Research Article

Industrial & Organizational Psychology Studies

Ind Organ Psychol Stud
Fall & Winter 2025, 12(2), 1-20
https://doi.org/10.22055/jiops.2025.49140.1476

Received: 01.03.2025
Received in revised form: 04.04.2025
Accepted: 12.05.2025

Estimating the Reliability & Validity of Assessment Center Ratings Using Generalizability Theory

Majid Salimi [1] ∗, Hassan Moshtaghian Abarghoei [2]

1. PhD of Measurement & Measurement, Allameh Tabatabaei University, Head of Measurement & Evaluation Center, National Iranian Oil Company, Tehran, Iran
2. PhD of Measurement & Evaluation, Allameh Tabatabaei University of Education, Shiraz, Iran

Citation: Salimi, M., & Moshtaghian Abarghoei, H. (2025). Estimating the reliability and validity of assessment center ratings using generalizability theory. *Industrial and Organizational Psychology Studies, 12*(2), 1-20. doi: 10.22055/jiops.2025.49140.1476

Abstract

Introduction

Over the past few decades, assessment centers have become a popular method of assessing behavior for recruitment, recognition, & development around the world (Krause, Rossberger, Dowdeswell, Venter, Joubert, 2011; Lowry, 1997; Thornton & Byham, 1982). This popularity can be attributed to several factors, including; Multiple methods used in assessment centers have been shown to have little adverse effect & to be able to predict a variety of performance measures (Thornton & Rupp, 2006), to have high predictive validity coefficients ranging from 0.37 to 0.52 (Gaugler, Rosenthal, Thornton & Bentson, 1987; Rupp Thornton & Gibbons, 2008; Thornton & Gibbons, 2009), to demonstrate high criterion-related validity (Schmitt, Gooding, Noe & Kirsch, 1984) while also having content validity (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Iles, 1992). In addition, simulations, which are an essential part of assessment centers, are more accurate in assessing candidates' job situations & conditions (Lievens & Schollaert, 2011; Thornton & Rupp, 2006). Although assessment centers are usually costly in terms of time, effort, & money, their excellent predictive validity & other strengths make them worthwhile (Krause, et al., 2006). Accordingly, there is general agreement that assessment center ratings predict job-relevant measures such as job performance or educational success (Hermelin, Lievens & Robertson, 2007; Sackett, Shewach, & Keiser, 2017). Undoubtedly, the use of the results of the ratings of applicants in the assessment center is acceptable when the collected data have adequate reliability & validity. Without an acceptable level of reliability, the question of the validity & appropriateness of the ratings is limited. On the other hand, without validity, the results of the assessment cannot be used (Whetzel, & Wheaton, 2007). Regardless of the purpose for which an assessment center is designed, evidence supporting validity in the main constructs measured is always relevant & meaningful (Thornton,

∗ **Corresponding Author:** Majid Salimi
Email: salimy28@yahoo.com

Authors retain the copyright and full publishing rights.
Published by Shahid Chamran University of Ahvaz. This article is an open access article licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0)

Rupp, & Hoffman, 2015). Of course, it is generally assumed that content validity, criterion validity, & construct validity & their estimation are related to each other, & it can be logically assumed that the existence of two of them necessarily proves the third (Arthur, Woehr, & Maldegen, 2000). Therefore, if a center has content validity & criterion validity, it can be logically assumed that this center also has construct validity. Although research has shown the content & criterion validity of assessment centers, the results regarding the construct validity of assessment centers have been less promising (Thornton, & Mueller-Hanson, 2004; Lievens & Christiansen, 2012; Jackson, Michaelides, Dewberry, Nelson, & Stephens, 2022). Accordingly, some researchers have focused on a variety of methodological approaches to address the construct validity problem of assessment centers. However, these efforts have failed to address the construct validity problem of assessment centers. Some researchers have suggested eliminating dimensions from assessment center studies & focusing only on exercises (Jackson, Barney, Stillman, & Kirkley, 2007; Neidig, & Neidig, 1984; Dewberry, 2024). Such a view is probably unacceptable because it contradicts the main purpose of assessment centers, which is to measure the position of assessees on various dimensions required for job performance (Bowler, & Woehr, 2006, 2009). According to some researchers, the lack of clear separation of sources of variance has led to the wandering of efforts to assess the construct validity of assessment centers. To establish construct validity, both convergent & discriminant validity are required (Campbell, & Fiske, 1959). Convergent validity is observed when scores on the same construct are consistent despite the use of different measurement methods (e.g., different raters, multiple measures, etc.) (Kraiger & Teachout, 1990; Marcoulides, 1989). Discriminant validity is observed when individuals are differentially rated on different trait measures of the measure (Kraiger & Teachout, 1990). Generalizability theory provides a way to examine both types of convergent/discriminant validity(Kraiger & Teachout, 1990 & Winfred, Woehr, & Maldegen, 2000). The main objective of this study was to use generalizability theory to examine the reliability & convergent/diagnostic validity of ratings from a government organization's assessment center.

Method

To study generalizability, data on 340 managers, who were rated on a 9-point scale (1 to 9) based on five competency dimensions (planning, leadership, problem recognition, innovation, & decision-making) by six raters using different methods (group exercise, group discussion, interview, & questionnaire), were extracted from the organization's Manager Competency Database & analyzed based on a completely randomized crossed design with 5 dimensions (D), 6 methods (E), & 340 managers (P) using the EduG program (Cardinet, Johnson, & Pini, 2010). In the present study, variability due to individuals does not contribute to the formation of error variance because individuals are the objects being measured & are expected to appear differently; therefore, a sufficiently large variance component for individuals indicates convergent validity (Kraiger & Teachout, 1990; Marcoulides, 1989), & the degree of consistency in the ratings of managers by different raters provides evidence of discriminant validity (Kraiger & Teachout, 1990).

Results

The variance component estimates showed that the main effect of dimensions accounted for 13.4% of the variance. This is evidence that performance varies across individuals in terms of the dimensions measured. While the value attributed to raters (methods) is about 0.3%, indicating that ratings made by raters vary very little across measurement methods (group exercises, interviews). The highest percentage of variance in the main effects belonged to the individual factor, so that this effect accounted for 29% of the total variance, which indicates that a high proportion of the variance in ratings is due to differences between individuals in terms of the competencies measured in this design. In other words, it can be said that the center's evaluators generally acted differently in evaluating managers in terms of the competencies measured & rated their performance differently, or that managers were different in terms of the competency under study. Since in this study, managers

constitute the object of measurement & not the error, such variability is desirable. Finally, a little more than 36% of the total variance is related to the residual or address. That is, the proportion of variance that is accounted for by the three-way interaction of individuals, raters (methods), & dimensions, or other systematic or unsystematic sources of variability that are not measured in this design. Therefore, there are other important sources of variance that are not accounted for by differences between managers, differences between raters (methods), differences between dimensions, or their two-way interaction, & are considered undesirable in this design. In accordance with the convergent-diagnostic validity approach using the components of generalizability theory (Kraiger & Teachout, 1990; Marcoulides, 1989; Winfred, et al., 2000), examination of the estimated variance components for the main effects & interaction effects yielded a relatively clear pattern in support of the validity of the ratings of the studied dimensions. The main effect related to dimensions accounted for a relatively high percentage of the total variance. This indicates that performance varies along the dimensions under study; also, the interaction effect between dimensions & individuals explained slightly more than 15% of the total variance. The variance associated with this interaction indicates the extent to which individuals' performance is segregated & different along the dimensions. This component essentially indicates the discriminant validity across dimensions; therefore, performance along the dimensions has a relatively high variability; That is, as expected, the ratings well reflect the latent individual differences in the dimensions under study (evidence of diagnostic validity). Also, the findings showed that 29% of the variance in the total ratings is due to the main effect of individuals. This effect indicates that the evaluators acted differentially in evaluating the performance of the evaluation center managers.

Discussion

According to the findings, the differential variance estimate (0.44) is much higher than the sum of the relative error variances (0.11) or the sum of the absolute error variances (0.16) of the instrumental aspects. Also, the values of the two relative & absolute G coefficients were obtained as 0.80 & 0.73, respectively. The value of the relative coefficient is around the conventional value of 0.8 (Cardinnet, 2010). Based on this coefficient, 80% of the observed variance in the average ratings is attributed to the true score & 20% is related to the interaction between individuals × dimensions × rater & unspecified aspects that reduce the accuracy of individual results; therefore, in general, the rating methods & measurement accuracy with this number of measurement dimensions & method/rater are relatively satisfactory & can provide a reliable rating of individuals' performance.

Keywords: Assessment center, generalizability theory, reliability, convergent/diagnostic validity

Authors retain the copyright and full publishing rights.
Published by Shahid Chamran University of Ahvaz. This article is an open access article licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0)

مقاله پژوهشی

مطالعات روان‌شناسی صنعتی و سازمانی

پاییز و زمستان ۱۴۰۴، ۱۲(۲)، ۲۰-۱
https://doi.org/10.22055/jiops.2025.49140.1476

تاریخ دریافت: ۱۴۰۳/۱۲/۱۰
تاریخ بازنگری: ۱۴۰۴/۰۱/۱۵
تاریخ پذیرش: ۱۴۰۴/۰۲/۲۲

برآورد پایایی و روایی درجه‌بندی کانون ارزیابی با بهره‌گیری از نظریه تعمیم‌پذیری

مجید سلیمی [۱] *، حسن مشتاقیان ابرقوئی [۲]

۱. دکتری سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، رئیس مرکز سنجش و ارزیابی شرکت ملی نفت ایران، تهران، ایران
۲. دکتری سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، آموزش و پرورش شیراز، ایران

چکیده

نتایج کانون ارزیابی و بهره‌برداری از این نتایج، وابستگی قطعی به پایایی و روایی درجه‌بندی‌های این کانون‌ها دارد. هدف این پژوهش، بررسی پایایی و روایی کانون ارزیابی یک سازمان دولتی با بهره‌گیری از نظریه تعمیم‌پذیری بود. نمونه آماری مشتمل بر داده‌های مربوط به ۳۴۰ مدیر (ارزیابی‌شونده) بود که براساس پنج بعد شایستگی و توسط شش ارزیاب، درجه‌بندی‌شده بود و براساس طرحی کاملاً متقاطع و تصادفی انتخاب شدند و بنابراین براساس ابعاد (D)، ارزیاب/روش (E) و ارزیابی‌شوندگان (P) مورد مطالعه قرار گرفتند. براساس یافته‌ها، مقدار دو ضریب G نسبی و مطلق به‌ترتیب ۰/۸۰ و ۰/۷۳ به‌دست آمد. همچنین براساس تحلیل واریانس، مؤلفه‌های مرتبط با ارزیابی‌شوندگان، ابعاد و تعامل این دو، مجموعاً ۵۸ درصد از واریانس کل را تبیین می‌کرد، درحالی‌که مؤلفه‌های مرتبط با ارزیابان (روش‌ها) و تعامل آن با افراد، تنها ۶ درصد از واریانس کل را تبیین می‌کرد؛ بنابراین مطابق با رویکرد روایی همگرا-تشخیصی، داده‌های این مطالعه از روایی کافی برخوردار بود.

کلیدواژگان: کانون ارزیابی، نظریه تعمیم‌پذیری، پایایی، روایی همگرا/تشخیصی

* **نویسنده مسئول:** مجید سلیمی
رایانامه: salimy28@yahoo.com

Authors retain the copyright and full publishing rights.
Published by Shahid Chamran University of Ahvaz. This article is an open access article licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0)

مقدمه

طی چند دهه گذشته کانون‌های ارزیابی (Assessment Centers) به‌عنوان یک روش محبوب برای ارزیابی رفتار به‌منظور استخدام، تشخیص و توسعه در سراسر جهان مورد توجه بوده است (Thornton & Byham, 1982; Lowry, 1997; Krause et al., 2011). این محبوبیت را می‌توان به عواملی چند نسبت داد از جمله؛ روش‌های چندگانه مورد استفاده کانون‌های ارزیابی که اثر نامطلوب کمی نشان می‌دهند و برخورداری از توانایی پیش‌بینی انواع معیارهای عملکرد (Thornton & Rupp, 2006)، داشتن ضریب روایی پیش‌بینی بالایی در محدوده ۰/۳۷ تا ۰/۵۲ (Gaugler et al., 1987; Rupp et al., 2008; Thornton & Gibbons, 2009)؛ نشان‌دادن روایی بالای مرتبط با ملاک (Schmitt et al., 1984) ضمن برخورداری از روایی محتوا (Gaugler et al., 1987; Iles, 1992). علاوه بر این، شبیه‌سازی‌ها که جز ضروری کانون‌های ارزیابی هستند، نسبت به موقعیت‌ها و شرایط شغلی داوطلبین، در ارزیابی دقیق‌ترند (Thornton & Rupp, 2006; Lievens & Schollaert, 2011). اگرچه کانون‌های ارزیابی معمولاً ازلحاظ زمان، تلاش و پول هزینه‌بر هستند، اما روایی پیش‌بینی‌کننده عالی آن‌ها و سایر نقاط قوتشان سبب ارزشمندی آن شده است (Krause et al., 2006). براین‌اساس توافق کلی وجود دارد که درجه‌بندی کانون ارزیابی معیارهای شغلی مربوطه مانند عملکرد شغلی یا موفقیت در آموزش را پیش‌بینی می‌کند (Hermelin et al., 2007; Sackett et al., 2017). کانون ارزیابی، فرایندی است که مشخصه آن استفاده از روش‌های سنجش چندگانه (Multiple Assessment Methods) است و این امکان را فراهم می‌آورد که چند ارزیاب، انجام تکالیف شغلی شبیه‌سازی‌شده توسط ارزیابی‌شوندگان در شایستگی‌های شغلی مرتبط (ابعاد) را درجه‌بندی نمایند (Gatewood et al., 2010). به عبارت ساده‌تر، در مرکز ارزیابی شوندگانی که مورد ارزیابی قرار می‌گیرند، در چند تمرین (موقعیت‌های شبیه‌سازی شغلی) شرکت می‌کنند و چندین ارزیاب، رفتار ارزیابی شوندگان را در هر تمرین براساس معیارهای از پیش تعریف شده در چند شایستگی (بعد) ارزیابی می‌کنند. بر این اساس، محور اصلی کانون ارزیابی، ماهیت تعاملی آن است که در آن علاوه بر ارزیاب و ارزیابی شونده، چندین تمرین وجود دارد که موقعیتهای مربوط به شغل را شبیه‌سازی می‌کنند و ارزیابان براساس آن ارزیابی شوندگان را در ابعاد مختلف درجه‌بندی می‌کنند (Sackett et al., 2017). تمرین‌ها و ابزارهای معمول و مورد استفاده در مرکز ارزیابی، عبارتند از ارائه‌های شفاهی (Presentation)، ایفای نقش (Role Play)، موردکاوی (Case Study)، روش کازیه (In-Basket Technique)، شبیه‌سازی‌ها (Simulations)، نمونه‌های کار، بحث گروهی بدون رهبری (Leaderless Group Discussions)، مصاحبه ساختاریافته، آزمون‌های شخصیت و توانایی و بازی‌های کسب و کار (Business Games) است. هریک از این فنون می‌تواند به‌خودی‌خود مورداستفاده قرار گیرد اما تنها زمانی بخشی از یک کانون ارزیابی محسوب می‌شوند که به‌صورت ترکیبی مورد استفاده قرار گیرد (International Task Force on Assessment Center Guidelines, 2015). نقش اصلی در مرکز ارزیابی را ارزیابان به عهده دارند، زیرا آن‌ها مسئول رتبه‌بندی‌های مرکز ارزیابی هستند، به عبارت دیگر پایه و اساس پایایی و روایی نتایج مرکز ارزیابی، کیفیت رتبه‌بندی و کار ارزیابان است (Kleinmann & Ingold, 2019). (2002) Kahneman & Frederick برای درک نقش ارزیابان در مرکز ارزیابی، سه نکته مهم که مشخصه ماهیت منحصربه‌فرد کار و الزامات شغلی آن‌هاست، را بیان کرده‌اند. نخست آنکه، مشاهده و ارزیابی یک یا چندسازه رفتاری از پیش تعریف شده که مشتمل بر ثبت رفتار، یادداشت‌برداری و قضاوت و رتبه‌بندی نهایی ابعاد (شایستگی‌ها) است، یک فعالیت شناختی است. دوم، تمرین‌های مرکز ارزیابی دارای مدت‌زمان از پیش تعریف شده هستند، به گونه‌ای که ارزیابان مدت‌زمان محدودی برای مشاهده و ارزیابی رفتارها در اختیار دارند و این الزام زمانی فشار مضاعفی بر ارزیابان وارد می‌کند. به بیان دیگر، آن‌ها فقط براساس نمونه‌های رفتاری محدود از تمرین‌های انجام‌شده باید سریع، تصمیم بگیرند. سوم، غالب روش‌های مرکز ارزیابی به گونه‌ای است که اولین تعامل ارزیاب با ارزیابی‌شوندگان هم‌زمان با فرایند ارزیابی است، به‌جز مصاحبه که در آن ارزیاب و ارزیابی‌شونده امکان تبادل اطلاعات را دارند، در نتیجه، اساس سنجش ابعاد در مراکز ارزیابی مبتنی بر فراهم آوردن یک

چارچوب نظری و پردازش اطلاعات حاصل از ارزیابی و رسیدن به رتبه‌بندی نهایی ابعاد است. به دلیل سرعت زیاد و بار شناختی بالا در کار ارزیابان، نظریه‌های چندی (پردازش خودکار، پردازش کنترل‌شده و مدل‌های دوگانه شناخت اجتماعی) برای تبیین این فرایند ایجاد شده‌اند (Fiske & Taylor, 2013). بی‌شک استفاده از نتایج حاصل از درجه‌بندی متقاضیان در کانون ارزیابی، هنگامی قابل‌قبول است که داده‌های گردآوری‌شده پایایی (Reliability) و روایی (Validity) مناسبی داشته باشند. بدون داشتن میزان قابل‌قبولی از پایایی، سؤال در مورد روایی و مناسب بودن درجه‌بندی‌ها با محدودیت همراه است. از سوی دیگر بدون روایی نمی‌توان نتایج حاصل از ارزیابی را به‌کار گرفت (Whetzel & Wheaton, 2007).

قطع‌نظر از اینکه کانون ارزیابی برای چه هدفی طراحی‌شده باشد، شواهد تأییدکننده روایی در سازه‌های اصلی سنجش شده، همواره مرتبط و معنادار است (Thornton et al., 2015). البته عمدتاً فرض بر این است که روایی محتوا، روایی ملاکی و روایی سازه (Content-Construct- & Criterion-Related Validity) و برآورد آن‌ها با یکدیگر مرتبط هستند و می‌توان منطقاً فرض کرد که وجود دو مورد آن‌ها، لزوماً سومی را اثبات می‌کند (Arthur et al., 2000)؛ بنابراین اگر کانونی از روایی محتوا و روایی ملاکی برخوردار بود، منطقاً می‌توان فرض کرد که این کانون دارای روایی سازه نیز می‌باشد. اگرچه پژوهش‌ها روایی محتوا و ملاکی کانون‌های ارزیابی را نشان داده‌اند؛ اما نتایج در مورد روایی سازه کانون‌های ارزیابی، کمتر امیدوارکننده بوده‌اند (Thornton & Mueller-Hanson, 2004 ;Lievens & Christiansen, 2012 ;Jackson et al., 2022). بر این اساس برخی محققان بر روی انواع رویکردهای روش‌شناختی جهت رفع مشکل روایی سازه کانون‌های ارزیابی تمرکز کرده‌اند. برای مثال Kleinmann et al. (1996) کاربرد روش‌های آماری همچون ماتریس چندصفتی-چندروشی ((Multitrait-(MTMM Multimethod) را برای برآورد بهتر روایی سازه کانون‌های ارزیابی پیشنهاد نمودند. برخی دیگر استفاده از مدل ساختاری ابعاد همبسته-تمرینات همبسته ((Correlated Dimensions Correlated Exercises(CDCE) با استفاده از تحلیل عاملی تأییدی را مطرح نمودند (Lance et al., 2004 ;Monahan, 2011). با این‌حال این تلاش‌ها در برطرف‌کردن مشکل روایی سازه کانون‌های ارزیابی ناکام مانده است. به‌طوری‌که برخی پژوهشگران خواستار حذف ابعاد از مطالعات کانون ارزیابی‌شده و تنها تمرکز بر تمرین‌ها را پیشنهاد کرده‌اند (Jackson et al., 2007 ;Neidig & Neidig, 1984 ;Dewberry, 2024). چنین دیدگاهی احتمالاً به دلیل موضع متضاد باهدف اصلی کانون‌های ارزیابی، یعنی اندازه‌گیری موقعیت ارزیابی‌شوندگان در ابعاد مختلف و موردنیاز برای عملکرد در شغل، مورد قبول نیست (Bowler & Woehr, 2006, 2009). بنا بر اظهارات برخی پژوهشگران عدم تفکیک روشن منابع واریانس، موجب سرگردانی تلاش‌های صورت‌گرفته جهت ارزیابی روایی مرتبط با سازه کانون‌های ارزیابی شده است. افزون بر آن، رویکردهای تحلیلی قبلی (به‌عنوان مثال، تحلیل‌های همبستگی و تحلیل عاملی تأییدی (Confirmatory Factor Analysis) و ماتریس‌های چندصفتی/چندروشی تنها بر واریانس مربوط به روش‌ها (تمرین) و ابعاد کانون‌های ارزیابی متمرکز هستند و تغییرپذیری احتمالی مربوط به دیگر منابع واریانس؛ افراد (ارزیابی‌شوندگان)، درجه‌بندی‌کنندگان (ارزیابان)، ابزارها و تعاملات مرتبط آن‌ها را نادیده می‌گیرند (Jones, 1992 ;Saal et al.,1980 ;Turnage & Muchinsky, 1982). Lievens (2001) سوگیری و عدم دقت (Bias & Inaccuracy) ارزیابان را به‌عنوان عاملی مهمی برای کاهش روایی کانون‌های ارزیابی قلمداد کرده و اضافه می‌کند که این امر می‌تواند نتیجه طرح ارزیابی ضعیف بوده و سبب واردشدن عاملی مزاحم در ارزیابی شود. از دیگر عوامل مداخله‌گر منجر به خطای ارزیابی می‌توان، تمایل به کسب مقبولیت اجتماعی، ترجیحات و ویژگی‌های شخصیتی ارزیاب و جذابیت شخصیتی و.... ارزیابی‌شونده نیز اشاره کرد. رویکرد کلیدی در ارزشیابی کیفیت کانون ارزیابی "درجه‌بندی‌های ویژگی میان تمرین (Across-Exercise Dimension Rating)"، در این رویکرد قابلیت اعتماد درجه‌بندی‌های چندین ارزیاب در مورد یک شایستگی بخصوص، یک داوطلب و یک تمرین خاص مورد بررسی قرار می‌گیرد (Lievens & Schollaert, 2011). با این‌حال، هنگام برآورد پایایی داده‌ها با استفاده از این روش و در چارچوب نظریه کلاسیک آزمون، چنین فرض می‌شود که پایایی نتایج برای همه زیرگروه‌های ارزیابان به یک اندازه است؛ اما اگر ارزیابان براساس ویژگی‌های جمعیت‌شناختی، دیدگاه متفاوتی نسبت به بعد (شایستگی) داشته باشند، این فرض

نقض می‌شود. اگرچه رویکردهای ارائه‌شده در مدل کلاسیک آزمون، روش‌های توانمندی برای مطالعه پایایی محسوب می‌شوند، اما به‌اندازه کافی انعطاف ندارند تا بتواند همه مشکلات مربوط به پایایی را که در اندازه‌گیری‌های چندگانه ایجاد می‌شوند، حل کند (Brennan, 2011).

یک راه نهایی برای بررسی قابلیت اعتماد درجه‌بندی‌های ارزیاب، این است که ببینیم اختلاف درجه‌بندی‌های ارزیاب تا چه اندازه بر اختلاف بین عملکرد کلی ارزیابی‌شونده، اثر می‌گذارد. این کار مستلزم تحلیل نظام‌مند واریانس درجه‌بندی‌های ویژگی است (Thornton et al., 2015). برای دست‌یابی به چنین نتیجه‌ای به یک مدل اندازه‌گیری نیاز است که بتواند سهم واریانس خطای منظم را برآورد نماید. نظریه تعمیم‌پذیری (Generalizability Theory(GT)) یک نظریه اندازه‌گیری است که با بهره‌گیری از طرح‌های تحلیل واریانس از چنین توانمندی برخوردار است؛ به‌عبارت‌دیگر، نظریه تعمیم‌پذیری یک نظریه آماری برای ارزیابی قابلیت اعتماد (پایایی) اندازه‌گیری‌های رفتاری است (Cronbach et al., 1972 ;Brennan, 2001 ;Shavelson & Webb, 1991). این نظریه رویکردی برای برآورد دقت اندازه‌گیری در شرایطی است که داده‌های جمع‌آوری‌شده در معرض منابع چندگانه خطا قرار دارند. نظریه تعمیم‌پذیری (GT) را می‌توان به‌عنوان بسط نظریه کلاسیک آزمون از طریق کاربرد روش‌های تحلیل واریانس، برای مسائل اندازه‌گیری در نظر گرفت. ارتباط بین نظریه تعمیم‌پذیری و تحلیل واریانس یک‌راهه همسو با ارتباط بین تحلیل واریانس ساده با تحلیل واریانس عاملی است. در آنوای ساده، معمولاً محققان، واریانس کل را به دو مؤلفه تحت عناوین واریانس "بین‌گروهی" و واریانس "درون‌گروهی" افراز می‌کنند. واریانس بین‌گروهی به‌عنوان یک واریانس سیستماتیک که گروه‌ها را از هم تفکیک می‌کند، در نظر گرفته می‌شود و واریانس درون‌گروهی به‌عنوان یک عامل تصادفی که به شکل خطا عمل می‌کند، مورد بررسی قرار می‌گیرد. به همین ترتیب، در نظریه تعمیم‌پذیری، واریانس کل را به واریانس "نمره واقعی" و واریانس "خطا" افراز می‌کند. واریانس نمره واقعی، یک واریانس سیستماتیک مرتبط با تفاوت‌های بین اشیای اندازه‌گیری است، درحالی‌که واریانس خطا به‌عنوان یک واریانس تصادفی نامرتبط با واریانس نمره واقعی عمل می‌کند (Shavelson & Webb, 1991). با استفاده از تحلیل واریانس عاملی به‌جای تحلیل واریانس ساده، محقق عوامل چندگانه‌ای را که در واریانس کل مشاهدات سهیم هستند، شناسایی و از این طریق آن‌ها را به اجزایی مرتبط با هر عامل، تعامل میان عوامل و خطای تصادفی افراز می‌کند. به‌طور مشابه، نظریه تعمیم‌پذیری تأثیرات چندگانه بر واریانس اندازه‌گیری را شناسایی می‌کند. این نظریه با برآورد مؤلفه‌های واریانس، شرایطی ایجاد می‌کند تا سهم نسبی عوامل تشکیل‌دهنده خطای سیستماتیک و خطای غیرسیستماتیک در نمره مشاهده‌شده به‌طور هم‌زمان تعیین گردد و همچنین با شناخت عوامل تشکیل‌دهنده خطا، بتوان طرح اندازه‌گیری مؤثری برای جمع‌آوری داده‌ها طرح‌ریزی نمود (Brennan, 2001). به‌طورکلی نظریه تعمیم‌پذیری به‌عنوان یک چارچوب مفید برای درک منابع واریانس در ارزیابی درجه‌بندی‌های کانون‌های ارزیابی (Brannick et al.,1989 ;Jones, 1992) و همچنین آزمون نمونه کار به‌شدت توصیه شده است، چراکه این نظریه قادر است ضمن شناسایی و تفکیک هریک از منابع مختلف تغییرپذیری، جداگانه هریک را برآورد کند (Webb & Shavelson, 2005).

معمولاً، جهت تحلیل کانون‌های ارزیابی با استفاده از تحلیل تعمیم‌پذیری سه جنبه شامل: ارزیابان (A) (Assessors)، تمرین‌ها (E) (Exercises) و ابعاد (D) (Dimensions) تحلیل و میزان واریانس خطای هر مؤلفه به‌طور جداگانه تعیین می‌شود. این سه جنبه کاملاً با یکدیگر در تعامل هستند و ارزیابی‌شوندگان (C) (Candidates) به‌عنوان هدف اندازه‌گیری مورد توجه قرار می‌گیرند (Lievens, 2001). مطابق تحقیقات قبلی، به‌طور خاص، واریانس مربوط به مؤلفه ارزیاب (اثر شخص) نشان‌دهنده سطح کلی توافق در رتبه‌بندی افراد در طول تمرین‌ها و ابعاد است. واریانس مربوط به تعامل اشخاص با ابعاد، نشان می‌دهد تا چه میزان عملکرد افراد در طول ابعاد تفکیک‌شده و متفاوت است. این مؤلفه در اصل، روایی تشخیصی در طول ابعاد را نشان می‌دهد. واریانس مربوط به اثر اصلی ارزیاب، نشان‌دهنده واریانس سیستماتیک مربوط به عملکرد

ارزیاب‌ها در درجه‌بندی ارزیابی‌شوندگان است (به‌عنوان مثال سوگیری ارزیاب). واریانس مربوط به تعامل ارزیابی‌شوندگان و تمرین نشان می‌دهد به چه میزان عملکرد افراد در تمرین‌ها متفاوت است. این مؤلفه نشان‌دهنده میزان قابل‌توصیف بودن عملکرد به‌وسیله تمرین (اثر موقعیت) است. درنهایت، واریانس مربوط به تعامل ابعاد در تمرین نشان‌دهنده میزانی است که ابعاد به شکل افتراقی (متفاوت) در تمرین‌های مختلف نشان داده می‌شود.

برای برقراری روایی سازه، وجود روایی همگرا و تشخیصی (Convergent & Discriminant Validity) لازم است (Campbell & Fiske, 1959). روایی همگرا هنگامی مشاهده می‌شود که علی‌رغم استفاده از روش‌های اندازه‌گیری مختلف (مثلاً ارزیابان متفاوت، اندازه‌های چندگانه و غیره)، نمره‌ها روی همان سازه نامتغیر باشند (Kraiger & Teachout, 1990; Marcoulides, 1989). روایی تشخیصی هنگامی مشاهده می‌شود که افراد به صورتی افتراقی روی اندازه‌های خصیصه‌های مختلفِ اندازه‌گیری، رتبه‌بندی شوند (Kraiger & Teachout, 1990). نظریه تعمیم‌پذیری شیوه‌ای را برای بررسی هر دو نوع روایی همگرا/تشخیصی فراهم می‌کند. هدف ثانویه مطالعه حاضر نیز توصیه و نشان دادن استفاده از تحلیل تعمیم‌پذیری برای ارزیابی روایی همگرایی/تشخیصی بود. استفاده از نظریه تعمیم‌پذیری برای بررسی هر دو نوع روایی همگرا/تشخیصی منحصربه‌فرد نیست به‌عنوان مثال Kraiger & Teachout (1990) بیان کرده‌اند که با استفاده از این نظریه، شواهد روایی مرتبط با سازه را می‌توان از میزان بزرگی واریانس سازه موردنظر (وجه اندازه‌گیری) نسبت به واریانس مربوط به شرایط اندازه‌گیری، استخراج نمود. این اثر نشان می‌دهد که واریانس اندازه‌گیری درواقع مربوط به سازه موردنظر است. در زمینه کانون‌های ارزیابی، روایی مرتبط با سازه از طریق مؤلفه نسبتاً بزرگ واریانس مربوط به ارزیابی‌شوندگان (اثر افراد)، سازه‌هایی که در آن افراد مورد ارزیابی قرار می‌گیرند (اثر ابعاد) و تعامل بین افراد و ابعاد (ابعاد× افراد) از یک طرف و مؤلفه نسبتاً کوچک واریانس مربوط به ارزیابان مرکز ارزیابی(اثر ارزیاب)، اثر تمرین‌ها، تعامل بین افراد و تمرین‌ها (اثر شخص×تمرین) و تعامل بین ابعاد و تمرین (ابعاد×تمرین) پشتیبانی می‌شود. با این‌حال، لازم به ذکر است که این الگوی اثرات است که شواهد مرتبط با روایی سازه را فراهم می‌کند و نه هریک از اثرات به‌تنهایی. اهمیت اولیه برای شواهد روایی مرتبط با سازه، واریانس مربوط به تعامل افراد با ابعاد است. به‌طور خاص، این مؤلفه نشان می‌دهد که چه میزان تمایز در میان افراد و ابعاد ایجاد می‌شود. ارزش پایین مؤلفه واریانس افراد × تمرین اشاره بر نامتغیر بودن درجه‌بندی افراد در طول تمرین‌ها دارد. هنگامی‌که عملکرد افراد در طول تمرین‌ها نسبتاً ثابت بماند شواهد روایی همگرا فراهم آمده است. شواهد روایی تشخیصی از مؤلفه واریانس افراد×ابعاد به‌دست می‌آید. ارزش بالای این مؤلفه بیانگر تفاوت‌های اساسی در درجه‌بندی افراد در طول ابعاد است (Kane, 1982; Kraiger & Teachout, 1990). تا به امروز، چندین مطالعه با استفاده از تئوری G به تجزیه و تحلیل داده‌های AC پرداخته‌اند که اگرچه به دلایل روش‌شناختی، به‌طور قابل توجهی در درجه‌ای که مسائل مربوط به اعتبار سازه AC را روشن می‌کنند، متفاوت هستند. به‌طور خاص، برخی از این مطالعات حاوی مسائل روش‌شناختی و فنی مورد توجه است. Winfred et al. (2000) در یک مطالعه تجربی به‌منظور نمایش و کاربرد نظریه تعمیم‌پذیری در بررسی روایی همگرا و تشخیصی کانون ارزیابی، داده‌های حاصل از درجه‌بندی ۱۴۹ آزمودنی با توجه به چهار بعد و سه نوع روش تمرین را بر مبنای یک طرح متقاطع با سه وجه (ابعاد، تمرین‌ها و ارزیاب) مورد تحلیل قرار داند. براساس نتایج پژوهش آن‌ها، درجه‌بندی‌های صورت‌گرفته از پایایی و روایی همگرا/تشخیصی کاملاً رضایت‌بخشی برخوردار بود. Arthur et al. (2000) در یک پژوهش با هدف بررسی روایی همگرا و واگرای ابعاد کانون ارزیابی، داده‌های حاصل از ۱۴۹ نفر از مدیران دولتی که طی ۸ سال در کانون ارزیابی شرکت کرده بودند را مورد بررسی قرار دادند، نتایج این مطالعه نشان دهنده پایایی بالای ارزیابی‌ها بود. Jackson et al. (2005) داده‌هایی را از یک AC به‌دست آوردند که برای انتخاب ۱۹۹ متقاضی برای پست خرده‌فروشی و خدمات مشتری در نیوزیلند استفاده شد. توسعه AC با تجزیه و تحلیل

شایستگی موقعیت هدف، از جمله مصاحبه با کارشناسان موضوع (SMEs)، انتخاب وظایف حیاتی و توسعه چک‌لیست‌های رفتاری برای این وظایف که شامل شاخص‌های رفتاری خاص برای عملکرد موفقیت‌آمیز برای استفاده در AC بود، آغاز شد. تجزیه و تحلیل شغل برای شناسایی ابعاد توسط تحلیل صفات آستانه هدایت شد. شرکت‌کنندگان در پنج بعد در سه تمرین رتبه‌بندی شدند. یک طرح کاملاً متقاطع استفاده شد که در آن هر بعد در هر تمرین ارزیابی شد. آن‌ها در تجزیه و تحلیل داده‌های خود ۷ مؤلفه واریانس منحصر به فرد را برای طراحی اندازه‌گیری کاملاً متقاطع محاسبه کردند. Putka, & Hoffman (2013) داده‌هایی را از سه پیاده‌سازی جداگانه (۱۵۳، ۱۹۸، ۲۹۸ =N) از یک AC با ریسک بالا در یک سازمان دولتی بزرگ در ایالات متحده به‌دست آوردند. هدف از این AC ارزیابی برای ارتقای نامزدهای داخلی برای پست‌های نظارتی خط اول بود. تمرین‌ها و مقیاس‌های رتبه‌بندی براساس داده‌های تحلیل شغل در یک سری کارگاه‌های ۲ روزه به رهبری روان‌شناسان صنعتی و سازمانی، آزمایش تمرین قبل از عملیات، استفاده از مقیاس‌های رتبه‌بندی با لنگرهای رفتاری متعدد و ارائه ۳۲ ساعت آموزش به هر ارزیاب بود. آن‌ها از تئوری تعمیم‌پذیری بهره گرفتند تا نتایج غیرقابل‌اعتماد ارزیاب (یعنی عدم قابلیت اعتماد ناشی از فقدان توافق) را از منابع قابل‌اعتماد واریانس در درجه‌بندی‌های ویژگی میان تمرین-که با نتایج تمرین نباید اشتباه گرفته شود-در سه مرکز مستقل ارزیابی، به‌گونه‌ای شفاف و قابل لمس نشان دهند. آن‌ها با استفاده از REML، توانستند ۱۵ منبع واریانس را در ACهای مورد مطالعه خود تجزیه کنند. Jackson et al. (2016) داده‌های به‌دست آمده از سه مدیریت یک AC عملیاتی پرمخاطره را تجزیه و تحلیل کردند که در مجموع ۶۹۸ نامزد شرکت داشتند. هدف این کانون، که در جنوب شرق آسیا اداره می‌شود، اطلاع‌رسانی درباره تصمیم‌گیری در مورد ارتقاء از سطح خط به سمت‌های مدیریت ارشد بود. همانند مطالعه پوتکا و هافمن، طراحی و اجرای AC از راهنمایی در مورد بهترین عملکرد پیروی کرد. برای مثال، تعداد ابعاد در هر تمرین به منظور کاهش بار شناختی ارزیابان به حداقل می‌رسد. حداقل از دو ارزیاب برای رتبه‌بندی هر داوطلب در هر تمرین استفاده می‌شود. ارزیابان به‌طور تصادفی به داوطلبان تقسیم شدند. انتخاب تمرینات مناسب با تجزیه و تحلیل شغل و مشاوره با کارشناسان موضوع اطلاع رسانی شد. ارزیابی‌کنندگان با توصیفات رفتاری مربوط به ابعاد مورد استفاده ارائه شدند. آیتم‌های رتبه‌بندی مستقیماً به لیست وظایف و اطلاعات تجزیه و تحلیل شغل قابل ردیابی بودند و ارزیابان به مدت دو روز توسط مشاوران مجرب و واجد شرایط آکادمیک آموزش دیدند. آن‌ها با استفاده از رویکرد بیزی برای تحلیل، یک مدل اثرات تصادفی خطی را به رتبه‌بندی‌های ارائه‌شده توسط ارزیاب‌ها برازش دادند و ۲۹ منبع واریانس (از جمله منابع مرتبط با آیتم‌ها و نمونه‌ها) را تجزیه کرد. Jackson et al. (2022) داده‌های به‌دست آمده از AC طراحی‌شده برای کمک به تصمیم‌گیری‌های ارتقاء در نیروی پلیس بریتانیا را بررسی کردند. ده نمونه از مجموع ۲۹۱۷ شرکت‌کننده به‌دست آمد. برای هر نمونه بین دو تا چهار تمرین (انتخاب از شش تمرین) استفاده شد. برای پنج تمرین نسبت ارزیاب به شرکت‌کنندگان ۲:۱ و در تمرین باقی‌مانده ۱:۱ بود. برای هر تمرین به ارزیابان یک روز آموزش توسط روان‌شناسان مجرب داده شد. این شامل آشنایی با روش ارزیابی، آگاهی از اشتباهات رایج ارزیابی‌کننده و آموزش چارچوب مرجع بود. آن‌ها با استفاده از یک رویکرد بیزی و یک مدل اثرات تصادفی خطی متقاطع برای تجزیه واریانس به‌دست آمده از هر نمونه به شش منبع برازش دادند. در مطالعات Jackson et al. (2016, 2022) و Putka & Hoffman (2013) پس از برآورد واریانس منحصربه‌فرد مرتبط با هر منبع، برآوردهای واریانس با استفاده از فرمول‌های مشخص‌شده در ادبیات تئوری G مجدداً مقیاس‌بندی شدند تا سهم هر مؤلفه در امتیازات سطح بعد و سطح نمره تمرین برآورد شود. همچنین نمرات کلی AC هنگامی که به نمرات سطح بعد تجمیع می‌یافتند (تکرار روش معمول میانگین نمرات در بین تمرین‌ها به‌منظور به‌دست آوردن مجموعه‌ای از

نمرات شایستگی برای هر داوطلب)، نسبت واریانس کلی AC مرتبط با همه منابع واریانس قابل‌اعتماد (یعنی بین داوطلبان) تفاوت در عملکرد کلی AC، در عملکرد مربوط به بعد، در عملکرد مربوط به تمرین و تفاوت در عملکرد که منعکس‌کننده تعامل بین تمرینات و ابعاد است، نشان داده شد و نتایج در رابطه با برآوردهای عملکرد کلی، عملکرد مرتبط با تمرین، عملکرد مرتبط با ابعاد، و تعامل بعد×تمرین مورد بحث قرار گرفت. Hoffman et al. (2010) کوشیدند با استفاده از تحلیل عامل دوم در "درجه‌بندی‌های تجمیع‌نشده ویژگی میان تمرین"، به پرسشی همانند پاسخ دهند. اگرچه در این مطالعات، از روش‌های متفاوتی استفاده شده است، هر دو رویکرد به شفاف‌سازی و بازشناسی واریانس مربوط به ارزیاب (یعنی اطلاعات منحصربه‌فرد یک ارزیاب به‌خصوص، واریانس خطا) از واریانس تمرین قابل اعتماد (یعنی اطلاعات مشترک بین ارزیابان) کمک می‌کند. نتایج هر دو مطالعه نشان داد که واریانس مربوط به ارزیاب، نمایانگر واریانس نسبتاً اندک در درجه‌بندی‌های مرکز ارزیابی است و این بدان معناست که الگوی نتایج که اغلب در تحلیل داده‌های مرکز ارزیابی می‌بینیم (یعنی ویژگی‌های به‌هم‌وابسته بین تمرین‌ها) پیش‌داوری ارزیاب را نشان نمی‌دهد؛ بلکه به‌جای آن، تفاوت‌های قابل‌اعتماد در عملکرد بین تمرین‌ها را منعکس می‌کند.

علی‌رغم پژوهش‌های متعدد در زمینه کانون ارزیابی با استفاده از نظریه تعمیم‌پذیری، این رویکرد تا کنون در ایران برای تحلیل کانون ارزیابی استفاده نشده است، ضمن آنکه کمتر مطالعه‌ای تمامی عوامل و تعامل‌های مرتبط در کانون ارزیابی را مورد بررسی قرار داده است؛ بنابراین انتخاب رویکرد نظریه تعمیم‌پذیری به‌عنوان هدف اولیه در مطالعه حاضر، به‌ویژه مناسب است؛ زیرا امکان تفکیک منابع مختلف واریانس سیستماتیک در درجه‌بندی کانون ارزیابی را فراهم می‌سازد. علاوه بر آن، امکان برآورد ضریب خلاصه (به‌عنوان مثال ضریب تعمیم‌پذیری) که نشان‌دهنده پایایی اندازه‌گیری است را میسر می‌سازد (Schneider & Schmitt, 1992; Turnage & Muchinsky, 1982). با توجه به مطالب فوق، هدف اصلی این پژوهش استفاده از نظریه تعمیم‌پذیری به منظور بررسی پایایی و روایی همگرا/تشخیصی درجه‌بندی‌های کانون ارزیابی یک سازمان دولتی بود.

روش

جامعه آماری، نمونه و روش نمونه‌گیری

داده‌ها به‌عنوان بخشی از عملیات کانون ارزیابی یک سازمان دولتی که در طی چند سال برای ارزیابی کارکنان توسعه‌یافته بود جمع‌آوری شد. برای مطالعه تعمیم‌پذیری، داده‌های مربوط به ۳۴۰ مدیر که براساس پنج بعد شایستگی (برنامه‌ریزی، رهبری، شناخت مسأله، نوآوری و تصمیم‌گیری) و توسط شش ارزیاب که با روش‌های متفاوت (تمرین گروهی، بحث گروهی، مصاحبه و پرسش‌نامه) در یک مقیاس ۹ درجه‌ای (۱ تا ۹) درجه‌بندی‌شده بودند از بانک اطلاعات شایستگی مدیران سازمان استخراج شد و براساس طرحی کاملاً متقاطع و تصادفی با ۵ مؤلفه ابعاد (D)، ۶ روش (E) و ۳۴۰ مدیر (P) با استفاده از برنامه EduG (Cardinet et al., 2010) مورد تحلیل قرار گرفت. جدول ۱ ابعاد و روش‌های تمرین مورد استفاده در طرح تعمیم‌پذیری را نشان می‌دهد. شش ارزیاب به شکلی انتخاب شد که بر روی گروه یکسانی از افراد و با توجه به ابعادی یکسان ولی با شیوه‌های متفاوتِ تمرین، عمل رتبه‌بندی را انجام داده باشند، برای مثال اگر ارزیاب یک، افراد را در برنامه‌ریزی، رهبری، شناخت مساله، نوآوری و تصمیم‌گیری مورد ارزیابی قرار داده، ارزیاب دو نیز همان افراد را در همان ابعاد مورد ارزیابی قرار داده باشد. تنها تمایز در این بود که هر ارزیاب، افراد را با یک روش ارزیابی متفاوت، مورد درجه‌بندی قرار داده بودند. یعنی ارزیاب یک افراد را با روش تمرین گروهی و ارزیاب دو افراد را با روش مصاحبه درجه‌بندی نموده بود. از آنجا که هر ارزیاب در درجه‌بندی هر فرد از یک نوع تمرین استفاده می‌کرد، امکان بررسی هم‌زمان اثر تمرین و اثر ارزیاب حتی

به‌صورت طرح آشیانی میسر نبود، لذا در این طرح اثر تمرین با اثر ارزیاب با یکدیگر آمیخته‌شده‌اند، به‌طوری‌که می‌توان اثر ارزیاب را متناظر با اثر تمرین دانست.

Table 1. Dimensions & methods studied in the generalizability design

Dimensions	Group Exercise	Discussion	Interview	Questionnaire
Planning	✓	✓	✓	
Decision making	✓	✓	✓	✓
Leadership	✓		✓	
Problem identification	✓	✓	✓	✓
Innovation	✓		✓	

در جدول ۱ یک طرح کاملاً متقاطع با سه مؤلفه: افراد (۳۴۰ نفر)، روش‌ها (تنها با دو سطح متناظر) و ابعاد (با پنج سطحِ برنامه‌ریزی، رهبری، شناخت مسأله، نوآوری و تصمیم‌گیری) مورد بررسی قرار گرفت. جدول ۲ طرح تعمیم‌پذیری مورد مطالعه را همراه با منابع تغییرپذیری آن‌ها نشان داده است. همان‌طورکه در جدول ۲ مشاهده می‌شود تغییرپذیری در درجه‌بندی مدیران به هفت مؤلفه واریانس افراز گردید. افراد (P)، ابعاد (D)، ارزیابان (روش‌ها) (E)، افراد در ابعاد (P×D)، افراد در ارزیابان (P×E)، ابعاد در ارزیابان (D×E) و افراد در ابعاد در ارزیابان (P×D×E). در این طرح مؤلفه‌های مدیران، ارزیابان و ابعاد اندازه‌گیری به‌عنوان متغیرهای تصادفی در نظر گرفته شد، زیرا سطوح هر عامل به‌عنوان معرفی از جامعه سطوح ممکن آن عامل، قضاوت می‌شود. به‌طورخاص Thornton & Byham (1982) فهرستی از ابعادی که در کانون‌های ارزیابی استفاده می‌شود را گزارش نموده‌اند. ابعاد مورد استفاده در کانون ارزیابی موردمطالعه به‌عنوان نماینده و معرف این فهرست محسوب می‌شود. همچنین ارزیابان کانون ارزیابی متشکل از روان‌شناسان و مشاوران مدیریت هستند؛ بنابراین این‌گونه قضاوت می‌شود که آن‌ها نماینده و معرف جمعیتی هستند که باید به‌عنوان ارزیاب به‌کار گرفته شوند (Spychalski et al., 1987).

Table 2. Sources of variability of two facets crossed design with five dimensional levels & two assessor levels (methods)

Source of Variance	Type of Variance	Component Description
Persons (P)	Universe score variance	The object of measurement
Dimensions (D)	Condition: Fixed effect for all persons due to behavioral inconsistency across dimensions	
Evaluators (Exercises) (E)	Fixed effect for all persons due to rater severity/leniency	
P × D	Instability in an individual's behavior across dimensions	
P × E	Instability in evaluator judgment of a specific individual's behavior	
E × D	Fixed effect for all persons due to differences in evaluator severity/leniency across dimensions	
P × D × E, Unspecified or Error	Residual including the interaction of all three facets, unmeasured facets, or random events	

در نظریه تعمیم‌پذیری، به‌دست‌آوردن برآورد واریانس برای یک طرح کاملاً متقاطع (طرحی که در آن تمام سطوح یک متغیر با تمام سطوح متغیر دیگر ترکیب شود) باعث افزایش قابلیت تعمیم نمرات می‌شود (Nunnally & Bernstein, 1994). علاوه بر این، یک طرح کاملاً متقاطع، به محقق این امکان را می‌دهد تا واریانس منحصربه‌فرد مرتبط با یک وجه را از واریانس خطا افراز کند، درنتیجه اطلاعات مفید بیشتری به‌دست می‌دهد؛ بنابراین برای مطالعه، یک طرح کاملاً تصادفی و متقاطع انتخاب شد. نمودار ون طرح مورد استفاده در این مطالعه در شکل ۱ نشان داده شده است.

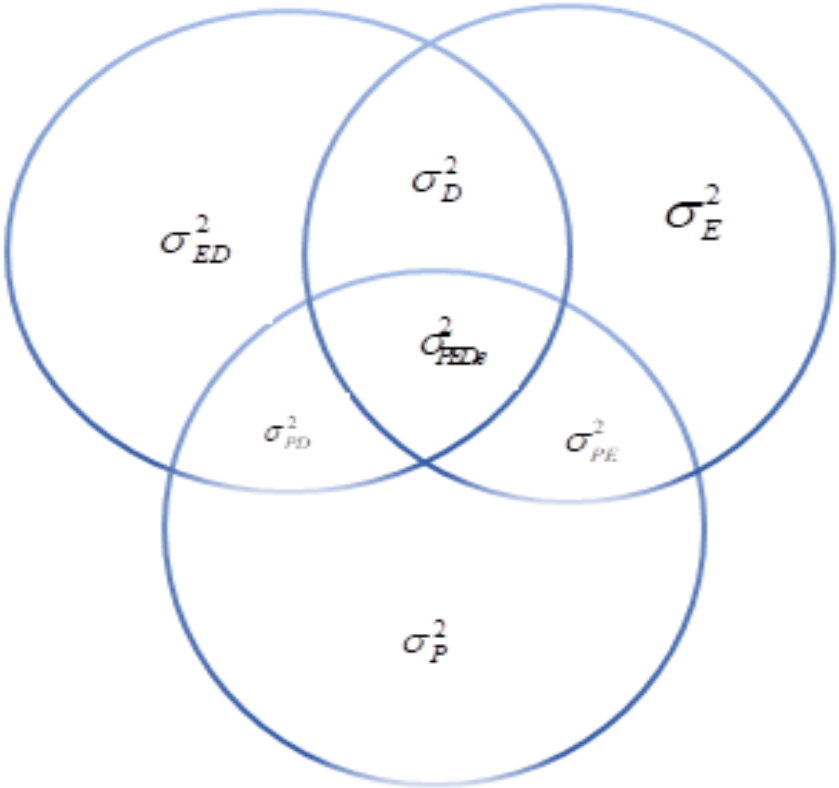


Figure 1. Venn diagram of the sources of variance of the generalizability design

به منظور بررسی روایی اندازه‌گیری‌ها، لازم است واریانس هر مؤلفه، جداگانه مورد ملاحظه قرار گیرد. در مطالعه حاضر، تغییرپذیری ناشی از افراد سهمی در تشکیل واریانس خطا ندارد؛ چراکه افراد شیء مورد اندازه‌گیری هستند و انتظار می‌رود آن‌ها به‌صورت متفاوت ظاهر شوند؛ بنابراین مؤلفه واریانسِ به اندازه کافی بزرگ برای افراد نشانگر روایی همگرا است (Kraiger & Teachout, 1990 ;Marcoulides, 1989) و میزان ثبات حاصل از درجه‌بندی مدیران توسط ارزیابان مختلف شواهدی از روایی تشخیصی فراهم می‌کند (Kraiger & Teachout, 1990). در طرح فوق اگر توانایی مدیران به‌عنوان یک سازه تک‌بعدی (Unidimensional) پذیرفته شود، یعنی تنها سازه‌ای که اندازه‌گیری می‌شود، آنگاه مقادیر کوچک‌تر مؤلفه واریانس افراد در ابعاد نشانگر روایی تشخیصی است و چنانچه توانایی مدیران به‌عنوان یک سازه چندبعدی درنظر گرفته شود، یعنی سازه‌ای که در طول ابعاد اندازه‌گیری می‌شود، آنگاه ارزش بزرگ‌تر برای این مؤلفه واریانس، بیانگر روایی تشخیصی است (Marcoulides, 1989). دیگر منابع تغییرپذیری (مؤلفه واریانس ارزیابان، افراد در ارزیاب، افراد در ارزیاب در ابعاد) در حمایت از روایی تشخیصی باید مقادیر کوچکی را به خود اختصاص دهند؛ چراکه آن‌ها می‌توانند در خطا مشارکت کنند و سبب نامعتبر بودن اندازه‌گیری توانایی مدیران شوند (Marcoulides, 1989).

جامعه آماری این مطالعه شامل کلیه مدیران شرکت ملی نفت ایران در سال ۱۴۰۱ بود. از این جامعه، داده‌های مربوط به ۳۴۰ مدیر که براساس پنج بعد شایستگی (برنامه‌ریزی، رهبری، شناخت مسأله، نوآوری و تصمیم‌گیری) و توسط شش ارزیاب که با روش‌های متفاوت (تمرین گروهی، بحث گروهی، مصاحبه و پرسش‌نامه) در یک مقیاس ۹ درجه‌ای (۱ تا ۹) درجه‌بندی‌شده بودند از بانک اطلاعات شایستگی مدیران شرکت براساس طرحی کاملاً متقاطع و تصادفی استخراج شد.

ابزار پژوهش

ابزارهای جمع‌آوری داده‌ها در این مطالعه عبارت بودند از تمرین گروهی، بحث گروهی، مصاحبه و پرسش‌نامه که هر کدام در یک مقیاس ۹ درجه‌ای پنج بعد (شایستگی) برنامه‌ریزی، رهبری، شناخت مساله، نوآوری و تصمیم‌گیری را مورد سنجش قرار می‌داد. ازآنجاکه فرض می‌شود این ابعاد مستقل از هم هستند، لذا تجمیع‌پذیر بوده و نمره افراد در همه ابعاد با یکدیگر جمع شده و نمره کلی عملکرد داوطلبین در تمرین‌ها بر این مبنا محاسبه گردید.

یافته‌ها

میانگین و انحراف استاندارد درجه‌بندی‌های مربوط به سطوح ابعاد/ارزیاب و هر بعد (در طول ارزیابان) و هر ارزیاب (در طول ابعاد) در جدول ۳ نشان داده شده است. همان‌طورکه مشاهده می‌شود، به‌طور متوسط میانگین سطوح مختلف ابعاد حدود یک واحد نمره تفاوت را نشان می‌دهد (به‌خصوص بین سطح رهبری و سطوح دیگر) درحالی‌که میانگین درجه‌بندی ارزیابان اختلاف کمتری را نشان می‌دهد.

Table 3. Mean & (standard deviation) of the dimensions' & assessors' (methods) facets

Dimension	1 (Group Exercise)	2 (Interview)	Mean	Standard deviation
Planning	4.01	4.08	4.04	1.19
Innovation	4.74	4.40	4.57	1.00
Leadership	3.86	3.11	3.48	1.31
Problem solving	4.41	4.26	4.34	1.04
Decision making	4.61	4.80	4.71	1.05
Mean	4.33	4.13		
Standard deviation	1.10	1.30		

میانگین کل، ۴/۲۳، واریانس کل، ۱/۴۶ و انحراف استاندارد کل، ۱/۲۱

برآورد مؤلفه‌های واریانس برای هر یک از سه اثر اصلی (ابعاد، روش‌ها/ ارزیابان و افراد) و نیز تعامل آن‌ها (ابعاد در ارزیاب‌ها، ابعاد در افراد، ارزیاب‌ها در افراد) و همچنین مؤلفه خطا (تعامل ابعاد در ارزیاب‌ها در افراد) در جدول ۴ ارائه شده است.

Table 4. Summary of ANOVA for two facets crossed design

Source of Variance	Variance component estimate	Percentage of total variance (%)	Standard error of measurement
Persons (P)	0.44	29.0	0.04
Methods / Evaluators (E)	0.01	0.3	0.02
Dimensions (D)	0.20	13.4	0.14
Persons × Methods / Evaluators (PE)	0.02	1.3	0.01
Persons × Dimensions (PD)	0.23	15.4	0.02
Methods / Evaluators × Dimensions (ED)	0.07	4.4	0.04
Persons × Methods × Dimensions (PED)	0.55	36.2	0.02

همان‌طورکه در جدول ۴ آمده است، بررسی برآورد مؤلفه‌های واریانس نشان داد که اثر اصلی مربوط به ابعاد ۱۳/۴ درصد واریانس را به خود اختصاص می‌دهد. این شاهدی است مبنی بر اینکه عملکرد از حیث ابعاد اندازه‌گیری‌شده، در بین افراد تغییر می‌کند. درحالی‌که مقدار اختصاص‌یافته به ارزیابان (روش‌ها)، حدود ۰/۳ درصد است که نشان می‌دهد درجه‌بندی‌های صورت گرفته توسط ارزیابان در روش‌های اندازه‌گیری (تمرین گروهی، مصاحبه) با تغییرات بسیار کمی روبه‌روست. بیشترین درصد واریانس اثرات اصلی متعلق به عامل افراد بود، به‌طوری‌که این اثر، ۲۹ درصد واریانس کل را شامل شد که نشان می‌دهد نسبت

بالایی از واریانسِ درجه‌بندی‌ها، ناشی از تفاوت‌های بین افراد از نظر شایستگی‌های مورد اندازه‌گیری در این طرح است، به‌عبارت‌دیگر می‌توان گفت ارزیابان کانون، در ارزیابی مدیران از نظر شایستگی‌های مورد اندازه‌گیری، به‌طورکلی افتراقی عمل نموده و عملکرد آن‌ها را متفاوت درجه‌بندی نموده‌اند و یا اینکه مدیران ازنظر شایستگی مورد بررسی متفاوت بوده‌اند. ازآنجاکه در این مطالعه، مدیران شیئی موردِاندازه‌گیری را تشکیل می‌دهند و نه خطا، چنین تغییرپذیری مطلوب است. در نهایت کمی بیش از ۳۶ درصد از واریانس کل مربوط به باقیمانده یا خطا بوده است. یعنی نسبتی از واریانس که به‌واسطه تعامل سه‌جانبه افراد، ارزیابان (روش‌ها) و ابعاد یا دیگر منابع سیتماتیک یا غیرسیستماتیک تغییرپذیری که در این طرح مورد اندازه‌گیری قرار نگرفته است، حاصل می‌شود. بنابراین منابع مهم واریانس دیگر موجودند که به‌وسیله تفاوت بین مدیران، تفاوت بین ارزیابان (روش‌ها)، تفاوت بین ابعاد یا تعامل دوطرفه آن‌ها، شکل نمی‌گیرد و در این طرح نامطلوب به‌حساب می‌آیند.

بررسی روایی همگرا-تشخیصی

مطابق با رویکرد روایی همگرا-تشخیصی با استفاده از مؤلفه‌های نظریه تعمیم‌پذیری (Kraiger & Teachout, 1990; Marcoulides, 1989) بررسی مؤلفه‌های واریانس برآوردشده برای اثرات اصلی و اثرات تعاملی (جدول ۴) الگوی نسبتاً روشنی در حمایت از روایی درجه‌بندی ابعاد مورد مطالعه به‌دست می‌دهد. اثر اصلی مربوط به ابعاد، درصد نسبتاً بالایی از واریانس کل را به خود اختصاص داد. این نشان می‌دهد که عملکرد در طول ابعاد مورد مطالعه متغیر است؛ همچنین اثر تعاملی بین ابعاد و اشخاص کمی بیش از ۱۵ درصد واریانس کل را توصیف می‌کرد. واریانس مرتبط با این تعامل، نشان می‌دهد تا چه میزان عملکرد افراد در طول ابعاد تفکیک‌شده و متفاوت است. این مؤلفه در اصل، روایی تشخیصی در طول ابعاد را نشان می‌دهد؛ بنابراین عملکرد در طول ابعاد تغییرپذیری نسبتاً بالایی دارد؛ یعنی همان‌طور که انتظار می‌رفت، درجه‌بندی‌های صورت‌گرفته، تفاوت‌های مکنون فردی در ابعاد مورد مطالعه را به‌خوبی نمایش می‌دهد (شاهدی مبنی بر روایی تشخیصی) همچنین جزئیات جدول ۴ نشان داد که ۲۹ درصد واریانسِ کل درجه‌بندی‌ها، ناشی از اثر اصلی افراد است. این اثر نشان می‌دهد که ارزیابان در ارزیابی عملکرد مدیران کانون ارزیابی افتراقی عمل نموده‌اند.

از سوی مقابل، مؤلفه مربوط به اثر اصلی ارزیابان (روش‌ها) که نشان‌دهنده واریانس سیستماتیک مربوط به عملکرد ارزیاب‌ها در درجه‌بندی ارزیابی‌شوندگان با شیوه‌های تمرین مختلف است، درصد بسیار ناچیزی از کل واریانس را به خود اختصاص داد که نشان می‌دهد بخش کوچکی از تغییرپذیری ناشی از ارزیابان یا شیوه‌های ارزیابی است. این بدان معنی است که بین دو ارزیاب (یا دو شیوه تمرین) ازنظر درجه‌بندی آزمودنی‌ها، تمایز نسبتاً کمی وجود دارد (عدم سوگیری ارزیاب) همچنین اثر تعامل بین افراد با روش‌ها/ ارزیاب‌ها نیز تنها ۱/۳ درصد واریانس کل را به خود اختصاص داد که چندان بزرگ نیست. واریانس مرتبط با این تعامل نشان می‌دهد که عملکرد افراد از یک ارزیاب (روش) به ارزیاب (روش) دیگر چندان تغییر نمی‌کند و نسبتاً باثبات (Stable) است (شاهدی بر روایی همگرا). در نهایت تعامل ابعاد و روش‌ها/ ارزیابان ۴/۴ درصد واریانس کل را به خود اختصاص داده که نشان می‌دهد به‌طورکلی همه ابعاد در هر دو روش (تمرین گروهی و مصاحبه) تقریباً به یک اندازه قابل سنجش هستند. به‌طور خلاصه، وجوه مرتبط با اشخاص، ابعاد و تعامل آن‌ها، مجموعاً حدود ۵۸ درصد از واریانس کل را به خود اختصاص می‌دهد، درحالی‌که وجوه مرتبط با روش‌ها/ارزیاب‌ها و تعاملات مربوطه، تنها ۶ درصد از واریانس کل را تبیین می‌کند؛ بنابراین به‌طورکلی می‌توان ادعا کرد که داده‌های این مطالعه از روایی همگرا/تشخیصی برخوردار است.

نتایج تعمیم‌پذیری

جدول ۵ نتایج مطالعه تعمیم‌پذیری را نشان می‌دهد. در این جدول، وجوه مختلف و تعامل بین آن‌ها در دو طبقه وجه افتراقی و وجه ابزاری به‌طور مجزا نشان داده شده است. همچنین واریانس‌های خطای نسبی و مطلق وجوه ابزاری در جدول منعکس شده است. بر طبق این جدول، برآورد واریانس افتراقی (۰/۴۴)، به مراتب بیشتر از مجموع واریانس‌های خطای نسبی (۰/۱۱) و یا مجموع واریانس‌های خطای مطلق (۰/۱۶) وجوه ابزاری است. دو ضریب G نسبی و مطلق در انتهای جدول ۵ شاخص‌های کلی تعمیم‌پذیری (Generalizability) و قابلیت اطمینان (Dependability) روش اندازه‌گیری است و نشانگر دقت اندازه‌گیری‌های انجام گرفته هستند. باید توجه داشت که ضریب تعمیم‌پذیری با ضریب پایایی قابل مقایسه است.

Table 5. Results of the generalizability study of two facets crossed design

Source of Variance	Differentiation variance	Source of variance	Relative error variance	Relative error	Absolute error variance	Absolute error
Persons (P)	0.44					
		Methods/ Evaluators (E)		0.00	1.6	
		Dimensions (D)		0.04	25.2	
Persons × Dimensions (PD)	0.05		0.05	41.9	0.05	28.9
Persons × Evaluators (PE)			0.01	8.7	0.01	6.0
		Evaluators×Dimensions (ED)		0.01	4.1	
Persons × Evaluators × Dimensions (PED)	0.05		0.05	49.4	0.05	34.1
Total Variance	0.44		0.11	100%	0.16	100%
Standard Deviation	0.66	Relative Error	0.33		Absolute Error	0.40
Relative G Coefficient	0.80					
Absolute G Coefficient	0.73					

در این مطالعه، مقدار دو ضریب G نسبی و مطلق به‌ترتیب ۰/۸۰ و ۰/۷۳ به‌دست آمد. مقدار ضریب نسبی حول‌وحوش ارزش قراردادی ۰/۸(کاردینت،۲۰۱۰) است. براساس این ضریب، ۸۰ درصد واریانس مشاهده‌شده در میانگین درجه‌بندی‌ها به نمره واقعی نسبت داده می‌شود و ۲۰ درصد آن به تعامل بین افراد× ابعاد× ارزیاب و وجوه تعیین‌نشده‌ای که دقت نتایج انفرادی را کاهش می‌دهد، مربوط است؛ بنابراین به‌طورکلی شیوه‌های درجه‌بندی و دقت اندازه‌گیری با این تعداد ابعاد اندازه‌گیری و روش/ارزیاب نسبتاً رضایت‌بخش است و می‌تواند یک درجه‌بندی پایا از عملکرد افراد فراهم کند. در ردیف پایین جدول ۳، ریشه مربع واریانس خطای (SEM) اندازه‌گیری نسبی و مطلق به‌ترتیب ۰/۳۳ و ۰/۴۰ گزارش شده است که می‌توان از آن برای تشکیل فاصله اطمینان، اطراف برآورد میانگین افراد یا روش‌ها و ابعاد، با فرض توزیع صفت زیربنایی استفاده کرد.

بحث و نتیجه‌گیری

هدف نخست این مطالعه، بررسی پایایی کانون ارزیابی یک سازمان دولتی با استفاده از نظریه تعمیم‌پذیری بود.

داده‌های مربوط به ۳۴۰ مدیر که بر اساس پنج بعد شایستگی (برنامه‌ریزی، رهبری، شناخت مساله، نوآوری و تصمیم‌گیری) و توسط شش ارزیاب که با روش‌های متفاوت ارزیابی (تمرین گروهی و مصاحبه) در یک مقیاس ۹ درجه‌ای (۱ تا ۹) درجه‌بندی‌شده بودند از بانک اطلاعات شایستگی مدیران سازمان استخراج شد و بر اساس طرحی کاملاً متقاطع و تصادفی با ۳ مؤلفه ابعاد (D)، ارزیابان(روش ها) (E) و افراد (P) موردمطالعه قرار گرفت. بر اساس یافته‌ها مقدار ضریب G نسبی ۰/۸۰ به دست آمد. که برابر با ارزش قراردادی کاردینت و همکاران(۲۰۱۰) است. بر اساس این ضریب، ۸۰ درصد واریانس مشاهده‌شده در میانگین درجه‌بندی‌ها، به نمره واقعی نسبت داده می‌شود و ۲۰ درصد آن به وجوه تعیین نشده‌ای که دقت نتایج انفرادی را کاهش می‌دهد، مربوط است؛ بنابراین به‌طورکلی شیوه‌های درجه‌بندی با این تعداد ابعاد اندازه‌گیری و ارزیاب نسبتا رضایت‌بخش است و می‌تواند یک درجه‌بندی پایا از عملکرد افراد فراهم کند.

یک هدف ثانویه این مطالعه توصیه و نشان دادن استفاده از تجزیه‌وتحلیل تئوری تعمیم‌پذیری برای ارزیابی روایی همگرا / تشخیصی بود. استفاده از نظریه تعمیم‌پذیری به این شیوه منحصربه‌فرد نیست (Kraiger & Teachout, 1990; Kane, 1982; Winfred, et al., 2000). شواهد روایی مرتبط با سازه را می‌توان از میزان بزرگی واریانس سازه موردنظر (وجه اندازه‌گیری) نسبت به واریانس مربوط به شرایط اندازه‌گیری، استخراج نمود. ارزش پایین مؤلفه واریانس افراد × روش‌ها اشاره بر نامتغیر بودن درجه‌بندی افراد در طول تمرین‌ها دارد. هنگامی‌که عملکرد افراد در طول تمرین‌ها نسبتا ثابت بماند شواهد روایی همگرا فراهم آمده است. شواهد روایی تشخیصی از مؤلفه واریانس افراد × ابعاد به دست می‌آید. ارزش بالای این مؤلفه بیانگر تفاوت‌های اساسی در درجه‌بندی افراد در طول ابعاد است (Kraiger & Teachout, 1990; Kane, 1982). بر اساس یافته‌ها، نه‌تنها اثر تعاملی بین ابعاد و اشخاص کمی بیش از ۱۵ درصد واریانس کل را توصیف می‌کرد، بلکه وجوه مرتبط با اشخاص و ابعاد، مجموعاً حدود ۵۸ درصد از واریانس کل را به خود اختصاص داد، درحالی‌که وجوه مرتبط با ارزیاب‌ها، تنها ۶ درصد از واریانس کل را تبیین می‌کرد؛ بنابراین مطابق با رویکرد روایی همگرا-تشخیصی و ازآنجاکه این الگوی اثرات است که شواهد مرتبط با روایی سازه را فراهم می‌کند و نه هریک از اثرات به‌تنهایی (Winfred, et al., 2000) داده‌های این مطالعه از روایی همگرا-تشخیصی کافی برخوردار است.

از جمله مهم‌ترین محدودیت‌های این تحقیق آن بود که هر ارزیاب در رتبه‌بندی هر فرد، از یک نوع تمرین استفاده می‌کرد، لذا امکان بررسی هم‌زمان اثر تمرین و اثر ارزیاب به‌صورت مجزا میسر نبود و درنتیجه طرح اثر تمرین با اثر ارزیاب با یکدیگر آمیخته‌شده بودند. بنابراین پیشنهاد می‌شود در طرح‌های تحقیقاتی آتی، این محدودیت موردتوجه قرار گیرد.

تشکر و قدردانی

نویسندگان بر خود لازم می‌دانند از کلیه افراد مشارکت‌کننده در این پژوهش قدردانی نمایند.

تضاد منافع

هیچ‌گونه تعارض منافعی در این جستار توسط نویسندگان بیان نشده است.

References

Arthur, J., Woehr, D. J. & Maldegen, R. (2000). Convergent & discriminant validity of assessment center dimensions: A conceptual & empirical re-examination of the assessment center construct-related

validity paradox. *Journal of Management*, *26*, 813–835. https://doi.org/10.1016/S0149-2063(00)00057-X

Bowler, M. C. & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension & exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*, 1114–1124. https://doi.org/10.1037/0021-9010.91.5.1114

Bowler, M. C. & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, *75*, 173–182. https://doi.org/10.1016/j.jvb.2009.03.008

Brannick, M. T., Michaels, C. E. & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology*, *74*, 957–963. https://doi.org/10.1037/0021-9010.74.6.957

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L. (2011). Generalizability theory & classical test theory. *Applied Measurement in Education*, *24*, 1–21. https://doi.org/10.1080/08957347.2011.532417

Campbell, D. T., & Fiske, D. W. (1959). Convergent & discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81. https://doi.org/10.1037/h0046016

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Taylor & Francis, Routledge. https://doi.org/10.4324/9780203866948

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

Dewberry, C. (2024). Assessment centers do not measure competencies: Why this is now beyond reasonable doubt. *Industrial & Organizational Psychology*, *17*, 154–175. https://doi.org/10.1017/iop.2024.5

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). Los Angeles, CA: Sage. https://doi.org/10.4135/9781446286395

Gatewood, R. D., Feild, H. S., & Barrick, M. (2010). *Human resource selection* (7th ed.). South-Western. https://www.scirp.org/reference/referencespapers?referenceid=1731673

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

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment center's for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection & Assessment*, *15*, 405–411. https://doi.org/10.1111/j.1468-2389.2007.00399.x

Hoffman, B. J., Melchers, K., & Kleinmann, M. (2010). Disentangling assessment center exercise from rater effects. Paper presented at the 25th annual conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.

Iles, P. (1992). Centers of excellence? Assessment & development centers, managerial competencies & human resource strategies. *British Journal of Management*, *3*(2), 79–90. https://doi.org/10.1111/j.1467-8551.1992.tb00037.x

International Task Force on Assessment Center Guidelines (2015). Guidelines & ethical considerations for assessment center operations. *Industrial & Organizational Psychology*, *17*, 243–253. https://doi.org/10.1111/j.1468-2389.2009.00467.x

Jackson, D. J. R., Michaelides, G., Dewberry, C., Nelson, J., & Stephens, C. (2022). Reliability in assessment centres depends on general & exercise performance, but not on dimensions. *Journal of Occupational & Organizational Psychology*, *95*(4), 739–757. https://doi.org/10.1111/joop.12398

Jackson, D. J. R., Michaelides, M., Dewberry, C., & Kim, Y. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*, *101*(7), 976–994. https://doi.org/10.1037/apl0000102

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, *18*(3), 213–241. https://doi.org/10.1207/s15327043hup1803_2

Jackson, D. R., Barney, A. R., Stillman, J. A. & Kirkley, W. (2007). When traits are behaviors: The relationship between behavioral responses & trait-based overall assessment center ratings. *Human Performance*, *20*, 415–432. https://doi.org/10.1080/08959280701522130

Jones, R. G. (1992). Construct validation of assessment center final dimension ratings: Definition & measurement issues. *Human Resource Management Review*, *2*(3), 195–220. https://doi.org/10.1016/1053-4822(92)90012-F

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics & Biases: The Psychology of Intuitive Judgment*, *49*, 81. https://doi.org/10.1017/CBO9780511808098.004

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160. https://doi.org/10.1177/014662168200600201

Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of assessment centers: A conceptual review. *Annual Review of Organizational Psychology & Organizational Behavior*, *6*, 349–372. https://doi.org/10.1146/annurev-orgpsych-012218-014955

Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review*, *45*, 67–68. https://doi.org/10.1111/j.1464-0597.1996.tb00849.x

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance*, *3*, 19–35. https://doi.org/10.1207/s15327043hup0301_2

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, C. G. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection & Assessment*, *14*(4), 360–371. https://doi.org/10.1111/j.1468-2389.2006.00357.x

Krause, D. E., Rossberger, R. J., Dowdeswell, K., Venter, N., & Joubert, T. (2011). Assessment center practices in South Africa. *International Journal of Selection & Assessment*, *19*, 262–275. https://doi.org/10.1111/j.1468-2389.2011.00555.x

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension & exercise variance components in assessment center post exercise dimension ratings. *Journal of Applied Psychology*, *89*, 377–385. https://doi.org/10.1037/0021-9010.89.2.377

Lievens, F. (2001). Assessor & use of assessment center dimension. *Journal of Organizational Behavior*, *22*, 203–221. https://doi.org/10.1002/job.65

Lievens, F., & Christiansen, N. D. (2012). *Core debates in assessment center research: Dimensions 'versus' exercises*. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The Psychology of Assessment Centers* (pp. 68–91). England: Routledge. https://psycnet.apa.org/record/2012-16855-004

Lievens, F., & Schollaert, E. (2011). Adjusting exercise design in assessment centers: Theory, practice, & research. In N. Povah & G. C. Thornton (Eds.), *Assessment centers & global talent management* (pp. 47–60). Surrey, UK: Gower Publishing Limited. chrome

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior & Personality*, *12*, 53–62. https://psycnet.apa.org/record/1998-10074-002

Marcoulides, G. A. (1989). Performance appraisal: Issues of validity. *Performance Improvement Quarterly*, *2*(2), 3–12. https://doi.org/10.1111/j.1937-8327.1989.tb00398.x

Monahan, E. L. (2011). *The search for the mythical assessment center dimension: Measurement artifact vs. substantive conclusion* (Unpublished master's thesis). University of Georgia, Athens, GA. https://openscholar.uga.edu/record/17939?v=pdf

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises & job relatedness. *Journal of Applied Psychology*, *69*, 182–186. https://doi.org/10.1037/0021-9010.69.1.182

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill. https://books.google.com/books/about/Psychometric_Theory_3E.html?id=_6R_f3G58JsC

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contributions of assesse-, dimension-, exercise-, & assessor-related effects to reliable & unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*, 114–133. https://doi.org/10.1037/a0030887

Rupp, D. E., Thornton, C. G., & Gibbons, A. M. (2008). The construct validity of the assessment center method & usefulness of dimensions as focal constructs. *Industrial & Organizational Psychology: Perspectives on Science & Practice*, *1*(1), 116–120. https://doi.org/10.1111/j.1754-9434.2007.00021.x

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413. https://doi.org/10.1037/0033-2909.88.2.413

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435. https://doi.org/10.1037/apl0000236

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 & 1982 & the investigation of study characteristics. *Personnel Psychology*, *27*(3), 407–422. https://doi.org/10.1111/j.1744-6570.1984.tb00519.x

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension & exercise constructs. *Journal of Applied Psychology*, *77*, 32–41. https://doi.org/10.1037/0021-9010.77.1.32

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage. https://doi.org/10.1037/10109-051

Spychalski, A. C., Quiñones, M. A., Gaugler, B. B., & Pohley, K. (1987). A survey of assessment center practices in organizations in the United States. *Personnel Psychology*, *50*(1), 71–90. https://doi.org/10.1111/j.1744-6570.1997.tb00901.x

Thornton III, G. C., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, & development*. Psychology Press. https://doi.org/10.4324/9781410617170

Thornton, G. C. III., Rupp, D. E., & Hoffman, B. J. (2015). *Assessment center perspectives for talent management strategies* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9781315746265

Thornton, G. C., & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review*, *19*(3), 169–187. https://doi.org/10.1016/j.hrmr.2009.02.002

Thornton, G. C., & Mueller-Hanson, R. A. (2004). *Developing organizational simulations*. Philadelphia, PA: Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410609458

Thornton, G., & Byham, W. (1982). *Assessment centers & managerial performance*. New York: Academic Press. https://doi.org/10.1016/B978-0-12-690620-2.50014-3

Turnage, J. J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. *Organizational Behavior & Human Performance*, *30*, 174–200. https://doi.org/10.1016/0030-5073(82)90217-3

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. https://doi.org/10.1002/0470013192.bsa703

Whetzel, D. L., & Wheaton, G. R. (2007). *Applied measurement: Industrial psychology in human resource management*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203936412