorder 7-item (gad-7) scale in a heterogeneous psychiatric sample. *Frontiers in psychology*, 10:1713, 2019.

T. J. Kalliath, M. P. O'Driscoll, and P. Brough. A confirmatory factor analysis of the general health questionnaire-12. *Stress and Health: Journal of the International Society for the Investigation of Stress*, 20(1):11–20, 2004.

A. Kamata and D. J. Bauer. A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1):136–153, 2008.

G. Karabatsos. Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4):277–298, 2003.

H. Kelderman. Multidimensional rasch models for partial-credit scoring. *Applied Psychological Measurement*, 20(2):155–168, 1996.

H. Kelderman. Loglinear multidimensional item response models for polytomously scored items. In *Handbook of modern item response theory*, pages 287–304. Springer, 1997.

H. Kelderman and C. P. Rijkes. Loglinear multidimensional irt models for polytomously scored items. *Psychometrika*, 59(2):149–176, 1994.

F. Kendel, M. Wirtz, A. Dunkel, E. Lehmkuhl, R. Hetzer, and V. Regitz-Zagrosek. Screening for depression: Rasch analysis of the dimensional structure of the phq-9 and the hads-d. *Journal of Affective Disorders*, 122(3):241–246, 2010.

S. Kertz, J. Bigda-Peyton, and T. Bjorgvinsson. Validity of the generalized anxiety disorder-7 scale in an acute psychiatric sample. *Clinical psychology & psychotherapy*, 20(5):456–464, 2013.

A. P. Keszei, M. Novak, and D. L. Streiner. Introduction to health measurement scales. *Journal of psychosomatic research*, 68(4):319–323, 2010.

B. T. Keum, M. J. Miller, and K. K. Inkelas. Testing the factor structure and measurement invariance of the phq-9 across racially diverse us college students. *Psychological assessment*, 30(8):1096, 2018.

S.-H. Kim, A. S. Cohen, and T.-H. Park. Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32(3):261–276, 1995.

D. W. King, R. J. Orazem, D. Lauterbach, L. A. King, C. L. Hebenstreit, and A. Y. Shalev. Factor structure of posttraumatic stress disorder as measured by the impact of event scale–revised: Stability across cultures and time. *Psychological Trauma: Theory, Research, Practice, and Policy*, 1(3):173, 2009.

T. Kline. *Psychological testing: A practical approach to design and evaluation*. Sage, 2005.

D. L. Knol and M. P. Berger. Empirical comparison between factor analysis and multidimensional item response models. *Multivariate behavioral research*, 26(3):457–477, 1991.

J. S. Krause, K. S. Reed, and J. J. McArdle. Factor structure and predictive validity of somatic and nonsomatic symptoms from the patient health questionnaire-9: a longitudinal study after spinal cord injury. *Archives of physical medicine and rehabilitation*, 91(8):1218–1224, 2010.

K. Kroenke, R. L. Spitzer, and J. B. Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

K. Kroenke, J. Wu, Z. Yu, M. Bair, J. Kean, T. Stump, and P. Monahan. Patient health questionnaire anxiety and depression scale: Initial validation in three clinical trials. copyright: American psychosomatic society. 2016.

G. F. Kuder and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2:151–160, 1937.

R. Langeheine. New developments in latent class theory. In *Latent trait and latent class models*, pages 77–108. Springer, 1988.

W.-Y. Lee, S.-J. Cho, R. W. McGugin, A. B. Van Gulick, and I. Gauthier. Differential item functioning analysis of the vanderbilt expertise test for cars. *Journal of vision*, 15 (13):23–23, 2015.

B. Lindsay, C. C. Clogg, and J. Grego. Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86(413):96–107, 1991.

C. H. Liu, E. Zhang, G. T. F. Wong, S. Hyun, et al. Factors associated with depression, anxiety, and ptsd symptomatology during the covid-19 pandemic: Clinical implications for us young adult mental health. *Psychiatry research*, 290:113172, 2020a.

Y. Liu and H. Liu. Detecting noneffortful responses based on a residual method using an iterative purification process. *Journal of Educational and Behavioral Statistics*, page 1076998621994366, 2021.

Y. Liu, Y. Cheng, and H. Liu. Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, 80(4):775–807, 2020b.

F. Lord. A theory of test scores. *Psychometric monographs*, 1952.

F. M. Lord and M. R. Novick. *Statistical theories of mental test scores*. IAP, 1968.

B. Löwe, K. Kroenke, W. Herzog, and K. Gräfe. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the patient health questionnaire (phq-9). *Journal of affective disorders*, 81(1):61–66, 2004.

B. Löwe, O. Decker, S. Müller, E. Brähler, D. Schellberg, W. Herzog, and P. Y. Herzberg. Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. *Medical care*, pages 266–274, 2008.

D. Magis, G. Raîche, S. Béland, and P. Gérard. A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, 11(4):365–386, 2011.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748, 1959.

A. Martin, W. Rief, A. Klaiberg, and E. Braehler. Validity of the brief patient health questionnaire mood scale (phq-9) in the general population. *General hospital psychiatry*, 28(1):71–77, 2006.

G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

R. Maunder. The experience of the 2003 sars outbreak as a traumatic stress among frontline healthcare workers in toronto: lessons learned. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1447):1117–1125, 2004.

D. Mavridis and I. Moustaki. The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, 18(4):1016–1034, 2009.

A. Maydeu-Olivares and H. Joe. Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, 100(471):1009–1020, 2005.

R. P. McDonald. Linear versus models in item response theory. *Applied Psychological Measurement*, 6(4):379–396, 1982.

R. P. McDonald. A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24(2):99–114, 2000.

T. Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, 105(1):156, 1989.

R. J. Mislevy and R. D. Bock. Biweight estimates of latent ability. *Educational and psychological measurement*, 42(3):725–737, 1982.

R. A. E. Muller, R. S. Ø. Stensland, R. S. van de Velde, et al. The mental health impact of the covid-19 pandemic on healthcare workers, and interventions to help them: A rapid systematic review. *Psychiatry research*, page 113441, 2020.

E. Muraki and J. E. Carlson. Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19(1):73–90, 1995.

B. Muthén. A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of educational statistics*, 10(2):121–132, 1985.

J. C. Nunnally. *Psychometric theory 3E*. Tata McGraw-hill education, 1994.

P. Osteen. An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research*, 1(2): 66–82, 2010.

N. Panchapakesan. *The simple logistic model and mental measurement.* PhD thesis, University of Chicago, Department of Education, 1969.

S. Pappa, V. Ntella, T. Giannakas, V. G. Giannakoulis, E. Papoutsi, and P. Katsaounou. Prevalence of depression, anxiety, and insomnia among healthcare workers during the covid-19 pandemic: A systematic review and meta-analysis. *Brain, behavior, and immunity*, 88:901–907, 2020.

J. M. Patton, Y. Cheng, M. Hong, and Q. Diao. Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, 44(3):309–341, 2019.

R. D. Penfield. Assessing differential item functioning among multiple groups: A comparison of three mantel-haenszel procedures. *Applied Measurement in Education*, 14(3): 235–259, 2001.

R. D. Penfield. Generalized full-information item bifactor analysis. *Psychological Methods*, 16(3):221–248, 2011.

J. Petrillo, S. J. Cano, L. D. McLeod, and C. D. Coon. Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value in Health*, 18(1):25–34, 2015.

M. Petropoulou, G. Salanti, G. Rücker, G. Schwarzer, I. Moustaki, and D. Mavridis. A forward search algorithm for detecting extreme study effects in network meta-analysis. *Statistics in medicine*, 40(25):5642–5656, 2021.

A. Picardi, D. Abeni, and P. Pasquini. Assessing psychological distress in patients with skin diseases: reliability, validity and factor structure of the ghq-12. *Journal of the European Academy of Dermatology and Venereology*, 15(5):410–417, 2001.

M. Piccinelli, G. Bisoffi, M. G. Bon, L. Cunico, and M. Tansella. Validity and test-retest reliability of the italian version of the 12-item general health questionnaire in general practice: a comparison between three scoring methods. *Comprehensive psychiatry*, 34 (3):198–205, 1993.

D. F. Polit. Assessing measurement in health: Beyond reliability and validity. *International journal of nursing studies*, 52(11):1746–1753, 2015.

P. Politi, M. Piccinelli, and G. Wilkinson. Reliability, validity and factor structure of the 12-item general health questionnaire among young males in italy. *Acta Psychiatrica Scandinavica*, 90(6):432–437, 1994.

M. W. Rabow, C.-H. S. Huang, G. E. White-Hammond, and R. O. Tucker. Witnesses and victims both: Healthcare workers and grief in the time of covid-19. *Journal of Pain and Symptom Management*, 2021.

G. Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

T. Raykov and G. A. Marcoulides. *Introduction to psychometric theory*. Routledge, 2011.

M. D. Reckase. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

D. Reid. The detection of psychiatric illness by questionnaire by dp goldberg.(pp. 156; illustrated;£ 3· 50.) oxford university press: London. 1972. *Psychological Medicine*, 3 (2):257–257, 1973.

S. P. Reise. The rediscovery of bifactor measurement models. *Multivariate behavioral research*, 47(5):667–696, 2012.

W. Revelle and T. Rocklin. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14(4): 403–414, 1979.

E. J. Richardson and J. S. Richards. Factor structure of the phq-9 screen for depression across time since injury among persons with spinal cord injury. *Rehabilitation Psychology*, 53(2):243, 2008.

M. W. Richardson. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1(2):33–49, 1936.

J. Rost. Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3):271–282, 1990.

J. Rost and M. von Davier. Mixture distribution rasch models. In *Rasch models*, pages 257–268. Springer, 1995.

V. Rousson, T. Gasser, and B. Seifert. Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in medicine*, 21(22):3431–3446, 2002.

L. A. Rutter and T. A. Brown. Psychometric properties of the generalized anxiety disorder scale-7 (gad-7) in outpatients with anxiety and mood disorders. *Journal of psychopathology and behavioral assessment*, 39(1):140–146, 2017.

F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.

A. Schimmenti. Dissociative experiences and dissociative minds: exploring a nomological network of dissociative functioning. *Journal of Trauma & Dissociation*, 17(3):338–361, 2016.

N. Schrnitz, J. Kruse, and W. Tress. Psychometric properties of the general health questionnaire (ghq-12) in a german primary care sample. *Acta Psychiatrica Scandinavica*, 100(6):462–468, 1999.

C. Schuster and K.-H. Yuan. Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, 36(6):720–735, 2011.

R. Shealy and W. Stout. A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dtf as well as item bias/dif. *Psychometrika*, 58(2):159–194, 1993.

Y. Sheng and C. K. Wikle. Bayesian irt models incorporating general and specific abilities. *Behaviormetrika*, 36(1):27–48, 2009.

M. Shevlin, N. Hunt, and I. Robbins. A confirmatory factor analysis of the impact of event scale using a sample of world war ii and korean war veterans. *Psychological Assessment*, 12(4):414, 2000.

A. B. Smith, L. J. Fallowfield, D. P. Stark, G. Velikova, and V. Jenkins. A rasch and confirmatory factor analysis of the general health questionnaire (ghq)-12. *Health and quality of life outcomes*, 8(1):1–10, 2010.

A. B. Smith, Y. Oluboyede, R. West, J. Hewison, and A. O. House. The factor structure of the ghq-12: the interaction between item phrasing, variance and levels of distress. *Quality of Life Research*, 22(1):145–152, 2013.

A. C. d. Souza, N. M. C. Alexandre, and E. d. B. Guirardello. Psychometric properties in instruments evaluation of reliability and validity. *Epidemiologia e Serviços de Saúde*, 26:649–659, 2017.

C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412107.

R. L. Spitzer, K. Kroenke, J. B. Williams, P. H. Q. P. C. S. Group, P. H. Q. P. C. S. Group, et al. Validation and utility of a self-report version of prime-md: the phq primary care study. *Jama*, 282(18):1737–1744, 1999.

R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.

J. Stochl, E. I. Fried, J. Fritz, T. J. Croudace, D. A. Russo, C. Knight, P. B. Jones, and J. Perez. On dimensionality, measurement invariance, and suitability of sum scores for the phq-9 and the gad-7. *Assessment*, page 1073191120976863, 2020.

D. L. Streiner. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, 80(1):99–103, 2003.

H. Suzuki, Y. Kaneita, Y. Osaki, M. Minowa, H. Kanda, K. Suzuki, K. Wada, K. Hayashi, T. Tanihata, and T. Ohida. Clarification of the factor structure of the 12-item general health questionnaire among japanese adolescents and associated sleep status. *Psychiatry research*, 188(1):138–146, 2011.

H. Swaminathan and H. J. Rogers. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4):361–370, 1990.

R. J. Tait, G. K. Hulse, and S. I. Robertson. A review of the validity of the general health questionnaire in adolescent populations. *Australian & New Zealand Journal of Psychiatry*, 36(4):550–557, 2002.

R. J. Tait, D. J. French, and G. K. Hulse. Validity and psychometric properties of the general health questionnaire-12 in young australian adolescents. *Australian & New Zealand Journal of Psychiatry*, 37(3):374–381, 2003.

Y. Takane and J. De Leeuw. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408, 1987.

J. A. Teresi, M. Kleinman, and K. Ocepek-Welikson. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in medicine*, 19(11-12):1651–1683, 2000.

C. B. Terwee, S. D. Bot, M. R. de Boer, D. A. van der Windt, D. L. Knol, J. Dekker, L. M. Bouter, and H. C. de Vet. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, 60(1):34–42, 2007.

A. Teymoori, A. Gorbunova, F. E. Haghish, R. Real, M. Zeldovich, Y.-J. Wu, S. Polinder, T. Asendorf, D. Menon, et al. Factorial structure and validity of depression (phq-9) and anxiety (gad-7) scales after traumatic brain injury. *Journal of clinical medicine*, 9 (3):873, 2020.

D. Thissen. Use of item response theory in the study of group differences in trace lines. *Test validity*, 1988.

D. Thissen and L. Steinberg. Data analysis using item response theory. *Psychological Bulletin*, 104(3):385, 1988.

D. Thissen, L. Steinberg, and H. Wainer. Detection of differential item functioning using the parameters of item response models. 1993.

S. Thoresen, K. Tambs, A. Hussain, T. Heir, V. A. Johansen, and J. I. Bisson. Brief measure of posttraumatic stress reactions: Impact of event scale-6. *Social psychiatry and psychiatric epidemiology*, 45(3):405–412, 2010.

L. L. Thurstone. A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16:433–451, 1925a.

L. L. Thurstone. A method of scaling psychological and educational tests. *Journal of educational psychology*, 16(7):433, 1925b.

E. Timminga and J. J. Adema. Test construction from item banks. In *Rasch Models*, pages 111–127. Springer, 1995.

W. J. Van Der Linden and R. K. Hambleton. Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory*, pages 1–28. Springer, 1997.

O. Vassend and A. Skrondal. The problem of structural indeterminacy in multidimensional symptom report instruments. the case of scl-90-r. *Behaviour Research and Therapy*, 37 (7):685–701, 1999.

W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

H. Wainer. Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, 8(2): 157–86, 1995.

H. Wainer and B. D. Wright. Robust estimation of ability in the rasch model. *Psychometrika*, 45(3):373–391, 1980.

A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3): 426–482, 1943.

M. I. Waller. Removing the effects of random guessing from latent trait ability estimates. *ETS Research Bulletin Series*, 1974(1):i–50, 1974.

D. S. Weiss. The impact of event scale: revised. In *Cross-cultural assessment of psychological trauma and PTSD*, pages 219–238. Springer, 2007.

U. Werneke, D. P. Goldberg, I. Yalcin, and B. Üstün. The stability of the factor structure of the general health questionnaire. *Psychological medicine*, 30(4):823–829, 2000.

R. J. Wherry and R. H. Gaylord. Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. *Psychometrika*, 9(4):237–244, 1944.

C. M. Woods. Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33(1):42–57, 2009.

L. Yao and R. D. Schwarz. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied psychological measurement*, 30(6):469–492, 2006.

K.-H. Yuan and B. Gomer. An overview of applied robust methods. *British Journal of Mathematical and Statistical Psychology*, 2021.

B. D. Zumbo, Y. Liu, A. D. Wu, B. R. Shear, O. L. O. Astivia, and T. K. Ark. A methodology for zumbo's third generation dif analyses and the ecology of item responding. *Language Assessment Quarterly*, 12(1):136–151, 2015. doi: 10.1080/15434303.2014.972559. URL https://doi.org/10.1080/15434303.2014.972559.

Appendix

Questionnaires (in Italian) administered to all participants are attached in this appendix.

They consist in the first-level evaluation with GHQ-12, IES-R and GAD-7 scales and in the second-level assessment (PHQ-9, DES and SCL-90 scales).

SEZIONE I – <u>COMPILAZIONE A CURA DEL MEDICO COMPETENTE</u>

1. **Data di compilazione:** ……./……./…….

2. **Nome:** ……
3. **Cognome:** ……
4. **Numero di matricola:** ……
5. **Numero di telefono:** ……
6. **E-mail:** ……

In caso emergesse l'indicazione di un sostegno o di un approfondimento preferisce essere contattata/o tramite:

❑e-mail ❑cellulare

7. **Mansione attuale:**

❑1 Dirigente Medico
❑2 Dirigente Medico Responsabile di UOC - UOSD
❑3 Dirigente Non Medico (Amministrativo, Professionale, Sanitario, Tecnico)
❑4 Collaboratore Professionale Sanitario (Infermiere, Ostetrico, Puericultrice, ecc.)
❑5 Responsabile Infermieristico di Unità Operativa (RIOU) o di Area (RIA)
❑6 Operatore Socio-Sanitario (OSS)
❑7 Personale Tecnico-Sanitario (Tec. Radio., Tec. Lab., Tec. Neurofis., ecc.)
❑8 Personale Tecnico Ausiliario (OTA, Ausiliario, Ass. Tecnico, Coll. Tecnico, Op. Tecnico, ecc.)
❑9 Personale Amministrativo

8. **Appartenente all'Unità Operativa di:**

……..……..…da: ..…. (mese) …… (anno)

9. **Da quanti anni lavora nella sua <u>attuale mansione</u>?** I__I__I anni

10. **Da quanti anni lavora presso <u>questo Ospedale</u>?** I__I__I anni

11. **Complessivamente, da quanti anni lavora nel <u>settore sanitario</u>?** I__I__I anni

12. **Attualmente lavora in area Covid?** □ SI □ NO

(Se sì): □ AREA INTENSIVA □ AREA SUB-INTENSIVA □ AREA BASSA INTENSITA'

Specificare la durata del servizio: lavora in area Covid da: ..… (giorno) …… (mese)

13. **<u>Se no</u>, ha lavorato precedentemente in area Covid?** □ SI □ NO

(Se sì): □ AREA INTENSIVA □ AREA SUB-INTENSIVA □ AREA BASSA INTENSITA'

Specificare la durata del servizio: ha lavorato in area Covid da: ..… (g) … (m) A …. (g) … (m)

14. **Soffre di una o più malattie croniche?** ❑1 SI ❑2 NO

Se sì, indichi quella/e principali:

1)_____

2)_____

3)_____

15. **Indicare le terapie farmacologiche abituali:**
<u>(Tra queste, sottolineare il nome di quelle iniziate da marzo 2020)</u>

❑2 Tranquillanti, ansiolitici, sonniferi, antidepressivi ❑3 Antidolorifici, analgesici, antinfiammatori ❑4 Diuretici	❑5 Cardiologici, antianginosi, ecc. (farmaci per il cuore) ❑6 Farmaci per la pressione alta ❑7 Vasodilatatori ❑8 Broncodilatatori	❑9 Antiacidi, antiulcera, gastroprot. ❑10 Vitamine ricostituenti ❑11 Antispastici, lassativi ❑12 Antibiotici ❑13 Altro: ……. ……..

SEZIONE II – <u>COMPILAZIONE A CURA DEL LAVORATORE</u>

16. Le seguenti domande riguardano le Sue condizioni psico-fisiche generali <u>nelle ultime due settimane</u>. Segni con una croce la Sua risposta.

	Nelle ultime due settimane:				
A	E' stata/o capace di concentrarsi su quello che stava facendo?	meglio del solito	come al solito	meno del solito	molto meno del solito
B	Ha perso molto sonno a causa di preoccupazioni?	per niente	non più del solito	più del solito	molto più del solito
C	Ha avuto la sensazione di giocare un ruolo utile in ciò che stava facendo?	più del solito	come al solito	meno del solito	molto meno del solito
D	Si è sentita/o in grado di prendere decisioni?	più del solito	come al solito	meno del solito	molto meno del solito
E	Si è sentita/o costantemente sotto pressione/stress?	per niente	non più del solito	più del solito	molto più del solito
F	Si è sentita/o di non poter superare le difficoltà?	per niente	non più del solito	più del solito	molto più del solito
G	E' stata/o in grado di godere delle sue normali attività quotidiane?	più del solito	come al solito	meno del solito	molto meno del solito
H	E' stata/o capace di far fronte ai suoi problemi?	più del solito	come al solito	meno del solito	molto meno del solito
I	Si è sentita/o infelice e depressa/o?	per niente	non più del solito	più del solito	molto più del solito
J	Ha avuto la sensazione di perdere fiducia in sè stessa/o?	per niente	non più del solito	più del solito	molto più del solito
K	Ha pensato a sè stessa/o come a una persona priva di valore?	per niente	non più del solito	più del solito	molto più del solito
L	Tutto considerato, si è sentita/o discretamente felice?	più del solito	come al solito	meno del solito	molto meno del solito

17. La seguente è una lista di difficoltà che le persone hanno talvolta in seguito ad eventi particolarmente stressanti e potenzialmente minacciosi per la vita, come l'emergenza Covid19.
Legga per favore ogni frase, indicando quanto l'ha coinvolta ognuna delle difficoltà in questione <u>negli ultimi sette giorni</u>. Se negli ultimi sette giorni non ci ha pensato, faccia una crocetta nella colonna contraddistinta da "Per niente".

	Per niente	Un poco	Moderatamente	Abbastanza	Estremamente
1. Ogni cosa che mi ricordava l'emergenza Covid19 mi faceva vivere emozioni relative ad essa	❑0	❑1	❑2	❑3	❑4
2. Ho avuto difficoltà a restare addormentato	❑0	❑1	❑2	❑3	❑4
3. Altre cose hanno continuato a farmi pensare all'emergenza Covid19	❑0	❑1	❑2	❑3	❑4
4. Mi sono sentito/a irritabile ed arrabbiato/a	❑0	❑1	❑2	❑3	❑4
5. Ho evitato di lasciarmi sconvolgere quando ho pensato all'emergenza Covid19 o mi è stata ricordata	❑0	❑1	❑2	❑3	❑4
6. Ho pensato all'emergenza Covid19 senza averne l'intenzione	❑0	❑1	❑2	❑3	❑4
7. Ho avuto la sensazione che l'emergenza Covid19 non fosse successa o non fosse reale	❑0	❑1	❑2	❑3	❑4
8. Ho cercato di evitare le cose che potevano ricordarmi l'emergenza Covid19	❑0	❑1	❑2	❑3	❑4
9. Le immagini dell'emergenza Covid19 mi entravano nella mente all'improvviso	❑0	❑1	❑2	❑3	❑4
10. Sono stato/a nervoso/a e mi sono spaventato/a facilmente	❑0	❑1	❑2	❑3	❑4
11. Ho cercato di non pensarci	❑0	❑1	❑2	❑3	❑4
12. Ero consapevole di avere ancora molte emozioni su l'emergenza Covid19, ma non sono riuscito/a a gestirle	❑0	❑1	❑2	❑3	❑4
13. Le mie emozioni riguardo all'emergenza Covid19 sono state una specie di intontimento	❑0	❑1	❑2	❑3	❑4
14. Mi sono ritrovato/a a comportarmi o a provare emozioni come se fossi ritornato/a indietro a quel momento	❑0	❑1	❑2	❑3	❑4

15. Ho avuto difficoltà ad addormentarmi	❑0	❑1	❑2	❑3	❑4
16. Ho provato ondate di forti emozioni relative all'emergenza Covid19	❑0	❑1	❑2	❑3	❑4
17. Ho cercato di rimuovere l'emergenza Covid19 dalla memoria	❑0	❑1	❑2	❑3	❑4
18. Ho avuto difficoltà a concentrarmi	❑0	❑1	❑2	❑3	❑4
19. Cose che me l'hanno fatta ricordare mi hanno provocato reazioni fisiche come sudorazione, difficoltà a respirare, nausea o accelerazione del cuore	❑0	❑1	❑2	❑3	❑4
20. Ho fatto sogni sull'emergenza Covid19	❑0	❑1	❑2	❑3	❑4
21. Mi sono ritrovato/a ad essere guardingo/a e vigilante rispetto all'ambiente o alle persone	❑0	❑1	❑2	❑3	❑4
22. Ho cercato di non parlarne	❑0	❑1	❑2	❑3	❑4

18. Le seguenti domande fanno riferimento a questioni inerenti la <u>salute e sicurezza psicologica</u> nel suo ambiente di lavoro. Indichi il suo grado di accordo o disaccordo con ciascuna frase:

	Totalmente in disaccordo	**In disaccordo**	**Né in accordo né in disaccordo**	**D'accordo**	**Totalmente d'accordo**
1. L'ospedale dimostra di sostenere la prevenzione dello stress con impegno ed interesse.	❑1	❑2	❑3	❑4	❑5
2. L'ospedale considera la salute psicologica dei lavoratori importante tanto quanto la produttività.	❑1	❑2	❑3	❑4	❑5
3. Nel mio ospedale posso parlare tranquillamente delle problematiche di salute e sicurezza psicologica che mi riguardano	❑1	❑2	❑3	❑4	❑5
4. Nel mio ospedale la prevenzione dello stress coinvolge tutti i livelli organizzativi.	❑1	❑2	❑3	❑4	❑5

19. ***Nelle ultime due settimane, con quale frequenza le ha dato fastidio ciascuno dei seguenti problemi?***

Nelle ultime due settimane:	**Mai**	**Alcuni giorni**	**Per oltre la metà dei giorni**	**Quasi ogni giorno**
1. Sentirsi nervoso/a, ansioso/a o teso/a	0	1	2	3
2. Non riuscire a smettere di preoccuparsi o a tenere sotto controllo le preoccupazioni	0	1	2	3
3. Preoccuparsi troppo per varie cose	0	1	2	3
4. Avere difficoltà a rilassarsi	0	1	2	3
5. Essere talmente irrequieto/a da far fatica a stare seduto/a fermo/a	0	1	2	3
6. Infastidirsi o irritarsi facilmente	0	1	2	3
7. Avere paura che possa succedere qualcosa di terribile	0	1	2	3

Sempre in riferimento all'emergenza Covid19, da marzo 2020 ad oggi:

20. **E' stata/o in quarantena**? ❑1 SI ❑2 NO
    Se si, indichi il periodo di inizio e fine della quarantena**:** da ….. (g) / ….. (m) A ….. (g) / ….. (m)

21. **E' risultata/o positiva/o al Covid19?** ❑1 SI ❑2 NO
    Se si, indichi il periodo in cui è stata/o positivo**:** da ….. (g) / ….. (m) A ….. (g) / ….. (m)

22. **Qualcuno dei colleghi con cui ha lavorato è risultato positivo al Covid19?** ❑1 SI ❑2 NO

23. **Qualcuno dei suoi familiari/conviventi è risultato positivo al Covid19**? ❑1 SI ❑2 NO

24. **Qualcuno dei suoi familiari/conviventi è stato ricoverato a causa del Covid19**? ❑1 SI ❑2 NO

25. **Qualcuno dei suoi familiari/conviventi è deceduto a causa del Covid19**? ❑1 SI ❑2 NO

26. **Si è preoccupata/o di poter contagiare o aver contagiato i suoi familiari?**
    ❑1 molto ❑2 abbastanza ❑3 poco ❑4 per nulla

27. **Al di fuori del contesto lavorativo, Le è capitato di sentirsi discriminato in quanto operatore sanitario?**
    ❑1 molto ❑2 abbastanza ❑3 poco ❑4 per nulla

28. **Al di fuori del contesto lavorativo, Le è capitato che qualcuno evitasse il contatto con lei in quanto operatore sanitario?**
    ❑1 sempre ❑2 spesso ❑3 ogni tanto ❑4 mai

29. **Durante l'emergenza Covid19, ha dovuto modificare le abitudini della sua famiglia?**
    ❑1 SI, molto ❑2 SI, abbastanza ❑3 SI, poco ❑4 NO, per nulla

30. **Dall'inizio dell'emergenza Covid19 ad oggi, ha pensato di cambiare lavoro?**
    ❑1 molto ❑2 abbastanza ❑3 poco ❑4 per nulla

31. **Dall'inizio dell'emergenza Covid19 ad oggi, ha temuto per la sua incolumità?**
    ❑1 molto ❑2 abbastanza ❑3 poco ❑4 per nulla

32. **Durante la pandemia da Covid19, è accaduto che alcuni colleghi siano stati costretti dalle circostanze emergenziali ad agire o prendere decisioni in maniera non coerente ai propri abituali standard etici e professionali; ciò ha causato successivamente in loro vissuti di colpa, vergogna o anche rabbia. Le chiediamo di indicarci se dall'inizio della pandemia ad oggi, è capitato anche a Lei di provare questi vissuti in relazione ad azioni di questo tipo:**

❑1 molto ❑2 abbastanza ❑3 poco ❑4 per nulla

33. **Ha avuto precedenti esperienze professionali nell'assistenza a persone con malattie infettive?**

❑1 SI ❑2 NO

34. **Ha ricevuto una formazione specifica sui Dispositivi di Protezione Individuale (DPI) connessi al rischio infettivo?** ❑1 SI ❑2 NO

35. **Ha mai usufruito di una o più risorse per la gestione dello stress presenti nella sezione intranet "Covid-19. Prendiamoci cura di noi"(es. supporto telefonico, pratiche di mindfulness, opuscolo informativo, ecc.)?** ❑1 SI ❑2 NO

Il questionario è terminato, la ringraziamo per la collaborazione.

Le ricordiamo inoltre che nella sezione intranet "Covid-19. Prendiamoci cura di noi" sono disponibili una serie di strumenti e risorse utili alla gestione dello stress.

1B. Il presente questionario è importante perché ci consente di fornirLe la miglior assistenza possibile. Le Sue risposte ci aiuteranno a capire i problemi che Lei può avere. La preghiamo, perciò, di rispondere con la massima precisione possibile.

Durante le <u>ultime due settimane</u>, per quanti giorni…	**Mai**	**Alcuni giorni**	**Più di metà dei giorni**	**Quasi tutti i giorni**
1. ha provato poco interesse o piacere nel fare le cose ?	0	1	2	3
2. si è sentito/a giù di morale, depresso/a, senza speranze ?	0	1	2	3
3. ha avuto problemi nell'addormentarsi, o nel rimanere addormentato, o ha dormito troppo ?	0	1	2	3
4. ha avuto sensazione di stanchezza o di poca energia ?	0	1	2	3
5. ha avuto poco appetito o mangiato troppo ?	0	1	2	3
6. si è sentito arrabbiato con se stesso, o di essere un fallito, o di avere danneggiato se stesso o la sua famiglia?	0	1	2	3
7. ha avuto difficoltà a concentrarsi su qualcosa, ad esempio leggere il giornale o guardare la TV ?	0	1	2	3
8. ha avuto movimenti o parole talmente lenti da poter essere stati notati dagli altri. O, al contrario, è stato talmente irrequieto e instancabile, da muoversi molto più del solito ?	0	1	2	3
9. ha pensato che sarebbe meglio essere morto, di farsi del male in qualche modo ?	0	1	2	3

10. Le chiediamo di indicarci quanto questi problemi sopra descritti abbiano reso difficile fare il suo lavoro, occuparsi delle sue cose a casa, stare insieme agli altri :

PER NIENTE DIFFICILE	ABBASTANZA DIFFICILE	MOLTO DIFFICILE	ESTREMAMENTE DIFFICILE
□	□	□	□

2B. In questo questionario ci sono 28 domande relative ad una serie di esperienze che possono capitare ad ogni persona nella vita di tutti i giorni. Vorremmo sapere **quanto spesso** anche Lei ha avuto queste esperienze.

Per rispondere segni con una crocetta (X) il punto che meglio corrisponde alla percentuale delle volte che ha avuto l'esperienza descritta nella domanda. Tenga presente che **0** significa **"questa cosa non mi è <u>mai</u> successa"** e **100**, all'estremo opposto, significa **"questa cosa mi succede <u>sempre</u>".** Se l'esperienza descritta nella frase Le capita "**qualche volta, ma non sempre"**, scelga un punto sulla linea **tra 0 e 100** che meglio corrisponde alla percentuale delle volte che ha avuto l'esperienza descritta nella domanda.

1. Alcune persone, mentre sono in macchina o sull'autobus o in metropolitana, improvvisamente si rendono conto di non ricordare cosa sia accaduto durante tutto il viaggio o parte di esso.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

2. Ad alcune persone accade di ascoltare qualcuno parlare, per poi accorgersi di non aver sentito tutta o parte di una conversazione.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

3. Ad alcune persone capita di trovarsi in un posto senza sapere come ci sono arrivati.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

4. Ad alcune persone capita di trovarsi vestiti con abiti che non ricordano di avere indossato.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

5. Ad alcune persone capita di trovare tra le proprie cose nuovi oggetti che non ricordano di aver comprato.

   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

6. Ad alcune persone capita di essere avvicinate da gente che non conoscono e che li chiamano con nomi diversi dal loro o che insistono nel dire di averli già incontrati.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

7. Alcune persone, a volte, hanno la sensazione di stare in piedi vicino a se stessi o guardarsi fare qualcosa e realmente si vedono come se stessero guardando un'altra persona.
   Quanto spesso è capitato anche a Lei?

   0 10 20 30 40 50 60 70 80 90 100

8. Ad alcune persone viene detto che a volte non riconoscono i propri amici o familiari.
   Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

9. Alcune persone scoprono di non aver alcun ricordo di eventi importanti della propria vita.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

10. Ad alcune persone accade di essere accusate di mentire mentre non pensano di averlo fatto.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

11. Ad alcune persone capita di guardarsi allo specchio e non riconoscersi.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

12. Ad alcune persone capita di avere la sensazione che le persone, gli oggetti e il mondo intorno non siano reali.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

13. Ad alcune persone capita di avere la sensazione che il proprio corpo non appartenga loro.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

14. Alcune persone a volte ricordano un avvenimento del passato in modo così vivo, come se lo stessero rivivendo nel presente.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

15. Alcune persone non sono sicure se gli avvenimenti che ricordano siano realmente accaduti oppure se li abbiano solamente sognati.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

16. Alcune persone, nonostante si trovino in luoghi familiari, li trovano strani e non li riconoscono.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

17. Alcune persone, mentre guardano la televisione o un film, sono così assorbiti dalla storia che stanno guardando, da non rendersi conto di ciò che sta accadendo intorno a loro.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

18. Alcune persone sono così coinvolte da una fantasia o da un sogno ad occhi aperti da credere che ciò che sta accadendo sia reale.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

19. Alcune persone si accorgono di essere capaci di non sentire il dolore.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

20. Alcune persone si accorgono di rimanere immobili con lo sguardo perso nel vuoto, senza pensare a niente, e senza essere consapevoli del tempo che passa.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

21. Alcune persone si accorgono di parlare ad alta voce con se stesse quando sono sole.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

22. Alcune persone si accorgono che in alcune situazioni si comportano così diversamente, tanto da sentirsi quasi due persone diverse.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

23. Alcune persone nel fare cose che di solito sono difficili per loro, talvolta si accorgono che sono capaci di farle con facilità e spontaneità sorprendenti (ad esempio sport, lavoro, situazioni sociali).

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

24. Ad Icune persone capita di non ricordare se hanno fatto una cosa oppure se hanno semplicemente pensato
di averla fatta (es. non sanno se hanno spedito una lettera oppure hanno solo pensato di farlo).

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

25. Alcune persone trovano le prove di aver fatto qualcosa che non ricordano di aver fatto.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

26. Ad alcune persone capita di trovare tra le proprie cose degli scritti, dei disegni, o degli appunti che sicuramente hanno fatto ma che non si ricordano assolutamente di aver fatto.

Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

27. Alcune persone a volte si accorgono di sentire delle voci nella propria testa che dicono loro di fare delle cose o che commentano quello che stanno facendo.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

28. Alcune persone, a volte, sentono come se stessero guardando il mondo attraverso una nebbia così che le persone e gli oggetti appaiono lontani o confusi.
Quanto spesso è capitato anche a Lei?

0 10 20 30 40 50 60 70 80 90 100

3B. Nella lista che segue, sono elencati problemi e disturbi che talvolta affliggono le persone. Per favore, la legga attentamente e cerchi di ricordare **se ne ha sofferto nella scorsa settimana, oggi compreso, e con quale intensità**. Risponda per favore alle domande facendo una crocetta nella casella corrispondente all'intensità di ciascun disturbo.

Nell'ultima settimana, in che misura soffre o ha sofferto di…	0	1	2	3	4
1. Mal di testa	Per niente	Un poco	Moderatamente	Molto	Moltissimo
2. Nervosismo o agitazione interna	Per niente	Un poco	Moderatamente	Molto	Moltissimo
3. Incapacità a scacciare pensieri, parole o idee indesiderate	Per niente	Un poco	Moderatamente	Molto	Moltissimo
4. Sensazione di svenimenti o vertigini	Per niente	Un poco	Moderatamente	Molto	Moltissimo
5. Perdita di interesse o del piacere sessuale	Per niente	Un poco	Moderatamente	Molto	Moltissimo
6. Tendenza a criticare gli altri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
7. Convinzione che gli altri possano controllare i tuoi pensieri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
8. Convinzione che gli altri siano responsabili dei tuoi disturbi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
9. Difficoltà a ricordare le cose	Per niente	Un poco	Moderatamente	Molto	Moltissimo
10. Preoccupazioni per la tua negligenza o trascuratezza	Per niente	Un poco	Moderatamente	Molto	Moltissimo
11. Sentirti facilmente infastidito o irritato	Per niente	Un poco	Moderatamente	Molto	Moltissimo

12. Dolori al cuore o al petto	Per niente	Un poco	Moderatamente	Molto	Moltissimo
13. Paura degli spazi aperti o delle strade	Per niente	Un poco	Moderatamente	Molto	Moltissimo
14. Sentirti debole o fiacco	Per niente	Un poco	Moderatamente	Molto	Moltissimo
15. Idee di toglierti la vita	Per niente	Un poco	Moderatamente	Molto	Moltissimo
16. Udire le voci che le altre persone non odono	Per niente	Un poco	Moderatamente	Molto	Moltissimo
17. Tremori	Per niente	Un poco	Moderatamente	Molto	Moltissimo
18. Mancanza di fiducia negli altri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
19. Scarso appetito	Per niente	Un poco	Moderatamente	Molto	Moltissimo
20. Facili crisi di pianto	Per niente	Un poco	Moderatamente	Molto	Moltissimo
21. Sentirsi intimidito nei confronti dell'altro sesso	Per niente	Un poco	Moderatamente	Molto	Moltissimo
22. Sensazione di essere preso in trappola	Per niente	Un poco	Moderatamente	Molto	Moltissimo
23. Paure improvvise senza ragione	Per niente	Un poco	Moderatamente	Molto	Moltissimo
24. Scatti d'ira incontrollabili	Per niente	Un poco	Moderatamente	Molto	Moltissimo
25. Paura di uscire da solo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
26. Avercela sempre con se stessi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
27. Dolori alla schiena	Per niente	Un poco	Moderatamente	Molto	Moltissimo
28. Senso di incapacità a portare a termine le cose	Per niente	Un poco	Moderatamente	Molto	Moltissimo
29. Sentirsi solo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
30. Sentirsi giù di morale	Per niente	Un poco	Moderatamente	Molto	Moltissimo
31. Preoccuparsi eccessivamente per qualsiasi cosa	Per niente	Un poco	Moderatamente	Molto	Moltissimo
32. Mancanza d'interesse	Per niente	Un poco	Moderatamente	Molto	Moltissimo
33. Senso di paura	Per niente	Un poco	Moderatamente	Molto	Moltissimo
34. Sentirsi facilmente ferito o offeso	Per niente	Un poco	Moderatamente	Molto	Moltissimo
35. Convinzione che gli altri percepiscano i suoi pensieri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
36. Sensazione di non trovare comprensione o simpatia	Per niente	Un poco	Moderatamente	Molto	Moltissimo
37. Sensazione che gli altri siano ostili o l'abbiano in antipatia	Per niente	Un poco	Moderatamente	Molto	Moltissimo
38. Dover fare le cose molto lentamente per essere sicuro di	Per niente	Un poco	Moderatamente	Molto	Moltissimo

	farle bene					
39.	Palpitazioni o sentirsi il cuore in gola	Per niente	Un poco	Moderatamente	Molto	Moltissimo
40.	Senso di nausea o mal di stomaco	Per niente	Un poco	Moderatamente	Molto	Moltissimo
41.	Sentimenti di inferiorità	Per niente	Un poco	Moderatamente	Molto	Moltissimo
42.	Dolori muscolari	Per niente	Un poco	Moderatamente	Molto	Moltissimo
43.	Sensazione che gli altri la guardino o parlino di lei	Per niente	Un poco	Moderatamente	Molto	Moltissimo
44.	Difficoltà ad addormentarsi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
45.	Bisogno di controllare ripetutamente ciò che fa	Per niente	Un poco	Moderatamente	Molto	Moltissimo
46.	Difficoltà a prendere decisioni	Per niente	Un poco	Moderatamente	Molto	Moltissimo
47.	Paura di viaggiare in autobus, nella metropolitana o in treno	Per niente	Un poco	Moderatamente	Molto	Moltissimo
48.	Sentirsi senza fiato	Per niente	Un poco	Moderatamente	Molto	Moltissimo
49.	Vampate di calore o brividi di freddo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
50.	Necessità di evitare certi oggetti, luoghi o attività perché spaventano	Per niente	Un poco	Moderatamente	Molto	Moltissimo
51.	Senso di vuoto mentale	Per niente	Un poco	Moderatamente	Molto	Moltissimo
52.	Intorpidimento o formicolio di alcune parti del corpo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
53.	Nodo alla gola	Per niente	Un poco	Moderatamente	Molto	Moltissimo
54.	Guardare al futuro senza speranza	Per niente	Un poco	Moderatamente	Molto	Moltissimo
55.	Difficoltà a concentrarsi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
56.	Senso di debolezza in qualche parte del corpo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
57.	Sentirsi teso o sulle spine	Per niente	Un poco	Moderatamente	Molto	Moltissimo
58.	Senso di pesantezza alle braccia o alle gambe	Per niente	Un poco	Moderatamente	Molto	Moltissimo
59.	Idee di morte	Per niente	Un poco	Moderatamente	Molto	Moltissimo
60.	Mangiare troppo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
61.	Senso di disagio quando la gente la guarda o parla di lei	Per niente	Un poco	Moderatamente	Molto	Moltissimo
62.	Avere dei pensieri che non sono suoi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
63.	Sentire l'impulso di colpire, ferire o fare del male a	Per niente	Un poco	Moderatamente	Molto	Moltissimo

qualcuno					
64. Svegliarsi presto al mattino senza riuscire a riaddormentarsi	Per niente	Un poco	Moderatamente	Molto	Moltissimo
65. Avere bisogno di ripetere lo stesso atto come toccare, contare, lavarsi le mani, ecc.	Per niente	Un poco	Moderatamente	Molto	Moltissimo
66. Sonno inquieto o disturbato	Per niente	Un poco	Moderatamente	Molto	Moltissimo
67. Sentire l'impulso di rompere o spaccare gli oggetti	Per niente	Un poco	Moderatamente	Molto	Moltissimo
68. Avere idee o convinzioni che gli altri non condividono	Per niente	Un poco	Moderatamente	Molto	Moltissimo
69. Sentirsi penosamente imbarazzato in presenza di altri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
70. Sentirsi a disagio tra la folla come nei negozi, al cinema, ecc.	Per niente	Un poco	Moderatamente	Molto	Moltissimo
71. Sensazione che tutto richieda uno sforzo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
72. Momenti di terrore e di panico	Per niente	Un poco	Moderatamente	Molto	Moltissimo
73. Sentirsi a disagio quando mangia o beve in presenza di altri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
74. Ingaggiare frequenti discussioni	Per niente	Un poco	Moderatamente	Molto	Moltissimo
75. Sentirsi a disagio quando è solo	Per niente	Un poco	Moderatamente	Molto	Moltissimo
76. Convinzione che gli altri non apprezzino nella giusta misura il suo lavoro	Per niente	Un poco	Moderatamente	Molto	Moltissimo
77. Sentirsi solo e triste anche in compagnia	Per niente	Un poco	Moderatamente	Molto	Moltissimo
78. Senso di irrequietezza tanto da non potere stare seduto	Per niente	Un poco	Moderatamente	Molto	Moltissimo
79. Sentimenti di inutilità	Per niente	Un poco	Moderatamente	Molto	Moltissimo
80. Il presentimento che debba accaderle qualcosa di molto spiacevole	Per niente	Un poco	Moderatamente	Molto	Moltissimo
81. Urlare o scagliare oggetti	Per niente	Un poco	Moderatamente	Molto	Moltissimo
82. Avere paura di svenire davanti agli altri	Per niente	Un poco	Moderatamente	Molto	Moltissimo
83. Impressione che gli altri possano approfittare di lei, se lo permette a loro	Per niente	Un poco	Moderatamente	Molto	Moltissimo
84. Pensieri sul sesso	Per niente	Un poco	Moderatamente	Molto	Moltissimo
85. Idea di dover scontare i propri peccati	Per niente	Un poco	Moderatamente	Molto	Moltissimo

86. Pensieri ed immagini di natura spaventosa	Per niente	Un poco	Moderatamente	Molto	Moltissimo
87. Pensiero di avere una grave malattia fisica	Per niente	Un poco	Moderatamente	Molto	Moltissimo
88. Non sentirsi mai vicino alle altre persone	Per niente	Un poco	Moderatamente	Molto	Moltissimo
89. Sentirsi in colpa	Per niente	Un poco	Moderatamente	Molto	Moltissimo
90. Idea che qualche cosa non vada bene nella sua mente	Per niente	Un poco	Moderatamente	Molto	Moltissimo

Il questionario è terminato, la ringraziamo per la collaborazione.

Le ricordiamo inoltre che nella sezione intranet "Covid-19. Prendiamoci cura di noi" sono disponibili una serie di strumenti e risorse utili alla gestione dello stress.