



Quality assurance for workplace health promotion. Validation of the quality criteria for WHP in the Austrian quality management system

Gert Lang¹ · Christoph Heigl² · Paulino Jiménez³

Received: 1 October 2018 / Accepted: 20 November 2018 / Published online: 8 December 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Aim Workplace health promotion (WHP) is being increasingly employed as a corporate strategy, being at its most effective when it is implemented in a high-quality and sustainable way. Based on normative principles, a quality management system for WHP was introduced in Austria in 2004. This article examines the measurement quality to validate this quality assurance system.

Subjects and methods From 2014 to 2017, WHP projects have been assessed rigorously by means of a standardized procedure using 15 quality criteria. The foundations for this are an application submitted by the company, and the decision as to whether a WHP quality certificate is awarded or not is based on the overall assessment. Data are available for the 1131 Austrian companies. Based on theoretical and methodological considerations, the measurement quality was investigated in relation to the overall quality of WHP projects with the help of structural equation models.

Results The confirmatory one-factor analysis revealed satisfactory construct validity ($\lambda \geq 0.40$) and high reliability for the overall scale ($\alpha = 0.87$), although the fit was not acceptable. The bi-factor analysis with a general factor (GF) and three specific residual factors resulted in good model fit. The GF explained most of the common variance (ECV = 63.9%); the overall scale was also characterized by its high reliability ($\alpha_{GF} = 0.90$, $\omega_{HGF} = 0.82$).

Conclusion The results justify the creation of an overall scale for assessing the quality of WHP. The mean varied by the year of submission, and there was a significant difference between both small/large enterprises and initial/renewal awards. The measurement tool can be considered a good screening instrument for awarding the WHP quality certificate.

Keywords Workplace health promotion · Quality assurance · Confirmatory factor analysis · Validity · Reliability

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10389-018-1005-6>) contains supplementary material, which is available to authorized users.

✉ Gert Lang
gert.lang@goeg.at; <http://www.fgoe.org>

Christoph Heigl
christoph.heigl@oogkk.at; <http://www.netzwerk-bgf.at>

Paulino Jiménez
paul.jimenez@uni-graz.at; <http://www.uni-graz.at>

¹ Austrian Health Promotion Fund, Austrian Public Health Institute, Vienna, Austria

² Upper Austrian Health Insurance Fund, Linz, Austria

³ Institute of Psychology, University of Graz, Graz, Austria

Abbreviations

λ	Factor loading
$(\sum \lambda)^2$	Sum of loadings squared
$(\sum \lambda^2)$	Sum of squared loadings
CBFA	Confirmatory bi-factor analysis
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CI	Confidence interval
df	Degrees of freedom
ECV	Explained common variance
ENWHP	European Network for Workplace Health Promotion
FGÖ	Fonds Gesundes Österreich
GF	General factor
IA	Initial award
ICC	Intra-class correlation
LE	Large enterprise
m	Arithmetic mean

MGCBFA	Multi-group confirmatory bi-factor analysis
N	Number of valid cases
ÖNBGF	Österreichisches Netzwerk Betriebliche Gesundheitsförderung
p	<i>p</i> value
QI	Quality indicator
RA	Renewal award
RF	Residual factor
RMSEA	Root mean standard error of approximation
s	Standard deviation
s^3	Skewness
s^4	Kurtosis
SE	Small enterprise
TLI	Tucker-Lewis index
WHP	Workplace health promotion
WLSMV	Weighted least square means and variance adjusted
α	Cronbach's alpha
χ^2	Chi-square
ω_H	Omega hierarchical
ω_S	Omega specific

Introduction

Workplace health promotion (WHP) has increased in both popularity and scope over the last 20 years. Although it principally emerged in medium-sized and large enterprises, in the meantime, smaller enterprises, too—despite having fewer resources and structures—are starting to appreciate the benefits of WHP (Pieper et al. 2015; Rosian-Schikuta et al. 2016) and are making use of this corporate strategy. The growing necessity for employers to increase their attractiveness for existing and future employees is also becoming a key incentive for introducing WHP (Lohaus and Rietz 2015; Winter and Grünwald 2016).

Effective and sustainable WHP does, however, require quality assurance as the processes, methods and measures applied in practice differ in relation to standards of quality. Kliche et al. (2004), for example, stated that a considerable number of programmes neglect central standards of quality such as documentation and/or the evaluation of processes and results. This, then, raises the question as to how the quality of future or current WHP projects can be assessed or, in other words, how high- and low-quality programmes can be distinguished from each other.

The quality of services can only be appraised when quality criteria have been specified (Jiménez et al. 2014). The assessment of quality often draws on the triad of structure, process and outcome proposed by Donabedian (1980)—occasionally supplemented by conceptual quality (Kliche et al. 2007). Structural quality relates to settings and parameters (e.g. qualifications, continuing professional development and in-service

training, methodological expertise, rooms and technical equipment) while procedural quality refers to the way in which components are delivered (e.g. process planning, diagnosis, goal setting, implementing measures). The quality of outcome describes the monitoring of success and satisfaction (e.g. the evaluation of efficacy and sustainability, knowledge transfer) while conceptual quality relates to theories, concepts and explanatory models that need to be taken into account (Donabedian 1980; Donabedian 2005 [1966]).

These quality criteria are deployed in various fields, for example, in health promotion and disease prevention (Ader et al. 2001). Here, quality criteria are commonly used to be able to assess specific characteristics in an objective way and can be employed consistently, economically and comparably as an inspection system.

As is also usual in other fields (e.g. organizational development), pre-existing norms in which assessment criteria are specified are taken as the basis for quality assessments. The result of the overall quality of a WHP project can be found in different outcomes. The idea of assessing this quality therefore was to include criteria that are covering the different aspects of WHP: In connection with WHP, the Luxembourg Declaration in the European Union (ENWHP 1997) takes on the main normative function. The fundamental principles formulated in it are taken as the accepted guidelines for basic standards of quality in WHP, including, in particular, (1) the integration of WHP in all areas of organizations, (2) the broad participation of employees in its conception and implementation, (3) a holistic concept of health and (4) systematic procedures that prevent the implementation of individual measures in isolation.

These principles form the foundations that safeguard the quality and effect of WHP, whether in the form of pilot projects or programmes that have already been sustainably integrated in corporate processes (workplace health management), and that, subsequently, allow WHP to be assessed (Jiménez et al. 2014). In this very spirit, various quality assurance systems have already taken root in German-speaking countries such as the Swiss seal of quality known as the “Friendly Work Space” (Gesundheitsförderung Schweiz 2018) or the “German Seal of Corporate Health” and the “German Corporate Health Award” (BKK Dachverband 2018a, b).

In conjunction with the Austrian Health Promotion Fund (Fonds Gesundes Österreich, FGÖ), the Austrian Network for Workplace Health Promotion (Österreichisches Netzwerk Betriebliche Gesundheitsförderung, ÖNBGF) already designed a quality assurance system in 2004 that was subsequently implemented nationwide and has been evolving ever since. Since 2013, the WHP quality certificate (BGF-Gütesiegel), which is awarded for a 3-year period, has been a standardized system of application coupled with a quantitative assessment procedure elaborating on the criteria laid down in the Luxembourg Declaration and applying a

catalogue of 15 different quality criteria (Heigl 2014). These quality criteria were drawn up with the aim of comprehensively reflecting the quality of WHP and are closely associated with the goal of maintaining the objectivity, transparency and stability of quality assurance whether WHP is being introduced for the first time or has been integrated sustainably in enterprises of any size.

As WHP quality can be seen as the fulfilment of all criteria in a holistic sense, the fundamental principles of the Luxembourg Declaration served as the frame for building the catalogue. As can be seen in the [Appendix](#), the quality indicators target these concepts, in particular the integration of WHP, participation, the holistic concept of health, and the distinctions among structure, process, outcome and concept (cf. Donabedian 1980; Kliche et al. 2007). Although WHP quality should be assessed by all possible aspects, the selected criteria should not be seen as an exhausting list but more as representing the universe of all possible aspects. Therefore, some quality indicators could be more overlapping than others.

Now that the quality assurance system has been in operation for several years, this article aims to explore the measurement quality of the WHP quality criteria as well as the application and assessment processes behind it. To do so, it is first necessary to specify the implicit theoretical and methodological assumptions that lie behind the conceptualization of quality in WHP. Based on this, hypotheses can be inferred and tested with the help of empirical data. Finally, this article will explain questions relating to potential implications for the continued development of this quality management system.

Study design and research methods

Theoretical and methodological foundations

To specify the concept behind the quality of WHP, it is necessary to set up theoretical and logical foundations as well as to define the methodological characteristics of an empirical examination of that concept of quality. Initially, it is important to establish which components make up the quality of WHP as a theoretical concept (in a disposing manner) and how the concept of quality contributes (functionally) to the rationale behind the theoretical construct of overall quality.

The concept can be specified on three levels (Weiber and Mühlhaus 2010): the primary goal on the object level is to assess the quality of WHP. We defined WHP quality as a latent factor that is being assessed in a holistic manner using a reflective measurement approach (Jarvis et al. 2003). The quality of WHP can be measured by ascertaining the extent to which the derived WHP quality criteria have been fulfilled. The attribute level, in contrast, defines which components should be used to reflect the quality of WHP. In accordance

with general considerations of quality (cf. Donabedian 1980), the components of WHP can be described in terms of their structure, process and outcome, each of which involves different aspects of quality. An attributive definition, thus, covers the overall quality of WHP, or a range of properties, which can be described in terms of quality criteria, while a functional definition sees the quality of WHP as a multi-attributitional construct of quality reflected by individual properties that can be defined by different quality indicators. Finally, on the subject level, it is necessary to clarify who is responsible for assessing the overall quality of WHP. In the Austrian Quality Management System, the application submitted by a company provides the basis for a quality certificate being awarded. To measure the quality of WHP, an inspection catalogue is used as a measurement operation.

Based on this theoretical specification of the concept, a latent construct is assumed, meaning that WHP—despite the theoretical concept consisting of several components and differing properties—is empirically unidimensional. The multi-attributive measurement instrument that this requires assumes that the overall quality of WHP is reflected by various quality indicators that can be used to create a quality scale.

The most common technique for determining dimensionality is factor analysis (DeVellis 2017). Not infrequently, however, such applications report that the response data do not reproduce the assumed unidimensional factor structure to a satisfactory extent. According to Reise et al. (2010), the problem occurs when the wrong default model is used. Particularly with multifaceted constructs and complex measurement instruments with heterogeneous manifestations, it has been pointed out in the literature that the (co-)variation of indicators can often not be ascribed to a general factor alone (Chen et al. 2012). In other words, the assumption of unidimensionality is too restrictive and can result in poor model fit.

A bi-factorial approach provides a possible solution and alternative modelling strategy (Chen et al. 2006: 190). Applications of bi-factor models for determining unidimensionality have become quite common in the meantime, helping to clarify the measurement structure of assessment scales. As Reise points out explicitly, “bi-factor modeling potentially provides a solid foundation for conceptualizing psychological constructs, constructing measures, and evaluating a measure’s psychometric properties” (Reise 2012: 667). The advantage of this type of analysis is that these models provide information on specific residual factors that go beyond the hypothetically assumed general factor (here: the overall quality of WHP). The analysis allows a comparison of the strength of the general quality factor in relation to the residual factors and, at the same time, the model estimates the factor loadings of all measurement indicators. In addition, the results of the analysis allow an estimation of whether the assumption of unidimensionality is met. If nothing else, the presence of a strong general factor with sufficiently high factor loadings justifies the

creation of an overall quality scale as is currently the case in the assessment process for awarding a WHP quality certificate.

Research questions and hypotheses

This article seeks to validate the measurement quality of the submission and assessment process of the Austrian Quality Management System using a theory-led approach. It investigates the applicability of the theory-driven WHP quality criteria and the measurement process used when describing and assessing the quality of WHP as implemented in Austrian enterprises. More specifically, it aims to answer empirically whether the quality criteria and indicators are suited as a valid and reliable measurement instrument to assess quality and justify the creation of a scale on the overall quality of WHP.

The investigation involves analysing the psychometric properties of the measurement instrument in several stages. As a first step, the data are described empirically; they comprise the individual assessments collected from a total of 1131 companies between 2014 and 2017. The quality construct of WHP is tested in the second stage using factor analysis; here several one- and bi-factor models are specified. The researched hypothesis verifies whether the 15 quality indicators can be subsumed empirically under one dimension of overall WHP quality, in other words whether satisfactory (construct) validity can be demonstrated for the individual indicators and reliability of the overall scale. In addition, the comparability of the overall quality is tested. In a third step, the properties of the overall scale for WHP quality are analysed descriptively, i.e. how the scale values of WHP quality differ in relation to small/large enterprises, year of submission and initial/renewal awards of the WHP quality certificate.

Submission and assessment process

On the basis of the Luxembourg Declaration (ENWHP 1997), seven WHP experts in the fields of health promotion and disease prevention from Austria and abroad reviewed the relevant literature and took part in several workshops before formulating 15 criteria with high content validity to evaluate the quality of WHP (including evaluation guidelines). These criteria were piloted in 178 Austrian companies in 2013 and then used regularly from 2014 onwards (cf. Heigl 2014).

Applications have to be submitted for the quality certificate to be awarded. Companies provide a comprehensive description of WHP projects or programmes that they have initiated by filling in a standardized application form. Applications for the quality certificate are collected at the coordination point and passed on to external experts who then examine them independently. The applications are used to evaluate the projects and programmes on the basis of the 15 criteria. To measure the quality of WHP, an inspection catalogue is used (as

the measurement operation) in which relevant questions and checklists are defined for every quality criterion and operationalized to ensure inter-reliability (ÖNBGF 2018, cf. Appendix).

In the assessment process, the inspection catalogue is used in a content-analytic procedure by experienced and specially trained experts from an independent, externally appointed institution. The individual criteria are evaluated using a uniform rating scale (0 = “quality criterion not fulfilled” – 3 = “quality criterion completely fulfilled”); the total number of points ranges between 0 and 45. The WHP quality certificate is not awarded when 0 to 19 points are obtained and is awarded automatically for 26 to 45 points. When 20 to 25 points are obtained, the application/company is dealt with in a meeting where the decision is made whether to award the certificate or not. For small enterprises (up to 49 employees), the points and thresholds are adjusted (Heigl 2014: 41).

The principle of objectivity states that the results of the measurement should not be affected by the individuals carrying them out. In unclear cases, the evaluation team makes use of double or triple verification. The quality of the 1131 overall assessments was compared with the 856 (75.7%) cases of optional self-assessment by the regional/service points of the Austrian Network for Workplace Health Promotion (ÖNBGF). Whether the coding was accurate or not (inter-rater reliability) was assessed with the help of a mixed, consistent, average intra-class correlation (ICC) coefficient (McGraw and Wong 1996). The calculated ICC = 0.682 (95% CI: 0.644–0.716) within the range of 0.60 to 0.74 (Cicchetti 1994) indicates that the assessments carried out by those involved in the process had a good degree of agreement with only a low measurement error and that the statistical power in particular was not reduced significantly for the subsequent analyses. The evaluations are therefore considered to be suitable for the hypothesis testing to follow.

Data basis

A total of $N = 1131$ applications submitted and evaluated between 2014 and 2017 are available to validate the submission and assessment process. Within this period, the annual number of applications increased steadily from 227 to 393.

The companies that applied represent a cross-section of enterprises in Austria in terms of the region they came from, the size of the enterprise and their WHP status: 469 applications (41.5%) came from enterprises in the densely populated eastern region of Austria, 466 (41.2%) from the equally densely populated western region and 196 (17.3%) from the relatively sparsely populated southern region of Austria. In terms of size, 531 applications (46.9%) were from companies with up to 49 employees (small enterprises, SE) and 600 applications (53.1%) were submitted by companies with 50 or more employees (large enterprises, LE); 659 companies (58.3%)

submitted applications to be awarded the WHP quality certificate for the first time (initial award, IA) and 472 enterprises (42.7%) did so to renew their WHP quality certificate for the first to the fourth time (renewal award, RA) (cf. Table 1).

Statistical data analysis

To answer the research questions and hypotheses, the data were analysed in several steps:

- (1) The variables were analysed descriptively in SPSS relating to the mean (m), standard deviation (s), skewness (s^3) and kurtosis (s^4). It was expected that the variables would have a significant dispersion around the mean (i.e. that they would differentiate) and that the distribution of the indicators would not be skewed ($s^3 < |2.0|$) and would be unimodal ($s^4 = |7.0|$) (Byrne 2012), in other words that there would be no floor or ceiling effects (Bortz and Döring 2016).
- (2) The measurement structure or quality was investigated using factor analysis as a means of examining the structure in Mplus 7 (Muthén and Muthén 2012). Baseline model 1 comprised a confirmatory one-factor analysis (CFA). This tests whether the theoretical preliminary considerations and the hypotheses derived from them apply to the factor structure. As the assumption of unidimensionality is too restrictive in some applications, this was followed by a confirmatory bi-factor analysis (CBFA) as an additional model 2—after a necessary exploratory intermediate stage (cf. Reise 2012). In addition to an estimation of the factor loadings on the general

quality factor (GF), it provides further information on residual factors (RF). This model also makes it possible to judge the acceptability of the assumption of unidimensionality. In addition, in a multiple group confirmatory bi-factor analysis (MGCBFA), the measurement structure for small and large enterprises as well as for initial and renewal awards was examined in a third model.

The parameter estimation of the models made use of the WLSMV method (weighted least squares means and variance adjusted). According to current simulation studies, using the WLSMV estimation method means that the factor loadings and standard errors for indicators with fewer than five categories can be estimated with less bias and more precision, especially with a larger number of cases and with a normally distributed latent dimension (Cheng-Hsien 2016; Rhemtulla et al. 2012).

To evaluate the model, standardized effect sizes (factor loadings $\lambda > 0.30$) were accepted that make a significant contribution to the factor (Bowling 2017) and the modification indices were inspected (Byrne 2012), e.g. meaningful measurement error covariances were taken into account. The level of significance was set at $p < 0.05$. Good model fit is shown by the following thresholds for the fit indices: χ^2 test¹ ($\chi^2/\text{degrees of freedom, } df < 3.0$), the root mean standard error of approximation (RMSEA < 0.08), the comparative fit index and the Tucker-Lewis index (CFI/TLI > 0.90) (Browne and Cudeck 1993; Hu and Bentler 1999; Kline 2011). To identify the content-related accuracy of the quality criteria, the reliability of the indicators was ascertained using Cronbach's alpha ($\alpha > 0.70$) (Nunnally and Bernstein 1994). Ascertaining the reliability of the bi-factor model usually makes use of the coefficient omega hierarchical ($\omega_H > 0.75$) in relation to the general factor saturation (Reise 2012) and omega specific (ω_S) for the saturation of the residual factors (McDonald 1999; Zinbarg et al. 2005). When checking the validity of bi-factor models, unidimensionality of the general factor is supported when it explains most of the variance of the indicators (explained common variance, e.g. ECV > 0.70) and, consequently, the explained variance is lower for the specified residual factors.

- (3) In the final stage of analysis, the properties of the overall quality scale are described in relation to the mean (95% confidence interval, CI) and standard deviation as well as mean differences ($p < 0.05$) according to groups (small vs. large enterprises, initial vs. renewal awards, year of submission).

Table 1 Description of cases ($N = 1131$)

Variable	Category	<i>N</i>	%
Total number		1131	100.0
Year	2014	227	20.1
	2015	249	22.0
	2016	262	23.2
	2017	393	34.7
Region (federal states)	Eastern Austria (B, NÖ, W)	469	41.5
	Western Austria (OÖ, S, T, V)	466	41.2
	Southern Austria (K, ST)	196	17.3
Size of enterprise	Small enterprise ^a	531	46.9
	Large enterprise ^b	600	53.1
Submissions	Initial award	659	58.3
	Renewal awards ^c	472	42.7

B Burgenland, *NÖ* Lower Austria, *W* Vienna, *OÖ* Upper Austria, *S* Salzburg, *T* Tyrol, *V* Vorarlberg, *K* Carinthia, *ST* Styria

^a Up to 49 employees

^b 50 employees or more

^c Includes submissions for the 1st, 2nd, 3rd or 4th renewal award

¹ A non-significant χ^2 test is preferable; however, the test tends to incorrectly reject models based on a large sample (Bollen 1989).

Results

Descriptive analysis

The scores for the 15 quality indicators are given in Table 2 below. The category used most frequently was “quality criterion almost fulfilled” (value 2) while the categories of “completely fulfilled” (3) or “not really fulfilled” (1) were used much less frequently and the category “not fulfilled” (0) was hardly used at all. The means ranged from $m = 1.96$ – 2.58 with standard deviations of $s = 0.52$ – 0.83 . This reveals that some quality indicators differentiated better [e.g. target group orientation, structure (of the project), individual-directed measures) than others (e.g. corporate principles/culture, responsibilities/contact person, general evaluation]. Despite the low number of measurement categories, the indicators exhibited unimodal distribution ($s^4 = -1.33$ – 0.01) and were only mildly skewed ($s^3 = -0.88$ – 0.00).

Construct validity and reliability

In the baseline model 1a, a confirmatory one-factor analysis (CFA) was run after combining categories 1 and 0 because of the small number of cases in the latter, and one factor loading was fixed to 1 to identify the model. While there were satisfactory coefficients for the support of validity of the individual indicators (factor loadings $\lambda \geq 0.40$ | $p < 0.001$) and the reliability of the overall scale (Cronbach's $\alpha = 0.87$), but no satisfactory model fit was achieved ($\chi^2/df = 12.2$ | RMSEA =

0.10 | CFI = 0.85 | TLI = 0.82). The modification indices indicated a series of measurement error covariances between individual indicators that needed to be specified. For good model fit, a total of five such modifications would have to be made in model 1b ($\chi^2/df = 4.7$ | RMSEA = 0.06 | CFI = 0.95 | TLI = 0.94), which suggests a more complex factor structure.

The exploratory bi-factor analysis (model 2a) carried out as an interim step pointed to three interpretable residual factors (RF) alongside the general quality factor (GF). Model 2a had excellent fit ($\chi^2/df = 2.0$ | RMSEA = 0.03 | CFI/TLI = 0.99).

When specifying the confirmatory bi-factor model 2b based on this information, one factor loading fixed to 1 was taken into account on the general factor (GF) and on the three residual factors (RF1-3). In addition, the analysis took account of covariances with the residual factors by the general factor and was therefore assumed to be orthogonal (Chen et al. 2012; Reise 2012). The confirmatory bi-factor model as specified indicated a good fit for the model in relation to the empirical data ($\chi^2/df = 5.1$ | RMSEA = 0.06 | CFI = 0.95 | TLI = 0.94). With its acceptable factor loadings ($\lambda_{GF} \geq 0.32$ | $p < 0.001$), the general factor (GF) did not only explain most of the common covariances among the indicators (ECV_{GF} = 63.9%) but was also characterized by the very high reliability of the overall scale ($\alpha_{GF} = 0.90$ | $\omega_{H_{GF}} = 0.82$).

The factor loadings for the residual factors were mostly acceptable ($\lambda_{RF} = 0.24$ – 0.70) and additionally explained between 9.0 and 13.4% of the variance, as the general factor did not do so entirely. In comparison with the general factor, the residual factors were less reliable, particularly in relation to

Table 2 Descriptive statistics for the quality indicators ($N = 1131$)

QI	Description	Score (in %)				m	s	s ³	s ⁴
		0	1	2	3				
1	Corporate principles/culture	0.0	1.3	56.8	41.7	2.40	0.52	0.10	-1.33
2	Structure (of the project)	0.0	7.7	49.9	42.3	2.35	0.62	-0.39	-0.67
3	Responsibilities/contact persons	0.0	3.2	35.5	61.2	2.58	0.55	-0.88	-0.26
4	Target group orientation	4.2	24.0	43.4	28.2	1.96	0.83	-0.36	-0.57
5	Diagnostic phase/tools/needs assessment	0.0	4.7	37.9	57.3	2.53	0.59	-0.81	-0.34
6	Employee orientation	0.0	11.6	49.5	38.7	2.27	0.66	-0.35	-0.75
7	Communication	0.0	5.6	47.9	46.2	2.41	0.60	-0.44	-0.63
8	Environment-directed measures	0.0	6.2	38.6	55.1	2.49	0.61	-0.78	-0.39
9	Individual-directed measures	0.0	19.1	46.6	34.2	2.15	0.72	-0.23	-1.03
10	Leadership	0.0	13.5	53.3	33.0	2.20	0.65	-0.23	-0.73
11	Quality of the formulation/scope of the goals	0.0	15.4	66.5	17.8	2.02	0.58	0.00	0.01
12	Monitoring of results and evaluation	0.0	4.2	37.8	57.8	2.54	0.58	-0.81	-0.35
13	Attainment of goals	0.1	4.0	53.6	42.2	2.38	0.57	-0.26	-0.60
14	Sustainability	0.1	3.6	51.8	44.3	2.41	0.57	-0.30	-0.63
15	General evaluation	0.0	4.9	62.6	32.3	2.27	0.55	0.05	-0.48

Range: 0 = quality criterion not fulfilled to 3 = quality criterion completely fulfilled

QI quality indicator, m mean, s standard deviation, s³ skewness, s⁴ kurtosis

the preferred coefficient omega specific ($\alpha_{RF} = 0.38\text{--}0.80$ | $\omega_{S_{RF}} = 0.33\text{--}0.60$).

Based on the residual factors, relevant sets of indicators could be identified. The first residual factor combined structural aspects of the WHP project (quality indicator i2), corporate principles or corporate culture (i1) and responsibility for WHP or the contact person (i3). The second residual factor included the diagnostic phase carried out in the needs assessment as well as diagnostic tools used (i5), the monitoring of results and evaluation carried out at the end of the project (i12) and the employee orientation which that implies (i6). The third residual factor clustered around the extent and quality of the goals that were set (i11) and those that were achieved (i13) in the WHP project (Table 3).

To ascertain whether the bi-factorial structure also held for relevant groups (configural invariance), a multiple group comparison was employed. The fit of group model 3a comparing small enterprises (SE) and large enterprises (LE) was just as acceptable ($\chi^2/df = 3.4$ | RMSEA = 0.07 | CFI = 0.92 | TLI = 0.92) as that of group model 3b comparing initial awards (IE) and renewal awards (RA) ($\chi^2/df = 3.4$ | RMSEA = 0.07 | CFI = 0.93 | TLI = 0.92) whereby the higher χ^2 values suggest

that the groups of small enterprises ($\chi^2_{SE} = 331.60$ vs. $\chi^2_{LE} = 324.53$) and initial awards ($\chi^2_{IA} = 343.16$ vs. $\chi^2_{RA} = 315.70$) had a slightly worse fit.

The reliability of the group models varied. It was best for the groups of renewal awards and large enterprises ($\alpha = 0.92/0.91$ | $\omega H = 0.87/0.84$ respectively), good for initial awards and small enterprises ($\alpha = 0.88/0.88$ | $\omega H = 0.78/0.79$ respectively) and much worse, in contrast, for the residual factors. The explained variance for the general factor was best for the group of renewal awards (ECV = 71.8%) and worst for the initial awards (ECV = 59.3%). The explained variance for the residual factors was never higher than 14.8%. The differences in the standardized factor loadings are given in Fig. 1. Table 4 provides an overview comparing the goodness of fit of the various models.

Scale properties

The unweighted additive overall scale had a mean of $m = 35.0$ (95% CI: 34.7–35.2), a significant standard deviation ($s = 4.8$) and minimal skewness ($s^3 = -0.13$) for the overall data. The distribution did not show any floor or ceiling effects because

Table 3 Standardized factor loadings from the one- and bi-factor analyses ($N = 1131$)

QI	Description	One factor (CFA)	Bi-factor (CBFA)			
			GF	RF1	RF2	RF3
1	Corporate principles/culture	0.56	0.54	0.33		
2	Structure (of the project)	0.49	0.45	0.66		
3	Responsibilities/contact persons	0.40	0.37	0.33		
4	Target group orientation	0.56	0.58			
5	Diagnostic phase/tools/needs assessment	0.55	0.46		0.65	
6	Employee orientation	0.62	0.60		0.24	
7	Communication	0.59	0.60			
8	Environment-directed measures	0.55	0.57			
9	Individual-directed measures	0.58	0.60			
10	Leadership	0.43	0.44			
11	Quality of the formulation/scope of the goals	0.45	0.39			0.70
12	Monitoring of results and evaluation	0.51	0.40		0.65	
13	Attainment of goals	0.39	0.32			0.70
14	Sustainability	0.72	0.73			
15	General evaluation	0.97	0.99			
	Sum of squared loadings ($\sum \lambda^2$)	4.96	4.67	0.65	0.91	0.97
	Sum of loadings squared ($\sum \lambda$) ²	70.04	64.21	1.75	2.38	1.93
	Cronbach's alpha (α)	0.87	0.90	0.55	0.38	0.80
	Omega hierarchical (ωH) or omega specific (ωS)		0.82	0.33	0.41	0.60
	Explained common variance (ECV)		63.9%	9.0%	12.6%	13.4%

All factor loadings are significant ($p < 0.001$)

QI Quality indicator (i1-15), GF general factor, RF residual factor

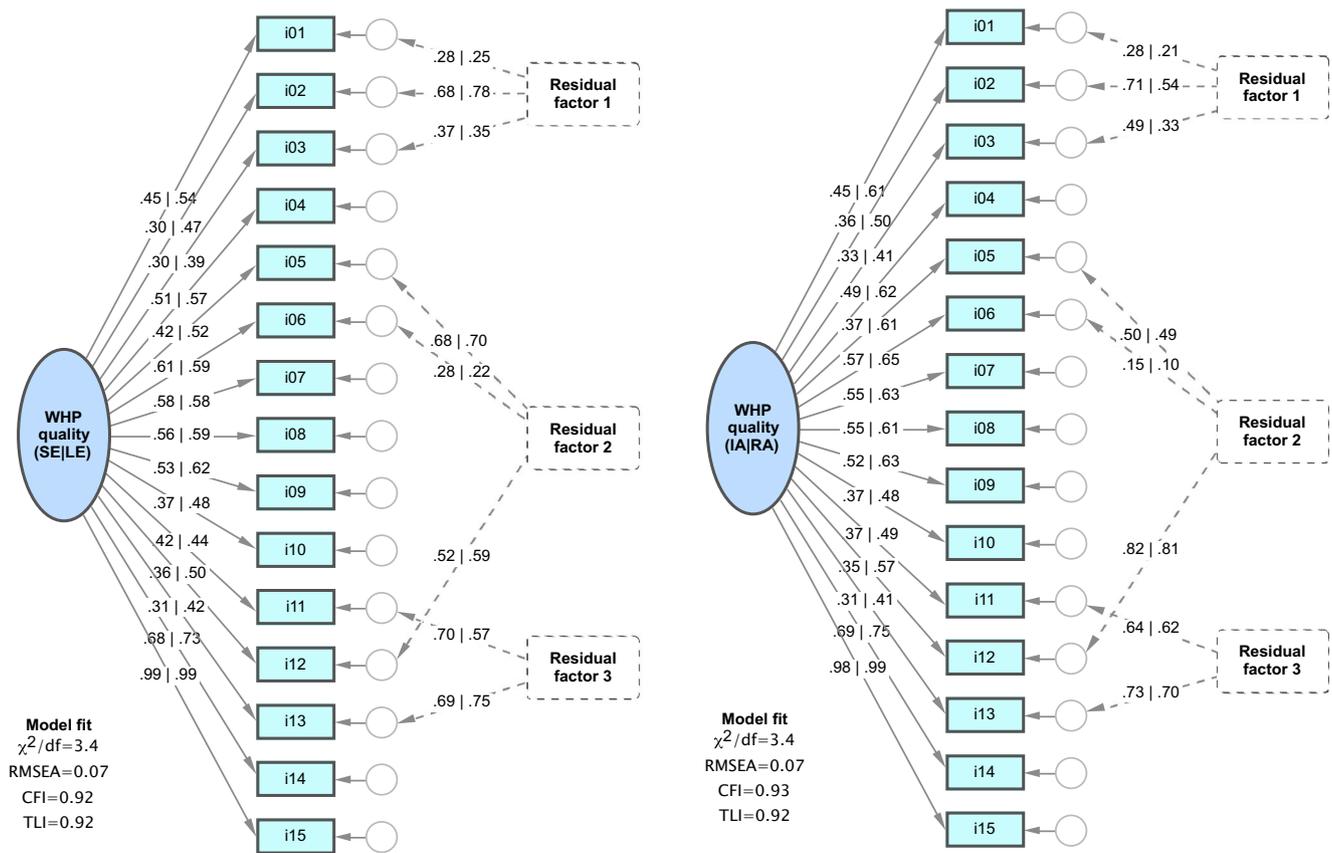


Fig. 1 Results of the multiple group comparison (MGCBFA, $N = 1131$) between small and large enterprises (left) and initial and renewal awards (right). SE = small enterprise, LE = large enterprise, IA = initial award,

RA = renewal award; χ^2 = chi-square test statistic, df = degrees of freedom, RMSEA = root mean standard error of approximation, CFI = comparative fit index, TLI = Tucker-Lewis index

only 0.3% had the minimum scale value of 21 and 1.2% the maximum scale value of 45 (Table 5).

The mean varied minimally but significantly ($p < 0.05$) depending on the year of submission. The bimodal distribution

(cf. Fig. 2) was due to the significant differences in mean between small and large enterprises ($m_{SE} = 34.4$ and $m_{LE} = 36.3$) as well as between initial and renewal awards ($m_{IA} = 34.0$ and $m_{RA} = 36.3$). In addition, the standard deviation was

Table 4 Comparison of models ($N = 1131$)

Model	χ^2	df	χ^2/df	RMSEA	90% CI		CFI	TLI
					Low	High		
One factor								
1a. CFA: one-factor model	1102.90	90	12.3	0.10	0.10	0.11	0.85	0.82
1b. CFA: incl. 5 measurement error covariances ^a	400.55	85	4.7	0.06	0.05	0.06	0.95	0.94
Bi-factor								
2a. EBFA: 1 GF + 3 RF	100.34	51	2.0	0.03	0.02	0.04	0.99	0.99
2b. CBFA: 1 GF + 3 RF	436.45	85	5.1	0.06	0.06	0.07	0.95	0.93
Bi-factor in group comparison								
3a. MGCBFA: SE and LE	656.13	194	3.4	0.07	0.06	0.07	0.92	0.92
3b. MGCBFA: IA and RA	658.86	194	3.4	0.07	0.06	0.07	0.93	0.92

GF general factor, RF residual factor, SE small enterprise, LE large enterprise, IA initial award, RA renewal award, CFA confirmatory factor analysis, EBFA exploratory bi-factor analysis, CBFA confirmatory bi-factor analysis, MGCBFA multi-group CBFA; χ^2 chi-square test statistic, df degrees of freedom, RMSEA root mean standard error of approximation, CI confidence interval, CFI comparative fit index, TLI Tucker-Lewis index

^a Includes five specific measurement error covariances (indicators: i11-i13, i5-i12, i2-i3, i5-i6, i12-i13)

Table 5 Comparison of the means of the overall WHP quality scale ($N = 1131$)

Variable	Category	N	m	95% CI		s	p <
				Lower bound	Upper bound		
Scale value	Total	1131	35.0	34.7	35.2	4.80	–
Year	2014	227	35.0	34.3	35.7	5.35	
	2015	249	34.3	33.7	34.9	4.60	
	2016	262	34.8	34.2	35.3	4.59	
	2017	393	35.5	35.0	35.2	4.69	0.018
Size of enterprise	Small enterprise ^a	531	33.4	33.1	33.8	4.27	
	Large enterprise ^b	600	36.3	35.9	36.7	4.85	0.001
Submissions	Initial award	659	34.0	33.7	34.3	4.31	
	Renewal award	472	36.3	35.8	36.8	5.13	0.001

m mean, *s* standard deviation, *CI* 95% confidence interval with its lower and upper bound

^a Up to 49 employees

^b 50 employees or more

higher for large enterprises and renewal awards than for small enterprises and initial awards.

Discussion

Although the one-factor model had to be rejected because of its poor model fit, the unidimensionality of the WHP quality construct and its indicators can be assumed. The results of the bi-factor model justify the creation of an overall scale to assess

the quality of WHP and thus confirm the first underlying hypothesis not only because all 15 quality indicators had a relevant contribution to make and therefore explained most of the total variance of the scale ($ECV_{GF} = 63.9\%$) but also because the overall scale was characterized by the high reliability of the measurement ($\omega_{H_{GF}} = 0.82$). For the intended screening purposes, the indicators measured the quality dimension of WHP very precisely and reliably.

The specific residual factors point to additional but non-independent individual dimensions ($ECV_{RF} = 9.0\text{--}13.4\%$ |

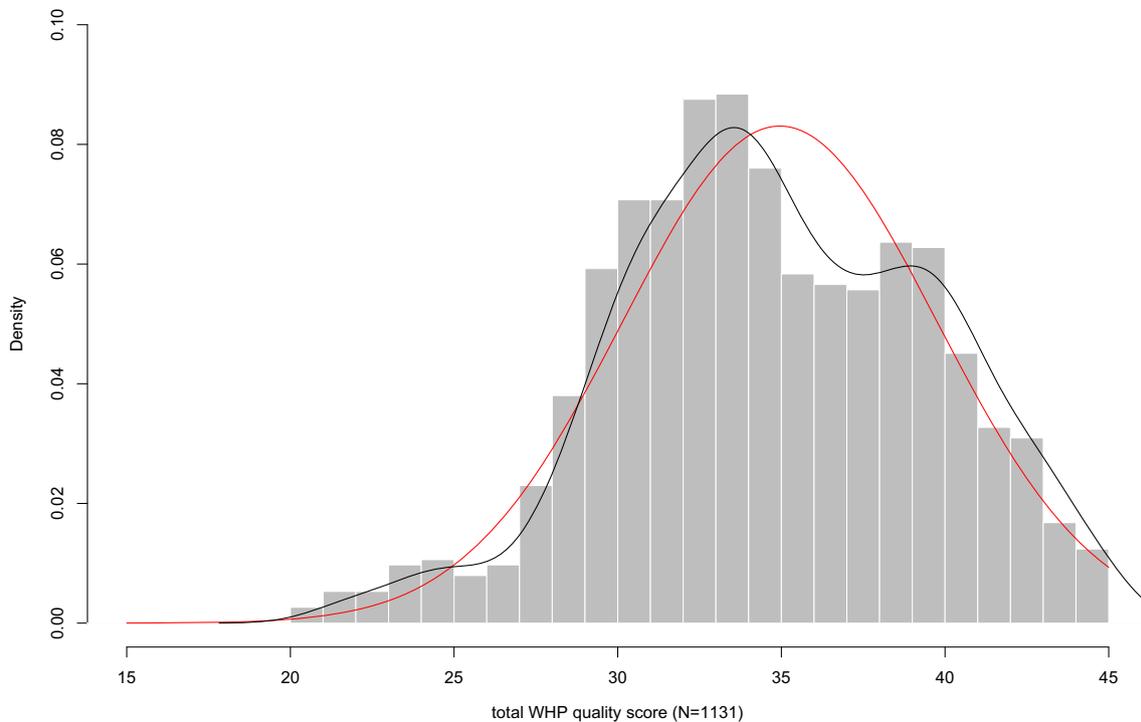


Fig. 2 Histogram of the WHP quality scale ($N = 1131$)

$\omega_{\text{RF}} = 0.33\text{--}0.60$) and a lack of discriminatory power for individual indicators. This is accounted for by the overlap between and difficulty of isolating individual aspects of the quality of WHP in relation to their structure, process and outcome on the one hand and by the methodological challenges of the complex data collection and evaluation process on the other. For example, the residual factor consisting of the indicators “corporate principles”, “structure of the project” and “responsibilities” indicates several particularities in relation to environment-directed aspects in the structure of WHP. “Diagnosis and needs assessment”, “employee orientation” and “monitoring of results and evaluation” indicate key processes in WHP, while the “quality of the formulation of goals” and “attainment of goals” relate to the effects and results of WHP, which are closely related to each other.

The quality of WHP can be represented as a general factor independently of its individual dimensions. Despite the dominance of the general factor, a few individual item variances that were not explained were substantial and have to be taken into account to represent the quality dimension of WHP in an adequate manner. The explicit consideration of error variances in the bi-factor model not only confirms that the quality of WHP is an independent dimension but also that several residual factors lie behind it. As the multiple group comparison revealed, this finding applied to differing extents when assessing the quality of both different-sized enterprises and those running WHP projects in different stages, from the pilot phase to a sustainable programme.

The strength of the scale is to be found in the multifaceted and multidimensional quality construct, which allows a comprehensive estimation of quality and, simultaneously, represents one latent variable. Despite the generally known limitations of self-assessment tools (see below for more details), the WHP quality scale and its 15 quality indicators provide a good foundation for determining the overall quality of and differences in quality in WHP projects. The values achieved on the scale are easy to interpret overall because the quality indicators represent relatively good measurements of the general factor of WHP quality and because they generally reflect one single dimension. In addition, the scale can differentiate between different levels of quality in WHP. The significant differences in the mean between small and large enterprises as well as between initial and renewal awards point to this characteristic of the scale. Thus, the overall scale is suitable for evaluating applications to decide whether a company can be awarded a quality certificate or not on the basis of a WHP project.

Limitations

In the analysis at hand, construct validity and internal consistency (reliability) could be confirmed. Future analyses should

furnish proof of criterion validity, for example, by identifying a match between WHP quality levels and employee satisfaction (concurrent validity) or the success of a company (predictive validity). To provide further proof of the reliability of the quality measurement, research designs with repeated measures are needed (test-retest reliability) (cf. Evers et al. 2013: 58 f.).

It is worth underlining again that the available data relate to companies that submitted an application to be awarded the WHP quality certificate and are thus not a representative cross section of all enterprises. Due to the selective nature of the data, it is not possible to generalize about the quality of WHP in all Austrian companies. Instead, the data represent so-called prototypical enterprises (Diamantopoulos 2005: 3) or examples of good WHP practice, as is apparent from the higher average scale points at $m = 35.0$ (95% CI: 34.7–35.2) and as is justified by the “WHP quality certificate” being awarded. Even if not included in the empirical data at hand, in principle the scale is also capable of capturing medium and poor quality WHP.

In a similar vein it needs to be noted that some companies possibly did undergo the process of evaluating their projects as the criteria are known and by self-evaluating they could have decided not to take part. This could be seen as a drawback for the award procedure but also as an opportunity to raise the quality. At this time there are no empirical data but this should be seen as an important research question for the future.

The 15 quality criteria or indicators do not need to be revised radically. All indicators contribute significantly towards explaining the general factor. Due to the above-mentioned limitation concerning the instrument itself, it should not be forgotten that such a complex measurement procedure cannot happen independently of the applicants or assessors. Despite the degree of reliability achieved, it is likely that the measurement includes both random and systematic imprecision. The literature frequently mentions causes such as potential distortions due to inconclusive indicators, the context of the measurement and the desire to achieve consistency and social desirability, to mention but a few (Schnell et al. 2013).

To minimize sources of measurement error, therefore, some measures already employed in the Austrian quality management system should be improved or the introduction of new measures contributing additionally to an improvement in construct validity should be discussed (e.g. anonymization/blind applications, a higher discriminatory power in their operationalization, on-going and in-depth training for assessors, reflection on the evaluation process). Another basis for optimizing the submission and assessment process concerns the residual factors that were identified as a source of common error variance. Here the recommendation is to consider the different levels (application form, enterprise as applicant, assessor) and aspects (clarity, economics, usefulness) (Jiménez et al. 2014).

One possible source of bias, for example, is the use of the same method and identical measurement tools (common-

method bias) (Podsakoff et al. 2003; Spector 2006; Van de Vijver 2013). Therefore, the assessors could be the source of systematic error covariances, causing correlations between the measurements. In this respect, it is worth mentioning a methodological review of the quality management system that is taking place at present in which the advantages and disadvantages of different techniques used to collect the data are being compared systematically (written application, telephone interview, participant observation/audit) to provide additional insights.

Conclusions

This article provides new information and insightful perspectives on the measurement quality of quality criteria to assess overall WHP. The quality of comprehensive WHP in line with the guiding principles of the Luxembourg Declaration can be described as a multi-attribitional construct. This is measured on the basis of how general areas of quality can be taken into account in relation to structure, process and outcome.

The analysis at hand confirms the construct validity and internal consistency (reliability) of the measurement. The existing 15 quality criteria provide a transparent and understandable basis for differentiating among low, medium and high quality WHPs. The measurement instrument is an eminently suitable tool for screening purposes when awarding the WHP quality certificate. The validation relates to the entire submission and assessment process because the evaluation can only be as good as the content-related criteria, submission/application procedure and associated processes.

The results also show that quality assurance can be achieved by awarding a quality certificate: As Heigl (2014) outlined, the WHP quality certificates are only awarded once a certain number of points have been assigned. In case of renewal awards, more points are achieved. This also implies that companies that resubmit strive to match the previous results or would like to improve their quality still further.

The results of this validation process reinforce efforts relating to quality assurance in WHP with the goal of creating a uniform system oriented towards international standards and of providing a quantifiable basis as a prerequisite for a high-quality and sustainable promotion of health in a company setting. The quality criteria are not only useful for assessors as an objective basis for deciding whether enterprises should be awarded a WHP quality certificate or not. Even before that stage, the quality criteria provide useful guidance for decision-makers within the company (e.g. project managers, prevention specialists, works council members) as well as external experts (e.g. consultants, evaluators) when planning, implementing and reviewing a WHP project or programme, not forgetting the employees in the enterprises who themselves benefit from such standards of quality.

Acknowledgements We are grateful for the feedback from Anita Bregenzer, which helped us improve the manuscript.

Compliance with ethical standards

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individuals included in the study.

Conflict of interest The authors declare that they have no conflict of interest.

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

References

- Ader M, Berensson K, Carlsson P, Granath M, Urwitz V (2001) Quality indicators for health promotion programs. *Health Promot Int* 16: 187–195. <https://doi.org/10.1093/heapro/16.2.187>
- BKK Dachverband (2018a) Deutscher Unternehmenspreis Gesundheit. www.deutscher-unternehmenspreis-gesundheit.de/. Accessed 17 July 2018
- BKK Dachverband (2018b) Deutsches Siegel Unternehmensgesundheit. <https://www.bkk-dachverband.de/gesundheitsfoerderung-selbsthilfe/betriebliche-gesundheitsfoerderung-bgf/deutsches-siegel-unternehmensgesundheit/>. Accessed 15 Sept 2018
- Bollen KA (1989) *Structural equations with latent variables*. Wiley, New York
- Bortz J, Döring N (2016) *Forschungsmethoden und evaluation für human- und sozialwissenschaftler*, 5th edn. Springer, Berlin. <https://doi.org/10.1007/978-3-642-41089-5>
- Bowling A (2017) *Measuring health: a review of subjective health and quality of life measurement scales*, 4th edn. McGraw Hill, London
- Browne MW, Cudeck R (1993) Alternative ways of assessing model fit. In: Bollen KA, Long JS (eds) *Testing structural equation models*. Sage, Newbury Park, pp 136–162
- Byrne BM (2012) *Structural equation modeling with Mplus. Basic concepts, applications, and programming*. Routledge, New York
- Chen FF, West SG, Sousa KH (2006) A comparison of bifactor and second-order models of quality of life. *Multivar Behav Res* 41: 189–225. https://doi.org/10.1207/s15327906mbr4102_5
- Chen FF, Hayes A, Carver CS, Laurenceau J-P, Zhang Z (2012) Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. *J Pers* 80:219–251. <https://doi.org/10.1111/j.1467-6494.2011.00739.x>
- Cheng-Hsien L (2016) Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res Methods* 48:936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6:284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- DeVellis RF (2017) *Scale development: theory and applications*, 4th edn. Sage, Thousand Oaks
- Diamantopoulos A (2005) The C-OAR-SE procedure for scale development in marketing: a comment. *Int J Res Mark* 22:1–9. <https://doi.org/10.1016/j.ijresmar.2003.08.002>
- Donabedian A (1980) *Explorations in quality assessment and monitoring*. Health Administration Press, Ann Arbor

- Donabedian A (2005 [1966]) Evaluations in quality of medical care. *Milbank Q* 83:691–729. <https://doi.org/10.1111/j.1468-0009.2005.00397.x>
- ENWHP (1997) Luxembourg declaration on workplace health promotion in the European union. European Network for Workplace Health Promotion
- Evers A, Hagemeister C, Høstmølingen A, Lindley P, Muniz J, Sjöberg A (2013) EFPA review model for the description and evaluation of psychological and educational tests: test review form and notes for reviewers version 4.2.6. European Federation of Psychologists Associations (EFPA). <http://www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4>. Accessed 28 Sept 2018
- Gesundheitsförderung Schweiz (2018) Friendly work space. www.friendlyworkspace.ch. Accessed 17 July 2018
- Heigl C (2014) Qualitätssicherung als laufende Entwicklung am Beispiel der Betrieblichen Gesundheitsförderung des Österreichischen Netzwerks für BGF (ÖNBGF). *Wirtschaftspsy* 16:35–43
- Hu L-t, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis. *Struct Equ Model* 6:1–55. <https://doi.org/10.1080/10705519909540118>
- Jarvis CB, MacKenzie SB, Podsakoff PM (2003) A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *J Consum Res* 30:199–218. <https://doi.org/10.1086/376806>
- Jiménez P, Eibel K, Dunkl A, Ratswohl A (2014) Entwicklung von Qualitätskriterien für AnbieterInnen von Dienstleistungen im Bereich der Gesundheitsförderung. *Wirtschaftspsy* 2:27–34
- Kliche T, Töppich J, Kawski S, Koch U, Lehmann H (2004) Die Beurteilung der Struktur-, Konzept- und Prozessqualität von Prävention und Gesundheitsförderung. Anforderungen und Lösungen. *Bundesgesundheitsbla* 47:125–132. <https://doi.org/10.1007/s00103-003-0771-6>
- Kliche T, Töppich J, Kawski S, Brunecker L, Önel J, Ullrich A, Koch U (2007) Professional expectations about quality assurance: a review-based taxonomy of usability criteria in prevention, health promotion and education. *J Public Health* 15:11–19. <https://doi.org/10.1007/s10389-006-0072-2>
- Kline RB (2011) Principles and practice of structural equation modeling, 3rd edn. Guilford Press, New York
- Lohaus D, Rietz C (2015) Arbeitgeberattraktivität. Der Stellenwert von Bekanntheit und Labels in der frühen Rekrutierungsphase. *Z Arb Organ* 59:70–84. <https://doi.org/10.1026/0932-4089/a000175>
- McDonald RP (1999) Test theory: a unified treatment. Erlbaum, Mahwah
- McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1:30–46
- Muthén LK, Muthén BO (2012) Mplus. Statistical analysis with latent variables. User's guide, 7th edn. Muthén & Muthén, Los Angeles
- Nunnally JC, Bernstein IH (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York
- ÖNBGF (2018) Die 15 Qualitätskriterien. <http://www.netzwerk-bgf.at/>. Accessed 17 July 2018
- Pieper C, Schöer S, Haupt J, Kramer I (2015) Wirksamkeit und Nutzen betrieblicher Gesundheitsförderung und Prävention - Zusammenstellung der wissenschaftlichen Evidenz 2006 bis 2012 (iga.Report 28). In: IGA (ed) Wirksamkeit und Nutzen betrieblicher Prävention, vol 28. Initiative Gesundheit & Arbeit, Berlin, pp 11–110
- Podsakoff PM, MacKenzie SB, Lee J-Y, Podsakoff NP (2003) Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J Appl Psychol* 88:879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Reise SP (2012) The rediscovery of bifactor measurement models. *Multivar Behav Res* 47:667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise SP, Moore TM, Haviland MG (2010) Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess* 92:544–559. <https://doi.org/10.1080/00223891.2010.496477>
- Rhemtulla M, Bruosseu-Liard PÉ, Savalei V (2012) When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol Methods* 17:354–373. <https://doi.org/10.1037/a0029315>
- Rosian-Schikuta I, Horvath I, Juraszovich B, Renner A-T, Langmann H, Atzler B (2016) Institutionelle Betriebliche Gesundheitsförderung und Prävention. Internationale Evidenz. Bestandsaufnahme Österreich. Perspektiven (Wissenschaftlicher Ergebnisbericht). Gesundheit Österreich, Wien
- Schnell R, Hill PB, Esser E (2013) Methoden der empirischen Sozialforschung, 10th edn. Oldenbourg Verlag, München
- Spector PE (2006) Method variance in organisational research: truth or urban legend? *Organ Res Methods* 9:221–232. <https://doi.org/10.1177/1094428105284955>
- Van de Vijver FJR (2013) Method bias. In: Keith K (ed) The encyclopedia of cross-cultural psychology. Wiley, Hoboken, pp 878–880. <https://doi.org/10.1002/9781118339893>
- Weiber R, Mühlhaus D (2010) Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS. Springer, Berlin
- Winter W, Grünewald C (2016) BGM als Stellschraube von Arbeitgeberattraktivität. In: Badura B, Ducki A, Schröder H, Klose J, Meyer M (eds) Fehlzeiten-Report 2016. Unternehmenskultur und Gesundheit-Herausforderungen und Chancen. Springer, Berlin, pp 225–235. <https://doi.org/10.1007/978-3-662-49413-4>
- Zinbarg RE, Revelle W, Yovel I, Li W (2005) Cronbach's α , Revelle's β , and McDonald's ω H: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* 70:123–133. <https://doi.org/10.1007/s11336-003-0974-7>