Revista Psicologia: Organizações & Trabalho (rPOT)
Psychology: Organizations and Work Journal
Revista Psicología: Organizaciones y Trabajo
ISSN 1984-6657 - https://doi.org/10.17652/rpot/2020.2.18100

Authentic Leadership: Development and Initial Validation of a Situational Judgment Test

Maria Isabel de Campos[1,*], Fabián Javier Marín Rueda[1,2]

[1] Universidade São Francisco (USF), Brasil
[2] Centro Universitário de Brasília (UniCEUB), Brasil

Submission: 22/04/2019
First Editorial Decision: 06/02/2020
Final Version: 05/03/2020
Accepted: 05/03/2020

Abstract

This study describes the development and the search for initial validity evidence (content and internal structure) of the Authentic Leadership Rating Scale (ALRS), a situational judgment test. The initial scale consisted of 16 items presenting challenges in leadership situations. It was evaluated by ten specialists, after which 13 items obtained Fleiss' kappa indices that were considered good. Internal structure validity evidence was obtained through the application of Cognitive Diagnosis Models on data collected from 532 Brazilian professionals. Good fit indices were obtained for the Generalized Deterministic-input, Noisy-and-gate Model (G-DINA). We concluded that the ALRS is promising for the continuity of Authentic Leadership investigation and presented suggestions for a research agenda.

Keywords: leadership, authenticity, selection tests.

Liderança Autêntica: Desenvolvimento e Validação Inicial de um Teste de Julgamento Situacional

Resumo

Este estudo descreve o desenvolvimento e as buscas de evidências de validade inicial (conteúdo e estrutura interna) da Escala de Avaliação do Líder Autêntico – EALA. Trata-se de um teste de julgamento situacional. A escala inicial contou com 16 itens expondo desafios em situações de liderança. Foi avaliada por dez especialistas, restando 13 itens que obtiveram índices Kappa de Fleiss considerados bons. As evidências de validade da estrutura interna foram obtidas com dados coletados de 532 profissionais brasileiros e por meio da aplicação de Cognitive Diagnosis Models. Foram obtidos bons índices de adequação para o Generalized Deterministic-input, Noisy-and-gate Model (GDINA). Concluiu-se que a EALA mostra-se promissora para a continuidade de pesquisas no campo da Liderança Autêntica e apresentaram-se sugestões para o prosseguimento das pesquisas.

Keywords: liderança, autenticidade, testes de seleção.

Liderazgo Auténtico: Desarrollo y Validación Inicial de una Prueba de Juicio Situacional

Resumen

Este estudio describe el desarrollo y las búsquedas de evidencias de validez inicial (contenido y estructura interna) de la Escala de Evaluación del Líder Auténtico - EALA. Se trata de una prueba de juicio situacional. La escala inicial contó con 16 ítems exponiendo desafíos en situaciones de liderazgo. Fue evaluada por diez especialistas, restando 13 ítems que obtuvieron índices Kappa de Fleiss considerados buenos. Las evidencias de validez de la estructura interna fueron obtenidas con datos recolectados de 532 profesionales brasileños y por medio de la aplicación de Cognitive Diagnosis Models. Se obtuvieron buenos índices de adecuación para el Generalized Deterministic-input, Noisy-and-gate Model (GDINA). Se concluyó que la EALA se mostró prometedora para la continuidad de investigaciones en el campo del Liderazgo Auténtico y se presentaron sugerencias para la continuación de las investigaciones.

Palabras-clave: liderazgo, autenticidad, pruebas de selección.

*** Informations about the main author:**
R. Waldemar César da Silveira, 105 - Jardim Cura D'ars, Campinas (SP), Brasil, CEP 13045-510.
E-mail: isabel.playit@gmail.com

How to cite this article:
Campos, M. I., & Rueda, F. J. M. (2020). Authentic Leadership: Development and Initial Validation of a Situational Judgment Test. *Revista Psicologia: Organizações e Trabalho, 20*(2), 1047-1056. https://doi.org/10.17652/rpot/2020.2.18100

Authentic leadership (AL) is a leadership style wherein leaders influence their followers based on an ethical behavior and acting in accordance with their personal values, which are always positive. This behavior is based on: 1) their abilities of self-awareness and awareness of others; 2) their capacity to act after listening and reflecting on others' visions; 3) their capacity to relate transparently to followers; 4) the way they act consistently with their core values and 5) by a genuine interest in the full development of their followers' potential (Avolio & Gardner, 2005; Avolio & Walumbwa, 2014; Gardner, Avolio, Luthans, May, & Walumbwa, 2005).

The first theorists in the field of AL were mainly motivated by the search for an adequate form of leadership. Not only a leadership able to achieve positive organizational results, but also that could maintain focus on positive results for other stakeholders, including their followers and society in general (Gardner, Cogliser, Davis, & Dickens, 2011). The initial theoretical assumptions proposed by those scholars were then corroborated. Researches showed that AL is related and/ or is a predictor, among others, of job satisfaction, psychological capital, confidence, creativity, work engagement, worker well-being, and performance of individuals and groups (Avolio & Walumbwa, 2014; Campos & Rueda, 2018b; Gardner et al., 2011).

Measurement of AL remains among the gaps in this research area, especially regarding the evolution of the construct and the AL theory itself (Banks, Gooty, Ross, Williams, & Harrington, 2017; Banks, McCauley, Gardner, & Guler, 2016; Campos & Rueda, 2018b; Gardner, Cogliser, Davis, & Dickens, 2011). The first instrument developed in this field of research was the Authentic Leadership Questionnaire (ALQ) (Walumbwa, Avolio, Gardner, Wernsing, & Peterson, 2008), since then, studies and debates involving the psychometric qualities of the instrument, as well as the nature of AL as a first — or second — order construct have been frequent (Campos & Rueda, 2018a).

Neider and Schriesheim (2011), for example, criticized the processes used by Walumbwa et al. (2008) to evidence the content validity and internal structure of the ALQ and developed the Authentic Leadership Inventory — ALI. They stated that the ALI presented psychometric qualities more adequate than those of the ALQ, but concluded that AL was shown as a first-order or second-order construct, depending on the leader under evaluation.

Both the ALQ and ALI were developed based on the theoretical definition proposed by Walumbwa et al. (2008). The authors defined AL as a multidimensional construct comprising of four factors: self-awareness (SA, ability to demonstrate understanding about oneself — strengths, weaknesses, values, beliefs, behavioral facets — and to know the impact these may cause on other people); balanced processing (BP, ability to solicit opinions from others, even if those challenge one's own, and to objectively analyze relevant information before making a decision); internalized moral perspective (IMP, ability to make decisions according to internal moral values rather than by group, organizational, or societal pressures); and relational transparency (RT, ability to be authentic in relationships, presenting a true self rather than a fake or distorted one).

Another similarity between these instruments concerns the strategy selected to measure the construct. ALQ and ALI are Likert scales that measure AL based on the perceptions that followers have on the leaders' behavior. For Neider & Schriesheim (2011) this is the correct strategy, since leadership attributes are clearly perceptual and therefore should be evaluated by observers. However, for Weiss, Razinskas, Backmann, and Hoegl (2017), while several aspects of leadership can and should be evaluated by followers, this should not be the case for AL, since external perspectives could not ensure that a leader acts consistently with their own thoughts (values and beliefs) and feelings (emotions), which would imply acting in accordance with the core attributes of AL.

Thus, even though these instruments were originally built and validated as hetero-reported, other researchers have used ALQ or ALI as self-reported instruments (Al-Moamary, Al-Kadri, & Tamim, 2016; Baron, 2016; Černe, Dimovski, Marič, Penger, & Škerlavaj, 2014; Fusco, O'Riordan, & Palmer, 2016; Kotzé & Nel, 2015; Monzani, Bark, van Dick, & Peiró, 2015, Pavlovic, 2015) — perhaps with the same understanding as Weiss et al. (2017). Most of these studies, however, did not find adequacy of these instruments for the measurement of AL.

In order to evaluate the construct instead of the quality of the instruments, Banks et al. (2016) performed a meta-analysis with data from 100 studies and more than twenty-five thousand participants; reaching the conclusion that although AL is theoretically differentiated from the Transformational Leadership construct, high magnitude correlations between both and the lack of incremental validity of either one over the other indicate redundancy between them. These findings stimulated new studies in the theoretical field and in the context of measurement instruments.

In the theoretical field, Sidani and Rowe (2018) proposed a reconceptualization of AL, suggesting that it should not be understood as a leadership style, but as a result of a process co-created by leader-follower interaction. However, Gardner and Coglisier (2018) argued that the definition of Walumbwa et al. (2008) remains, in fact, sufficiently good and that rather than deconstructing the mainstream theory, what should be better investigated are the existing barriers to the flourish of AL.

Regarding the quality of psychometric instruments in the context of AL, Levesque-Côté, Fernet, Austin and Morin (2017) stated that the multidimensional nature of AL and the high correlations among its components were, generally, neglected in ALQ — and ALI — based surveys, leading to the different order models for AL. They applied Exploratory Structural Equation Modeling (ESEM) to analyze data collected with the ALQ and ALI in two samples comprised of professionals and found that both instruments were inadequate in correctly capturing the multidimensionality of AL. The authors proposed a new instrument by integrating ALQ and ALI items, leading to a model with adequate psychometric qualities. Thus, they created the Authentic Leadership Integrated Questionnaire (AL-IQ), found construct and criterion validity evidence for it, and demonstrated its invariance for gender and for professionals in distinct industries.

Although the AL-IQ showed encouraging results for the continuity of its use, both in research and practice (Avolio, Wernsing, & Gardner, 2018; Levesque-Coté et al., 2017), the fact that it is a Likert scale that hetero-perceptively measures AL, maintains the gap for its applicability in certain recurring processes in organizational practice; such as selection, training and development interventions (T&D), promotion into leadership roles or even layoffs (Campos & Rueda, 2018a). In this sense, a popular type of instrument, typically used in selection processes that could also be applied to the other mentioned organizational processes, is the Situational Judgment Test (SJT). SJTs are used to evaluate candidates' or employees' judgment regarding hypothetical scenarios (called stem items), which represent realistic situations and challenges in the work environment. A respondent must score the appropriateness of the available solutions (called response items) for each scenario; we inforce that each set formed by one scenario and its available solutions composes one item of the instrument (McDaniel & Nguyen, 2001).

SJTs are inherently challenging when it comes to showing validity, specially construct validity, mainly because SJTs are not homogenous in situations, as well as in responses; since those might require multiple types of knowledge, skills or traits (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). Such characteristics can produce multidimensional results in SJTs, therefore, the items within the instrument can also be considered multidimensional (Schmitt & Chan, 2006). Under these circumstances, statistical models that enable multidimensional item analysis — to distinguish the effect of multiple latent traits — have been advocated as essential to surveys applying SJTs (Weekley, Hawkes, Guenole, & Ployhart, 2015). In order to contribute on this issue, recent studies have applied Cognitive Diagnosis Models (CDM), obtaining positive results in the search for validity evidence of SJTs (García, Olea, & de la Torre, 2014; Sorrel et al., 2016).

CDM are probabilistic multidimensional confirmatory models of latent variables with a simple or complex load structure. They are suitable for modeling observable categorical response variables and contain unobservable (i.e., latent) predictor variables (Rupp & Templing, 2008). For this reason, CDM are suitable for assessing a respondent's mastery, or non-mastery, regarding a given competence (knowledge, ability, attitudes, and others). Competencies are the latent variables and are commonly referred to in the CDM literature as attributes (Sorrel et al., 2016).

For the purpose of CDM, an item typically requires more than one attribute, which leads to complex load structures, in which each item requires distinct competencies to be answered correctly. The CDM, like the CFA, are confirmatory in nature, because the attributes need to be defined a priori, according to a substantial theory regarding a construct; and by means of a Q-matrix, which constitutes the load structure of the CDM (Ravand & Robitzsch, 2015). A Q-matrix is usually created based on the opinion of experts. The number of rows corresponds to the number of test items and the number of columns to the number of attributes required to answer that test. Thus, the Q-matrix constructed on the basis of theory and expert understanding can be tested in conjunction with actual data (Ravand, 2016).

Broadly, CDM can be grouped into compensatory, non-compensatory, and general models. In compensatory models, for example, deterministic-input, noisy-or- gate-model, or DINO (Templin and Henson, 2006), the mastery of one of the attributes required for correctly answering an item can compensate the non-mastery of others. In non-compensatory models, for example, deterministic-input, noisy-and-gate model, or DINA (Junker & Sijstma, 2001), lack of mastery in an attribute cannot be compensated in terms of performance by mastery in other attributes. Generalized models such as Generalized DINA, or GDINA (de la Torre, 2011), are comprehensive with respect to the other two model types and allow compensatory and non-compensatory relationships to be applied in the same test at the item level. According to Ravand and Robitzsch (2015), CDM have been applied both in performing post-hoc analyses of previously existing non-diagnostic tests, and in the process of designing a set of items or tasks from inception, in order to achieve a diagnosis. To these authors, CDM are at the intersection between cognitive psychology and statistical analysis.

Given the brief reviews of AL theory, SJTs and CDM, we present the objectives of this article. The first one is to report the development, based on theory and empirical data, of an SJT to measure AL. This is the Authentic Leader Rating Scale (ALRS). The second objective is to report the initial searches for content and internal structure validity evidence, the last being investigated through the application of CDM. These objectives adhere to the understanding of Weiss et al. (2017) when they affirm that the attributes of AL cannot be correctly and completely perceived from an external perspective.

We believe that measuring AL through an SJT can be helpful in reducing gaps in this field of research (see Banks et al., 2016), stimulating the development of new theories and researches on incremental validity regarding other leadership constructs (e.g. transformational leadership). We also believe that, once validated, the instrument will be useful for application in organizational practice, enabling leaders to be assessed through their own perceptions and not just through hetero-perceptions; thus, allowing organizations to assess AL in processes such as selection, training and development, and promotions. The currently validated AL instruments are not yet accomplishing such applications and assessments (Campos & Rueda, 2018a).

Method

In order to meet the objectives of the study we conducted three distinct steps: step 1 - constructing the Authentic Leadership Rating Scale – ALRS; step 2 - searching for content validity evidences and step 3 - searching for internal structure validity evidences. Each of these steps was supported by distinct samples, instruments, data collection and analysis procedures.

Participants

Step 1. The sample consisted of ten Brazilian professionals, seven male and three female, with experience in leadership roles ranging from 3 to 41 years (M = 19.40, SD = 13.48). Four of the participants worked in the service sector, two in industry, one in commerce, two in the public sector and one in the nonprofit sector.

Step 2. This step relied on ten participants, six of them with expertise in psychology (PhD or MA) and in assessment development. Two hold PhDs in management, one is a PhD in science with published studies on AL and knowledge of measurement instruments, and one holds an MA in communication with 28 years of experience in leadership roles. Such a distribution aligns with the proposal by Weekley, Ployhart & Holtz (2006), which considers that the content analysts should gather theoretical and practical expertise on the context under analyses. Three extra subjects (sex = M) participated in a pilot collection. They had experience in leadership roles ranging from 5 to 30 years, working in different sectors (service, industry and nonprofit sectors).

Step 3. A total of 532 Brazilian professionals participated, of which 304 were male and 228 were female. The level of schooling ranged from elementary school (one participant), to doctorate (24 participants). Participants' age ranged from 18 to 78 years (M = 41.41, SD = 11.21). Of the total, 98 subjects reported never having held a leadership position, while 290 reported being in a leadership position at that time. Among these 290, 74 were presidents or business owners, 4 vice presidents, 40 directors, 85 managers, 62 coordinators or supervisors and 25 held other positions. The length of leadership experience ranged from less than 3 to more than 20 years. Participants were residents of all regions of Brazil, with a predominance from the Southeast region (424); and 444 worked in private companies, 51 in the public sector, 36 in nonprofit organizations, and one did not respond to the question.

Instruments

Step 1. We used the 10-question structured interview developed by Campos and Rueda (2019). An example of question

is: "Considering your daily routine and the most varied experiences you go through, what criteria do you most often use to evaluate situations that require you to make a decision as a leader?"

Step 2. We applied the initial version of the ALRS with 16 items. Pilot study participants also answered a four-question socio-demographic survey.

Step 3. We applied a socio-demographic questionnaire and the ALRS. In this step the ALRS consisted of the 13 items.

Data Collection Procedures and Ethical Considerations

The project was approved by an ethics committee under protocol CAAE 51356515.1.0000.5514. All participants, at each stage of collection, offered their agreement and accepted their respective Terms of Free and Clarified Agreement (TFCA).

Step 1. The participants were invited by e-mail and through social networks by one of the researchers. Interviews were answered online (Google Forms), or through the exchange of Word documents via electronic messages. Response time was not controlled. Each participant chose a place and time to provide their information at their convenience.

Step 2. The pilot collection group of participants was invited to answer the ALRS. They provided comments and suggestions about its content, concerning the text and ease of comprehension. They were also asked to indicate any doubts that might arise during the response process. The judges received a manual with definitions of AL, explanations regarding SJTs, and guidelines for analyzing the instrument. They were invited to evaluate two distinct aspects for each item: 1) which factor was measured by the item (situation and responses); and 2) the level of AL in each response item, on a scale of 1 to 4 — 1 indicating least authentic and 4 indicating most authentic.

Step 3. Participants were invited through the social networks Facebook, LinkedIn and WhatsApp. For the collection, an ad hoc tool was developed on KNBS' Prospektor platform (www.knbs.net.br). Participants were instructed to select two of the solutions for each ALRS situation, according to the following classifications, but using each classification only once: 1 = unlikely or impossible (indicating low or no possibility that the respondent would behave according to that option); 4 = very possible or certainly (indicating a high possibility or certainty that the respondent would behave according to that option).

Data Analysis Procedures

Step 1. Empirical data were submitted to theoretical thematic analysis (Braun & Clarke, 2006), consisting of six steps: familiarization with the data, initial coding, search for themes, theme definition, theme naming, and findings report production. We then associated real situations described by the leaders with the behavioral factors that identify an authentic leader, based on AL theory and developed the content for ALRS.

Step 2. We applied the criteria recommended by Giannarou and Zervas (2014) to evaluate the level of agreement between judges: (a) $SD < 1{,}5$; (b) level of agreement between judges $\geq 51\%$; (c) Interquartile range difference (Q3 - Q1) < 1.0; and (d) the absolute difference between the median value and the level of AL hypothesized for the item by the test developers when constructing the instrument - (Med - Hyp) < 0.5. The content evaluation of each response item should meet at least three of these four criteria simultaneously. We also verified the agreement between the judges for the AL level in the response options by means of Fleiss' kappa. The judges also provided their insights as to which factor would be evaluated by each item. We used the answers to evaluate the level of agreement between them, and to formulate two of the Q-matrices to be used in the searches for internal structure validity evidence of the ALRS. We then validated the most appropriate Q-matrix to be applied with the CDM following the available recommendations (de la Torre & Chiu, 2016; García et al., 2014; Ravand & Robitzsch, 2015; Sorrel et al. al., 2016).

Step 3. The SPSS V.21 package was used to analyze the descriptive statistics. In order to find the appropriate statistical model, the R software version 3.4.4 was used with the CDM package (Robitzsch, Kiefer, George, & Uenlue, 2014) and the GDINA package (Ma & de la Torre, 2018). These packages allow us to investigate the adequacy of the Q-matrix that defines the attributes required for good performance in each test item, or absolute fit, and the adequacy of the hypothesized model for the test (e.g. compensatory, non-compensatory, or generalized), or relative fit. We followed the method proposed by Chen, de la Torre and Zangh (2013). The analysis was also aligned with the guidelines found in de la Torre and Chiu (2016), García et al. (2014), Ravand and Robitzsch (2015), and Sorrel et al. (2016).

We first split the participants' answers to the 13 items of the ALRS into two distinct databases: 1) the selections of the least probable item to be executed by the respondent (least responses), and 2) the selections of the most probable item to be executed by the respondent (most responses). Then, with the CDM package, we investigated the existence of absolute fit for the Q-matrices (see step 2) under the criteria for the GDINA model using the databases most responses and least responses.

Next, the GDINA package was used to obtain recommendations for improvements that could fit one of the provisional Q-matrices, in order to find a model that demonstrated suitability to the data. We only tested recommendations provided by the software package when they were adherent to the judges' analysis and/or to the theoretical model of AL. Eventually, the final Q-matrices (good fit) for the most responses and least responses were found. We then run comparisons between the GDINA, DINA and DINO models to verify the relative fit and to select the appropriate model for the ALRS.

Results

As a result of step 1 — Constructing the Authentic Leadership Rating Scale — we developed the initial ALRS. It consisted of 16 items (situations and responses for them). Table 1 shows the content of items 1 and 13 as samples.

Each ALRS item consisted of one situation, or scenario (the stem item), and four response options (the response items). Hypothetically, the ALRS would be able to measure the four factors of AL: SA, BP, IMP and RT. In order to demonstrate the veracity of these hypotheses, we first searched for content validity evidence (step 2 of this study). Table 2 shows the results obtained for the 16 original scale items in relation to the judges' assessment and the observations received in the pilot collection, focusing on response items (alternatives a, b, c, d).

Table 2 shows that three items of the instrument (1, 4, 13), i.e. the set of stem item (situation) and response items (answering options), met the criteria for all response items; nine items (3, 5, 6, 8, 9, 10, 11, 12, 16) met the criteria for three response items; one item (2) met the criteria for two response items; and three items (7, 14, 15) met the criteria for only one response item. The 16 initial items were then reduced to 13.

We decided to keep items marked M1 in Table 2 based on: 1) items considered as least authentic and most authentic were validated, allowing the development of directives that asked

Table 1
Examples of ALRS items

Item	Description
Situation item	Carlos is the manager of a production area in a large industry. Two team leaders, who report directly to him, got into a conflict. It started to negatively impact the teams' performance on their tasks, because these two leaders are responsible for sectors of which results are immediately linked to one another. Carlos talked about the case in a meeting with his superior and his peers. They found a possibility to transfer one of the team leaders to another sector, in which there was an available position, but he would have to work in another shift. If you were Carlos, what would you do?
Response items	a) I would ask the HR department to relocate any of the team leaders to the available position, communicating with him directly. However, only after finding a suitable substitute for that position. b) In a meeting with both of them, I would highlight their responsibilities regarding their teams' outcomes, giving them a deadline for solving the situation. I would not mention the available position, leaving room for a future alternative. c) In a meeting with both of them, I would emphasize the need to maintain their professionalism, highlighting their responsibilities regarding their teams' outcomes. I would mention the available position in the other shift, explaining that one of them can ask for the transfer, if desired. d) I would listen to each team leader individually about the situation and, according to my understanding on the conflict, I would choose one of them to be transferred. After that, I would select a substitute for the position, and then personally inform the decision to the chosen one.
Situation Item	Jaime is the XPTO owner, a provider of software solutions. His team developed a new application, which is leveraging the business. A competitor, which owns 70% of market share and is a member of a group with great financial strength, made a hostile bid to acquire XPTO, under the threat that if it is not accepted, they will destroy their presence in the market. The offer implies good immediate financial results. Jaime knows that the team is very motivated, but they are facing a technical limitation that impairs a rapid expansion of XPTO. If you were Jaime, what would you do?
Response Items	a) I would reject the offer, because I believe in the team and as the competitor is so interested, it means that our future potential gain is high. b) I would accept the offer, realizing the immediate financial gain and getting rid of the technical limitation. c) I would gather the team to explain the situation and understand every person's view. From this I would decide whether to sell or not. D) I would secure the competitor through a negotiation and, without scaring the team about the possible sale, would require a quick solution to the limitation.

Table 2
Results for response items

Item	Alt	*DP*	Q3–Q1	Freq %	Dif	Hyp	Action	Item	Alt	*DP*	Q3–Q1	Freq %	Dif	Hyp	Action
1	a	0.68	0.25	80	0		M	9	a	0.85	1.25	70	0		M
	b	0.47	0.00	80	0				b	0.42	0.25	80	0		
	c	0.63	0.00	90	0				c	0.52	1.00	60	0		
	d	0.67	0.50	60	0				d*	1.08	1.50	40	0		
2	a	0.70	1.00	80	0		M1	10	a	0.68	1.00	50	0		M1
	b*	0.88	2.00	30	-1	3			b*	0.82	1.25	50	-0.5	3	
	c	0.00	0.00	100	0				c	0.48	1.00	70	0		
	d*	0.52	1.00	50	-1	2			d	0.63	0.00	90	0		
3	a	1.16	2.00	70	0		M1	11	a	0.67	0.50	60	0		M1
	b	0.97	0.25	80	0				b	0.97	1.00	70	0		
	a	0.67	0.50	60	0				d	0.97	1.00	70	0		
	d*	1.06	1.50	50	0	3			d*	0.94	2.00	40	-1	3	
4	a	0.42	0.25	80	0		M	12	a	0.97	1.00	70	0		M1
	b	0.42	0.25	80	0				b	0.84	1.00	50	0		
	c	0.84	1.25	60	0				c	0.00	0.00	1	0		
	d	0.57	0.25	70	0				d*	0.53	1.00	50	-0.5	2	
5	a*	1.08	2.00	50	0	3	E**	13	a	0.67	0.50	60	0		M
	b	0.71	1.00	60	0				b	0.71	1.00	60	0		
	c	0.68	0.25	80	0				c	0.42	0.25	80	0		
	d	0.79	0.25	70	0				d	0.32	0.00	90	0		
6	a	0.63	1.00	60	0		M1	14	a*	0.74	1.25	50	0	3	E
	b	0.32	0.00	90	0				b*	0.97	1.25	20	-1.5	4	
	c*	1.05	2.00	50	-1	3			c*	1.16	2.25	30	0	2	
	d	0.00	0.00	100	0				d	1.14	2.00	60	0		
7	a*	0.94	1.25	50	0	3	R	15	a*	0.74	1.25	50	0	2	E
	b*	0.52	1.00	40	-1	2			b*	0.53	1.00	50	-0.5	4	
	c*	1.32	3.00	40	-1	4			c*	0.70	1.00	50	0.5	1	
	d	0.32	0.00	90	0				d	0.52	1.00	60	0		
8	a	0.32	0.00	90	0		M1	16	a	0.71	1.00	60	0		M1
	b	0.84	1.00	50	0				b	0.70	1.00	70	0		
	c*	1.14	2.25	40	0	3			c	0.97	0.25	80	0		
	d	1.06	1.25	60	0				d*	0.70	1.00	50	-0.5	3	

Dif = Med- Hyp; Hyp = hypothetical value; * response items that did not meet at least 3 criteria simultaneously; M = item maintained; E = item eliminated; R = item reformulated to improve its content due to pilot collection results;
M1= items in which one or two of the responses did not meet at least 3 criteria simultaneously were maintained in this first phase when the unvalidated responses were different from 1 (less authentic) and 4 (more authentic).
E** item 5 was eliminated due to its high similarity to item 9, purposely created at the time of development and commented on by judges and pilot participants.

respondents to mark two of the four response options: one as least likely and one as most likely to be the action they would take; 2) more information about the quality of response items was expected from the next steps of the validity evidence search.

We then evaluated the resulting 13-item scale by means of Fleiss' kappa. According to Fleiss, Levin and Paik (2013), kappa values above 0.75 can be considered excellent, between 0.4 and 0.75, good, and below 0.4 poor. The values obtained, all with $p < 0.001$ were: 0.71 for level 1 (least authentic); 0.46 for level 2; 0.32 for level 3; and 0.68 for level 4 (most authentic). The overall scale kappa was 0.54. Thus, the responses were considered indicative that the least and the most authentic levels reached a content validity criteria that can be interpreted as good, as well as the scale in its overall assessment.

Table 3 shows the percentages of agreement among judges about which factor would be assessed by each item of the instrument, and the respective Q-matrix generated based on these results. It also shows the hypothesis matrix that we built, covering the primary factor expected to be measured, and a combination of the two matrices based on judges' agreement levels with values equal to or greater than 30%.

Data in Table 3 shows that in 10 of the 13 ALRS items, the highest percentage of agreement among judges was coincident with our initial hypothesis. In only three of them, though, this agreement was equal to, or greater than, 80%. These values, which are considered targets in content validity studies that follow the Classical Theory of Tests (CTT), are not mandatory to achieve when working with SJT, given the possibility that an item can demand several attributes to be answered correctly. At this point of the study, we proceeded to step 3.

As results of step 3 — internal structure validity — Table 4 shows the absolute fit indices obtained for each of the Q-matrices shown in Table 3. This is for both the most responses and the least responses databases.

Table 4 data indicate that none of the Q-matrices presented good absolute fit indices in the GDINA model. Max (X^2) is the maximum value of all $\chi 2ij$ statistics and, although the corresponding p values obtained by the Holm procedure (Ravand & Robitzsch, 2015) were not significant, and the SRMSR values were satisfactory, the p values for the abs(fcor) indices were not significant. Abs(fcor) represents the absolute value of the deviations of the transformed Fisher correlations and is one of the indices proposed by Chen et al. (2013) to verify the absolute fit of a CDM.

Due to the inadequacy of the three Q-matrices, the process of searching for suitable matrices for the data was initiated using the GDINA package. This is an exploratory process, since each matrix suggested from one of the initial matrices was contrasted with the opinion previously provided by the judges and also with the theoretical concepts of AL. After a sequence of tests of new matrices with the databases of most responses and least responses, it was possible to find solutions with good absolute fit. Table 5 presents the two final Q-matrices and the absolute fit indices for them.

For the most responses Q-Matrix there were four recommendations for including factors necessary to fit the model

Table 3
Evaluations of the judges on the AL factors and corresponding Q-matrix

Item	Percentage – specialists				Q-matrix – specialists				Main Attribute (hypothesis)				Combination (30% or more)			
	SA	BP	IMP	RT	AS	BP	IMP	RT	SA	BP	IMP	RT	AS	BP	IMP	RT
1	0	70	0	30	0	1	0	1	0	0	0	1	0	1	0	1
2	0	20	30	50	0	1	1	1	0	0	1	0	0	0	1	1
3	10	0	30	60	1	0	1	1	0	0	0	1*	0	0	1	1
4	0	20	20	60	0	1	1	1	0	0	0	1*	0	0	0	1
5	10	0	80	10	1	0	1	1	0	0	1*	0	0	0	1	0
6	0	0	90	10	1	0	1	1	0	0	1*	0	0	0	1	0
7	40	0	60	0	1	0	1	0	1	0	0	0	1	0	1	0
8	10	0	90	0	1	0	1	0	0	0	1*	0	0	0	1	0
9	10	30	30	10	1	1	1	1	0	1*	0	0	0	1	1	0
10	40	20	20	20	1	1	1	1	1*	0	0	0	1	0	0	0
11	70	0	10	20	1	0	1	1	1*	0	0	0	1	0	0	0
12	10	70	10	10	1	1	1	1	0	1*	0	0	0	1	0	0
13	30	10	10	50	1	1	1	1	0	0	0	1*	1	0	0	1

Note. * items in which the factor hypothesized as the one that would be the main evaluated received the highest percentage of agreement by the experts.

Table 4
Absolute fit indices of the GDINA model — initial Q-matrices

Q-Matrix	Most responses						
	Max(X^2)	p	abs(fcor)	P	SRMSR	loglike	Npars
Specialists	2.82	1	0.08	1	0.028	-3698.84	127
Hypothesis	8.64	0.26	0.13	0.13	0.041	-3779.80	37
Combination	5.63	1	0.1	0.68	0.038	-3773.67	45
Q-Matrix	Least responses						
	Max(X^2)	p	abs(fcor)	P	SRMSR*	Loglike	Npars
Specialists	4.66	1	0.1	0.82	0.035	-3769.34	127
Hypothesis	9.44	0.17	0.13	0.08	0.041	-3839.86	37
Combination	10.28	0.1	0.14	0.05	0.039	-3839.61	45

*Note.**SRMSR = standardized mean square root of squared residuals; loglike = log-likelihood; Npars = number of estimated parameters.

Table 5
Final Q-Matrices and absolute fit indices of the GDINA model

Item	Most Responses Q-Matrix				Least Responses Q-Matrix			
	AC	PB	PMI	TR	AC	PB	PMI	TR
1	0	1	0	1	0	1	0	1
2	0	0	1	1	0	0	1	1
3	0	0	1	1	0	0	1	1
4	1*	1	0	1	0	1	0	1
5	0	0	1	0	0	0	1	0
6	1*	0	1	1	0	0	1	1
7	1	1*	1	0	1	1*	1	0
8	0	0	1	0	0	0	1	0
9	0	1	1	1	0	1	1	1
10	1	0	1	0	1	0	1	0
11	1	1*	1	1	1	0	0	1
12	0	1	0	0	0	1	0	0
13	1	1	0	1	1	1	0	1
Fit Indexes								
Max(X^2) - (p)	11.56 - (0.05)**				8.94 - (0.22)			
abs(fcor) - (p)	0.15 - (0.02)				0.14 - (0.04)			
SRMSR	0.038				0.039			
Loglike	-3739.97				-3811.85			
Npars	89				69			

Note. * factors that were included in the analysis and which were not predicted by the specialists. ** p = 0.05257232

(items 4, 6, 7 and 11). For the least responses Q-Matrix there was just one recommendation (item 7). Table 5 also shows the absolute fit of the GDINA model obtained through the matrices for the most responses and least responses.

Since other types of CDMs (e.g. DINA and DINO) are nested under the GDINA model, it is necessary to verify which of them is most suitable for the ALRS. One way to do this is to apply the Likelihood Ratio Test (LR). Table 6 shows the results of the model comparison through the LR test.

The LR test verifies whether observed differences between the models are statistically significant by comparing the log-likelihood of the models. The p values in Table 6 were significant for the LR tests considering the DINA and the DINO models for both Q-matrices. This indicates that the generalized model (GDINA) was significantly more fit to the data than the reduced models.

ranked, identified from 0000 to 1111 where each 0 indicates non-mastery and each 1 indicates mastery regarding the factor. The cut-off point used for this differentiation was 0.5. That is, scores smaller than 0.5 were taken as indicating non-mastery and greater than or equal to 0.5 as indicating mastery. The order of the factors in the class composition is SA, BP, IMP and RT.

It is possible to note that 8 respondents did not master any of the factors for the most authentic responses (0000), and that 23 respondents showed no mastery in any of the factors for least authentic responses. The class 1000 indicates mastery only on the SA factor for one of the respondents regarding the most authentic responses and for 13 respondents regarding least authentic responses. The rationale continues until 1111, which indicates mastery in the four factors (SA, BP, IMP, RT) by 123 respondents regarding the most authentic responses, and on the part of 52 respondents regarding least authentic responses.

Table 6
Relative fit indices (model comparison)

Model	Most Matrix						
	loglike	*deviation*	Npars	*N*	LR	*df*	*p*
GDINA	-3739.97	7479.94	89	532			
DINA	-3779.23	7558.46	37	532	78.52	52	<0.05
DINO	-3783.47	7566.95	37	532	87.00	52	<0.01
Model	Least Matrix						
	loglike	*deviation*	Npars	*N*	LR	*df*	*p*
GDINA	-3811.85	7623.71	69	532			
DINA	-3851.10	7702.19	37	532	78.48	32	< 0.001
DINO	-3813.22	7626.45	37	532	77.74	32	<0.001

*Note.**loglike = *log-likelihood*; Npars = number of estimated parameters; LR = likelihood ratio; *df* = degrees of freedom

One characteristic of the CDM is to provide a way of scoring and ranking test respondents, indicating whether the respondent does or does not master the elements of the evaluated components. Since the ALRS measures the four components of AL (k = 4), there are 16 possible classes (2k). Table 7 shows the distribution of the participants, according to the classes in which they were

Table 7
Classes and distribution of the respondents

Class	N_Most*	N_Least**	Class	N_Most*	N_Least**
0000	8	23	1000	1	13
0001	51	7	1001	70	18
0010	16	1	1010	4	0
0110	31	1	1011	79	0
0100	13	0	1100	4	148
0101	3	0	1101	23	75
0110	17	94	1110	82	72
0111	7	28	1111	123	52

Note. *N_Most = number of people in the class, according to the selection of the most authentic answers.
**N_Least = number of people in the class, according to the selection of the least authentic answers.

Discussion

This article aimed to describe the development of the Authentic Leadership Rating Scale (ALRS) in a situational judgment test format, and the search for initial evidence of its content and internal structure validity through the application of Cognitive Diagnosis Models. We proceeded with this development stimulated by literature that indicated gaps not covered by previously developed instruments (Banks et al., 2016; Campos & Rueda, 2018a; Levesque-Coté et al., 2017).

We observed that regarding the ALRS, the process of integrating a theoretical approach and an empirical strategy differs from those described as more usual in the literature when developing an SJT (Weekley et al., 2015, Weekley et al., 2006) and it constitutes an innovation in the development process. In the present study, this differentiation took place in two aspects: the search for adherence to the theoretical construct in the content of the situation, as well as in the options of answers; and the development of the response options carried out by us, based on a theoretical context — instead of having solution options provided by specialists in the subject, usually from organizational practice. In a next stage of instrument improvement, experts in leadership may be invited to provide new response options, as well as to review the scores attributed to the options initially created.

As noted by Weekley et al. (2006), however, it is the response option generation by specialists that inherently promotes the heterogeneous nature of a SJT. Anyway, what can be inferred from the judges' evaluation presented in step 2, and from the evidence of internal structure validity shown by the CDM application in step 3 is that, in the case of this work, the characteristic heterogeneity of the SJT may be attributed to the AL construct itself.

We believe our understanding is consistent with Levesque-Coté et al. (2017), when they address the fundamentally multidimensional nature of AL and the high correlations between its components as defined by Walumbwa et al. (2008); which were present in several studies that applied both the ALQ and the ALI. In developing and presenting the initial validity evidence for the AL-IQ, the authors sought to reduce the incidence of these high correlations through the new instrument, formulating it based on items that were more homogeneous and compatible with what is expected of measures which follow the CTT.

However, this strategy does not reduce the heterogeneity of the AL concept. In a way and indirectly, the matter was also addressed by Levesque-Côté et al. (2017), who exemplified it by using an ALQ item that is part of the self-awareness factor, showing that the attitude described in that item could also require relational transparency and balanced processing by the leader under assessment. The assessments made by the experts participating in step 2 of this study apparently demonstrated this aspect of heterogeneity relative to the situations and responses in 11 of the 13 items that passed through the first sieve, presenting good Fleiss' kappa indices (Fleiss, Levin, & Paik, 2013) for levels 1, 2 and 4.

These results indicated partial content validity and the possibility to perform the data collection in order to search for evidence of internal structure validity, based on the choice of the least authentic and of the most authentic for each situation. However, in the typical context of instruments that follow the CTT, such results would have been perceived as unsatisfactory in terms of content validity and, indeed, the question of whether SJTs are a method of measurement and not a method to measure a specific construct (Schmitt & Chan, 2006) remains unanswered.

In this sense, by finding internal structure validity evidence for the ALRS through the CDM application — which resulted in satisfactory fit indices — this work joined several surveys referenced by Weekley et al. (2015). While using validation methods other than CDM application, researches indicated the feasibility of applying SJT to measure constructs, such as: goal orientation, initiative, knowledge of the team role, dimensions of a leadership model, integrity and emotional intelligence, among others. However, this work also corroborates the assertion made by the same authors that, to date, there is no convincing evidence that an SJT can be developed a priori to provide homogeneous scores for a specific construct of interest.

With regard to the preceding point, items 5 and 12 of the ALRS showed characteristics of homogeneity, the first by evaluating internalized moral perspective and the second by evaluating balanced processing. This occurred for both the most authentic and the least authentic subscales and this homogeneity is compatible with our initial hypotheses for these items.

These results may indicate that CDM, as an analyses method applied to SJTs, are also promising in regard to exploring the viability of developing and validating SJTs that might be homogeneous while integrating situations and response options. This is a challenge that was amply discussed by Weekley et al. (2006) and Schmitt and Chan (2006).

In the context of classifications, the ALRS presented a good discrimination capacity among the respondents, that were distributed through all 16 evaluation classes for the case of the most responses and through 10 of them for the case of the least responses. This result is also promising for the research field because, as indicated by Weekley et al. (2006), the scoring forms of an SJT are commonly determined by the test developers. It would be interesting to investigate whether scoring through statistical methods (i.e. CDM) is more accurate than scores obtained through subjective definitions (i.e. scale developers'), and the application of the ALRS in such studies could also contribute to the SJT research area.

Due to the findings and analyses discussed, we suggest further

studies with the ALRS, especially to search for criterion validity. To contribute more effectively, these studies should be multi-method, enabling professionals who are performing leadership roles to respond to the ALRS — while their teams can evaluate them using one of the other three existing instruments, preferably two of them, given that the AL-IQ (Levesque-Cotê et al., 2017) is also a recent instrument. Other possibilities for future studies could incorporate some of the most studied constructs in conjunction with AL, among them: psychological capital, job satisfaction, engagement and work commitment (Campos & Rueda, 2018b).

There is also a potential research path to demonstrate the reliability of the ALRS. In addition to researching these indices by means of test and retest, as suggested by Weekley et al. (2015) and by Weekley et al. (2006), it is possible to apply data simulation, as used by Sorrel et al. (2016).

Finally, studies that could apply the ALRS in organizational practices such as selection and T&D, and compare results with those obtained for other constructs of interest (e.g. work engagement, creativity, psychological capital) would also be welcomed. In addition, they would help to verify if one of the initial objectives for its development can be achieved, that is, enabling AL to be adopted and to bring benefits to organizational practice (Campos & Rueda, 2018b). In T&D situations, longitudinal studies could also contribute to verify the possible impact of interventions aimed at the development of the authentic leader.

Through this analysis, this article fulfilled its objectives: to present the development and to search for initial validity evidence for the ALRS. The scale showed good indices for content validity, with possibilities for improvement in future versions regarding response options that, for example, would allow different directives — given that in this first version respondents had to opt for solutions most and least representative of their actions.

Regarding internal structure validity, verified through the application of CDM, the scale obtained good fit indices, indicating the feasibility of measuring the four components of Authentic Leadership: self-awareness, balanced processing, internalized moral perspective and relational transparency, as defined by Walumbwa et al (2008) and restated by Gardner e Coglisier (2018). The ALRS also made it possible to discriminate respondents through the 16 classes at two distinct levels, which inform whether the respondent recognizes both how to act and how not to act in situations that stimulate or require the presence of AL.

Thus, this work contributes to increasing the diversity of AL measurement instruments, making it possible to measure it not only through the perception of followers, but based on information provided by the leader being evaluated. The SJT format selected for the ALRS also adds another contribution to the fields of research and practice: it may eventually reduce the risk of inherent bias in self-reporting, as found by Černe et al. (2014) when comparing results obtained by applying the ALI with followers and their respective leaders. The work also opens up a vast field for new research, since it will be necessary to confirm the validity of the ALRS by investigating its convergence and its capacity to discriminate in relation to other constructs of interest. In addition, it will be interesting to see if any discrimination between Authentic Leadership and Transformational Leadership can be found with the application of this instrument.

The ALRS is also promising for organizational practice. Once the evidence of its validity multiplies, it can cover the applicability gaps left by the hetero-reported instruments; this way, it will enable the assessment of the authentic leader's behavior in selection, training and development processes, promotions, and other areas of interest to organizations. In this context, it should also be remembered that, since the beginning of research in the AL field, several theorists have been positioning the development of authentic leaders as one of the main motivations of the area (Avolio & Gardner, 2005; Avolio & Walumbwa, et al., 2005, Gardner et al., 2011). Thus, an instrument that allows evaluating the leader in moments before and after T&D interventions, without the need to consult their followers, becomes contributory.

It is important to mention the limitations of this work. The first limitation concerns the restricted number of leaders interviewed when performing step 1 and the small base of existing instrument items (16 on the ALQ and 14 on the ALI, both of which are very similar in content). This has restricted the possibilities of creating extreme situations of measurement. In a way, this may have contributed to the heterogeneity demonstrated by 11 of the ALRS items. The second limitation concerns the parallel execution of the pilot collection and the evaluation of the judges. If the pilot collection had been done in advance, some of the changes made to the items would have undergone assessment by the experts and could eventually reflect in the results of the content validation. These are native limitations to this project and mentioning them may help other researchers, who wish to follow similar methods, to reflect on these aspects at the time of planning.

There are still two other important limitations worth mentioning. Suggestions for improvements to the initial Q-matrices were obtained through the GDINA package (Ma & de la Torre, 2018). Because it is a new package, one should consider the possibility that the Q-matrices selected in this work may not be the best possible for the ALRS. Finally, the GDINA model was superior to the DINA and DINO models, but no comparisons were made with several other types of CDM.

Despite these limitations, results obtained are promising. They indicate the feasibility of pursuing research that will secure the availability of the ALRS for application in organizational practice. Therefore, we believe the work carried out and the results obtained are a contribution, increasing the current body of knowledge in the AL field. The results also provide encouragement, so that studies on the construct will not be abandoned due to possible overlap between AL and TL. Investigations with the new style of measurement may eventually lead to different findings, since, according to Weekley et al. (2015) and Weekley et al. (2006), SJTs have demonstrated incremental validity over other typical predictors.

References

Al-Moamary, M. S., Al-Kadri, H. M., & Tamim, H. M. (2016). Authentic leadership in a health sciences university. *Medical teacher, 38*(1), 19-25. https://doi.org/10.3109/0142159X.2016.1143092

Avolio, B. J., & Gardner, W. L. (2005). Authentic leadership development: Getting to the root of positive forms of leadership. *Leadership Quarterly, 16*(3), 315-338. https://doi.org/10.1016/j.leaqua.2005.03.001

Avolio, B.J., & Walumbwa, F.O. (2014), Authentic leadership theory, research and practice: steps taken and steps that remain. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 331-356). New York, NY: Oxford University Press.

Avolio, B. J., Wernsing, T., & Gardner, W. L. (2018). Revisiting the Development and Validation of the Authentic Leadership Questionnaire: Analytical Clarifications. *Journal of Management, 44*(2), 399-411. https://doi.org/10.1177/0149206317739960

Banks, G. C., Gooty, J., Ross, R. L., Williams, C. E., & Harrington, N. T. (2017). Construct redundancy in leader behaviors: A review and agenda for the future. *The Leadership Quarterly, 29*(2018), 236-251. https://doi.org/10.1016/j.leaqua.2017.12.005

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly, 27*(4), 634-652. https://doi.org/10.1016/j.leaqua.2016.02.006

Baron, L. (2016). Authentic leadership and mindfulness development through action learning. *Journal of Managerial Psychology, 31*(1), 296-311. https://doi.org/10.1108/JMP-04-2014-0135

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Campos, M. I., & Rueda, F. J. M. (2019). Authentic leadership: A theoretical thematic analysis of the contemporary Brazilian leader's speech. *Paidéia (Ribeirão Preto), 29*, e2924. https://doi.org/10.1590/1982-4327e2924

Campos, M. I., & Rueda, F. J. M. (2018a). *Authentic Leadership Measures: a Literature Review.* Manuscript submitted for publication.

Campos, M. I., & Rueda, F. J. M. (2018b). Evolução do construto liderança autêntica: uma revisão de literatura. *Revista Psicologia Organizações e Trabalho, 18*(1), 291-298. https://doi.org/10.17652/rpot/2018.1.13473

Černe, M., Dimovski, V., Marič, M., Penger, S., & Škerlavaj, M. (2014). Congruence of leader self-perceptions and follower perceptions of authentic leadership: Understanding what authentic leadership is and how it enhances employees' job satisfaction. *Australian Journal of Management, 39*(3), 453-471. https://doi.org/10.1177/0312896213503665

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123-140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C. -Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika,* 1-21. https://doi.org/10.1007/s11336-015-9467-8

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

Fusco, T., O'Riordan, S., & Palmer, S. (2016). Assessing the efficacy of authentic leadership group-coaching. *International Coaching Psychology Review, 11*(2), 118-128.

García, P. E., Olea, J., & De la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*(3), 372-377. https://doi.org/10.7334/psicothema2013.322

Gardner, W. L., Avolio, B. J., Luthans, F., May, D. R., & Walumbwa, F. O. (2005). "Can you see the real me?" A self-based model of authentic leader and follower development. *The Leadership Quarterly, 16*(3), 343-372. https://doi.org/10.1016/j.leaqua.2005.03.003

Gardner, W. L., Cogliser, C. C., Davis, K. M., & Dickens, M. P. (2011). Authentic leadership: A review of the literature and research agenda. The Leadership Quarterly, 22(6), 1120-1145. https://doi.org/10.1016/j.leaqua.2011.09.007

Giannarou, L., & Zervas, E. (2014). Using Delphi technique to build consensus in practice. *International Journal of Business Science and Applied Management, 9*(2), 65-82.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Levesque-Côté, J., Fernet, C., Austin, S., & Morin, A. J. (2017). New wine in a new bottle: Refining the assessment of authentic leadership using exploratory structural equation modeling (ESEM). *Journal of Business and Psychology,* 1-18. https://doi.org/10.1007/s10869-017-9512-y

Ma, W. & de la Torre, J. (2018). GDINA: T*he generalized DINA model framework. R package version 2.1.* Recovered from https://CRAN.R-project.org/package=GDINA

McDaniel, M. A., & Nguyen, N. T. (2001). Situational Judgment Tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*(1/2), 103-113. https://doi.org/10.1037/0021-9010.86.4.730

Monzani, L., Bark, A. S. H., van Dick, R., & Peiró, J. M. (2015). The synergistic effect of prototypicality and authenticity in the relation between leaders' biological gender and their organizational identification. *Journal of Business Ethics, 132*(4), 737-752. https://doi.org/10.1007/s10551-014-2335-0

Neider, L. L., & Schriesheim, C. A. (2011). The authentic leadership inventory (ALI): Development and empirical tests. *The Leadership Quarterly, 22*(6), 1146-1164. https://doi.org/10.1016/j.leaqua.2011.09.008

Pavlovic, N. (2015). Authentic Leadership in Educational Institutions. *International Journal for Quality Research, 9*(2), 309-322. http://www.ijqr.net/journal/v9-n2/10.pdf

Ravand, H., & Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research & Evaluation, 20*(11), 1-12.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799. https://doi.org/10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2014). *CDM: Cognitive Diagnosis Modeling. R Package Version 4.1.* Recovered from http://CRAN.Rproject.org/package=CDM

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219-262. https://doi.org/10.1080/15366360802490866

Schmitt, N., & Chan, D. (2006). Situational Judgment Tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp.135-155). New Jersey: Lawrence Erlbaum Associates.

Sidani, Y. M., & Rowe, W. G. (2018). A reconceptualization of authentic leadership: Leader legitimation via follower-centered assessment of the moral dimension. *The leadership quarterly, 29*(6), 623-636. https://doi.org/10.1016/j.leaqua.2018.04.005

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506-532. https://doi.org/10.1177/1094428116630065

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305. https://doi.org/10.1037/1082-989x.11.3.287

Walumbwa, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S., & Peterson, S. J. (2008). Authentic leadership: Development and validation of a theory-based measure. *Journal of Management, 34*(1), 89–126. https://doi.org/10.1177/0149206307308913

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scalling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests: Theory, measurement, and application* (pp. 1-10). New Jersey: Lawrence Erlbaum Associates.

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational. Psychology, 2*(1), 295-322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Weiss, M., Razinskas, S., Backmann, J., & Hoegl, M. (2017). *Authentic leadership and leaders' mental well-being: An experience sampling study. The Leadership Quarterly.* In press. https://doi.org/10.1016/j.leaqua.2017.05.007