Personality and Individual Differences 120 (2018) 159–165

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Development of a new abbreviated form of the Junior Eysenck Personality Questionnaire-Revised

Daiana Colledani*, Egidio Robusto, Pasquale Anselmi

Department FISPPA, Section of Applied Psychology, University of Padova, Italy

ARTICLE INFO

Keywords:
Junior Eysenck Personality Questionnaire Revised-Abbreviated
2PL
Gender DIF.

ABSTRACT

This work aims to develop a new version of the Junior Eysenck Personality Questionnaire-Revised-Abbreviated (JEPQR-A) with improved measurement properties. Two studies were carried out and various analyses performed. In the first study, the 89 items of the full version of the questionnaire were administered to a sample of participants ($N = 549$) and the data analysed in order to select the 24 items with the best metric properties. An investigation of the parameters of the 2PL model, DIF statistics and item fit measures allowed the selection of 24 items that were unbiased, well discriminating and well fitting. In the second study, the reliability, factor structure and convergent validity of the new abbreviated questionnaire were evaluated on a different sample of participants ($N = 234$). The results suggest that the new version of the JEPQR-A outperforms the previous form.

1. Introduction

Eysenck's questionnaires are well known in personality psychology, both the adult and junior forms being still widely used for clinical, scientific and professional purposes (e.g., Bowden, Saklofske, Van de Vijver, Sudarshan, & Eysenck, 2016; Chan et al., 2016; Manandhar et al., 2017; Saigh et al., 2016; Tao, Zhang, & Qu, 2017). Eysenck's instruments have been refined through many years of work and the contributions of many experts. Research efforts have been devoted to improving the psychometric properties of the instruments and refining their underlying theory. However, these refinements have not only improved the psychometric characteristics of the questionnaires, they have also increased their length. The last revision of the adult form (EPQ-R; Eysenck & Eysenck, 1991) assesses the four PEN-L traits (Psychoticism, Extraversion, Neuroticism-Lie) through 100 items, while the Junior form uses 89 (JEPQ-R; Corulla, 1990).

The need for instruments that are valid and reliable but at the same time easy to manage, is a well-known concern in psychology. Hence, over the years many suggestions have been offered for devising brief versions of Eysenck's questionnaires. Short and abbreviated forms have been developed for both adult and junior scales, and, because of their usefulness and satisfactory psychometric properties, have received reasonable approval in the scientific field (e.g., Forrest, Lewis, & Shevlin, 2000; Karanci, Dirik, & Yorulmaz, 2006; Maltby & Talley, 1998; Scholte & De Bruyn, 2001).

The short questionnaires contain 48 items (Corulla, 1990; Eysenck, Eysenck, & Barrett, 1985), while abbreviated forms assess all the PEN-L traits using only 24 items (Francis, 1996; Francis, Brown, & Philipchalk, 1992; Francis & Pearson, 1988). The abbreviated version of the Junior Eysenck Personality Questionnaire-Revised (JEPQR-A; Francis, 1996), for instance, contains four scales of six items each, and was developed from the 48 items of the short form of the questionnaire (Corulla, 1990). The author reported satisfactory psychometric properties of each of the scales, with validity coefficients ranging between 0.88 and 0.92 (assessed in relation to JEPQR-S scores; Corulla, 1990), and reliability coefficients ranging from 0.57 to 0.70 (respectively 0.61, 0.66, 0.70 and 0.57 for the PEN-L scales). Cross-cultural studies have confirmed these results and support the four-factor structure of the instrument (e.g., Maltby & Talley, 1998; Scholte & De Bruyn, 2001). These results are satisfactory given that each scale contains six items only. However, the reliability coefficients of the P and L scales remain rather low.

Further improvements could also be made to control for gender DIF (Differential Item Functioning). Many studies have highlighted gender effects in Eysenck's scales, at both the item and scale levels (e.g., Escorial & Navas, 2007; Eysenck et al., 1985; Eysenck & Eysenck, 1991; Forrest et al., 2000; Karanci et al., 2006). Gender DIF is a serious weakness for psychological assessment tools that should be accurately evaluated and reduced, especially in short and abbreviated instruments.

Abbreviated instruments should be developed by selecting items that cover the entire trait *continuum* and excluding those that are biased, unclear or ambiguous. Item response theory (IRT) is a valuable tool for this purpose. IRT models help to detect biased and misfitting items, as well as latent traits levels that are not adequately measured by the items being used (Baker & Kim, 2004; Bartholomew & Knott, 1999;

* Corresponding author at: FISPPA, Section of Applied Psychology, Via Venezia 14, 35131 Padova, Italy.
E-mail address: daianacolledani@gmail.com (D. Colledani).

http://dx.doi.org/10.1016/j.paid.2017.08.037
Received 25 April 2017; Received in revised form 22 August 2017; Accepted 23 August 2017
Available online 20 September 2017

Bortolotti, Tezza, de Andrade, Bornia, & de Sousa Júnior, 2013; Devlieger, Mpofu, Hawkins, & Prince, 2010; Zanon, Hutz, Yoo, & Hambleton, 2016). To date, these models have been only rarely employed in the refinement of Eysenck's scales (e.g., Escorial & Navas, 2007; Ferrando, 2001; Ferrando & Chico, 2001; Maij-de Meij, Kelderman, & Flier, 2008).

The aim of the present paper is to further improve the psychometric properties of the abbreviated JEPQ-R. To this end, a new version of the instrument was developed by selecting items without gender DIF and with adequate measurement properties.

The work is divided into two main studies. In the first, 24 items (six for each scale) with the best psychometric properties were selected from the 89 items of the JEPQ-R (Corulla, 1990). In the second, the psychometric characteristics of the new abbreviated form were tested on a new sample of participants. The reliability, validity and factor structure of the new questionnaire were evaluated.

2. Study 1

2.1. Participants and procedure

The sample consisted of 549 adolescents aged between 13 and 17 (mean age = 15.65, *SD* = 1.01; 222 males). All participants were native Italian speakers. They completed the JEPQ-R (Corulla, 1990) during school hours in the presence of teachers and researchers. The headmasters gave consent for the study, and all participants were informed that their participation was anonymous and voluntary. All standards for research with human subjects were respected.

2.2. Instruments

The Junior Eysenck Personality Questionnaire–Revised (Corulla, 1990) is an 89-item, self-report questionnaire used to assess young people on three dimensions of personality, namely psychoticism (25 items; e.g., Do you seem to get into a lot of fights? Would you feel very sorry for an animal caught in a trap?); extraversion (25 items; e.g., Have you got lots of friends? Would you rather sit and watch than play at parties?); and neuroticism (20 items; e.g., Do you worry for a long while if you feel you have made a fool of yourself? Do you worry about awful things that might happen?). The questionnaire also includes a Lie scale devised to detect dissimulation and socially desirable responding (19 items; e.g., Do you always say you are sorry when you have been rude? Have you ever said anything bad or nasty about anyone?). The response to each item is dichotomous (yes/no).

The questionnaire showed satisfactory psychometric characteristics, and the P scale performed better in this version than in previous forms. In this study, the Italian version of the tool was used.[1]

2.3. Analysis strategy

In order to select the 24 items with the best psychometric properties, a two-parameter logistic model (2PL) was separately estimated for the data of each of the four scales of the JEPQ-R. All analyses were performed using the packages ltm (Rizopoulos, 2012), and difR (Magis, Beland, Raiche, & Magis, 2015) for the open-source statistical environment (R Core Team, 2016).

[1] The scale was translated from English to Italian, and then back-translated by a native English speaker. Two items of the P scale were reworded. Specifically, Item 8 (Do you think sniffing glue is dangerous?) was translated as 'Do you think sniffing chemical substances, that may have "strange" effects, is dangerous?', whereas Item 87 (Do you think football hooligans are bad people?) was translated as 'Do you think extreme football supporters are bad people?' These items were reworded because 'hooligans' and 'glue sniffing' are not common concepts in an Italian context. Moreover, the terms 'child-children' were translated as 'young people' because it seemed more suitable for a questionnaire intended for adolescents.

In the 2PL model, the probability that a subject will endorse an item (i.e., provide a 'yes' response) is a function of the level of the latent trait of the subject (parameter θ), the level of 'endorsability' of the item (ease of providing a 'yes' response; item easiness; parameter b), and the effectiveness of the item in detecting the trait level (item discrimination; parameter a). Let us consider, for example, data from the P scale: the larger the value of parameter θ, the greater the level of psychoticism of the subject; the larger the value of parameter b, the greater the ease of responding 'yes' to the item (i.e., of providing a response that is indicative of the presence of psychoticism); and the larger the value of parameter a, the more effective the item is at differentiating between subjects with high and low levels of psychoticism.

2.3.1. Item selection

2.3.1.1. Fit indices. Three item fit indices were considered in order to detect misfitting items, which were then flagged for exclusion. Specifically, two chi-square statistics were calculated: 'item infit' (information-weighted, mean square residual goodness of fit statistic) and 'item outfit' (outlier-sensitive, mean square residual goodness of fit statistic). The first represents an 'information-weighted' fit statistic, sensitive to unexpected responses close to a subject's trait level, whereas the latter is an unweighted statistic, useful for detecting unexpected responses far from a person's trait level. These statistics are useful for identifying items that may be ambiguous or unclear. In the present study, items showing infit and/or outfit larger than 1.4 (Wright & Linacre, 1994) were considered for exclusion.

Item fit was also analysed using the test suggested by Bock (1972), which enables items that do not conform to model predictions to be identified. To conduct this test, subjects were grouped into four categories on the basis of their parameter θ, and the proportion of subjects endorsing the item for each group compared with the predicted proportion (Bock, 1972; Reise, 1990). Items showing a medium ($0.3 \leq \Phi < 0.5$) or large ($\Phi \geq 0.5$) effect size (Cohen, 1988) were considered as misfitting and not selected for inclusion in the new questionnaire.

2.3.1.2. Differential item functioning. Both uniform and non-uniform DIF were taken into account. Uniform DIF occurs when the probability of a particular response option is systematically greater for one group, whereas non-uniform DIF represents a non-systematic bias with a different impact on groups across trait levels. In order to detect DIF, a logistic regression model was computed, including trait level, group membership (uniform DIF), and the interaction between trait level and group membership (non-uniform DIF) as covariates (Magis, Beland, Tuerlinckx, & De Boeck, 2010; Swaminathan & Rogers, 1990). The statistical significance of the effects pertaining to group membership and group-score interaction was verified by Wald test. The effect size was calculated as ΔR^2 (the difference between the R^2 of nested models; Gomez-Benito, Dolores Hidalgo, & Padilla, 2009; Nagelkerke, 1991), and the results interpreted using the Jodoin and Gierl scale (Jodoin & Gierl, 2001). Items showing moderate (values higher than 0.035) to strong (values over 0.07) effect size were not selected (Magis et al., 2015).

2.3.1.3. Parameters a and b. Items that, according to the aforementioned criteria, exhibited DIF or misfit were excluded. The parameters a and b of the 2PL model were used to select those with the best psychometric properties from the remaining items. A key assumption in IRT is that in order to properly measure the different levels of latent trait that can be observed in the population, the questionnaire should include items that cover the entire *continuum* of the latent trait. The location of an item on the latent trait is expressed by its parameter b. For each of the four scales of the JEPQ-R, items with different values of parameter b were chosen in order to cover the entire continuum. In addition, items with the largest parameter a were chosen because they allow for good discrimination between subjects with