



# Structural validity and classification performance of the Italian Short Negative Acts Questionnaire: A Structural Equation Modeling approach for building ROC curves

Enrico Perinelli<sup>1</sup> · Cristian Balducci<sup>2</sup> · Franco Fraccaroli<sup>1</sup>

Accepted: 2 September 2022 / Published online: 17 September 2022  
© The Author(s) 2022

## Abstract

We investigated the structural (internal) validity and classification performance of the Italian *Short Negative Acts Questionnaire* (SNAQ), a 9-item self-report instrument assessing bullying at work. Consistent with recent attention of researchers to control measurement error in predictive models (Jacobucci & Grimm, *Perspectives on Psychological Science*, 15(3), 809–816 2020), classification performance was investigated through a proposed novel procedure that uses Structural Equation Modeling for building ROC curves. Participants included 357 workers (females = 50.4%) from various sectors. Our results showed that (a) the Italian SNAQ demonstrates adequate levels of structural validity; (b) its classification performance (in terms of self-labeled bullying) is outstanding; and (c) the ROC curves estimated by means of Structural Equation Modeling outperform those estimated with classical observed-variable approaches. In conclusion, we provided further evidence regarding the good psychometric properties of the Italian SNAQ and we also offered a novel approach for estimating ROC curves that does not neglect the issue of measurement quality.

**Keywords** Short Negative Acts Questionnaire · Workplace bullying · Self-labeled bullying · Latent variables · Structural equation modeling · ROC curves

## Introduction

Bullying at work denotes a class of “situations where an employee repeatedly and over a prolonged time period is exposed to harassing behavior from one or more colleagues

(including subordinates and leaders) and where the targeted person is unable to defend him-/herself against this systematic mistreatment” (Nielsen & Einarsen, 2018, p. 73). Scholars are largely in agreement about two characteristics of bullying at work, namely it “is *repeated* and *systematic* negative social behavior [and it] endures *over a longer period of time*” (Notelaers & Van der Heijden, 2021, p. 370). Thus, workplace bullying<sup>1</sup> is a widely studied phenomenon in organizations (Nielsen & Einarsen, 2018) for two main reasons: First, it has been widely attested that workplace bullying has severe negative consequences for both work organization and the target’s mental health (Balducci et al., 2020); second, it is not a rare phenomenon, since a meta-analysis of prevalence rates has shown that approximately 15% of workers (on a global basis) are exposed to some level of bullying at work (Hershcovis et al., 2015; Nielsen et al., 2010).

---

An earlier version of this paper was presented (a) at the 37th Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), Seattle, WA, USA (April 28, 2022), and (b) at the 15th Conference of the European Academy of Occupational Health Psychology (EAOHP), Bordeaux, France (July 7, 2022), thanks to a Starting Grant for young researchers awarded by the University of Trento to Enrico Perinelli (Title: Big data and machine learning approaches in the workplace: Advancing the quality of industrial/organizational research).

---

✉ Enrico Perinelli  
enrico.perinelli@unitn.it

<sup>1</sup> Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31, 38068 Rovereto, TN, Italy

<sup>2</sup> Department of Psychology, University of Bologna, Bologna, Italy

<sup>1</sup> According to Nielsen and Einarsen (2018, p. 73), workplace bullying has been conceptualized through different labels, such as abusive supervision, incivility, bullying/mobbing, harassment, victimization, interpersonal deviance, emotional abuse, ostracism, and social undermining. Here we use the term “workplace bullying” because it is most often used by the developers of the SNAQ (Notelaers et al., 2019).

That said, an easy-to-use self-report instrument for assessing workplace bullying is of pivotal importance for researchers and practitioners in order to (a) study this phenomenon in large national surveys, (b) enhance the investigation of *personal x environmental* antecedents, *health and organizational* consequences, and *moderators/mediators* affecting those relationships (Balducci et al., 2020, 2021; Howard et al., 2020; Reknes et al., 2019; Van den Brande et al., 2016), and (c) help organizations and their managers through initial assessments and (eventually) subsequent monitoring of the course of interventions.

Across the years, many self-report instruments have been developed for assessing various typologies of work-related abusive behaviors (see Table A1 in Appendix, for an overview). Among them, the 22-item Negative Acts Questionnaire – Revised (NAQ-R; Einarsen et al., 2009) proved to be a valid and reliable instrument for assessing workplace bullying, and its recent 9-item short version (SNAQ; Notelaers & Einarsen, 2008; Notelaers et al., 2019) has the advantage of being easily utilized by researchers and practitioners. In the Italian context, a first validation of the scale was provided by Balducci et al. (2010); here we aimed at expanding their work by adding an investigation of its structural validity and classification performance. Moreover, given that the SNAQ scores may be affected by a certain amount of measurement error (as the majority of self-report instruments), consistent with the recent attention of researchers to control for measurement errors in predictive models (e.g., in machine learning analytic strategies; Jacobucci & Grimm, 2020), classification performance was investigated through a novel procedure that uses Structural Equation Modeling for building ROC curves (the gold standard instrument for evaluating the classification performance of a continuous variable; Fawcett, 2006; Kuhn & Johnson, 2020).

In what follows, we outlined the rationale and procedures adopted in this contribution to investigate the SNAQ's structural validity and classification performance.

## Investigating the SNAQ's structural validity

Structural validity regards the internal characteristics of a measurement instrument (Bleidorn & Hopwood, 2019; Loevinger, 1957). We investigated the structural validity of the Italian SNAQ by means of a SEM approach for categorical data, given that each item is rated by participants by means of a 5-point frequency scale. More in detail, we investigated measurement invariance and reliability through a series of categorical confirmatory factor analyses (CFAs). Indeed, measurement invariance is among the most important latent variable approaches to prove that measurement properties of latent variables are stable and thus, to ensure that the meaning of the construct

being assessed is consistent across groups or time (Little, 2013; Millsap, 2012; Newsom, 2015; van de Schoot et al., 2012, 2015). We chose to investigate gender invariance because it is one of the most studied types of invariance in organizational measures (Vandenberg & Lance, 2000), in that it allows verifying the “generalizability of scale properties across gender” (Vandenberg & Lance, 2000, p. 22), which is a fundamental component of the internal robustness of the test (see Ock et al., 2020). In conclusion, given that gender differences in workplace bullying have been widely studied (Rosander et al., 2020) and we are interested in estimating classification performance across gender, the Italian SNAQ must show an acceptable degree of invariance before being used for these purposes.

## Investigating the SNAQ's classification performance

### Self-labeling bullying as a classifier

Classification performance is the ability of a continuous variable or instrument to correctly predict a qualitative (usually dichotomous) variable (James et al., 2021). In this case, the aim is to attest how well the SNAQ can classify people who felt they were bullied vs. those who did not feel they were bullied at their workplace. In doing so, we adopted a *self-labeling approach* (Notelaers & Van der Heijden, 2021), which consists of providing a thorough definition of workplace bullying along with a list of characteristics, and then ask workers if they perceive to have been bullied or not. This approach has been consistently used in workplace bullying research since the seminal introduction by Einarsen and Skogstad (1996), but with some refinements. For example, a meta-analysis by Nielsen et al. (2010) showed the importance of including a definition of workplace bullying when using a self-labeling approach, given that the prevalence of reported workplace bullying may be overestimated if no definition is provided. Indeed, they found a workplace bullying rate of 11.3% for studies using a “self-labeling with definition” approach, 14.8% for studies using a “behavioral experiences” approach (e.g., self-report instruments like the SNAQ), and a rate of 18.1% for studies using a “self-labeling without definition” approach. Hence, they concluded that the best method to investigate workplace bullying is the use of both “self-labeling with definition” and “behavioral experiences” approach. This conclusion is echoed in more recent contributions (Nielsen et al., 2020; Notelaers & Van der Heijden, 2021), in which the authors pointed out the importance of estimating “whether respondents were exposed to systematic harassment at their workplace with the behavioural experience method [as well as to] ask them whether they perceive themselves as victimized (by

this exposure) with the self-labelling method” (Nielsen et al., 2020, p. 256).

From an empirical perspective, there are a number of studies that have used the self-labeling approach (see Notelaers & Van der Heijden, 2021, for a review). For example, Bonde et al. (2016), in a longitudinal study on 7502 public service and private sector Danish employees, showed that self-labeled bullying is negatively associated with a number of health-related variables (e.g., self-rated health, sleep quality and mood symptoms) and that self-labeled bullying persisted even after 4 years for 20–40% of the sample. Again, Rosander and Blomberg (2019) found a good degree of overlap between self-labeled victimization and the NAQ-R cut-off score provided by Notelaers and Einarsen (2013), but with some discrepancies.

That said, the above findings attest that self-labeling bullying may be a valid classifier for SNAQ validation purposes, given that it showed good external validity (Blonde et al., 2016), good stability across time (Blonde et al., 2016), and it overlaps sufficiently with behavioral experience methods (Rosander & Blomberg, 2019).

### Building ROC curves with a SEM approach

The classification performance of the Italian SNAQ was investigated through a two-step approach. In the first step, we used the previous final (i.e., best fitting) CFA model for predicting a dichotomous variable of *self-labeled bullying*, which was used as a classifier according to the aforementioned reasons. This SEM was used for gathering the predicted values (in terms of factor scores) of the effect of SNAQ (predictor) on *self-labeled bullying* (outcome). In the second step, predicted values (factor scores) were extracted and then used for analyzing the classification performance of the SNAQ through a widely used machine learning tool, namely the ROC curve. As Cali and Longobardi (2015, p. 395) put it: “a ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. By considering all possible values of the cut-off  $c$ , the ROC curve can be constructed as a plot of sensitivity (TPR) versus  $1 - \text{specificity}$  (FPR)”, where TPR stands for True Positive Rate, and FPR stands for False Positive Rate. Hence, following Cali and Longobardi’s (2015) notation, a ROC curve is defined as

$$\text{ROC}(\cdot) = \{\text{FPR}(c), \text{TPR}(c), c \in (-\infty, +\infty)\},$$

that may also be written as

$$\text{ROC}(\cdot) = \{(t, \text{ROC}(t)), t \in (0, 1)\},$$

where the ROC function maps  $t$  to  $\text{TPR}(c)$ , and  $c$  is the cut-off corresponding to  $\text{FPR}(c) = t$  (Cali & Longobardi, 2015,

p. 395). Then, the most informative index that establishes the classification performance of a continuous variable is the Area Under Curve (AUC) of the  $\text{ROC}(t)$  function, that ranges from 0 to 1 and can be interpreted as a summary (in terms of percentage) of “the discriminatory accuracy of a test” (Gonçalves et al., 2014, p. 5; see also Fawcett, 2006). More formally, AUC is defined by

$$\text{AUC} = \int_0^1 \text{ROC}(t) dt$$

However, prediction values for performing a ROC curve are usually gathered from analyses using “observed” predictors, such as logistic regression and random forest (Lantz, 2019; Zúmel & Mount, 2020), that do not control for their degree of measurement error (Jacobucci & Grimm, 2020). In fact, Jacobucci and Grimm (2020) contended that the consequences of measurement error are widely known in traditional psychological science (e.g., Bollen, 1989; Cole & Preacher, 2014; McNeish & Wolf, 2020), while in machine learning or predictive models those consequences have not been sufficiently addressed (see also Brandmaier & Jacobucci, in press). In organizational psychology, it is rare to deal with perfectly reliable tools, and the SNAQ is surely an instrument with a certain degree of measurement error (Notelaers & Einarsen, 2008; Notelaers et al., 2019). Hence, in this study we show a method for attesting the classification performance of an instrument through a ROC curve estimated by a model (i.e., a SEM) that takes into account the measurement error of the predictor variable (i.e., SNAQ factor scores) as well as its structural validity (i.e., gender invariance). This strategy allows for an AUC value that is free from the effect of measurement error, consistent with Jacobucci and Grimm’s (2020).

### The present study

To summarize, the present study has three aims. First, to attest whether the Italian SNAQ is a reliable, gender-invariant, and valid instrument in classifying those who self-labeled themselves as bullied and those who did not. Second, to provide a data analytic strategy for building ROC curves that takes into account the measurement quality (i.e., degree of measurement error and gender invariance parameters) of the predictor variable, thus maximizing predictive values, that in turn allow to better estimate AUC. Third, to compare the results obtained with our approach (i.e., ROC curves estimated with best fitting SEM obtained from structural validity routine) with those obtained with more classical approaches for estimating ROC curve (that use observed variables), namely logistic regression and random forest.

## Method

### Procedure

In the first half of 2018, the *Provincial Antimobbing Committee*<sup>2</sup> of an Italian province asked the local University to develop a survey project on the state of workplace bullying in the province. In order to do this, during 2019, the first and third author (University of Trento) collaborated with the local branch of the national statistical center, which is a public institution officially appointed to collect data, in full compliance with national regulations to protect the privacy and anonymity of respondents (indeed, the procedures adopted have been validated by a public guarantee body, namely their research ethics committee). In more detail, a two-phase sampling strategy was carried out, selecting 23 municipalities that were chosen starting from a subdivision into three groups on the basis of the distribution of employees: (a) so-called “representative municipalities” – or the 5 most populous municipalities – with a number of employees exceeding 4500; (b) municipalities with a number of employees between 1000 and 4500; (c) municipalities with a number of employees less than 1000. Subsequently, the statistical offices of the sampled municipalities were asked to extract 4 random samples (one “base” sample and three “back-up” samples) consisting of 1060 families each. Each family consisted of at least one worker that received a mail, delivered at home, which included (a) a brief description of the project; (b) an invitation to a link on a website to complete an anonymous self-report battery; (c) specific requirements to fill in the battery: Be at least 18 years old; be currently employed (not being on leave, on layoffs, or retired); be an employee or assimilated worker (not a freelancer, not a self-employed, not an entrepreneur); and (d) a username and password (distinctive for each household) that would allow starting the battery and ensuring that only people from selected households could complete the battery (the survey started in July 2019). After the survey ended (November 2019), the local branch of the national statistical center removed the username and password from the database and returned the anonymized dataset to the researchers.

### Participants

Participants included 357 workers (i.e., ~8% of the total 4240 invitations and ~33% of the 1060 target sample size) from various sectors (see Table A2 in Appendix) that took

part in the above-mentioned provincial survey. Males were 177 (49.6%) and females 180 (50.4%). In regards to age, 66 were aged 18–34 (18.5%), 200 were aged 35–54 (56%), and 91 were aged 55-more (25.5%). Tenure ranged from 1 to 42 years ( $M = 14.90$ ,  $SD = 11.22$ ). Regarding education, 29 (8.1%) had no certificate or had an elementary/junior-high school degree, 192 (53.8%) had a vocational/high school degree, 136 (38.1%) had a bachelor or higher degree. Concerning the type of contract, 52 (14.6%) had a fixed-term contract and 305 (85.4%) had a permanent contract. Regarding sector, 230 (64.4%) worked in the private sector, while 127 (35.6%) worked in the public sector. As for the job position, 72 (20.2%) were laborers, 213 (59.6%) were clerks, 55 (15.4) were *quadri* (a specific Italian job classification, ranked between clerks and managers), and 17 (4.8%) were managers.

### Measures

**Short-Negative Acts Questionnaire (SNAQ)** We used the Italian version of the SNAQ (Notelaers & Einarsen, 2008; Notelaers et al., 2019) validated in Italy by Balducci et al. (2010). The SNAQ measures the extent to which a worker feels they have been exposed to workplace bullying behaviors. Items describe common negative acts at work, such as work-related bullying (item example, “*Someone withholding information which affects your performance*”), personal bullying (item example “*Having insulting or offensive remarks made about your person, attitudes or your private life*”), and social isolation/exclusion (item example “*Being ignored or excluded*”). The SNAQ consists of 9 items anchored with a 5-point frequency scale that expresses frequency of exposure (anchors: 1 = *never*, 2 = *once or occasionally*, 3 = *monthly*, 4 = *weekly*, 5 = *daily*). The introduction to the scale reads “*Please, indicate how often you have experienced each of the following behaviors in your workplace, over the past six months*”.

**Self-Labeled Bullying** In order to specifically measure whether each participant perceived to be bullied at their workplace or not, we followed an approach used by the Workplace Bullying Institute (2017). First, we provided a specific definition and description of workplace bullying, reported as follows:

Workplace bullying is defined as a set of aggressive behaviors, carried out by one or more people (called “Bully”), which are protracted over time and directed towards a worker (victim), whose purpose is to affect one or more of the following aspects

- the victim’s reputation
- his/her ability to communicate

<sup>2</sup> This institution is provided for by provincial regulations. Note that in the Italian context, workplace bullying is usually indicated with the word “mobbing”. *Antimobbing* can be translated as “against mobbing”.

- his/her social relations
- his/her employment quality (e.g., no or only insignificant tasks are assigned)
- his/her health and well-being

Then, according to the provided definition of workplace bullying, participants were asked to respond *Yes* (coded as 1) or *No* (coded as 0) to the following question: “*I have experienced it now or have experienced it in the last year*”.

**Measures used for external validity** In order to preliminarily attest the external validity of the SNAQ, we used six measures of constructs belonging to three areas that proved to be significantly correlated with workplace bullying: *Organizational climate*, namely organizational social climate (Balducci et al., 2010) and unethical climate (Bulutlar & Öz, 2009); *job attitudes*, namely burnout (Conway et al., 2021b), job satisfaction (Rodríguez-Muñoz et al., 2009), and intention to quit (Bentley et al., 2021); and *personality*, namely self-efficacy beliefs (Fida et al., 2018).

*Organizational social climate* ( $\alpha = .77$ ) was measured by means of 4 items gathered from Vartia (1996) and Balducci et al. (2010, p. 152). The Likert scale ranged from 1 (*disagree*) to 5 (*agree*). Three items were reversed in order to measure positive organizational climate (e.g., “*There is envy at my workplace*”, “*Workmates compete with each other*”; a straight item example is, “*In my workplace, there is a good level of cooperation and agreement among workers*”).

*Unethical climate* ( $\alpha = .88$ ) was measured by means of 4 items gathered from the *self-interest scale* subscale of the *ethical climate questionnaire* by Cullen et al. (1993; Italian version: Pagliaro et al., 2018). The Likert scale ranged from 1 (*disagree*) to 5 (*agree*). An item example includes, “*There is no room for one’s own personal morals or ethics in my workplace*”.

*Burnout* ( $\alpha = .83$ ) was measured by means of 4 items gathered from the *Copenhagen Burnout Inventory* by Kristensen et al. (2005; Italian validation: Avanzi et al., 2013). The Likert scale ranged from 1 (*never*) to 5 (*always*). An item example is, “*I feel worn out at the end of the work day*”.

*Job satisfaction* was measured with one item gathered from the *Satisfaction with Job – General* scale by Dubinsky and Hartley (1986; Barbaranelli et al., 2010), that reads, “*Generally speaking, I am very satisfied of my job*”, and has a Likert scale ranging from 1 (*disagree*) to 5 (*agree*).

*Intention to quit* ( $\alpha = .90$ ) was measured by means of 3 items gathered from Landau and Hammer (1986, p. 404). The Likert scale ranged from 1 (*disagree*) to 5 (*agree*). An item example includes, “*I am seriously thinking about quitting my job*”.

*Self-efficacy beliefs in managing negative emotions at work* ( $\alpha = .81$ ) were measured by means of the six-item

scale used by Alessandri et al. (2018). This scale assesses the degree to which a worker perceives how well he/she is able to manage negative emotions due to adverse or stressful work-related events. The Likert scale ranged from 1 (*not well at all*) to 5 (*very well*). The introductory question was: “*At work, how well can you?*” and an item example is, “*Get over irritation quickly after the experience of a failure?*”.

## Data analyses

All analyses were conducted using the statistical open source software R (Version 4.0.2; R Core Team, 2020) and proceeded through the following steps.

**Descriptive statistics and external validity** After examining the correlations between SNAQ items – which we expected to be of at least medium size ( $\geq .30$ ) given that the tool has been used as a unidimensional measure – we investigated the external validity of the SNAQ by inspecting a series of zero-order correlations between the SNAQ and six external constructs mentioned in the “Measures used for External Validity” paragraph. First, we built a composite score for each of the seven constructs (computing the mean of all items) and we expected that the SNAQ correlates (a) significantly and positively with unethical climate, burnout, intention to quit; and (b) significantly and negatively with organizational social climate, job satisfaction, self-efficacy beliefs in managing negative emotions at work. Data wrangling and descriptive statistics were conducted with the R package `dplyr` (which is part of the `tidyverse` ecosystem; Wickham et al., 2019; Wickham & Grolemund, 2017) and `apaTables` (Stanley, 2018).

**Structural validity** Structural validity was investigated through inspection of reliability indices and gender invariance. Both analyses were conducted by taking into account the categorical nature of SNAQ items and through the R packages `lavaan` (Rosseel, 2012) and `semTools` (Jorgensen et al., 2020).

First, structural validity was investigated through the analyses of two reliability coefficients, namely the nonlinear SEM reliability coefficient ( $\rho_{NL}$ ; Yang & Green, 2015) and the Average Variance Extracted (AVE; Fornell & Larcker, 1981). These indices were computed for the overall scale and for males/females separately. Both indices were computed by means of the `reliability()` function of the `semTools` package, so as to take into account the ordinal nature of the scale (Flora, 2020). Indeed, these reliability indices were computed on the basis of a confirmatory factor model (composed by one latent variable and the nine SNAQ items) estimated in `lavaan` using the WLSMV estimator, which is the most used estimator for structural equation models with

categorical indicators (Finney & DiStefano, 2013). Good reliability indices were  $> .70$  for  $\rho_{NL}$  and  $> .50$  for AVE.

Second, structural validity was investigated by probing the tenability of measurement invariance across gender. In particular, we conducted measurement invariance for categorical variables following Wu and Estabrook's (2016) approach, through the routine provided by Svetina et al. (2020). Thus, gender invariance is attested if two conditions are met: (a) constraining thresholds<sup>3</sup> to be equal across gender (*Equal Thresholds Model*) does not worsen the fit of the *Baseline Model* and, (b) constraining loadings to be equal across gender (*Equal Thresholds and Loadings Model*) does not worsen the fit of the *Equal Thresholds Model*. The non-significant worsening of a model is attested if (a) the scaled chi-square difference test is not significant, (b) the CFI-scaled does not decrease more than 0.01, and (c) the RMSEA-scaled does not increase more than 0.01 (see Svetina et al., 2020). In order to ensure the robustness of findings (see Svetina et al., 2020) – and to provide scripts with a more familiar software for latent variable users – we also estimated measurement invariance models with *Mplus* Version 8.4 (Muthén & Muthén, 1998–2017).

**Classification performance** Classification performance of the SNAQ was investigated by performing a *receiver operating characteristic* (ROC) curve using the dichotomous variable *self-labeled bullying* as the outcome and the SNAQ as the independent variable. In order to take into account the measurement error of the SNAQ (Jacobucci & Grimm, 2020), the ROC curve was estimated using predicted probabilities (in regards of *being* vs. *not being* bullied) gathered from a structural equation model (SEM). In particular, we estimated two nested SEMs: In the first model (*Conditional and Unconstrained model*), we specified the same measurement structure as an *Equal Thresholds and Loadings Model* and added a path from the latent variable SNAQ to the dichotomous variable *self-labeled bullying*; in the second model (*Conditional and Constrained model*), we constrained that path to be equal across gender and tested whether the model significantly worsened. Then, we used the best fitting

model to extract factor scores to be used in estimating the area under the ROC curve. Moreover, in order to compare models' performance and consistency (see Lantz, 2019), ROC curves were estimated using predicted probabilities gathered from two of the most popular predictive models, namely, a logistic regression model and a random forest model.

We evaluated the validity of the SNAQ as a classifier of workplace bullying by inspecting the rate of the area under the ROC curve (for the above three models and for males and females, separately), according to the following criteria outlined by Lantz (2019, p. 333): Outstanding classifier = 0.9 to 1.0; excellent/good classifier = 0.8 to 0.9; acceptable/fair classifier = 0.7 to 0.8; poor classifier = 0.6 to 0.7; no discrimination = 0.5 to 0.6; classifier with no predictive value  $< 0.50$ .

In order to improve the visualization of results, we also provide double density plots for each model (Zumel & Mount, 2020). The above new analyses were conducted with the R packages *pROC* (for ROC curve; Robin et al., 2011), *randomForest* (for random forest model; Liaw & Wiener, 2002), and *WVPlots* (for double density plots; Mount & Zumel, 2020).

**Comparison of ROC curves** In order to verify whether our latent variable approach for the estimation of ROC curves outperforms traditional observed-variable approaches, we compared the ROC curves obtained with the SEM approach with those obtained by logistic regression and random forest models by means of the `roc.test()` function of the *pROC* package (Robin et al., 2011), using bootstrap as the specified method (`resampling = 2000`). This method estimates a statistics  $D$  where  $D = \frac{AUC1 - AUC2}{s}$ ,  $s$  is the standard deviation of the bootstrap differences, and  $AUC1$  and  $AUC2$  the  $AUC$  of the two (original) ROC curves. The  $D$  statistic is  $z$ -distributed and if the resulting  $p$ -value is significant (we fixed our alpha-level to .05, thus if  $p < .05$ ), then there is evidence for the significant difference between  $AUC1$  and  $AUC2$ .

## Results

### Descriptive statistics and external validity

In Table 1 we reported descriptive statistics and correlation matrices<sup>4</sup> (for the whole sample and separated for gender, given that we subsequently ran a multiple-group analysis) for the SNAQ items. As can be seen, all correlations were

<sup>3</sup> Kline (2016) provided a brief but comprehensive description of threshold: In the “continuous/categorical variable methodology (Muthén, 1984) [...], each ordinal indicator is associated with a latent response variable, which is the underlying amount of a continuous and normally distributed continuum that is required to respond in a certain way on the corresponding indicator. For dichotomous items, this amount or threshold is the point on the latent variable where one response option is given (e.g., true), if the threshold is exceeded. It is also the point where the other response option is given (e.g., false), if the threshold is not exceeded. Dichotomous items have a single threshold, but the number of thresholds for polytomous items with three or more response categories equals the number of categories minus one” (Kline, 2016, p. 324).

<sup>4</sup> For easiness of interpretation we have reported Pearson correlation coefficients, but given the ordinal nature of the variables we also computed Kendall's rank (tau-b) correlations and Polychoric correlations, which are available upon request to the first author.

**Table 1** Descriptive statistics and zero-order correlations for SNAQ items

Sample (Size)	Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9
Whole ( <i>N</i> = 357)	1. snaq1	1.92	1.13	1								
	2. snaq2	1.64	0.93	.49**	1							
	3. snaq3	1.69	1.04	.60**	.49**	1						
	4. snaq4	1.38	0.77	.47**	.67**	.56**	1					
	5. snaq5	1.58	0.82	.44**	.41**	.44**	.52**	1				
	6. snaq6	2.10	0.89	.29**	.35**	.28**	.41**	.37**	1			
	7. snaq7	1.72	0.91	.61**	.51**	.58**	.47**	.47**	.34**	1		
	8. snaq8	1.57	0.82	.48**	.59**	.46**	.55**	.48**	.48**	.53**	1	
	9. snaq9	1.14	0.52	.39**	.43**	.44**	.56**	.38**	.32**	.32**	.47**	1
Male ( <i>n</i> = 177)	1. snaq1	1.94	1.17	1								
	2. snaq2	1.72	1.05	.54**	1							
	3. snaq3	1.67	1.12	.64**	.54**	1						
	4. snaq4	1.45	0.88	.53**	.73**	.57**	1					
	5. snaq5	1.58	0.86	.43**	.52**	.40**	.54**	1				
	6. snaq6	2.2	1.06	.30**	.32**	.30**	.40**	.39**	1			
	7. snaq7	1.73	0.96	.64**	.56**	.49**	.50**	.44**	.30**	1		
	8. snaq8	1.59	0.91	.50**	.61**	.47**	.63**	.53**	.45**	.51**	1	
	9. snaq9	1.16	0.59	.43**	.44**	.52**	.56**	.43**	.38**	.42**	.62**	1
Female ( <i>n</i> = 180)	1. snaq1	1.90	1.09	1								
	2. snaq2	1.56	0.79	.42**	1							
	3. snaq3	1.71	0.97	.56**	.43**	1						
	4. snaq4	1.30	0.63	.39**	.55**	.56**	1					
	5. snaq5	1.58	0.79	.46**	.26**	.48**	.51**	1				
	6. snaq6	1.99	0.67	.27**	.37**	.26**	.40**	.35**	1			
	7. snaq7	1.71	0.86	.57**	.45**	.69**	.44**	.51**	.43**	1		
	8. snaq8	1.54	0.73	.46**	.57**	.44**	.43**	.43**	.54**	.55**	1	
	9. snaq9	1.11	0.43	.34**	.41**	.31**	.55**	.33**	.20**	.18*	.23**	1

\*\*  $p < .01$ . \*  $p < .05$

significant at  $p < .01$  and ranged from .28 (item 6 with item 3) to .67 (item 4 with item 2) for the whole sample; from .30 (item 6 with item 1, item 6 with item 3, item 7 with item 6) to .73 (item 4 with item 2) for the male sample; and from .18 (item 9 with item 7) to .69 (item 7 with item 3) for the female sample. Overall, the size of the correlations ranged from medium to high (as expected for a one-dimension variable), with only 7 zero-order correlations below |.30| (which is a medium effect size, according to Cohen, 1992), 2 for the whole sample and 5 for the female sample. Interestingly, 5 (out of 7) regard Item 6 (“Repeated reminders of your errors or mistakes”). This finding might attest that the content of this item is related less to the others, but at the same time there are no reasons to eliminate it, given that correlations are in any case positive and significant. In Table 2, we reported the external validity analysis along with descriptive statistics for each composite score. All zero-order correlations were in the expected direction and significant, with the exception of the correlation with self-efficacy beliefs in managing negative emotions at work in the female sample.

This finding might suggest that organizational climate and job attitudes are more related to workplace bullying than a personal-oriented variable like self-efficacy (Balducci et al., 2021), while such a low effect size as far as self-efficacy was concerned was not expected (Fida et al., 2018). Nonetheless, overall, the external validity of the scale was supported.

### Structural validity

Before running structural validity analyses, we replaced the value of ‘5’ with ‘4’ in all items, given that we noticed an imbalance of cells across gender (i.e., two zero-count cells in the female sample: cell ‘5’ for item 8, cell ‘4’ for item 9). This “collapsing” procedure is suggested when a dataset includes few (or zero) responses in extreme categories (DiStefano et al., 2021). Hence, subsequent models have 3 thresholds instead of 4.

The first step of structural validity was the examination of the above-mentioned reliability indices. First, we ran two SEMs for the measurement model of the scale: A model

**Table 2** Zero-order correlations for SNAQ composite score (external validity analyses) and descriptive statistics for each composite score

Construct	Whole sample ( <i>N</i> =357)			Male sample ( <i>n</i> =177)			Female sample ( <i>n</i> =180)		
	<i>r</i>	<i>M</i>	<i>SD</i>	<i>r</i>	<i>M</i>	<i>SD</i>	<i>r</i>	<i>M</i>	<i>SD</i>
SNAQ	-	1.64	0.63	-	1.67	0.71	-	1.60	0.55
Organizational social climate	-.54***	3.16	0.99	-.56***	3.12	1.01	-.52***	3.19	0.98
Unethical climate	.50***	2.72	1.25	.54***	2.80	1.27	.47***	2.65	1.24
Burnout	.48***	2.49	0.79	.52***	2.45	0.80	.46***	2.53	0.79
Job satisfaction	-.40***	3.84	1.20	-.42***	3.93	1.15	-.41***	3.75	1.25
Intention to quit	.38***	1.99	1.31	.39***	1.95	1.25	.38***	2.04	1.37
em_eff_w	-.13*	3.55	0.62	-.15*	3.67	0.64	-.14 <sup>n.s.</sup>	3.44	0.57

em\_eff\_w = Self-efficacy beliefs in managing negative emotions at work

<sup>n.s.</sup>  $p > .05$ . \*  $p < .05$ . \*\*\*  $p < .001$

for computing reliability indices for the whole sample and a multiple group SEM for computing reliability indices for male and female samples. The models fit the data well, with the exception of scaled (or robust) RMSEA only.<sup>5</sup> For the whole sample, the unscaled (or standard) indices of fit were: WLSMV-based  $\chi^2(27) = 51.884$ ,  $p = .003$ ; CFI = 0.995; TLI = 0.994; RMSEA = 0.051; while the scaled (robust) indices of fit were: WLSMV-based  $\chi^2(27) = 93.554$ ,  $p < .001$ ; CFI = 0.976; TLI = 0.968; RMSEA = 0.083. For the multiple group model, the unscaled (or standard) indices of fit were: WLSMV-based  $\chi^2(54) = 87.526$ ,  $p = .003$ ; CFI = 0.995; TLI = 0.993; RMSEA = 0.059; while the scaled (robust) indices of fit were: WLSMV-based  $\chi^2(54) = 150.014$ ,  $p < .001$ ; CFI = 0.970; TLI = 0.960; RMSEA = 0.100. Hence, from these models we extracted reliability indices through the `reliability()` function of the `semTools` package: For the whole sample,  $\rho_{NL} = .90$  and AVE = .59; for the male sample,  $\rho_{NL} = .91$  and AVE = .61; for the female sample,  $\rho_{NL} = .90$  and AVE = .58. All indices were above the recommended thresholds; thus, we can conclude that the SNAQ has a good degree of reliability.

The second step of structural validity analysis was the examination of measurement invariance across gender. Results of measurement invariance for categorical variables (Svetina et al., 2020; Wu & Estabrook, 2016) were reported in Table 3. A model comparison demonstrated that both thresholds and loadings are invariant across gender, without worsening the model (all delta chi-square tests were not significant, and both CFI and RMSEA did not worsen more than 0.01). Parameter estimates from *Equal Thresholds and Loadings Model* are reported in Table 4.

<sup>5</sup> There is little consensus on the superiority of standard fit indices over robust ones, as well as on acceptable levels of fit indices in categorical SEM (see Xia & Yang, 2019). For this reason, we reported both scaled/robust and unscaled/standard indices of fit. We interpreted the indices obtained as acceptable, as most of the indices are in line with the results of Xia and Yang (2019).

## Classification performance

In the Appendix 1, we reported descriptive statistics (Appendix Table A3) and double density plots (Appendix Fig. A1) for SNAQ composite scores as a function of gender and *self-labeled bullying*.

The first step of classification performance analysis was the estimation and comparison of the *Conditional and Unconstrained model* and the *Conditional and Constrained model*. A representation of both models is reported in Fig. 1, in which the only difference was that parameter  $\beta_{21}$  was free in the first model and constrained to be equal across gender in the second. Both models fit the data well [*Conditional and Unconstrained model*: WLSMV-based  $\chi^2(87) = 183.628$ ,  $p < .001$ ; CFI = 0.973; TLI = 0.972; RMSEA = 0.079. *Conditional and Constrained model*: WLSMV-based  $\chi^2(88) = 201.340$ ,  $p = .003$ ; CFI = 0.968; TLI = 0.967; RMSEA = 0.085], but constraining the path  $\beta_{21}$  to be equal across groups significantly worsened the fit of the model (Scaled- $\Delta\chi^2 = 7.6138$ ,  $\Delta df = 7$ ,  $p = .006$ ). Thus, we gathered predictive estimates from the *Conditional and Unconstrained model*, in which  $\beta_{21}$  was not constrained to be equal across gender models (for the male model:  $\beta_{21} = 0.774$ ,  $se = 0.078$ ,  $z = 9.890$ ,  $p < .001$ ;  $R^2 = 0.600$ ; for the female model:  $\beta_{21} = 1.134$ ,  $se = 0.138$ ,  $z = 8.228$ ,  $p < .001$ ;  $R^2 = 0.826$ ). After performing a logistic regression model and a random forest model, we computed the areas under the curve obtained through estimates from the three models. In Fig. 2, we reported ROC curve and percentage of area under the curve for each model and separated for gender. Following the aforementioned criteria (Lantz, 2019), the classification performance of the SNAQ was outstanding for the SEM model, excellent (male) and outstanding (female) for the logistic regression model, and excellent for the random forest model.

## Comparison of ROC curves

For the male sample, the AUC computed on the basis of the ROC curve built using SEM (90.3%) outperformed the



**Table 3** Measurement invariance results

Software	Model	WLSMV-based $\chi^2$	df	p	CFI	TLI	RMSEA	Scaled- $\Delta\chi^2$	$\Delta df$	p	$\Delta CFI$	$\Delta RMSEA$
R (Lavaan)	Baseline	150.014	54	<.001	.970	.960	.100	-	-	-	-	-
	Equal Thresholds	165.795	63	<.001	.968	.963	.096	13.514	9	.141	.002	-.004
	Equal Thresholds and Loadings	158.085	71	<.001	.973	.972	.083	5.140	8	.743	-.005	-.013
Mplus	Baseline	150.791	54	<.001	.970	.960	.100	-	-	-	-	-
	Equal Thresholds	165.590	63	<.001	.968	.964	.096	13.933	9	.125	.002	-.004
	Equal Thresholds and Loadings	152.627	71	<.001	.975	.974	.080	5.049	8	.752	-.007	-.016

WLSMV = Weighted Least Squares Mean- and Variance-adjusted; df = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation;  $\Delta CFI$  = decrease in CFI;  $\Delta RMSEA$  = increase in RMSEA

AUC computed by logistic regression (87.3%,  $D = 1.9773$ ,  $p = .048$ ), but was not significantly better than the one computed on the basis of random forest (87.6%,  $D = 1.2438$ ,  $p = .214$ ). For the female sample, instead, the AUC computed on the basis of the ROC curve built using SEM (96.6%) outperformed both that of logistic regression (94.5%,  $D = 2.6164$ ,  $p = .009$ ) and random forest (84.4%,  $D = 2.7108$ ,  $p = .007$ ). In sum, 3 out of 4 comparisons attested that AUC computed with our approach outperformed those obtained by observed-variable approaches.

### Discussion

This contribution had a three-fold aim, that is, to further confirm the good psychometric properties of the Italian version of the SNAQ, to show a method for computing AUC values from ROC curves that are not biased from measurement errors (unlike more traditional approaches), and to compare the AUC obtained after controlling for measurement error with the AUC computed by means of more traditional (observed-variable) approaches.

Our findings showed that the Italian SNAQ holds optimal reliability values and gender-invariant parameters, thus supporting the structural validity of the instrument. In particular, gender-invariance analysis showed the consistency of latent variable parameters across gender. This implies that this scale is able to (a) assess the same construct without being affected by potential gender-related differences (Ock et al., 2020; Vandenberg & Lance, 2000) and (b) detect parameter differences across gender since “the assumption that observed scores on a scale accurately reflect respondents’ standings on a measured construct” (Ock et al., 2020, p. 657) has been supported. Given that no previous study has investigated SNAQ gender invariance, this finding is an important contribution to the literature on the psychometric properties of the SNAQ. Following, ROC curves showed that the classification performance of the Italian SNAQ was outstanding (according to Lantz, 2019, p. 333) for both male (90.3%) and female (96.6%) samples; hence, the SNAQ is a good and brief instrument for classifying those who self-labeled themselves as bullied and those who did not. In conclusion, the Italian SNAQ showed very satisfactory psychometric properties that ensures its validity.

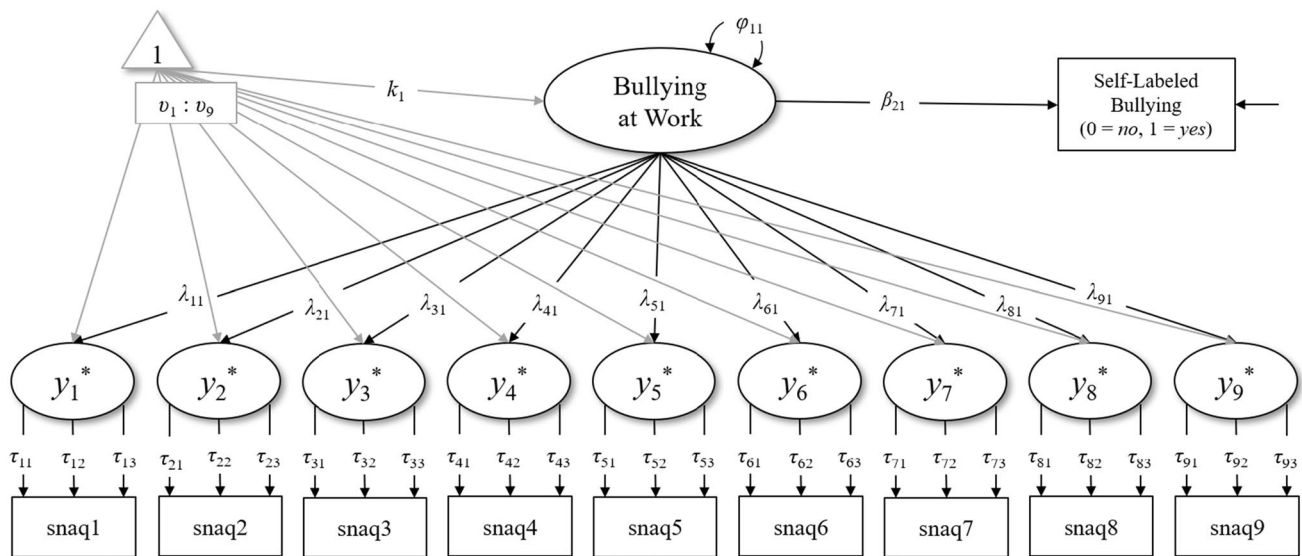
Another important point is that the definition we used for the self-labeled measure included a clear *intent to harm* (that is, the bully does the listed examples on purpose).<sup>6</sup> There is disagreement with regards to the inclusion of intent to harm among the features of bullying at work, in particular in the European tradition (see Einarsen et al., 2020, for an in-depth discussion). However, as Einarsen et al. (2020) put

<sup>6</sup> We are grateful to an anonymous reviewer for this observation.

**Table 4** Parameters of interest from the ‘Equal Thresholds and Loadings Model’ estimated in lavaan

Unconstrained parameters				Constrained parameters		
Parameter		Male model	Female model	Parameter		Male/Female model
Intercepts	$v_1$	0	0.042	Factor loadings	$\lambda_{11}$	0.794
	$v_2$	0	-0.114		$\lambda_{21}$	0.875
	$v_3$	0	0.216		$\lambda_{31}$	0.789
	$v_4$	0	-0.099		$\lambda_{41}$	0.924
	$v_5$	0	0.183		$\lambda_{51}$	0.671
	$v_6$	0	-0.106		$\lambda_{61}$	0.530
	$v_7$	0	0.156		$\lambda_{71}$	0.765
	$v_8$	0	0.078		$\lambda_{81}$	0.805
	$v_9$	0	-0.026		$\lambda_{91}$	0.830
Explained variances	$R^2_{snaq1}$	0.630	0.538	Thresholds	$\tau_{11} \tau_{12} \tau_{13}$	-0.092, 0.822, 1.139
	$R^2_{snaq2}$	0.765	0.597		$\tau_{21} \tau_{22} \tau_{23}$	0.083, 1.107, 1.403
	$R^2_{snaq3}$	0.622	0.619		$\tau_{31} \tau_{32} \tau_{33}$	0.254, 1.166, 1.444
	$R^2_{snaq4}$	0.854	0.738		$\tau_{41} \tau_{42} \tau_{43}$	0.552, 1.309, 1.756
	$R^2_{snaq5}$	0.450	0.489		$\tau_{51} \tau_{52} \tau_{53}$	0.249, 1.242, 1.675
	$R^2_{snaq6}$	0.281	0.366		$\tau_{61} \tau_{62} \tau_{63}$	-0.749, 0.700, 1.145
	$R^2_{snaq7}$	0.586	0.728		$\tau_{71} \tau_{72} \tau_{73}$	0.093, 1.002, 1.388
	$R^2_{snaq8}$	0.649	0.623		$\tau_{81} \tau_{82} \tau_{83}$	0.231, 1.204, 1.591
	$R^2_{snaq9}$	0.689	0.549		$\tau_{91} \tau_{92} \tau_{93}$	1.245, 1.900, 2.082
Latent variance	$\varphi_{11}$	1	0.648	Latent mean	$k_1$	0

Parameters are reported in unstandardized form

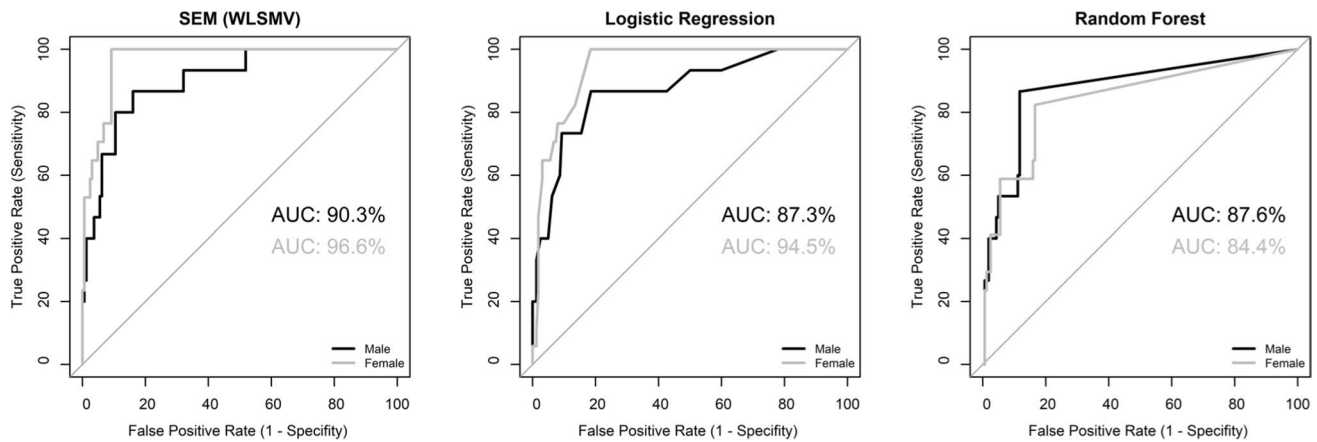


**Fig. 1** Specified structural equation model. *Note.* Asterisks indicate latent continuous variables assumed to underlie the observed categorical indicators. Given that we collapsed category 5 within category

4 (due to cells imbalance across gender), each item has 3 thresholds. Residual variance of each  $y_i^*$  is fixed to be zero. Paths for mean-level structure are reported in grey for sake of clarity

it, “whereas intent may be a controversial feature of bullying definitions, there is little doubt that perception of intent is important as to whether an individual decides to label their experience as bullying or not” (Einarsen et al., 2020,

p. 12–13). Our classification performance analysis supports the above reasoning, given that the SNAQ appears to be an optimal classifier of a self-labeled measure that includes “perceived” intent to harm, and the prevalence rate we found



**Fig. 2** ROC curves estimated from three different models. *Note.* The line along the diagonal represents the hypothetical performance of a random classifier (AUC=50%). SEM (WLSMV)=Structural Equation Model estimated with WLSMV estimator; AUC=Area Under the Curve

is similar to those commonly reported in the literature (i.e., from 6 to 11%; Zapf et al., 2020).

In this contribution, we also provided a data analytic strategy for investigating the classification performance of a continuous variable that may have a certain degree of measurement error (Jacobucci & Grimm, 2020). Our results are consistent with Jacobucci and Grimm (2020), in that the comparison of ROC curves highlighted the importance of taking into account measurement error when calculating AUC values, in order to evaluate the classification performance of an imperfectly reliable instrument (in our case, the SNAQ). Indeed, in all cases, the AUCs computed from SEM were higher than those computed with observed-variable techniques, and in three out of four cases, this difference was significant. Furthermore, we should also point out that our data analytic approach allows verifying whether the path linking a predictor variable to its classifier ( $\beta_{21}$  in Fig. 1) can be fixed to be equal or not across groups, and accordingly, to decide whether to build an overall ROC curve or separate ROC curves. Indeed, if the likelihood ratio test attests that the aforementioned path cannot be constrained to equality across groups, then ROC curves should be estimated separately, because this means that the classification performance of the instrument may vary across groups. Given that this information requires performing a preliminary multiple-group analysis along with a likelihood ratio test analysis, it would not be possible using a more traditional observed-variable approach. This is another advantage of using SEM when predictive values are of interest (Jacobucci & Grimm, 2020).

Related to the above point, Appendix Fig. A1 (see also Appendix Table A3) appeared to explain the gender difference in the parameter  $\beta_{21}$ : It seems that the SNAQ distribution for self-labeling bullied is different between males and females, in fact males reported a higher SNAQ mean but a

wider distribution, as also evidenced by a standard deviation approximately double that found in the female sample. Accordingly, results from the ROC curves showed that the Italian SNAQ is more reliable in classifying self-labeled bullying among females than males. In fact, as Appendix Fig. A1 shows, females who responded *yes* to the *self-labeled bullying* single-item were mostly concentrated in the right part of the SNAQ scores distribution, while this was not the case for men. This finding is consistent with Rosander et al. (2020), that found (a) how gender moderates workplace bullying across different measurement methods (behavioral experience vs. self-labeling) and (b) that men are more reluctant to admit being the victim of bullying. These findings again reinforce the need to previously attest gender measurement invariance of behavioral experience methods when studying workplace bullying.

### Limitations and future research

The present study has several limitations that should be acknowledged. First of all, data are self-report and cross-sectional. While we believe that the expressed purposes of the paper match with the design – and thus justify the use of a cross-sectional and self-report dataset, with retrospective assessment of work-related bullying experiences (see Spector, 2019) – we also acknowledge that the use of longitudinal data and/or objective measures of *self-labeled bullying* may significantly enhance analysis on the validity of the SNAQ.

Second, in predictive analytic models (i.e., machine learning, data mining, big data approaches), models are both trained and tested, in order to prevent overfitting and maximize predictive accuracy. In this study, we only trained data for two reasons: First, the low sample size and the low frequencies of self-labeled bullied (see Appendix Table A3) made it difficult to create  $k$ -folds and cross-validate results;

second, to our knowledge, there are no contributions or guidelines on procedures for test-and-train datasets when using Structural Equation Models (e.g., for CFAs design – such as categorical CFA – and measurement invariance routine). Hence, we decided to use classical approaches for estimating SEM, but we hope that future contributions may provide specific guidelines on training/testing data and using *k*-fold cross validation in such scenarios (e.g., Brandmaier & Jacobucci, *in press*).

Third, given the presence of some zero-cell counts, we collapsed categories before testing measurement invariance. While this procedure is advised by some scholars (DiStefano et al., 2021), it is also criticized by others (Rutkowski et al., 2019); again, we hope that future studies will resolve this matter.

Fourth, ROC relies upon “a gold standard” (Streiner & Cairney, 2007). In the Introduction, we reported important reasons for testing the classification performance of the SNAQ by means of the self-labeling method. However, as Notelaers and Van der Heijden (2021) put it, while in medical science the gold standard is in most of the cases objective in nature (e.g., having or not having a disease) in psychology (in particular, in organizational psychology), the standard is less strong and subjective to researcher’s decision. Therefore, “in workplace bullying and harassment research, an objective standard is often not available, as also in this scholarly field the empirical evidence is less strong” (Notelaers & Van der Heijden, 2021, p. 404). Thus, in future research alternative outcomes could be used to attest the classification performance of the Italian SNAQ. For example, Notelaers and Einarsen (2013) relied on diagnostic criterion that identified the need for psychiatric treatment to prevent depression by means of the Hopkins Checklist (Derogatis et al., 1974). While future studies could adopt a similar “gold standard” (as well as a multi-item self-labeling measure of workplace bullying; see Conway et al., 2021a), our contribution adds to the literature a way of taking into account measurement error in SNAQ (and in other instruments with a non-negligible degree of measurement error) when building ROC curves.

Fifth, given that we used a categorical latent variable approach, it was not possible to extract a specific cut-off, as in Notelaers and Einarsen (2013). However, this approach is useful for the computation of an overall classification performance for validity purposes (as in this case).

Sixth, as a constraint on the generalizability of results (Simons et al., 2017), our findings may be generalized to a specific Italian province, while a country-level generalization could be done only when data are collected on a representative workforce of the whole country (e.g., Notelaers & Einarsen, 2013).

Seventh, we used a six-month time frame for the SNAQ (according to Balducci et al., 2010), while a one-year time frame for self-labeled bullying due to constraints on the purposes of the whole project. However, given the high stability of the self-labeling victimization (attested even after four years; see Rosander & Blomberg, 2019), results should not be significantly affected by this slight difference. Yet, in future research it would be preferable to use the same time frame for both methods.

Finally, we ran a series of comparisons between ROC curve estimated with a latent variable approach and those estimated with an observed variable approach. Albeit our findings are promising (three out of four AUCs are significantly higher), we recognize that in order to attest differences between these methods a large-scale simulation is needed.

## Conclusion

In conclusion, our study demonstrated the validity of the Italian version of the SNAQ, hence providing further support to the utility of this instrument in assessing (with only 9 items) workplace bullying. Moreover, we provided new insights for the organizational research methodology literature by providing a novel analytic strategy for building ROC curves through a Structural Equation Modeling approach, that performed significantly better than classical observed-variable approach in the analyses of the classification performance of an instrument that is not perfectly reliable – as it is often the case with self-report instruments commonly used in organizational psychology and management. Finally, we hope that our contribution may be a step forward in the ongoing integration between predictive modeling approaches and latent variable models (e.g., Brandmaier & Jacobucci, *in press*; Jacobucci & Grimm, 2020), in particular with regard to the classification performance of measures with a non-negligible degree of measurement error.

## Appendix 1

**Table A1** Self-report instruments for assessing bullying at work and related phenomena (e.g., Ostracism)

Source	Items	Scale	Field
International Context			
Dilek and Aytolan (2008)	33	Workplace psychologically violent behaviours instrument	Nursing
Einarsen et al. (2009)	22	Negative Acts Questionnaire-Revised (NAQ-R)	Organizational Psychology
Ferris et al. (2008)	10	Workplace Ostracism Scale	Organizational Psychology
Notelaers et al. (2019)	9	Short Negative Acts Questionnaire (SNAQ)	Organizational Psychology
Ozturk et al. (2008)	60	Mobbing Scale for Academic Nurses	Nursing
Spector and Jex (1998)	4	Interpersonal Conflict at Work Scale	Organizational Psychology
Steffgen et al. (2019)	5	Luxembourg Workplace Mobbing Scale (LWMS)	Organizational Psychology
Zapf et al. (1996)	38	Leymann Inventory of Psychological Terrorization (LIFT)	Organizational Psychology
Italian Context			
Giorgi et al. (2011)	17	Negative Acts Questionnaire-Revised (NAQ-R)	Organizational Psychology
Balducci et al. (2010)	9	Short Negative Acts Questionnaire (SNAQ)	Organizational Psychology

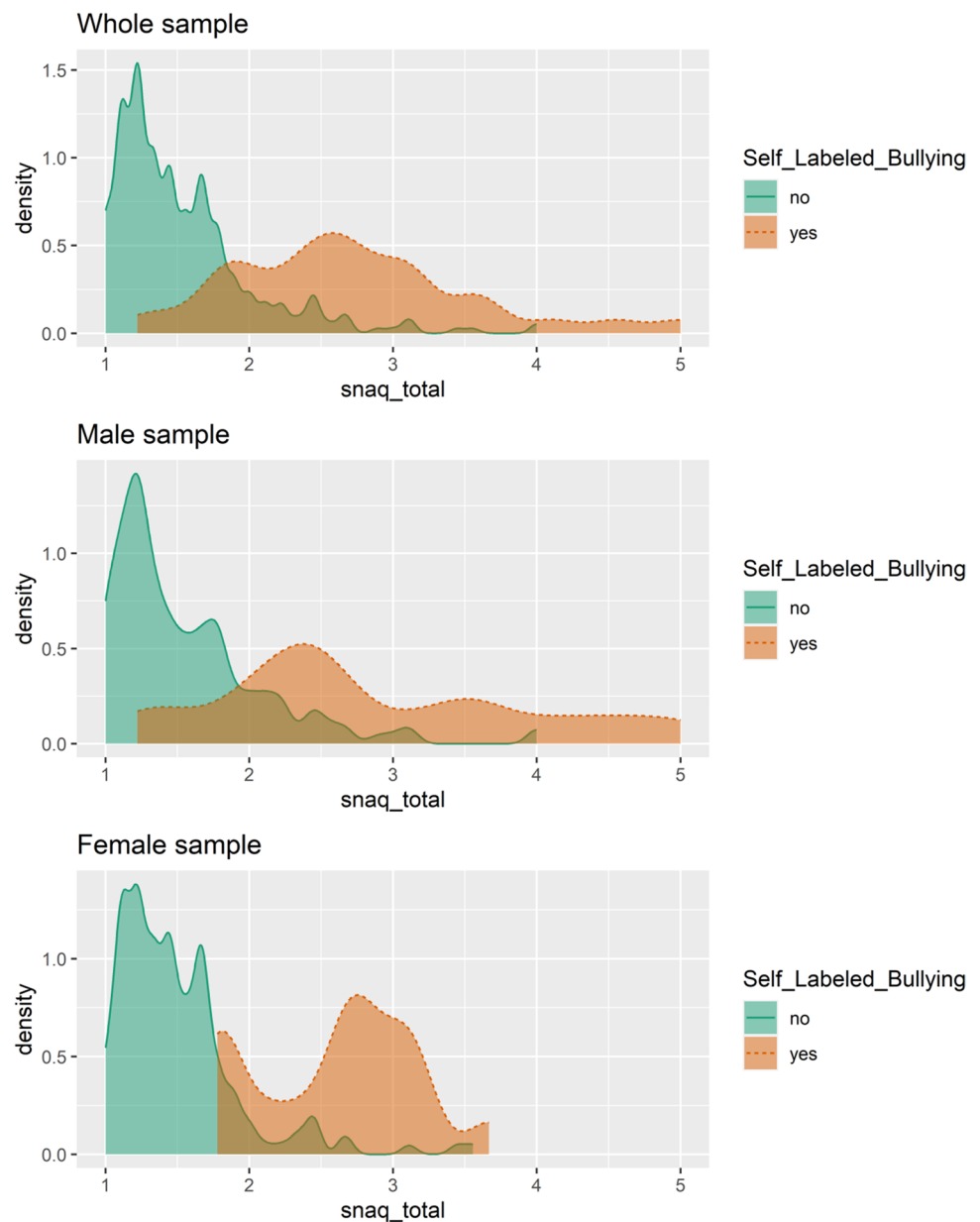
**Table A2** Frequencies and percentages of job sectors

Job sector	Frequency	Percentage
Agriculture	10	2.80%
Banks and insurance companies	25	7.00%
Commerce and public exercises	46	12.89%
Building	12	3.36%
Energy, gas, water, telephony	10	2.80%
Law enforcement / military	3	0.84%
Industry	51	14.29%
Public administration	47	13.17%
Health	32	8.96%
School/University	52	14.57%
Business services	24	6.72%
Services to people	35	9.80%
Transportation	10	2.80%

**Table A3** Descriptive statistics for SNAQ composite score and self-labeled bullying ( $N=357$ )

Sample	Self-labeled bullying	Frequency	Percentage	$M_{\text{SNAQ}}$	$SD_{\text{SNAQ}}$
Whole	No	325	91.04%	1.53	0.49
	Yes	32	8.96%	2.72	0.86
Male	No	162	45.38%	1.56	0.55
	Yes	15	4.20%	2.85	1.11
Female	No	163	45.66%	1.50	0.44
	Yes	17	4.76%	2.59	0.57

**Fig. A1** Double density plots: Distribution of SNAQ scores (snaq\_total) conditioned to Self-Labeled Bullying



**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s12144-022-03741-4>.

**Acknowledgements** The authors are grateful to the Trento branch of the national statistical center for their pivotal role in sampling and data collection.

**Funding** Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement. Enrico Perinelli was supported by funds from the University of Trento (WeBeWo LAB and Department of Psychology and Cognitive Science) and by a grant from the Autonomous Province of Trento (Project: *Mobbing in the Trentino Area*).

**Data availability** The dataset analyzed during the current study are not publicly available due to privacy restrictions. However, R codes and *Mplus* syntaxes are available in Supplementary Material.

## Declarations

**Ethical approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

**Conflict of interest** All authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alessandri, G., Perinelli, E., De Longis, E., Schaufeli, W. B., Theodorou, A., Borgogni, L., Caprara, G. V., & Cinque, L. (2018). Job burnout: The contribution of emotional stability and emotional self-efficacy beliefs. *Journal of Occupational and Organizational Psychology*, *91*(4), 823–851. <https://doi.org/10.1111/joop.12225>
- Avanzi, L., Balducci, C., & Fraccaroli, F. (2013). Contributo alla validazione italiana del Copenhagen Burnout Inventory (CBI) [A contribution to the Italian validation of the Copenhagen burnout inventory]. *Psicologia Della Salute*, *2*, 120–135. <https://doi.org/10.3280/PDS2013-002008>
- Balducci, C., Baillien, E., Broeck, A. V. D., Toderi, S., & Fraccaroli, F. (2020). Job demand, job control, and impaired mental health in the experience of workplace bullying behavior: A two-wave study. *International Journal of Environmental Research and Public Health*, *17*(4), Article 1358. <https://doi.org/10.3390/ijerph17041358>
- Balducci, C., Conway, P. M., & van Heugten, K. (2021). The contribution of organizational factors to workplace bullying, emotional abuse and harassment. In P. D'Cruz et al. (Eds.), *Pathways of job-related negative behaviour. Handbooks of workplace bullying, emotional abuse and harassment* (Vol. 2, pp. 3–28). Springer. [https://doi.org/10.1007/978-981-13-0935-9\\_1](https://doi.org/10.1007/978-981-13-0935-9_1)
- Balducci, C., Spagnoli, P., Alfano, V., Barattucci, M., Notealers, G., & Fraccaroli, F. (2010). Valutare il rischio mobbing nelle organizzazioni: Contributo alla validazione italiana dello Short Negative Acts Questionnaire (S-NAQ) [Assessing the mobbing risk in organizations: Contribution to the Italian validation of the Short Negative Acts Questionnaire (S-NAQ)]. *Psicologia Sociale*, *5*(1), 147–167. <https://doi.org/10.1482/32024>
- Barbaranelli, C., Bortone, I., & Di Matteo, F. (2010). La misura della soddisfazione lavorativa: Un contributo empirico [Measuring job satisfaction: An empirical contribution]. *Giornale Italiano Di Psicologia*, *37*(1), 159–180. <https://doi.org/10.1421/32175>
- Bentley, T. A., Teo, S. T. T., Nguyen, D. T. N., Blackwood, K., Catley, B., Gardner, D., Forsyth, D., Bone, K., Tappin, D., D'Souza, N., & Port, Z. (2021). Psychosocial influences on psychological distress and turnover intentions in the workplace. *Safety Science*, *137*, Article 105200. <https://doi.org/10.1016/j.ssci.2021.105200>
- Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. <https://doi.org/10.1177/1088868318772990>
- Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley, Hoboken. <https://doi.org/10.1002/9781118619179>
- Bonde, J. P., Gullander, M., Hansen, Å.M., Grynderup, M., Persson, R., Hogh, A., Willert, M. V., Kaerlev, L., Rugulies, R., & Kolstad, H. A. (2016). Health correlates of workplace bullying: A 3-wave prospective follow-up study. *Scandinavian Journal of Work, Environment & Health*, *42*(1), 17–25. <https://doi.org/10.5271/sjweh.3539>
- Brandmaier, A. M., & Jacobucci, R. (in press). Machine-learning approaches to structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed.). Guilford.
- Bulutlar, F., & Öz, E. Ü. (2009). The effects of ethical climates on bullying behaviour in the workplace. *Journal of Business Ethics*, *86*(3), 273–295. <https://doi.org/10.1007/s10551-008-9847-4>
- Calì, C., & Longobardi, M. (2015). Some mathematical properties of the ROC curve and their applications. *Ricerche Di Matematica*, *64*(2), 391–402. <https://doi.org/10.1007/s11587-015-0246-8>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. <https://doi.org/10.1037/a0033805>
- Conway, P. M., Burr, H., Rose, U., Clausen, T., & Balducci, C. (2021a). Antecedents of workplace bullying among employees in Germany: Five-year lagged effects of job demands and job resources. *International Journal of Environmental Research and Public Health*, *18*(20), Article 10805.
- Conway, P. M., Høgh, A., Balducci, C., & Ebbesen, D. K. (2021b). Workplace bullying and mental health. In P. D'Cruz et al. (Eds.), *Pathways of job-related negative behavior. Handbooks of workplace bullying, emotional abuse and harassment* (Vol. 2, pp. 101–128). Springer. [https://doi.org/10.1007/978-981-13-0935-9\\_5](https://doi.org/10.1007/978-981-13-0935-9_5)
- Cullen, J. B., Victor, B., & Bronson, J. W. (1993). The ethical climate questionnaire: An assessment of its development and validity. *Psychological Reports*, *73*(2), 667–674. <https://doi.org/10.2466/pr0.1993.73.2.667>
- Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral Science*, *19*(1), 1–15. <https://doi.org/10.1002/bs.3830190102>
- DiStefano, C., Shi, D., & Morgan, G. B. (2021). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 237–249. <https://doi.org/10.1080/10705511.2020.1803073>
- Dilek, Y., & Aytolan, Y. (2008). Development and psychometric evaluation of workplace psychologically violent behaviours instrument. *Journal of Clinical Nursing*, *17*(10), 1361–1370. <https://doi.org/10.1111/j.1365-2702.2007.02262.x>
- Dubinsky, A. J., & Hartley, S. W. (1986). A path-analytic study of a model of salesperson performance. *Journal of the Academy of Marketing Science*, *14*(1), 36–46. <https://doi.org/10.1007/BF02722111>
- Einarsen, S., Hoel, H., & Notealers, G. (2009). Measuring exposure to bullying and harassment at work: Validity, factor structure and psychometric properties of the Negative Acts Questionnaire-Revised. *Work & Stress*, *23*(1), 24–44. <https://doi.org/10.1080/02678370902815673>
- Einarsen, S. V., Hoel, H., Zapf, D., & Cooper, C. L. (2020). The concept of bullying and harassment at work: The European tradition. In S. V. Einarsen, H. Hoel, D. Zapf, & C. L. Cooper (Eds.), *Bullying and harassment in the workplace: Theory, research and practice* (3rd ed., pp. 3–53). CRC Press.
- Einarsen, S., & Skogstad, A. (1996). Bullying at work: Epidemiological findings in public and private organizations. *European Journal of Work and Organizational Psychology*, *5*(2), 185–201. <https://doi.org/10.1080/13594329608414854>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Ferris, D. L., Brown, D. J., Berry, J. W., & Lian, H. (2008). The development and validation of the Workplace Ostracism Scale. *Journal of Applied Psychology*, *93*(6), 1348–1366. <https://doi.org/10.1037/a0012743>
- Fida, R., Laschinger, H. K. S., & Leiter, M. P. (2018). The protective role of self-efficacy against workplace incivility and burnout in

- nursing: A time-lagged study. *Health Care Management Review*, 43(1), 21–29. <https://doi.org/10.1097/HMR.0000000000000126>
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A second course* (2nd ed., pp. 439–492). Information Age Publishing.
- Flora, D. B. (2020). Your coefficient Alpha is probably wrong, but which coefficient Omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501. <https://doi.org/10.1177/2515245920951747>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.1177/002224378101800104>
- Giorgi, G., Arenas, A., & Leon-Perez, J. M. (2011). An operative measure of workplace bullying: The negative acts questionnaire across Italian companies. *Industrial Health*, 49(6), 686–695. <https://doi.org/10.2486/indhealth.MS1287>
- Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). ROC curve estimation: An overview. *REVSTAT - Statistical Journal*, 12(1), 1–20. <https://www.ine.pt/revstat/pdf/rs140101.pdf>
- Hershcovis, M. S., Reich, T. C., & Niven, K. (2015). *Workplace bullying: Causes, consequences, and intervention strategies*. Society for Industrial and Organizational Psychology, UK, London.
- Howard, M. C., Cogswell, J. E., & Smith, M. B. (2020). The antecedents and outcomes of workplace ostracism: A meta-analysis. *Journal of Applied Psychology*, 105(6), 577–596. <https://doi.org/10.1037/apl0000453>
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809–816. <https://doi.org/10.1177/1745691620902467>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). *semTools: Useful tools for structural equation modeling. R package version 0.5–3*. Retrieved from <https://CRAN.R-project.org/package=semTools>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Kristensen, T. S., Borritz, M., Villadsen, E., & Christensen, K. B. (2005). The Copenhagen Burnout Inventory: A new tool for the assessment of burnout. *Work & Stress*, 19(3), 192–207. <https://doi.org/10.1080/02678370500297720>
- Kuhn, M., & Johnson, K. (2020). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.
- Landau, J., & Hammer, T. H. (1986). Clerical employees' perceptions of intraorganizational career opportunities. *Academy of Management Journal*, 29(2), 385–404. <https://doi.org/10.5465/256194>
- Lantz, B. (2019). *Machine learning with R: Expert techniques for predictive modeling* (3rd ed.). Packt.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News* 2(3), 18–22. [https://cran.r-project.org/doc/Rnews/Rnews\\_2002-3.pdf](https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf)
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635–694. <https://doi.org/10.2466/pr0.1957.3.3.635>
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52(6), 2287–2305. <https://doi.org/10.3758/s13428-020-01398-0>
- Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.
- Mount, J., & Zume, N. (2020). *WVPlots: Common Plots for Analysis. R package version 1.2.7*. <https://CRAN.R-project.org/package=WVPlots>
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132. <https://doi.org/10.1007/BF02294210>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide. Eighth Edition*. Muthén & Muthén.
- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.
- Nielsen, M. B., & Einarsen, S. V. (2018). What we know, what we do not know, and what we should and could have known about workplace bullying: An overview of the literature and agenda for future research. *Aggression and Violent Behavior*, 42, 71–83. <https://doi.org/10.1016/j.avb.2018.06.007>
- Nielsen, M. B., Matthiesen, S. B., & Einarsen, S. (2010). The impact of methodological moderators on prevalence rates of workplace bullying. A meta-analysis. *Journal of Occupational and Organizational Psychology*, 83(4), 955–979. <https://doi.org/10.1348/096317909X481256>
- Nielsen, M. B., Notelaers, G., & Einarsen, S. V. (2020). Methodological issues in the measurement of workplace bullying. In S. V. Einarsen, H. Hoel, D. Zapf, & C. L. Cooper (Eds.), *Bullying and harassment in the workplace: Theory, research and practice* (3rd ed., pp. 235–265). CRC Press.
- Notelaers, G., & Einarsen, S. (2008). *The construction and validity of the Short-Negative Acts Questionnaire*. Paper presented at the 6th International Conference on Bullying and Harassment in the Workplace, Montreal, Canada.
- Notelaers, G., & Einarsen, S. (2013). The world turns at 33 and 45: Defining simple cutoff scores for the Negative Acts Questionnaire-Revised in a representative sample. *European Journal of Work and Organizational Psychology*, 22(6), 670–682. <https://doi.org/10.1080/1359432X.2012.690558>
- Notelaers, G., & Van der Heijden B. I. J. M. (2021). Construct validity in workplace bullying and harassment research. In P. D'Cruz, E. Noronha, G. Notelaers, & C. Rayner (Eds.), *Concepts, approaches and methods. Handbooks of workplace bullying, emotional abuse and harassment* (Vol. 1, pp. 369–424). Springer. [https://doi.org/10.1007/978-981-13-0134-6\\_11](https://doi.org/10.1007/978-981-13-0134-6_11)
- Notelaers, G., Van der Heijden, B., Hoel, H., & Einarsen, S. (2019). Measuring bullying at work with the short-negative acts questionnaire: Identification of targets and criterion validity. *Work & Stress*, 33(1), 58–75. <https://doi.org/10.1080/02678373.2018.1457736>
- Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The practical effects of measurement invariance: Gender invariance in two Big Five personality measures. *Assessment*, 27(4), 657–674. <https://doi.org/10.1177/1073191119885018>
- Ozturk, H., Sokmen, S., Yilmaz, F., & Cilingir, D. (2008). Measuring mobbing experiences of academic nurses: Development of a mobbing scale. *Journal of the American Academy of Nurse Practitioners*, 20(9), 435–442. <https://doi.org/10.1111/j.1745-7599.2008.00347.x>
- Pagliaro, S., Lo Presti, A., Barattucci, M., Giannella, V. A., & Barreto, M. (2018). On the effects of ethical climate(s) on employees' behavior: A social identity approach. *Frontiers in Psychology*, 9, Article 960. <https://doi.org/10.3389/fpsyg.2018.00960>
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>
- Reknes, I., Einarsen, S. V., Gjerstad, J., & Nielsen, M. B. (2019). Dispositional affect as a moderator in the relationship between role



- conflict and exposure to bullying behaviors. *Frontiers in Psychology*, 10, Article 44. <https://doi.org/10.3389/fpsyg.2019.00044>
- Robin, X., Turk, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, Article 77. <https://doi.org/10.1186/1471-2105-12-77>
- Rodríguez-Muñoz, A., Baillien, E., De Witte, H., Moreno-Jiménez, B., & Pastor, J. C. (2009). Cross-lagged relationships between workplace bullying, job satisfaction and engagement: Two longitudinal studies. *Work & Stress*, 23(3), 225–243. <https://doi.org/10.1080/02678370903227357>
- Rosander, M., & Blomberg, S. (2019). Levels of workplace bullying and escalation – a new conceptual model based on cut-off scores, frequency and self-labelled victimization. *European Journal of Work and Organizational Psychology*, 28(6), 769–783. <https://doi.org/10.1080/1359432X.2019.1642874>
- Rosander, M., Salin, D., Viita, L., & Blomberg, S. (2020). Gender matters: Workplace bullying, gender, and mental health. *Frontiers in Psychology*, 11, Article 2683. <https://doi.org/10.3389/fpsyg.2020.560178>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rutkowski, L., Svetina, D., & Liaw, Y.-L. (2019). Collapsing categorical variables and measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5), 790–802. <https://doi.org/10.1080/10705511.2018.1547640>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Spector, P. E. (2019). Do not cross me: Optimizing the use of cross-sectional designs. *Journal of Business and Psychology*, 34(2), 125–137. <https://doi.org/10.1007/s10869-018-09613-8>
- Spector, P. E., & Jex, S. M. (1998). Development of four self-report measures of job stressors and strain: Interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory. *Journal of Occupational Health Psychology*, 3(4), 356–367. <https://doi.org/10.1037/1076-8998.3.4.356>
- Stanley, D. (2018). *apaTables: Create American Psychological Association (APA) Style Tables*. R package version 2.0.5. <https://CRAN.R-project.org/package=apaTables>
- Steffgen, G., Sischka, P., Schmidt, A. F., Kohl, D., & Happ, C. (2019). The Luxembourg Workplace Mobbing Scale: Psychometric properties of a short instrument in three different languages. *European Journal of Psychological Assessment*, 35(2), 164–171. <https://doi.org/10.1027/1015-5759/a000381>
- Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry*, 52(2), 121–128. <https://doi.org/10.1177/070674370705200210>
- Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 111–130. <https://doi.org/10.1080/10705511.2019.1602776>
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zonder-van-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, 6, Article 1064. <https://doi.org/10.3389/fpsyg.2015.01064>
- Van den Brande, W., Baillien, E., De Witte, H., Vander Elst, T., & Godderis, L. (2016). The role of work stressors, coping strategies and coping resources in the process of workplace bullying: A systematic review and development of a comprehensive model. *Aggression and Violent Behavior*, 29, 61–71. <https://doi.org/10.1016/j.avb.2016.06.004>
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70. <https://doi.org/10.1177/109442810031002>
- Vartia, M. (1996). The sources of bullying—psychological work environment and organizational climate. *European Journal of Work and Organizational Psychology*, 5(2), 203–214. <https://doi.org/10.1080/13594329608414855>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... Kuhn, M. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.
- Workplace Bullying Institute (2017). *U.S. Workplace Bullying Survey*. Retrieved from <https://workplacebullying.org/download/2017-wbi/>
- Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, 81(4), 1014–1045. <https://doi.org/10.1007/s11336-016-9506-0>
- Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51(1), 409–428. <https://doi.org/10.3758/s13428-018-1055-2>
- Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology*, 11(1), 23–34. <https://doi.org/10.1027/1614-2241/a000087>
- Zapf, D., Knorz, C., & Kulla, M. (1996). On the relationship between mobbing factors, and job content, social work environment, and health outcomes. *European Journal of Work and Organizational Psychology*, 5(2), 215–237. <https://doi.org/10.1080/13594329608414856>
- Zapf, D., Escartín, J., Scheppa-Lahyani, M., Einarsen, S. V., Hoel, H., & Vartia, M. (2020). Empirical findings on prevalence and risk groups of bullying in the workplace. In S. V. Einarsen, H. Hoel, D. Zapf, & C. L. Cooper (Eds.), *Bullying and harassment in the workplace: Theory, research and practice* (3rd ed., pp. 105–162). CRC Press.
- Zumel, N., & Mount, J. (2020). *Practical data science with R* (2nd ed.). Manning Publications.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.