Journal of Work and Organizational Psychology 29 (2013) 13-20

Journal of Work and Organizational Psychology

www.elsevier.es/rpto

Trade-Offs Between Assessor Team Size and Assessor Expertise in Affecting Rating Accuracy in Assessment Centers

Andreja Wirz[a]*, Klaus G. Melchers[b]*, Filip Lievens[c], Wilfried De Corte[c], and Martin Kleinmann[a]

[a]*Universität Zürich, Switzerland*
[b]*Universität Ulm, Germany*
[c]*Ghent University, Belgium*

ARTICLE INFORMATION

Manuscript received: 15/10/2012
Revision received: 30/1/2013
Accepted: 30/1/2103

Keywords:
Rating accuracy
Assessment center
Assessor expertise
Assessor background
Assessor training
Assessor team

ABSTRACT

We investigated the effects of assessor team size on the accuracy of ratings in a presentation exercise as it is commonly used in assessment centers and compared it to the effects of two factors related to assessor expertise (assessor training and assessor background). On the basis of actual ratings from a simulated selection setting (N = 383), we sampled assessor teams of different sizes and with different expertise and determined the accuracy of their ratings in the presentation exercise. Of the three factors, assessor training had the strongest effect on rating accuracy. Furthermore, in most conditions, using larger assessor teams also led to more accurate ratings. In addition, the use of larger assessor teams compensated for having not attended an assessor training only when the assessors had a psychological background. Concerning assessor background, we did not find a significant main effect. Practical implications and directions for future research are discussed.

Trade-offs entre tamaño del equipo evaluador y pericia del evaluador y su efecto sobre la precisión de la valoración en los assessment centers

Palabras clave:
Precisión de la calificación
Assessment center
Pericia de los evaluadores
Historial de los evaluadores
Entrenamiento de los evaluadores
Equipo evaluador

RESUMEN

Investigamos los efectos del tamaño del equipo evaluador sobre la precisión de las valoraciones en un ejercicio de presentación tal como es habitualmente utilizado en los AC y lo comparamos con los efectos de dos factores relacionados con la pericia del evaluador (entrenamiento e historial). Sobre las valoraciones en una situación simulada de selección (N = 383), muestreamos equipos de evaluadores de diferente tamaño y con diferente pericia y determinamos la precisión de sus valoraciones en el ejercicio de presentación. De los tres factores, el entrenamiento de evaluadores tuvo el efecto más fuerte sobre la precisión de la valoración. Además, en la mayoría de las condiciones, usar equipos con mayor número de evaluador también da lugar a valoraciones más precisas. También, el uso de equipos mayores compensó la falta de asistencia de un valorador al entrenamiento cuando los evaluadores tenían formación psicológica. En relación con esto último, no encontramos un efecto principal significativo. Se comentan las implicaciones para la práctica y la investigación futura.

Assessment centers (ACs) enjoy popularity in both the private and public sector, where they play an important role for personnel selection and employee development. ACs are criterion valid (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2004; Hermelin, Lievens, & Robertson, 2007) and they explain incremental variance in job or training performance over and above other procedures such as cognitive ability tests or personality inventories (e.g., Dilchert & Ones, 2009; Krause, Kersting, Heggestad, & Thornton, 2006; Melchers & Annen, 2010; Meriac, Hoffman, Woehr, & Fleisher, 2008). However, ACs are relatively expensive. Hence, an important issue for companies is how to reduce costs for ACs while still ensuring the accuracy of the performance evaluations obtained.

*Correspondence regarding this article should be sent to Andreja Wirz, Psychologisches Institut, Universität Zürich, Binzmühlestrasse 14/12, CH-8050 Zürich, Switzerland, or to Klaus Melchers, Institut für Psychologie und Pädagogik, Abteilung Arbeits- und Organisationspsychologie, Albert-Einstein-Allee 47, D-89069 Ulm, Germany. E-mail: a.wirz@psychologie.uzh.ch or klaus.melchers@uni-ulm.de.

1576-5962/$ - see front matter © 2013 Colegio Oficial de Psicólogos de Madrid. Todos los derechos reservados

DOI: http://dx.doi.org/10.5093/tr2013a3

Recent surveys revealed that there is considerable variability in the design and implementation of ACs (cf. Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Thornton, 2009). However, currently only limited empirical evidence is available concerning the potential trade-offs between different design factors that are related to both the costs of ACs and the accuracy of the performance evaluations from these ACs.

Therefore, in the present research we considered three AC design factors that are of importance in this regard: Assessor team size, assessor training, and assessor background. Assessor training and assessor background are related to the expertise of the assessors and they have both been shown to affect rating accuracy in ACs (e.g., Lievens, 2001a). In contrast to this, increasing the size of the assessor team might serve as a potential means to compensate for lack of expertise. However, to our knowledge, so far no study has investigated the effects of assessor team size on rating accuracy in AC exercises. Thus, it remains unknown whether increasing the size of the assessor team is indeed a viable way to improve the accuracy of the evaluations from ACs in comparison to factors related to assessor expertise. Therefore, we evaluated whether increasing the size of the assessor team can compensate for lack of expertise. Specifically, we aimed to investigate the effects of assessor team size on the accuracy of ratings in an AC exercise and to compare its effects to those of assessor training and assessor background. At a practical level, our result shall provide AC users and developers with guidance concerning these design issues.

Assessor Team Size

We expect the size of the assessor team to be related to the accuracy of their ratings. When multiple assessors rate a candidate and when ratings from these assessors are aggregated, this should lead to more accurate ratings as compared to ratings from single assessors because the aggregation over multiple measurements is designed to improve behavioral prediction. That is, aggregation of ratings over judges "reduces error of measurement associated with the idiosyncrasies of different judges" (Epstein, 1983, p. 368). More precisely, for aggregated ratings psychometric theory states that error components are divided by the number of assessors, which results in a larger proportion of true variance in comparison to ratings from a single assessor (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Thus, we posit that enlarging assessor teams and aggregating their ratings should serve as a potential means to improve rating accuracy in ACs.

Assessor Expertise

Meta-analyses have shown that assessor characteristics such as expertise moderate AC validity so that ratings provided by assessors with more expertise show better criterion- and construct-related validity (e.g., Gaugler et al., 1987; Woehr & Arthur, 2003). These results are in line with the expert model underlying ACs (Lievens & Klimoski, 2001). According to this model, expert assessors benefit from well-established cognitive structures when observing and evaluating candidates, whereas assessors without expertise do not. These well-established structures guide the attention, categorization, integration, and recall of observed behavior and enable expert assessors to better cope with the high cognitive demands of the rating task. Consequently, expert assessors are able to provide more reliable and accurate ratings than assessors with lower expertise, which results in higher AC validity for the former group. Two important factors that contribute to expertise include (a) assessor training and (b) assessor background.

Assessor training. Several training approaches for improving rating accuracy have been suggested (Bernardin, Buckley, Tyler, & Wiese, 2000; Woehr & Huffcutt, 1994). For example, behavior observation training (BOT), which is based on the assumption that inaccurate ratings stem from a lack of behavioral information, focuses on the improvement of the observation process (i.e., detection, perception, and recall of relevant behavior). In BOT, assessors are instructed to distinguish between observation and evaluation. Furthermore, BOT stresses the importance of being a good observer, of *focusing on* actual behavior, and of taking *notes of* behaviors that are observed.

Conversely, the major purpose of frame-of-reference (FOR) training consists of imposing a common performance theory on raters, thereby establishing a common evaluation standard among assessors. In FOR training, assessors learn to identify behavioral aspects related to the dimensions of interest and to assign observed behavior to the appropriate performance level. Hence, FOR training should particularly foster the correct utilization and evaluation of behavioral cues when providing dimension ratings.

Meta-analyses confirmed that rater training in general has a positive effect on rating accuracy (Roch, Woehr, Mishra, & Kieszczynska, 2012; Woehr & Huffcutt, 1994) and AC validity (Gaugler et al., 1987; Woehr & Arthur, 2003). After BOT or FOR training, assessors provide more accurate ratings than after control training. However, rating accuracy is better after FOR training than after BOT (Lievens, 2001a; Woehr & Huffcutt, 1994). In addition, as compared to untrained assessors, FOR-trained assessors provide more valid and more dimensionally distinct ratings (Schleicher, Day, Mayes, & Riggio, 2002). Research by Schleicher and Day (1998) showed that the improved rating accuracy of FOR-trained assessors is particularly due to reduced idiosyncratic representations of candidates' performance. Similarly, Gorman and Rentsch (2009) found that rating accuracy after FOR training was higher when assessors' performance theory corresponded more closely to the performance theory taught in the training.

Assessor background. In operational ACs, line managers, HR specialists, and psychologists (Eurich et al., 2009; Krause & Gebert, 2003; Krause & Thornton, 2009) typically serve as assessors. It can be assumed that assessors with different backgrounds have different work experience and, therefore, also have different experience with the performance domain. Zedeck (1986) argued that experienced managers have established schemata of job performance that facilitate the evaluation of AC candidates. In line with this, previous performance appraisal research (e.g., Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Kozlowski, Kirsch, & Chao, 1986) confirmed that raters with experience in the performance domain (who thus hold appropriate performance schemata) provide more accurate ratings. For example, personnel administrators provided more accurate ratings of managers' performance in appraisal interviews than MBA students, who, in turn, were more accurate than undergraduates (Cardy et al., 1987). Similarly, in the AC domain, Lievens (2001a) found that managers provided more accurate ratings than psychology students for candidates in an AC exercise (although the former distinguished less between the dimensions than the latter).

Trade-Offs Between Assessor Team Size versus Assessor Expertise

Taken together, it can be assumed that assessor team size, assessor training, and assessor background impact rating accuracy. This means that in the composition of assessor teams, AC users and developers have to carefully decide (a) how many assessors should be in each assessor team, (b) whether (and if so, what kind of) assessor training should be provided, and (c) what background assessors should have. Each of these design decisions has consequences not only for rating accuracy, but also for AC administration and implementation costs. Concerning the number of assessors, it is obvious that multiple assessors are more expensive than single assessors. Accordingly, the larger the assessor team, the higher the costs. With regard to assessor training, costs arise across different stages of the training process and for different requirements, such as trainer fees, equipment, facilities, and material (Noe, 2002).

Finally, regarding assessor background, managerial assessors are relatively expensive as compared to (internal) psychologists or HR professionals. This is because the task of being an assessor is not necessarily part of a manager's job. Hence, in contrast to (internal) psychologists or HR professionals, managers who participate in assessor training or in an AC might invoke indirect extra costs.

Depending on assessor expertise, different numbers of assessors might be needed to reach a particular level of rating accuracy. For example, a larger number of untrained assessors might be able to reach a similar degree of rating accuracy as a smaller team of trained assessors. Thus, increasing the number of assessors in an assessor team might serve to compensate for lack of assessor expertise. Conversely, expertise developed through appropriate assessor training or a specific assessor background might reduce the need for larger assessor teams. Therefore, for AC users and developers, a relevant question is how to weigh the size of the assessor team against assessor expertise so that rating accuracy can be ensured, thereby preventing unnecessary increases in AC costs. To provide AC users with guidance in this regard, we examined the effects of assessor team size on rating accuracy in a common AC exercise. Moreover, we compared the effects of assessor team size to those of assessor expertise factors. Accordingly, we extend previous research on the effects of AC design factors on rating accuracy.

Method

To examine the effects of assessor team size on rating accuracy and to answer the question of whether increasing the size of the assessor team can compensate for missing assessor expertise and vice versa, data obtained in a setting with the following characteristics are required: First, groups of raters with differing expertise provide ratings of candidates' performance. Second, all raters evaluate the same candidates in the same exercise and on the same predefined dimensions to obtain dimension ratings that can be aggregated across assessors with identical expertise. And third, team size varies across a sufficiently large range of values so that it is possible to adequately compare its effect to the effects of assessor expertise.

Data obtained in a setting that fulfills the first two characteristics will enable us to determine rating accuracy for raters with different expertise and for assessor teams of different sizes that is not influenced by the evaluated candidates, dimensions, exercise, or level of candidates' performance. Making use of a simulated selection situation, in which large numbers of raters can be asked to evaluate the same candidate in an exercise (for example, when using standardized videotapes of candidates' performance), seems to be a suitable way to ensure that these two requirements are met.

However, because of the third requirement, an excessive number of participants would be necessary if all three factors are varied experimentally with a sufficiently large sample size in each cell. As we aimed at comparing rating accuracy of assessor teams that consisted of more than two or three assessors, we decided to use a simulation approach in which we simulated teams of differing team size on the basis of data from actual raters.

Concerning the selection of a suitable AC exercise for the present study, we intended to use a type of exercise that is commonly used in ACs and in which the influence of variables other than assessor expertise and candidates' behavior on dimension ratings is reduced to a minimum. In doing so, we intended to ensure that potential effects of assessor expertise are not distorted by effects of other potentially influencing factors (e.g., other candidates) and that error variance in the data is minimized. Therefore, we decided to use an exercise in which assessors have the opportunity to concentrate on a single candidate and in which they do not face stimuli provided by other persons than the candidate (e.g., discussion partners in a group discussion). This should keep the cognitive demands posed on assessors comparably low and thus restrict potential effects of variables on the assessor side related to cognitive ability. Furthermore, the amount of stimuli that might erroneously affect the evaluation of candidate's performance (for example, reactions of interaction partners in a group discussion or an interview, respectively) is minimized. The presentation exercise fulfills these requirements and, furthermore, represents one of the most popular exercises used in ACs (Eurich et al., 2009; Krause & Thornton, 2009).

Existing data from a study by Lievens (2001a) were suitable for the present purpose. In his study, Lievens explored the effects of two factors that contribute to assessor expertise (assessor training and assessor background). He collected data in a simulated selection situation setting where he used standardized stimulus materials. Specifically, all assessors evaluated the same videotaped candidates on the same predefined dimensions in the same presentation exercise and these candidates differed in their performance. Thus, taken together, the available data fulfilled the first two requirements described above.

On the basis of the actual ratings obtained by Lievens (2001a), we simulated assessor teams of different sizes and with different expertise. Specifically, we used these data and extended the 3 (Assessor training: BOT, FOR training, or control training) × 2 (Assessor background: managers or I/O psychology students) design with the third factor that was the main focus of the present study, namely assessor team size. More precisely, to fulfill the third requirement described above, we determined rating accuracy for single assessors and for teams of two to ten assessors with different expertise and thus assessor team size had ten levels (i.e., between 1 and 10 assessors). This led to a 3 × 2 × 10 design. On the basis of the results obtained with this design, we were able to draw conclusions that substantially extend those from Lievens (2001a).

Sample and Procedure

Data from 390 participants were available. Seven participants were excluded from our analyses because of missing data. Thus, the final sample consisted of 225 advanced I/O psychology students (130 women, 95 men) and 158 managers (35 women, 123 men) enrolled in an executive MBA program. More than half of the students had been attending university for at least four years and all students had a Bachelor's degree in psychology. Although some students had experience in psychological consulting firms or in a company's personnel departments, all were inexperienced as assessors. The managers had more than eleven years of working experience on average in different functional backgrounds. However, none of them had served as an assessor in the past (cf. Lievens, 2001a).

Participants were told to assume the role of assessors for the selection of a district sales manager. Then, they received general information about ACs and a description of the job of the district sales manager and the organization. Afterwards, participants were randomly assigned to one of three training conditions: BOT, FOR, or control training.

In the BOT condition, participants were taught to distinguish between observation and evaluation, and to improve the processes of observing and recording behavior. First, they were instructed to make behavioral instead of non-behavioral descriptions of candidates' behavior. Then, participants learned to classify behavior into dimensions on the basis of the dimension definitions. Next, the trainer instructed participants to provide dimension ratings according to the amount of behavioral observations made. Participants practiced recording, classifying, and rating with a videotaped candidate. Afterwards, the behaviors that were used to provide dimension ratings were discussed and discrepancies among ratings were clarified. Finally, participants received feedback pertaining to their ratings.

In FOR training, the aim was to establish a common frame-of-reference for evaluating AC candidates. To this end, the trainer presented definitions of the dimensions and gave examples of normative behaviors for different levels of performance. Afterwards, participants completed a written exercise in which they had to assign behavioral incidents to one of three dimensions and to one of three performance levels. Then, the answers were discussed and feedback was provided to participants. Next, participants practiced the rating task with a videotaped candidate. Their dimension ratings were discussed and discrepancies among ratings were clarified. Finally, the trainer provided participants with feedback regarding their ratings.

Participants in the control condition were told that they were expected to watch videotaped AC candidates, to take notes if necessary, and to evaluate the candidates. Then, participants observed and evaluated a videotaped candidate. However, their ratings were not discussed and no feedback was provided. Hence, participants in the control condition did not get a specific preparation for rating AC candidates and thus were untrained.

After their respective training, participants observed four videotaped candidates who had to deliver a sales presentation (also see Lievens, 1999, for additional information concerning the development of the videotapes). All participants were unfamiliar with the specific presentation exercise. In this sales presentation, candidates had to present an analysis of the buyer's needs and argue which of three software systems was most appropriate. The presentation was given to a panel of decision makers who asked questions to challenge the candidate. The candidates were semiprofessional actors who performed according to pre-specified scripts. The scripts were written on the basis of predefined performances on three dimensions. The predefined performances were later used as true scores to determine rating accuracy as described in the following paragraph. The three dimensions were problem analysis and solving, interpersonal sensitivity, and planning and organization. After every videotaped presentation, each participant had to rate the candidate on the three dimensions using a five-point scale (ranging from 1 = *poor* to 5 = *excellent*). For more details of the procedure we refer to Lievens (2001a).

Rating Accuracy

To investigate the effects of assessor team size, we used the ratings provided by the participants and sampled a total of 1000 assessor teams (with replacement after each team was sampled) for all cells of the study design with between two to ten assessors. For example, we randomly drew 1000 teams of ten assessors in such a way that each assessor could be sampled in multiple teams. Afterwards, for each of the 1000 teams, we calculated average ratings for the three rating dimensions.

Then, we determined rating accuracy for single assessors and for teams between two to ten assessors. Rating accuracy refers to deviations between the assessors' ratings and comparison scores (cf. Sulsky & Balzer, 1988). Therefore, we compared the dimension ratings obtained with the predefined performances on which the scripts for the candidates were based to determine Borman's differential accuracy (BDA: Borman, 1977). BDA reflects the correlation between ratings and true scores (i.e., the predefined performances that the scripts were based on) across candidates, averaged across dimensions. Hence, BDA is a correlational measure of rating accuracy that represents an index of rater validity (Sulsky & Balzer, 1988) and can be expressed by the following equation:

$$BDA = 1/d \sum_{j=1}^{d} (T_{rt})_j$$

where d refers to the number of dimensions and T_{rt} refers to the correlation between ratings r and true scores t for a particular dimension j (Sulsky & Balzer, 1988). Before computing BDA, all correlations T_{rt} are transformed to Z scores.

Results

Effects of Assessor Team Size, Assessor Training, and Assessor Background

Figure 1 shows the mean rating accuracy for each cell of the study design. To determine the effects of assessor team size, assessor training, and assessor background on rating accuracy, we conducted a three-way analysis of variance (ANOVA) with BDA as the dependent variable. In order to keep the cell sizes balanced, we only used cells with a sample size of 1000 for the analyses, that is, only cells with two to ten assessors per assessor team. This resulted in a 3 × 2 × 9 design for the ANOVA.

In line with psychometric theory, assessor team size had a significant main effect on BDA, $F(8, 53946) = 906.76$, $p < .01$, $\eta^2 = .07$. According to conventional standards (cf. Cohen, 1988), this reflects a moderate effect size. The significant effect of assessor team size indicates that rating accuracy usually increased with an increasing number of assessors in the assessor team (cf. Figure 1).

Furthermore, the three-way ANOVA yielded a main effect for assessor training on BDA, $F(2, 53946) = 6824.31$, $p < .01$, $\eta^2 = .14$, with a large effect size. As shown in Figure 1, assessors who had taken part in FOR training provided the most accurate ratings and untrained assessors provided the least accurate ratings.

Surprisingly, the main effect for assessor background was not significant, $F < 1$, indicating that in general there was no difference between advanced psychology students and managers. However, this does not mean that assessor background did not influence rating accuracy because all interaction effects involving assessor background were significant. Specifically, the interaction between assessor background and assessor training was significant and had a moderate effect on BDA, $F(2, 53946) = 4216.62$, $p < .01$, $\eta^2 = .08$, indicating that the effect of assessor training differed between managers and psychology students. In addition, several interaction effects including assessor team size were significant: The interaction between assessor team size and assessor background had a large effect on BDA, $F(2, 53946) = 1853.47$, $p < .01$, $\eta^2 = .15$, indicating that the effect of increasing the size of the assessor team differed between managers and psychology students. Furthermore, both the interaction between assessor team size and assessor training, $F(16, 53946) = 73.72$, $p < .01$, $\eta^2 = .01$, and the three-way interaction between assessor team size, assessor training, and assessor background, $F(16, 53946) = 73.70$, $p < .01$, $\eta^2 = .01$, had small but significant effects.

To further explore the source of the interaction effects between the investigated factors, we conducted additional analyses. One-way ANOVAs with assessor training as the independent variable revealed that the training effect was larger for managers than for psychology students (Table 1). Thus, managers benefited more from assessor training than psychology students. Furthermore, the effect for assessor training became more pronounced with a larger assessor team. This means that the difference in rating accuracy between untrained and trained assessors increased with an increasing size of the assessor team, especially for managers (also see Figure 1).

Concerning assessor team size, a larger team was associated with a higher rating accuracy in general. However, there was one noteworthy exception from this general pattern. Specifically, untrained managers soon reached asymptotic values of rating accuracy and then did not show additional improvements with increasing size of the assessor team. In line with this, the one-way ANOVAs with assessor team size as the independent variable showed that the effect of assessor team size was not even half as large for managers in the control condition as it was in any of the other cells (Table 2), indicating that for untrained managers the effect of increasing the assessor team size was limited. Furthermore,

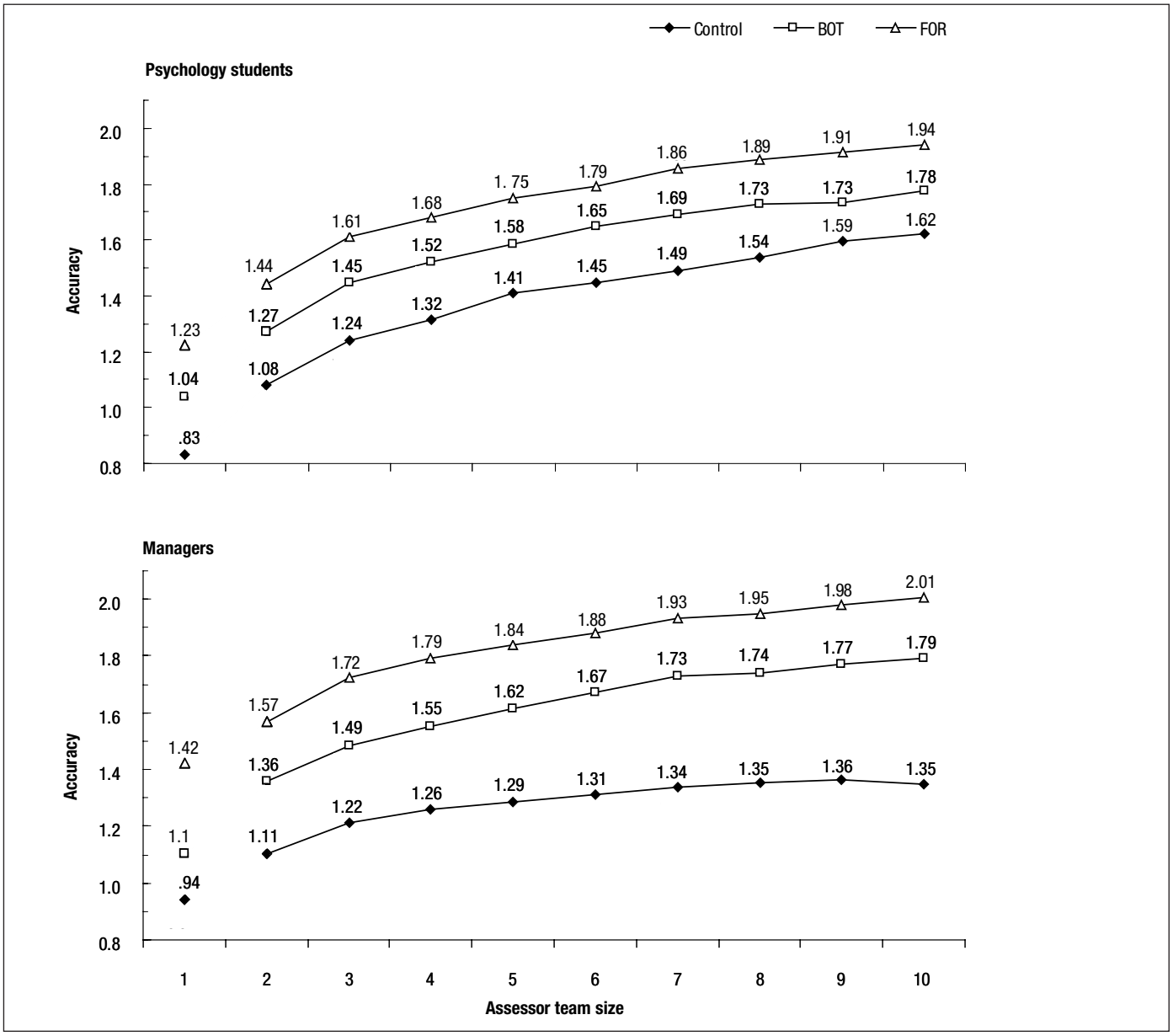

Figure 1. Average rating accuracy (Borman's differential accuracy, BDA) by assessor team size and by training condition. Higher scores indicate better accuracy. Cell-specific *n* for single psychology students (Control, *n* = 86; BOT, *n* = 73; FOR, *n* = 66) and for single managers (Control, *n* = 45; BOT, *n* = 61; FOR, *n* = 52). *n* = 1000 for all cells with a team size of ≥ 2.

compared to untrained managers, rating accuracy for untrained psychology students improved continuously with increasing size of the assessor team. Therefore, larger teams of untrained psychology students provided more accurate ratings than untrained managers even though single managers were more accurate than single students (Figure 2). This resulted in a significant interaction between assessor team size and assessor background when we conducted an Assessor team size x Assessor background ANOVA in the control condition, $F(8, 17982) = 33.47$, $p < .01$, $\eta^2 = .01$.

Examination of Trade-Offs

First, we evaluated whether increasing the size of the assessor team can compensate for missing assessor training and vice versa. Concerning psychology students, increasing the size of the assessor team was a means to compensate for missing BOT and FOR training. For example, on average, two untrained students reached the accuracy level of a single student with BOT and three untrained students reached the accuracy level of a single student with FOR training (Figure 1). In contrast to psychology students, increasing the size of the assessor team consisting of managers could only partly compensate for missing BOT, and it was not a suitable means to compensate for a lack of FOR training. Specifically, to reach the average accuracy of a single manager with BOT, two untrained managers sufficed. However, untrained managers were not able to reach the accuracy level of a single FOR-trained manager, even when ratings were aggregated within teams of ten assessors. Similarly, large numbers of untrained managers were also not able to outperform two managers with BOT.

Second, concerning assessor background, increasing the size of the assessor team could compensate for using assessors with a suboptimal background (i.e., psychology students) in all training conditions. Specifically, in all three training conditions, ratings from two and three psychology students were at least as accurate as ratings from one and two managers, respectively. In the BOT and FOR conditions, managers were more accurate than psychology students,

Table 1
Results From one-way ANOVAs with Assessor Training as the Independent Variable for Each Combination of Assessor Team Size and Assessor Background

Assessor team size	Psychology students		Managers	
	$F(2, 2997)$	η^2	$F(2, 2997)$	η^2
2	141.47**	.086	218.07**	.127
3	146.03**	.089	283.62**	.159
4	229.22**	.133	533.13**	.262
5	197.41**	.116	674.32**	.310
6	256.35**	.146	725.80**	.326
7	270.34**	.153	876.44**	.369
8	293.23**	.164	1049.50**	.412
9	239.34**	.138	1157.36**	.436
10	251.15**	.144	1491.10**	.499

Note. ** $p < .01$.

Table 2
Results From one-way ANOVAs with Assessor Team Size as the Independent Variable for Each Combination of Assessor Background and Assessor Training

Assessor background	FOR		BOT		Control	
	$F(8, 8991)$	η^2	$F(8, 8991)$	η^2	$F(8, 8991)$	η^2
Psychology students	217.64**	.162	169.10**	.131	193.00**	.147
Managers	146.60**	.115	154.30**	.121	57.49**	.049

Note. ** $p < .01$.

irrespective of the size of the assessor team, but the difference in accuracy between trained psychology students and trained managers decreased continuously with an increasing assessor team size. Conversely, with an increasing assessor team size, relations between rating accuracy of psychology students and managers changed in the control condition as already noted above. Thus, even though single untrained managers were more accurate than single untrained psychology students, untrained students provided more accurate ratings than untrained managers in the case of teams of three or more assessors.

Discussion

The present study is the first that examined the effects of assessor team size on rating accuracy in an exercise that is commonly used in ACs (a presentation exercise) and it also compared them with effects of two factors associated with assessor expertise (assessor training and assessor background). Thus, this study extends previous research on the effects of assessor expertise on rating accuracy and allows drawing conclusions concerning the trade-offs between assessor expertise and assessor team size in an AC context. At a practical level, this study provides evidence-based guidance regarding decisions about these three factors for the design and administration of operational ACs.

Of the three factors, assessor training had the largest main effect on rating accuracy in the presentation exercise. In line with psychometric theory, rating accuracy improved when ratings were aggregated across multiple assessors. Concerning assessor expertise, we did not find a significant main effect. Apart from these main effects, we also looked at trade-offs between these factors. In particular, we examined whether increasing the size of the assessor team could compensate for missing assessor training. Along these lines, an important finding of the present study is that rating accuracy only improved to a limited degree with an increasing size of the assessor team consisting of untrained managers; even teams of ten untrained managers were unable to reach the same level of rating accuracy as a single FOR-trained manager. A possible explanation for the limited effect of increasing the number of untrained managers is that single untrained managers had difficulty differentiating between dimensions (see also Lievens, 2001a; 2001b) and rated candidates holistically. Our results suggest that inaccuracies of ratings due to such a holistic rating approach can be reduced only to a limited degree by aggregating multiple ratings. However, as compared to increasing the number of untrained managers, assessor training –

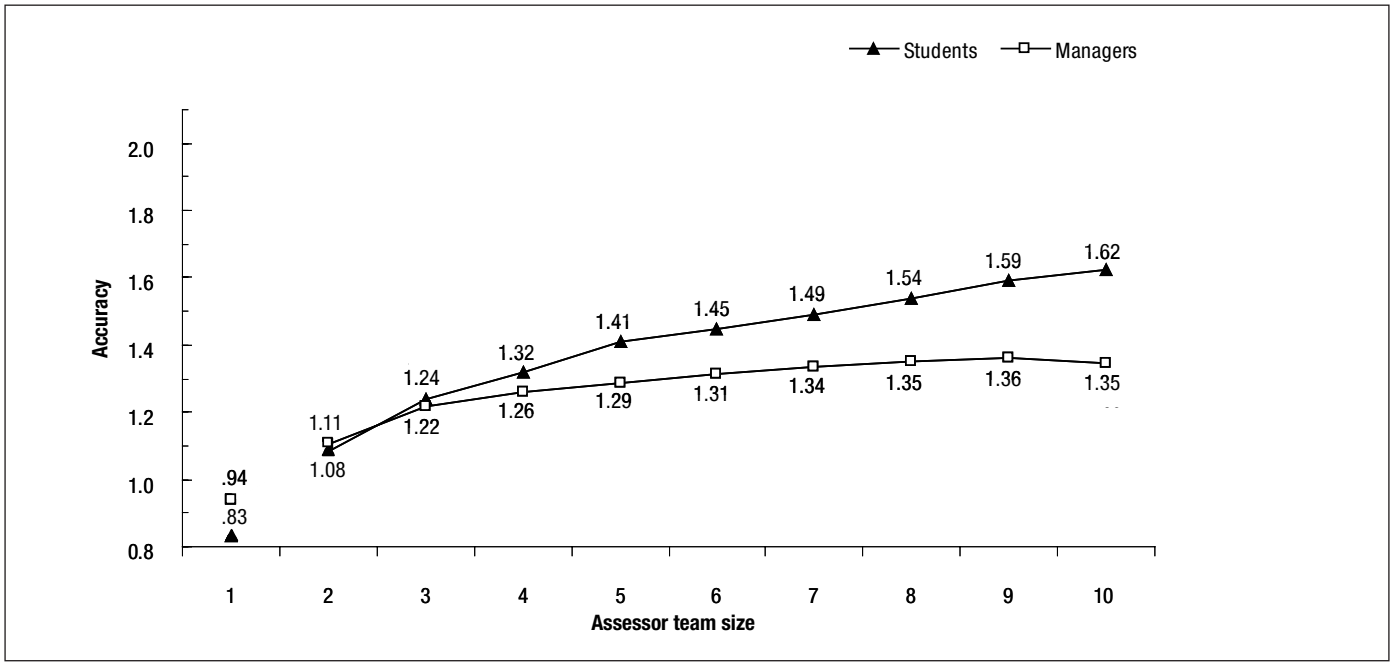

Figure 2. Average rating accuracy (Borman's differential accuracy, BDA) by assessor team size and by assessor background. Higher scores indicate better accuracy. Cell-specific *n* for single assessors (Students, $n = 86$; Managers, $n = 45$). $n = 1000$ for all cells with a team size of ≥ 2.

and especially FOR training – seems to be an effective means to overcome a holistic rating approach and to improve managers' rating accuracy.

In contrast to untrained managers, rating accuracy of untrained psychology students continuously improved with an increasing size of the assessor team. However, given that assessor training had the largest main effect of the three independent variables on rating accuracy, large teams of untrained psychology students were needed to reach the level of rating accuracy of a smaller team of trained psychology students, especially when the latter had taken part in FOR training. For example, to reach the accuracy of one, two, or three FOR trained psychology students, three, six, or ten untrained psychology students were needed, respectively. Thus, with respect to AC costs, increasing the size of the assessor team as a means to compensate for a lack of FOR training might triple personnel costs. Furthermore, it is unrealistic to assume that more than three or four assessors are used in operational ACs to evaluate a candidate's performance in an exercise (cf. Arthur & Day, 2010; Krause & Thornton, 2009) – either because of increased AC costs or because of decreased feasibility with an increasing number of assessors.

Finally, we investigated whether increasing the size of the assessor team could compensate for assessor expertise. Our results suggest that using larger teams of assessors can indeed compensate for missing expertise related to assessor background. More specifically, just two and three psychology students reached the level of rating accuracy of one and two managers, respectively. Thus, using a somewhat larger number of assessors with a suboptimal background might be a viable way to ensure rating accuracy, for example, under conditions when not enough assessors with sufficient expertise are available. At the same time, such moderate increases in the assessor team size might also help to keep AC costs under control.

As argued above, we used a presentation exercise because cognitive demands posed on assessors are assumed to be lower than in other types of AC exercises. Findings from previous studies imply that increased cognitive demands are associated with lower accuracy of the ratings obtained (cf. Melchers, Kleinmann, & Prinz, 2010). Therefore, in exercises with higher cognitive demands, assessor expertise as well as increasing the number of assessors and aggregating their ratings might be even more important than in a presentation exercise.

Practical Implications

The present study provides at least three pieces of advice to AC users and designers. First, although assessor training is associated with higher costs for ACs, it should be an inherent part of an AC program because it is an effective means to improve rating accuracy. A training in which a common frame of reference for evaluating AC candidates is imposed on assessors is especially recommended (see also Lievens, 2001a; Roch et al., 2012; Schleicher et al., 2002; Woehr & Huffcutt, 1994). When taking into account that assessors can use competencies gained through assessor training each time they are employed as assessors again, the benefit of appropriate assessor training will probably outweigh training costs in the long term. Moreover, recent research has shown that beneficial effects of assessor training can also transfer to other performance appraisals (Macan et al., 2011) and thus go beyond the improvement of rating accuracy in ACs.

Second, if no training is provided to assessors, increasing the size of the assessor team can improve rating accuracy in an AC to some degree. However, increasing the size of the assessor team is a means to compensate for missing FOR training only if assessors have a psychological background, but not if they are managers. Yet, even if assessors have a psychological background, appropriate assessor training is probably more cost-efficient in the long-term than using larger groups of untrained assessors. Furthermore, as noted above, the effect of appropriate assessor training can transfer to other contexts such as performance appraisals (Macan et al., 2011), whereas the effect of increasing the size of the assessor team is limited to a single AC. However, when an AC is only administered once or twice, increasing the size of the assessor team might be cheaper than providing extensive assessor training.

Third, the present results do not allow us to conclude whether it is more advantageous to use managers versus individuals with a psychological background as assessors in an AC. Rather, our results imply that assessors with differing backgrounds have different perspectives that might both contribute to a valuable evaluation of candidates' performance (e.g., Damitz, Manzey, Kleinmann, & Severin, 2003). As the different perspectives due to differing assessor backgrounds "are expected and welcomed as a part of the principle of multiple assessors" (Thornton & Rupp, 2006, p. 42), we recommend using trained assessors with diverse backgrounds (see also the guidelines of the International Task Force on Assessment Center Guidelines, 2009).

Concerning the composition of assessor teams in ACs, rating accuracy, cost aspects, and sustainability might be decisive factors. However, in addition to these factors, AC users and designers should also take "political" issues at the organizational level into account. On the one hand, for example, organizational guidelines might prescribe that specific organizational members (e.g., line managers) serve as assessors. On the other hand, the composition of assessor teams might also have consequences for the acceptance of decisions that are based on AC ratings.

Limitations and Suggestions for Future Research

Some limitations of this study should be mentioned. First, the analyses were based on data from a simulated selection situation setting. On the one hand, this enabled us to sample assessor teams of large sizes, which is not possible in an applied setting. On the other hand, the nature of the stimulus materials did not allow us to determine construct- and criterion-related validity of the dimension ratings. Therefore, we focused on rating accuracy as the dependent variable. However, previous findings suggest that factors that improve rating accuracy usually also lead to improvements in construct- and criterion-related validity (Gaugler & Thornton, 1989; Lievens, 2001a; Melchers et al., 2010; Schleicher et al., 2002). Therefore, we assume that assessor team size, assessor training, and assessor background have similar effects on the construct- and criterion-related validity of ACs as they have on rating accuracy. Nevertheless, future research is needed to confirm this assumption.

Second, in the present study managers and I/O psychology students served as assessors, respectively. Hence, it is unclear to what degree our results generalize to professional psychologists (or experienced HR professionals in general) who have specialized in conducting ACs and who regularly serve as assessors. In light of previous findings (e.g., Gaugler et al., 1987; Sagie & Magnezy, 1997; Woehr & Arthur, 2003), we would expect professional psychologists to generally outperform managerial assessors in terms of rating accuracy. Furthermore, the effect of assessor training might be less pronounced for professional psychologists because professional psychologists might hold more appropriate performance schemata in the first place due to their background as well as experience.

Finally, the exercise used in this study was a presentation exercise in which assessors observed only one candidate. When assessors have to observe multiple candidates simultaneously as, for example, in a group discussion, cognitive demands increase and thus inaccuracies in ratings are also likely to increase (cf. Melchers et al., 2010). Therefore, as mentioned above, assessor expertise as well as increasing the number of assessors and aggregating their ratings might be particularly important in exercises with increased cognitive

demands. Future research with regard to this issue is needed. Furthermore, future research might examine the effects of assessor team size and factors related to assessor expertise in different types of AC exercises and compare them against each other with regard to their impact on the quality of the ratings from the different exercises as well as of the overall assessment rating for the entire AC.

Conflicts of interest

The authors of this article declare no conflicts of interest.

Financial support

The research reported in this paper was supported by a grant from the Swiss National Science Foundation (Schweizerischer Nationalfonds; Grant 100014-117917) to Klaus Melchers and Martin Kleinmann.

References

Arthur, W., Jr., & Day, E. A. (2010). Assessment centers. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (Vol. 2, pp. 205-235). Washington, DC: APA.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi: 10.1111/j.1744-6570.2003.tb00146.x

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A reconsideration of strategies for rater training. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 18, pp. 221-274). Greenwich, CT: JAI Press.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252. doi: 10.1016/0030-5073(77)90004-6

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational and Organizational Psychology, 60*, 197-205. doi: 10.1111/j.2044-8325.1987.tb00253.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsday, NJ: Lawrence Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology: An International Review, 52*, 193-212. doi: 10.1111/1464-0597.00131

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*, 254-270. doi: 10.1111/j.1468-2389.2009.00468.x

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51*, 360-392. doi: 10.1111/j.1467-6494.1983.tb00338.x

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology, 24*, 387-407. doi: 10.1007/s10869-009-9123-3

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi: 10.1037/0021-9010.72.3.493

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618. doi: 10.1037/0021-9010.74.4.611

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*, 1336-1344. doi: 10.1037/a0016476

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion-related validity: A meta-analytic update*. Unpublished manuscript.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi: 10.1111/j.1468-2389.2007.00399.x

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi: 10.1111/j.1468-2389.2009.00467.x

Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: An exploration. *Journal of Applied Psychology, 71*, 45-49. doi: 10.1037/0021-9010.71.1.45

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center practices in organizations in German-speaking regions and the United States. *International Journal of Selection and Assessment, 11*, 297-312. doi: 10.1111/j.0965-075X.2003.00253.x

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371. doi: 10.1111/j.1468-2389.2006.00357.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi: 10.1111/j.1464-0597.2008.00371.x

Lievens, F. (1999). Development of a simulated assessment center. *European Journal of Psychological Assessment, 15*, 117-126. doi: 10.1027//1015-5759.15.2.117

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264. doi: 10.1037/0021-9010.86.2.255

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*, 203-221. doi: 10.1002/job.65

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245-286). Chichester, UK: Wiley.

Macan, T., Mehner, K., Havill, L., Meriac, J. P., Roberts, L., & Heft, L. (2011). Two for the price of one: Assessment center training to focus on behaviors can transfer to performance appraisals. *Human Performance, 24*, 443-457. doi: 10.1080/08959285.2011.614664

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105-115. doi: 10.1024/1421-0185/a000012

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in assessment center group discussions. *International Journal of Selection and Assessment, 18*, 329-341. doi: 10.1111/j.1468-2389.2010.00516.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi: 10.1037/0021-9010.93.5.1042

Noe, R. A. (2002). *Employee training and development* (2nd ed.). New York: McGraw-Hill.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*, 370-394. doi: 10.1111/j.2044-8325.2011.02045.x

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103-108. doi: 10.1111/j.2044-8325.1997.tb00634.x

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101. doi: 10.1006/obhd.1998.2751

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. doi: 10.1037/0021-9010.87.4.735

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506. doi: 10.1037/0021-9010.73.3.497

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi: 10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi: 10.1111/j.2044-8325.1994.tb00562.x

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in organizational behavior, 8*, 259-296.