

• 研究方法(Research Method) •

评价中心的构想效度谜题：测量维度还是活动？^{*}

卞 冉 高 钦 车宏生

(应用实验心理北京市重点实验室, 北京师范大学心理学院, 北京 100875)

摘 要 评价中心是一种高保真度的情境模拟, 它被设计用来在多种与工作相关的活动中测量多项维度。30 年来的大量研究发现, 评价中心具备良好的内容效度和效标关联效度, 但构想效度却始终不理想, 评价中心评分反映的总是由活动而非预先设想的维度带来的效应。这一评价中心“构想效度谜题”吸引了大量研究关注, 并逐步形成了维度中心取向、活动中心取向及交互作用取向三种主要观点, 分别主张控制各种误差因素以改善维度测量、放弃维度而转向活动或任务以及关注维度与活动的共同作用。未来研究应在传统的维度中心取向之外给予活动中心取向足够重视, 并重点发展交互作用取向。

关键词 评价中心; 构想效度; 维度中心取向; 活动中心取向; 交互作用取向

分类号 B849:C91; B841

1 引言

评价中心(Assessment Centers, ACs)应用于人力资源管理领域已有超过 50 年的历史(Thornton & Gibbons, 2009)。作为一种非常流行的人事测评技术, 它被广泛用于各类组织并服务于人事选拔、晋升、诊断及发展等目的(Thornton & Rupp, 2006)。一般而言, 评价中心是一种高保真度的情境模拟, 它将个体置于多种与工作相关的活动(exercises)中, 由多名训练有素的评分者(assessors)观察和评价作答者在刺激情境下真实的行为, 并在多项维度(dimensions)上对其表现进行评分(Lievens, Tett, & Schleicher, 2009; International Task Force on Assessment Center Guidelines, 2009)。以往大量研究证明, 评价中心是一种相当有效的测评手段, 它不仅具有良好的内容效度(Woehr & Arthur, 2003)、效标关联效度(Gaugler, Rosenthal, Thornton, & Bentson, 1987; Arthur, Day, McNelly, & Edens, 2003)、递增效度(Meriac, Hoffman, Woehr, & Fleisher, 2008)、评分

者信度(Lievens, 2002), 还在应用中展现出很好的效用指标(Hoffman & Thornton, 1997)和较好的公平性(Petrides, Weinstein, Chou, Furnham, & Swami, 2010), 并能引发积极的应聘者反应(applicant reaction)(Hausknecht, Day, & Thomas, 2004)等。

然而, 除了这些有利证据, 在评价中心效度研究领域, 有一个问题一直困扰着众多研究者, 即评价中心构想效度相关证据的缺乏(Arthur, Woehr, & Maldegen, 2000; Woehr & Arthur, 2003; Lievens, Tett et al., 2009; Lance, 2008a)。评价中心的基本设计理念是对作答者在不同活动或情境下的绩效行为表现按照维度进行评价, 并形成活动后维度评定(postexercise dimension rating, PEDR)。这些维度代表了个体稳定的、跨情境一致的特征。所以, 评价中心是被设计用来测量以行为或特质为基础的维度的。但是, 众多研究却发现活动后维度评定的变异中, 占主导地位的并不是我们想要测的维度, 而是被视为方法的活动; 研究一再发现, 评价中心的维度评分缺乏聚合效度(convergent validity)与区分效度(discriminant validity)等构想效度证据(Lance, 2008a, 2008b; Lievens, Tett et al., 2009)。研究者普遍认为, 评价中心具有良好的内容效度和效标关联效度, 但同时又缺乏内部结构相关的构想效度。这样的结果

收稿日期: 2012-07-02

^{*} 教育部人文社会科学研究青年基金项目(12YJC880001)。

通讯作者: 高钦, E-mail: qingao@mail.bnu.edu.cn

显然是与整体效度观念(unitarian view of validity) (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society for Industrial and Organizational Psychology, 2003; Binning & Barrett, 1989)相违背的, 所以被称为评价中心的“构想效度悖论”(construct validity paradox) (Arthur et al., 2000; Woehr & Arthur, 2003; Lance, Foster, Nemeh, Gentry, & Drollinger, 2007; Lance, 2008a)或“构想效度谜题”(construct validity puzzle) (Lance, et al., 2000; Lievens, 2002)。

近年来, 在整体效度观和人事测量构想导向(construct-oriented) (e.g., Arthur & Villado, 2008; Huffcutt, Conway, & Stone, 2001)的影响下, 人们不再仅仅满足于关注评价中心的预测效度, 更对其测量的实质即构想效度研究提出了进一步的要求: 也许我们已经知道评价中心对于选拔评价是有效的, 但它可能并没有按照我们预先设计的那样去产生效果。那么, 它为什么有效? 评价中心到底测量到的是什么是维度还是活动(方法)? 如果是活动, 它又为什么能预测工作绩效? 本文试图在前人研究的基础上, 结合最新的研究成果, 对评价中心构想效度相关证据进行总结, 并对此问题的成因及影响因素进行深入探讨。

2 评价中心构想效度相关研究回顾

一直以来, 对于评价中心与构想相关的(construct-related)构想效度的验证, 研究者基本都采用多特质多方法(multitrait-multimethod, MTMM)模型(Campbell & Fiske, 1959)或因素分析的方法。1982年, Sackett与Dreher在一项研究中使用这两种方法, 首次报告了评价中心构想效度不足的结果。

因为评价中心采用多种活动(方法)测量多种维度(特质), 在多种活动测得的多个维度的MTMM相关矩阵中, 我们应该假设同一维度在不同活动之间的相关(同质异法)较高, 而不同特质在同种活动内的相关(异质同法)较低。这分别反映了评价中心的聚合效度证据和区分效度证据(Campbell & Fiske, 1959)。但是, Sackett与Dreher(1982)却得到了完全相反的结论。他们对取自三个不同组织的评价中心活动后维度评定数据进行分

析后发现, 三组数据的维度评分都表现出同质异法相关较低(聚合效度差)而异质同法相关较高(区分效度差); 进一步使用探索性因素分析发现, 数据更支持抽取活动因子而非维度因子, 即评价中心维度评分的变异更应来自于活动而不是预先设计的维度。这样的结果令人困扰, 而Sackett与Dreher更是表示, “这说明评价中心有严重的问题: 方法变异(而不是维度变异)占据了主导地位”。

在此之后, 很多研究者开始重视这一问题并开展大量研究试图重复及修正Sackett和Dreher(1982)的结果, 从而找到有力的评价中心构想效度证据。但令人意外的是, 无论被试来源于何种组织以及评价中心服务于何种目的, 几乎所有的研究都得到了与前述类似的结果, 即评价中心是以活动为导向而不是以维度为导向。这样的结论, 甚至在各种文化背景下的相关研究中都普遍存在, 包括美国(Harris, Becker, & Smith, 1993; Joyce, Thayer, & Pond, 1994; Kudisch, Ladd, & Dobbins, 1997; Reilly, Henry, & Smither, 1990; Sackett & Dreher, 1982; Schneider & Schmitt, 1992; Silverman, Dalessio, Woods, & Johnson, 1986)、英国(Anderson, Lievens, van Dam, & Born, 2006; Petrides et al., 2010; Robertson, Gratton, & Sharpley, 1987)、德国(Kleinmann & Koller, 1997; Kleinmann, Kuptsch, & Koller, 1996)、比利时(Lievens & Van Keer, 2001)、法国(Rolland, 1999)、澳大利亚(Atkins & Wood, 2002)、新西兰(Jackson, Stillman, & Atkins, 2005)、新加坡(Chan, 1996)以及中国(吴志明和张厚粲, 2001; 王博, 田效勋, 邵燕萍, 车宏生, 2010)等等。而只有极少数研究得到了有利于评价中心构想效度的证据(e.g., Arthur et al., 2000)。

基于大量相关研究的结果, 近年来, 有几个元分析对评价中心构想效度问题进行了总结(Woehr & Arthur, 2003; Lievens & Conway, 2001; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Bowler & Woehr, 2006)。Woehr与Arthur(2003)搜集了32个研究报告的评价中心评分数据, 并对其中31组MTMM相关矩阵进行了分析, 结果发现其平均的同质异法相关为0.34, 平均异质同法相关却达到0.55(均为样本加权后); 对16个验证性因素分析结果进行总结也发现, 同时含有维度因子和活动因子的模型拟合最优, 且活动因子

的载荷均大于维度因子。

Lievens 和 Conway (2001)对搜集到的 34 个评价中心维度评分的 MTMM 相关矩阵进行了再分析。与 Woehr 和 Arthur (2003)不同的是,他们将这 34 个矩阵重新使用验证性因素分析进行了模型拟合,并估计维度与活动各自的变异解释量。结果发现,虽然同时含有维度因子和活动变异(在模型中设定为残差相关)的模型得到了最佳拟合,但活动不再显得比维度更重要:二者解释的变异量平均下来均为 34%。另外,维度因子之间的平均相关却也很高,达到 0.71。Lievens 与 Conway 的研究结论似乎让人们看到了一线曙光,但很快就有研究者对他们的结论进行质疑。Lance, Lambert 等人(2004)认为,活动与维度解释了相同的变异量,这一结果应归结于人为统计误差: Lievens 与 Conway 错误使用了相关特质-相关残差(correlated trait-correlated uniqueness, CTCU)模型从而系统高估了维度的变异。鉴于此, Lance, Lambert 等人(2004)又再次分析了 Lievens 和 Conway (2001)研究中的 34 个 MTMM 矩阵,结果仍然发现同时含有维度因子和活动因子的模型拟合最优,但不同的是此模型只包含一个总的维度因子和多个相关的活动因子;活动因子比总维度因子解释了更多的变异(52% vs 14%)。

不难看出,这几个元分析面临着一个共同的问题:它们都是先对个别研究数据重新分析,之后再对结果进行整合。这样难免使最终结果受所选研究样本量、维度和活动的不同等因素影响。为了解决这一问题,在最近的一个研究中, Bowler 与 Woehr (2006)使用元分析方法将 35 个 MTMM 矩阵合并成一个矩阵,并以此为基础进行验证性因素分析。结果表明,获得最佳拟合的是同时含有相关维度因子和相关活动因子的模型;活动因子仍然解释了大部分维度评分变异(33%),维度因子只解释了小部分变异(22%);维度因子之间的相关仍很高,平均达到了 0.79。

总体上来看,无论是个别研究还是元分析,似乎都“顽固”地指向评价中心构想效度不佳的结论。其中原因究竟是什么呢?为什么评价中心的维度评分并没有按照我们预先设计的那样去表现出跨情境的一致性(cross-situational consistency)呢?我们究竟在测什么?是维度还是活动?从 Sackett 与 Dreher (1982)的研究发表至今,有很多

理论和研究试图对这一“谜题”进行破解,从不同角度来阐释这一现象的成因及影响因素。目前,这一领域主要存在三种相互对立但又紧密联系的观点来分析和解释评价中心构想效度问题:维度中心取向(dimension-based approach)、活动中心取向(exercise-based approach)以及交互作用取向(interactionist approach)。

3 维度中心取向

虽然大量研究表明评价中心以维度为基础的评分构想效度欠佳,很多研究者还是坚定地认为评价中心仍然是以维度为基础的(Arthur, Day, & Woehr, 2008; Brannick, 2008; Melchers & König, 2008; Schuler, 2008; Rupp, Thornton, & Gibbons, 2008)。原因在于,很多评价中心构想效度研究都是借用多质多法模型研究范式,将维度看作特质,活动看作方法,并假设维度是稳定的、跨情境一致的。所以,维度变异被视为真分数变异,而活动变异被视为误差变异;活动变异过大就会导致评价中心构想效度不佳。基于这一理论假设,维度中心取向的研究认为,之所以出现构想效度证据不足的情况,根本原因在于评价中心测量过程中可能存在诸多因素影响导致测量误差(即活动变异)增大而维度变异减小。因此,该取向的研究致力于探索哪些因素会对评价中心构想效度产生影响,并进一步针对这些因素予以控制从而改进对维度的测量。长期以来,研究所关注的影响因素主要包括评价中心的设计与实施、统计方法、维度界定等三个方面(Lievens, 1998; Sackett & Tuzinski, 2001; Arthur et al., 2000; Woehr & Arthur, 2003; Lievens & Conway, 2001; Lance, 2008a; Gibbons & Rupp, 2009; 王忠军, 龙立荣, 2006)。

3.1 评价中心的设计与实施因素

在回答为什么评价中心构想效度证据不足这一问题时,大部分研究者首先想到的是设计与实施环节可能出了问题。这种观点假设,评价中心普遍存在着一些设计与实施方面的缺陷,如让评分者同时评定大量维度、对评分者培训不足等,都会引起评分出现系统偏差(如晕轮效应),进而导致活动变异占据主导(Lance, 2008a; Lievens & Klimoski, 2001)。正是基于这种思路,很多研究着力于找到并修正这些潜在的缺陷,以改善评价中心的构想效度。在较长一段时间内,这种“修正设

计”(design fix)取向的研究在评价中心研究领域内处于主流地位,研究采取的修正思路包括改善评分者图式、降低评分者认知负荷或改变评定方法等几种。

一种观点认为,评价中心构想效度问题可能源于评分者使用了不正确的图式(schema)来分类和加工观察到的信息。相应地,有两种评价中心设计因素与之有关,包括从评分者自身入手的评分者类型(type of assessor)和从外部入手的评分者培训(assessor training)。一方面,对于评分者类型,研究者一般认为,在评价中心常用的心理学家和经理两类评分者中,心理学家评分更为客观有效,因为其教育和训练背景保证了他们能更有效地观察、记录和评价受测者的行为(Gaugler et al., 1987)。Sagie 和 Magnezy (1997)曾分析了心理学家和经理两类评分者在评价中心中各自评分的因素效度,因素分析的结果发现心理学家评分的结构与预想维度结构一致,经理评分则不是;Lievens 和 Conway (2001)的元分析发现,与经理不同,心理学家评分中维度变异要显著高于活动变异;Woehr 和 Arthur (2003)的元分析结果也证明,心理学家评分比经理评分表现出更多支持性的构想效度证据。另一方面,研究者也认为良好的培训有助于改善评分者图式进而提高评价中心构想效度(Lievens, 1998, 2001a, 2002)。元分析发现,有评分者培训相对于无评分者培训,评价中心构想效度结果更好(Woehr & Arthur, 2003)。但具体到培训时间、培训方法等因素上,研究结论不尽一致:有研究发现增加培训时间并不能改善评分者评分(Dugan, 1988),也有研究发现评分者培训时间越长确实越能改善评价中心构想效度(Woehr & Arthur, 2003),还有研究却发现短时培训(1 天或更短)比长时培训(超过 1 天)带来的评分中维度变异比例更大(Lievens & Conway, 2001);而针对培训方法,研究者参考绩效评价领域研究,认为参考框架法(frame-of-reference)是最为有效的(Woehr & Huffcutt, 1994),实证研究证明参考框架培训方法确实在一定程度上能够提高评价中心的构想效度(Lievens, 2001a; Schleicher, Day, Mayes, & Riggio, 2002)。

除了从改善评分者图式入手,研究者还尝试从降低评分者认知负荷角度来改进评价中心对维度的测量。这种取向认为,评分者的信息加工能

力是有限的,而评价中心过程对评分者认知要求又很大(如同时对多个受测者的行为进行观察和记录、在多个维度上对行为进行分类、加工和评价等等),在这种情况下易出现认知超载(cognitive overload)而导致评分出现偏差。所以,研究尝试从多处设计入手进行改进以尽量降低评分者认知负荷,包括使用行为检核表(behavioral checklist)、减少维度数量、控制受测者-评分者比例(participant-to-assessor ratio)等。Reilly 等人(1990)的研究发现,使用行为检核表确实可以显著提高评价中心的聚合和区分效度,原因在于行为检核表可以帮助评分者更有效提取和分类信息以降低观察过程中的认知负荷;而另外一些研究对此也得到了部分支持性的结论(Donahue, Truxillo, Cornwell, & Gerrity, 1997; Hennessy, Mabey, & Warr, 1998; Lievens & Conway, 2001)。对于评价过程中维度数量的控制,研究结论较为一致。很多研究证明,在同一活动内降低评价维度的数量,有助于降低评分者的认知负荷并提高维度之间的区分,从而提高维度测量的质量(Gaugler & Thornton, 1989; Schneider & Schmitt, 1992; Lievens & Conway, 2001; Woehr & Arthur, 2003)。而对于受测者-评分者比例,研究者认为减少评分者在活动中同时观察和评价的受测者数量,可以降低其认知负荷并减少评分偏差出现,进而提高构想效度。虽然这一假设并未在 Woehr 和 Arthur (2003)的元分析中得到证实,但最近的研究对此还是得出了支持性的结论(Melchers, Kleinmann, & Prinz, 2010)。

第三种修正思路关注评分者采用何种策略对维度进行评定:活动内方法(within-exercise approach)还是维度内方法(within-dimension approach)。活动内方法是在每个活动完成后即对受测者各个维度进行评定,而维度内方法则是在所有活动完成后再对各个维度进行跨活动的评定。研究者认为,正是因为常用的活动后维度评定(PEDR)即活动内评定方法使评分者过多关注活动特异信息,才造成夸大活动内不同维度相关以致评价中心构想效度不足(Haaland & Christiansen, 2002; Arthur et al., 2008),所以要提高构想效度,必须转而使用维度内评定方法。Silverman 等人(1986)比较了这两种评定方法对评价中心构想效度的影响,结果发现使用维度内方

法可以显著提高聚合效度与区分效度;而 Harris 等人(1993)的类似研究却没有得出支持性的结论。Robie, Osburn, Morris, Etchegaray 和 Adams (2000)则尝试使用实验方法对这一问题进行研究,结果发现活动内评定方法确实会对评价中心构想效度产生不利影响。而 Woehr 和 Arthur (2003)的元分析也进一步证实使用维度内评定方法的评价中心比使用活动内评定方法的评价中心展示出更高的构想效度。

3.2 统计方法因素

相对于修正设计取向将活动变异占优问题归因于设计缺陷导致的无关变异增大,另一些研究者则认为问题出在评价中心构想效度研究所使用的统计方法上:活动变异大是因为方法带来的虚假变异或统计方法的误用(Kolk, Born, & van der Flier, 2002; Lance et al., 2000; Lance, Lambert et al., 2004; Lance, 2008a; Arthur et al., 2000)。

一种质疑声音认为,我们使用 MTMM 方法对评价中心构想效度进行分析,实际上高估了活动引起的变异,这才造成维度变异“看起来”比代表系统误差的活动变异小(Sackett & Dreher, 1982)。在 MTMM 方法中,评价中心评分的系统变异被分成两部分:维度和活动,分别对应特质和方法。但如此操作得到的“活动变异”不仅包含由活动引起的变异,同时还包含着由评分者带来的变异,我们实际上把这二者混淆在了一起,所以才会造成活动变异占据主导的结果。有两项研究对此假设进行了检验(Kolk et al., 2002; Robie et al., 2000)。在这两项研究中,研究者让每位评分者在各项活动中都只对某一维度进行评定,从而控制评分者变异。但结果均与假设相反,异质同法相关依然高于同质异法相关,活动因素仍占主导地位。

另外一些质疑直接指向了评价中心构想效度研究中常用的因素分析方法。对于探索性因素分析,一些观点认为很多研究使用正交旋转策略是存在问题的,因为评价中心的维度本身应该具有一定相关(Arthur et al., 2003);而对于验证性因素分析,虽然在测量领域有比较成熟的范式应用这种技术在 MTMM 框架下对测验构想效度进行探讨,但在统计模型上如何正确设定代表方法的活动因子及代表特质的维度因子仍是众多研究争议不休的问题,包括到底该使用相关特质-相关方法

模型(CTCM)还是相关特质-相关残差模型(CTCU)、活动因子之间到底该不该设定相关、模型易出现不恰当解(inadmissible solution)等问题都会对评价中心构想效度的估计产生极大影响(Kleinmann & Koller, 1997; Sagie & Magnezy, 1997; Lievens & Conway, 2001; Lance et al., 2004; Lance, Woehr, & Meade, 2007; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011)。

正是因为这样的问题,近来一些研究放弃了基于 CTT 的 MTMM 及因素分析统计方法,转而使用基于现代测量理论的概化理论(generalizability theory)技术对评价中心构想效度进行分析(Arthur et al., 2000; Lievens, 2001b, 2002; Jackson et al., 2005; Bowler & Woehr, 2009; 王博,田效勋,邵燕萍,车宏生,2010)。这些研究尝试从维度、活动、评分者等不同侧面对评分变异进行分解,同时探索各侧面之间可能存在的交互作用。尽管到目前为止研究使用这种方法得到的构想效度支持证据还不多(e.g., Arthur et al., 2000; Bowler & Woehr, 2009),但这种尝试为评价中心构想效度领域研究提供了新的思路(Woehr, Putka, & Bowler, 2012)。

3.3 维度界定因素

最后一种解释聚焦于维度本身。研究者认为,在评价中心开发与应用中,人们对测量目标也就是维度的界定往往是比较模糊、随意的,有时甚至是想当然的,缺乏统一标准。如此做法也是导致评价中心构想效度不佳的原因之一(Jones & Born, 2008; Thornton & Gibbons, 2009; Hoffman et al., 2011; Arthur et al., 2008; Howard, 1997; Howard, 2008)。

任何一种测量,对其所要测量的构想(construct)如何界定和操作化是其首先必须解决的问题,而这个问题对评价中心来说,似乎从来没有得到很好的解决。评价中心测得的各种“维度”到底是什么,它与构想、特质的关系又是什么,一直以来很少有研究进行清楚地阐述(Jones & Born, 2008)。由于评价中心的应用性与实践性,它要测的维度一般来自于工作分析或胜任特征模型,通常包含各种特质、技能、态度、行为、动机等,这些维度的提出往往带有一定的随意性、特殊性,其背后缺乏坚实的测量学基础。Arthur 等人(2003)曾就不同评价中心试图测量的维度进行总结,结果发现在 34 个评价中心研究中总共提出了 168 个

不同的维度,这对于描述和解释人在工作中的行为来说显然过于冗杂。在这样的情况下,很可能出现不同名称的维度其相关重叠很大,难以区分;同样名称的维度其含义却不尽一致,难以会聚,从而造成构想效度不佳的结果。为解决这一问题,一些研究者试图通过改变维度的宽度,将含义相近的一些狭窄维度(narrow dimension)组合成为宽泛的维度因子(broad dimension factor),形成概括性更高、更能区分的元维度(meta-dimension)(Hoffman et al., 2011; Arthur et al., 2003; Kolk, Born, & van der Flier, 2004; Howard, 2008)或总体绩效因子(general performance factor)(Thornton & Gibbons, 2009)。Hoffman 等人(2011)最新的研究尝试在评价中心验证性因素分析模型设定中引入此类因子,并与传统模型进行比较。结果发现,虽然各模型活动因子解释变异量仍然最大,但当模型中同时包含宽泛维度因子、总体绩效因子和活动因子时,模型对评价中心评分的代表和解释最好。这也为评价中心构想效度提供了新的证据。

总之,维度中心取向想方设法从各个方面入手试图提高评价中心维度测量的质量,不管是设计上的细节修正,还是改善统计方法,抑或改变维度的界定等等。我们不禁要问,这样一些控制和改变到底有没有用?目前来看,这个问题还很难回答。总体上,无论是个别研究还是元分析研究,得到的结论多是混合的。一方面,并不是所有的改进都被证明有效;另一方面,那些被认为有效的改进措施,研究得到的结论也并不完全一致。而且,即使修正措施确实改善了构想效度,效果总体来说仍比较弱:有时即便研究者采用各种策略对多种因素进行细致谨慎的控制,仍然不能完全得到支持维度测量的构想效度证据(Schneider & Schmitt, 1992; Chan, 1996)。这也就是说,维度中心取向的研究虽然取得了一些效果,但并没有从根本上改变评价中心评分中活动变异优于维度变异的基本模式;对于还要控制哪些因素以及如何控制以提高评价中心维度测量构想效度,可能还需要更多的研究来进行进一步探索。

4 活动中心取向

相对于维度中心取向仍坚持认为评价中心测量的构想是维度,另一些研究者则提出了不同见解。他们认为,构想效度不理想很有可能是因为

我们错误假设或设定了评价中心的测量构想(Lance et al., 2000; Lance, Lambert et al., 2004; Lievens & Conway, 2001; Woehr & Arthur, 2003; Lance, 2008a)。Russell 和 Domm (1995)曾指出,“评价中心评分对反映某些构想来说肯定是有效的,只不过我们不知道这些构想到底是什么”;Sackett 和 Tuzinski (2001)也提出,“评价中心并不缺乏构想效度,只是在它到底测了什么构想这一问题上研究者缺乏共识……关键问题是,以往对于评价中心到底测量了何种构想的观点是否正确”。由这种构想错误设定(construct misspecification)观点出发,一些研究者开始假设,既然如此多研究都一致发现评价中心中活动变异总是优于维度变异,那么我们有理由认为评价中心测量的应该是活动(Robertson et al., 1987; Howard, 1997; Lance, Lambert et al., 2004; Lance, 2008b)。于是,活动中心取向应运而生。持该观点的研究者提出了情境特异性理论(situational specificity interpretation)来对评价中心构想效度问题进行解释,并主张把维度从评价中心框架中去除,而把焦点聚焦在活动或任务上,设计基于任务的评价中心(task-based assessment centers)来取代传统的基于维度的评价中心(dimension-based assessment centers)(Lance, 2008a, 2008b)。

4.1 情境特异性理论

传统的评价中心理论借用多质多法模型研究范式,将活动变异视作与构想无关的、由测量方法带来的系统误差,它的增大会带来方法偏差(method bias),并对维度(特质)的测量产生“污染”(contamination)。所以,套用多质多法框架来分析评价中心的构想效度,研究者很容易得出这样的结论:维度变异是我们所期望的、与真分数相关的,需要在测量中最大化;活动变异是我们不想要的,需要尽量避免;活动变异过大就会导致评价中心构想效度不佳(Lievens & Conway, 2001; Lance et al., 2000)。

一些研究者对此提出了质疑,认为正是以往研究不假思索直接套用多质多法范式,才让我们对评价中心的活动变异产生错误理解和解释,使得研究者从一开始在提出“评价中心到底测了何种构想”这一问题时就出现了方向性的偏误,这才会产生无法解释的“构想效度悖论”现象(Lance, Lambert et al., 2004; Lance, 2008a; Lance, Baranik,

Lau, & Scharlau, 2009)。所以, 与传统的将维度视为测量构想、活动效应视为方法偏差的研究不同, Lance 等人提出情境特异性理论对评价中心构想效度问题进行解释。他们认为, 活动效应反映的是受测者跨情境(活动)的特异性表现, 它实际上代表着我们想要测量的真分数变异, 而不是测量方法偏差(Lance et al., 2000; Lance, 2008a; Gibbons & Rupp, 2009)。这种观点来源于交互作用理论, 认为不同的评价中心活动对人的行为表现的要求是不一样的, 体现在任务特征、情境要求以及情境要求与受测者胜任特征的交互上, 而正是这样一些活动间的变化带来了与真分数相关的受测者跨情境特异性表现(Russell & Domm, 1995; Lievens & Conway, 2001; Lance et al., 2000; Lance, 2008a; Hoffman & Meade, 2012)。

对这一理论已有一些实证研究方面的支持。

Lance 等人(Lance et al., 2000; Lance, Foster, Gentry, & Thoresen, 2004; Lance et al., 2007)在三个研究中对六组评价中心数据分别进行基于多质多法模型的验证性因素分析, 并着重探讨活动因子与其他外部变量(如工作绩效、认知能力、人格特质等)的相关。研究假设, 如果活动因子反映的是方法偏差, 它应该与其他变量不相关; 而如果活动因子代表与真分数相关的跨情境特异表现, 它与其他变量之间就应该存在一定关系。研究结果发现, 大部分活动因子与外部变量之间的相关显著, 从而证实了情境特异性假设。Lievens, Dilchert 和 Ones (2009)尝试在一项研究中同时收集评价中心多个方面的效度证据, 研究使用相关特质-相关方法(correlated trait-correlated method, CTCM)验证性因素分析发现, 与维度因子相比, 活动因子不仅解释了更多的评价中心评分变异, 还对工资这一效标显示出更强的预测作用。而最近 Hoffman 等人(2011)的研究在活动因子与外部效标之间的相关上也得到了类似的结果。

4.2 基于任务的评价中心

在情境特异性理论的启发下, 一些研究者开始尝试将维度从评价中心中去除, 转而把注意力仅仅放在活动上, 设计基于活动或任务的评价中心。实际上, 早在 20 世纪 80 年代就有研究者提出过这样的建议(Sackett & Dreher, 1984; Sackett & Harris, 1988; Goodge, 1988), 但却没有受到足够的重视。Goodge (1988)最早在他的研究中对基

于任务的评价中心进行了描述, 并尝试比较它与传统评价中心的区别。随后, Lowry 又通过一系列研究对基于任务的评价中心的特点及开发流程进行系统论述(Lowry, 1995, 1997), 他首先界定并区分了两类评价中心: 一类是传统的“维度特定”(dimension-specific)评价中心, 另一类是“任务特定”(task-specific)评价中心。Lowry (1997)认为, 这两种评价中心都是在工作分析的基础上建立起来的, 但区别在于前者以胜任特征维度为设计框架, 后者则更依靠从工作分析中得到的岗位关键任务; 前者将活动视为获得维度测量的方法(Arthur & Villado, 2008), 后者则直接将活动中的角色或任务表现作为测量对象。任务特定或基于任务的评价中心以岗位关键任务为蓝本, 构建一系列工作样本(work sample)模拟, 并使用行为检核表等方法对受测者在各个活动内完成任务的行为表现进行评价(Lowry, 1997; Lance, Lambert et al., 2004; Lance, 2008a)。

尽管基于任务的评价中心提出已有 20 多年, 并在实践领域中得到一定的应用, 但我们仍对它知之甚少, 实证研究非常缺乏(Lievens, 2008)。仅有的一些研究中, Lowry (1995)曾得到一些初步的证据, 证明任务特定的评价中心有令人满意的测量学指标; Jackson 等人(2005)对同一批被试采用基于维度和基于任务两种不同的评分方式, 结果显示它们的测量学指标相近, 而基于任务的评价中心数据对模型的拟合更好; Jackson, Barney, Stillman 与 Kirkley (2007)还对基于任务的行为评分与基于维度的总体评分(OARs)之间的关系进行探讨, 结果发现评价中心中基于任务的评分能解释近 90%的维度总体得分, 这表明维度总体得分的实质仍体现的是受测者基于任务的行为表现; 而 Jackson, Stillman 和 Englert (2010)最近的研究则开始从系统模型的角度, 探讨受测者在基于任务的评价中心中评分的影响因素。总体上来看, 基于任务的评价中心与传统的评价中心相比有何优势, 它到底测了什么、是否有效等等, 这些问题都亟待更多的实证研究进一步探讨。

4.3 争议: 维度还是活动?

评价中心到底测量了什么? 维度取向和活动取向的研究各执一词, 争议不断。从大量有关基于维度的评价中心构想效度研究来看, 活动变异占主导这一普遍性结果已不容忽视, 传统的维度

取向观点受到极大挑战。活动取向认为我们应就此抛弃维度, 转而针对活动或任务对评价中心框架进行重新设计, 面对这样的观点, 研究者也提出了很多争论性意见。

首先, 维度不是评价中心特有的, 而是存在于绩效评价等很多人力资源管理功能中的“通用语言”(Rupp et al., 2008)。从理论角度, 维度之所以被广泛应用, 根本原因在于研究者普遍认为人的绩效表现(performance)是一个多维结构(Campbell, McCloy, Oppler, & Sager, 1993); 而从应用角度, 保留维度可以更好地使评价中心与其他人力资源管理过程有效衔接与整合。从这个意义上来看, 维度的存在是有价值和依据的。

其次, 基于任务的评价中心是否能涵盖工作复杂性的全貌仍受质疑(Connelly, Ones, Ramesh, & Goff, 2008; Howard, 2008)。由于去掉了维度, 基于任务的评价中心测的仅仅是受测者行为的样本, 而不是更抽象、概括的个人潜在特征, 并由这些行为样本预测将来的工作绩效。这样可能带来的后果是, 所选活动或任务情境有限, 不能有效代表实际工作的复杂性, 进而导致预测力下降。这样的质疑也使一些活动取向的研究者开始认识到, 在取消维度后, 还应更深入地去研究情境的特征, 以提高活动情境对工作的代表性(Lievens, Tett et al., 2009)。

最后, 基于任务的评价中心反馈方式也是备受争论的问题(Howard, 1997)。传统的基于维度的反馈比较明确, 而基于任务或活动的反馈对受测者来说就比较模糊、不自然, 缺乏概括性, 对于评价中心培训和发展的目的来说不够有效。但也有研究者认为, 基于任务的行为式反馈与基于维度的反馈相比, 更能够被受测者适应和接受(Jackson et al., 2010; Lance, 2008a)。二者孰优孰劣尚待更多的实证研究证实。

5 新的观点: 交互作用取向

前两种观点一直在争论评价中心中维度和活动谁更重要, 第三种观点则把注意力放在受测者行为表现及其影响因素上, 以此来解释评价中心构想效度问题。交互作用取向观点认为, 评价中心基于维度的构想效度不高, 受测者行为表现出跨情境的不一致性, 这是由维度与活动共同的交互作用影响而导致的(Lievens & Conway, 2001;

Hoffman & Meade, 2012; Lievens, De Koster, & Schollaert, 2008; Gibbons & Rupp, 2009)。也就是说, 活动不再被视为平行的测验, 不同的活动可能给受测者呈现了不同的情境要求(Neidig & Neidig, 1984; Howard, 2008), 从而引发了受测者在相同维度上不尽相同甚至是很不一致的行为表现, 最终导致维度变异降低而活动变异增高。这种观点来源于人格与社会心理学领域的交互作用观(interactionist perspective), 特别是人格心理学领域对行为解释的特质-情境之争。交互作用观认为人的行为是由特质和情境共同作用导致的, 并由此来解释行为跨情境不一致的现象。应用在评价中心领域, 交互作用取向也将受测者的行为归于个体差异与活动特征共同影响的结果, 而不是单独归于维度或者活动, 并试图回答二者如何影响受测者跨活动表现的一致性。虽然已有少数研究沿此思路从活动特征角度探讨评价中心构想效度问题(e.g., Neidig & Neidig, 1984; Highhouse & Harris, 1993; Schneider & Schmitt, 1992), 但这些研究缺乏明确的理论框架。近来, 研究者开始借用更多的理论为依据指导交互作用取向的研究, 这其中, 认知情感人格系统理论(Cognitive-affective Personality System theory, CAPS)与特质激活理论(Trait Activation Theory, TAT)最受研究者所重视(Gibbons & Rupp, 2009; Jansen, Lievens, & Kleinmann, 2011; Haaland & Christiansen, 2002; Lievens, Chasteen, Day, & Christiansen, 2006; Lievens et al., 2008; Lievens, Tett et al., 2009; Hoffman & Meade, 2012)。

Mischel 和 Shoda (1995)提出的认知情感人格系统理论对行为的跨情境不一致性进行了详细解释。这种理论认为, 所谓人格, 即是个体在面对不同情境时表现出的一系列类似“如果...那么...”行为脚本的稳定行为模式。当人面临情境时, 情境特征会激发个体一系列情感与认知心理表征, 在此基础上特定的行为模式(即行为脚本)被激活, 个体行为产生。而 CAPS 理论假设, 行为之所以会产生跨情境的不一致, 这是因为人对情境特征的知觉是存在个体差异的。CAPS 理论故此区分了名义情境(nominal situation)与心理情境(psychological situation)两个概念, 前者指情境中被大家统一理解或界定的特征, 后者则是会引发不同个体不同理解或解释的情境特征。即使两个

情境在名义上是相同的,个体在心理上对它们的理解也可能不同,从而造成行为的差异。这一理论应用于评价中心领域颇有意义,可以为受测者维度评分为何存在跨活动不一致提供另一种解释。根据 CAPS 理论,我们可以假设,不同活动间行为的一致性只有当受测者在不同活动上知觉到相同情境要求时才可能实现。Kleinmann (1993) 的研究就发现,当活动的透明度(transparency)高,即受测者能准确识别不同活动所测量的同一维度时,维度评分的聚合效度最好。Jansen 等人(2011)的研究也发现,在评价中心中,只有当受测者准确识别出活动的情境要求时,与之相关的特质才会被触发;这种所谓的效标识别能力(ability to identify criteria) (Melchers, Klehe, Richter, Kleinmann, König, & Lievens, 2009; Kleinmann, Ingold, Lievens, König, Melchers, & Jansen, 2011; König, Melchers, Kleinmann, Richter, & Klehe, 2007)在个人特质与相关维度表现之间可以起到调节作用,只有在准确识别到情境要求的受测者中,宜人性才会与评价中心中与之相应的合作性维度得分显著相关。

另一方面,特质激活理论也开始在解释评价中心构想效度问题上发挥重要作用(Lievens et al., 2008; Lievens, Tett et al., 2009)。特质激活理论与 CAPS 理论有很多相似之处,它将交互作用观应用于工作领域,对特质如何影响工作行为进行解释。TAT 认为人的行为是对情境中与特质有关的线索的反应;它的核心假设在于,这些能激活人行为的情境特征,可以由特质激活潜力(trait activation potential)来描述,特质激活潜力越高,一个情境能激发的相关行为的差异性就越大(Tett & Guterman, 2000; Tett & Burnett, 2003)。Tett 和 Guterman (2000)认为,特质激活潜力主要是由情境的特质相关性(trait relevance)与情境强度(strength)两个因素决定;前者指一个情境包含能引发某种特质相关行为的线索的程度,后者则体现了情境对行为要求的清晰度,情境越弱,对行为要求就越模糊,行为的差异性就越大。由此, TAT 为我们解释评价中心构想效度问题提供了有效的理论框架。基于 TAT 我们可以假设,只有在不同活动包含相似的特质相关情境线索,即不同活动都在某特质上具有高特质激活潜力时,维度评分才可能在活动间具有高的聚合效度。Haaland

与 Christiansen (2002)率先通过对一个小样本评价中心数据的分析来对此假设进行验证,结果发现,在那些高特质激活潜力的活动之间,维度评分的聚合效度确实更高。此后, Lievens 等人(2006)又通过 30 个评价中心数据的分析再次验证了这一假设。

6 总结与展望

自从 Sackett 和 Dreher (1982)发现评价中心构想效度问题以来,研究者不断从各个方面探寻评价中心构想效度证据。然而结果一次又一次证明,评价中心评分对于预测受测者未来表现上确实有效,但它却始终没有按照我们预先设想的那样以维度为基础实现有效测量;评价中心评分变异中,活动变异总是大于维度变异。30 年来,研究者一直尝试对此谜题进行破解,提出了很多不同的看法:从维度中心取向坚持认为是误差因素导致评价中心维度测量出现问题,到活动中心取向开始认识到活动不只是方法,它们也有独特的贡献,再到交互作用取向越来越重视维度和活动二者共同的作用,研究者对这一现象的理解和解释经历着不停的变化与发展。虽然目前评价中心构想效度谜题仍没有统一的答案,但也许正像 Ployhart (2006)所描述的那样,“我们离真正解决评价中心构想效度问题已经不远了”。

实际上,如果跳出评价中心研究领域,我们可以发现,这种所谓的构想效度问题并不为评价中心所独有,类似这种情境变异(或方法变异、评分者变异等)大于构想变异的结果在很多其他情境化测评的研究中也屡有发现,如人格测验(Biesanz & West, 2004)、结构化面试(Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004)、情境判断测验(Westring et al., 2009)、绩效评估(Lance, Hoffman, Gentry, & Baranik, 2008)等等。例如, Van Iddekinge 等人(2004)就发现在结构化面试中,不同形式的结构化面试方法(情境式面试与行为式面试)测量同一维度的得分之间相关较低; Westring 等人(2009)则发现在用多个情境测量多个成就动机特质的情境判断测验中,情境变异是特质变异的三倍。

所以,从一个更深层次的角度来理解,评价中心构想效度谜题表面上是维度与活动之争的问题,本质却反映了“人的行为背后是特质还是情

境起决定作用”这一问题;表面上是一个心理测量学问题,本质上却反映了“人的行为的跨情境一致性是否存在、何时存在”这一人格与社会心理学问题。或许,从这个角度出发,我们能看到维度中心、活动中心与交互作用这三种不同取向观点的共同之处,即它们都认为评价中心中受测者的行为是个人因素与情境因素共同作用的结果,而区别在于各自强调或聚焦的点不同。也许,未来评价中心构想效度的研究不再关注“测量维度还是活动”的问题,而更关注“维度与活动(任务)如何共同作用”的问题;研究重视的不再是这三种取向孰优孰劣,而是如何融合不同取向的研究使评价中心在现有的基础上发挥出更强的有效性。

从这个意义上来说,未来评价中心构想效度的研究应在传统的维度中心取向之外给予活动中心取向足够重视,并重点发展交互作用取向。对于维度中心取向,我们不能因为构想效度问题的存在而简单把它丢弃,相反,未来的研究仍需在评价中心设计修正的道路上继续深入探索,借助各种测量理论与统计方法,在一定范围内不断改善维度测量的质量;另外,研究需更加重视对维度本身的探讨,回答什么样的维度适合作评价中心测量构想的问题,并进一步建立更为系统的、多层级的维度框架体系,以规范对维度的操作和界定。对于活动中心取向,未来它应该获得更多的研究重视,研究重点将放在对活动或任务自身的分析和探讨上:活动或任务应该包含哪些情境特征、我们如何对情境特征进行界定和分类等,以此为基础并借鉴工作样本研究来探讨基于任务的评价中心是否有效以及如何有效等等。此外,基于任务的反馈其效果如何也是研究者需要进一步关注的问题。最后,交互作用取向将在评价中心研究领域扮演越来越重要的角色,未来的研究需在 TAT 及 CAPS 等理论指引下,不仅探讨评价中心中出现受测者跨活动表现不一致的原因,更要分析在什么条件下、何种因素会影响到这种不一致;不仅从理论角度出发针对构想效度问题进行事后分析,更要将交互作用观融合于评价中心设计和操作的实践中去。

参考文献

- 王博,田效勋,邵燕萍,车宏生.(2010).评价中心结构效度的多元概化理论分析研究.《心理学探新》,30,66-71.
- 王忠军,龙立荣.(2006).评价中心的结构效度研究.《心理

科学进展》,14,426-432.

吴志明,张厚粲.(2001).评价中心的构想效度和结构模型.《心理学报》,33,372-378.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D. C.: American Educational Research Association.

Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology*, 91, 555-566.

Arthur, W., 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.

Arthur, W., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 105-111.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, 93, 435-442.

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, 26, 813-835.

Atkins, P. W. B., & Wood, R. E. (2002). Self-versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology*, 55, 871-904.

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the big five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, 72, 845-876.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74, 478-494.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, 91, 1114-1124.

Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, 75, 173-182.

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 131-133.

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, J. P., McCloy, R., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.
- Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology*, 69, 167–181.
- Connolly, B. S., Ones, D. S., Ramesh, J., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 121–124.
- Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality*, 12, 85–108.
- Dugan, B. (1988). Effects of assessor training on information use. *Journal of Applied Psychology*, 73, 743–748.
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72, 493–511.
- 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.
- Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*, 35, 1154–1180.
- Goodge, P. (1988). Task based assessment. *Journal of European Industrial Training*, 12, 22–27.
- Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, 55, 137–163.
- Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology*, 78, 675–678.
- Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, 57, 639–683.
- Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment centre observation procedures: An experimental comparison of traditional, checklist and coding methods. *International Journal of Selection and Assessment*, 6, 222–231.
- Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology*, 23, 140–155.
- Hoffman, B. J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise rating. *International Journal of Selection and Assessment*, 20, 82–97.
- Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology*, 64, 351–395.
- Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, 50, 455–470.
- Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality*, 12, 13–52.
- Howard, A. (2008). Making assessment centers work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 98–104.
- Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 80, 897–913.
- 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.
- Jackson, D. J. R., Barney, A. R., Stillman, J. A., & Kirkley, B. (2007). When traits are behaviors: The relationship between behavioral responses and trait-based overall assessment center ratings. *Human Performance*, 20, 415–432.
- Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, 18, 213–241.
- Jackson, D. J. R., Stillman, J. A., & Englert, P. (2010). Task-based assessment centers: Empirical support for a systems model. *International Journal of Selection and Assessment*, 18, 141–154.
- Jansen, A., Lievens, F., & Kleinmann, M. (2011). Do individual difference in perceiving situational demands moderate the relationship between personality and assessment center dimension ratings? *Human Performance*, 24, 231–250.
- Jones, R. G., & Born, M. P. (2008). Assessor constructs in use as the missing component in validation of assessment center dimensions: A critique and directions for research. *International Journal of Selection and Assessment*, 16, 229–238.
- Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions. *Personnel Psychology*, 47,

- 109–121.
- Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, 78, 988–993.
- Kleinmann, M., Ingold, P. V., Lievens, F., König, C. J., Melchers, K. G., & Jansen, A. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review*, 1, 128–146.
- Kleinmann, M., & Koller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality*, 12, 65–84.
- Kleinmann, M., Kuptsch, C., & Koller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review*, 45, 67–84.
- Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance*, 15, 325–337.
- Kolk, N. J., Born, M. P., & van der Flier, H. (2004). A triadic approach to the construct validity of the assessment center: The effect of categorizing dimensions into feeling, thinking, and power taxonomy. *European Journal of Psychological Assessment*, 20, 149–156.
- König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. -C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment*, 15, 283–292.
- Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior and Personality*, 12, 129–144.
- Lance, C. E. (2008a). Why assessment centers don't work the way they're supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 84–97.
- Lance, C. E. (2008b). Where have we been, how did we get there, and where shall we go? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 140–146.
- Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis) application of the multitrait-multimethod methodology in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in organizational and social research* (pp. 337–360). New York: Routledge.
- Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, 89, 22–35.
- Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, A. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance*, 20, 345–362.
- Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review*, 18, 223–232.
- Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, 89, 377–385.
- Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent crosssituational specificity, not method bias. *Human Performance*, 13, 323–353.
- Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, 10, 430–448.
- Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment*, 6, 141–152.
- Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, 86, 255–264.
- Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, 22, 203–221.
- Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, 87, 675–686.
- Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 117–120.
- Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*, 86, 1202–1222.
- Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied*

- Psychology*, 91, 247–258.
- Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance*, 22, 375–390.
- Lievens, F., De Koster, L., & Schollaert, E. (2008). Current theory and practice of assessment centers: The importance of trait activation. In S. Cartwright & C. L. Cooper (Eds.), *The Oxford Handbook of Personnel Psychology* (pp. 215–233). New York: Oxford University Press.
- Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 245–286). Chichester: John Wiley & Sons, Ltd.
- Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Bingley: JAI Press.
- Lievens, F., & Van Keer, E. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organizational Psychology*, 74, 373–378.
- Lowry, P. E. (1995). The assessment center process: Assessing leadership in the public sector. *Public Personnel Management*, 25, 443–450.
- Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality*, 12, 53–62.
- Melchers, K. G., & König, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 125–127.
- Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment center candidates on rating quality. *International Journal of Selection and Assessment*, 18, 329–341.
- Melchers, K. G., Klehe, U. -C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, 22, 355–374.
- Meriac, J. P., Hoffman, B. H., 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.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268.
- Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology*, 69, 182–186.
- Petrides, K. V., Weinstein, Y., Chou, J., Furnham, A., & Swami, V. (2010). An investigation into assessment centre validity, fairness, and selection drivers. *Australian Journal of Psychology*, 62, 227–235.
- Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, 32, 868–897.
- Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology*, 43, 71–84.
- Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centers: Dimensions into exercises won't go. *Journal of Occupational Psychology*, 60, 187–195.
- Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance*, 13, 355–370.
- Rolland, J. P. (1999). Construct validity of in-basket dimensions. *European Review of Applied Psychology*, 49, 251–259.
- Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 116–120.
- Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology*, 68, 25–47.
- Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, 67, 401–410.
- Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology*, 69, 187–190.
- Sackett, P. R., & Harris, M. J. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business Psychology*, 3, 214–229.
- Sackett, P. R., & Tuzinski, K. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.), *How people evaluate others in organizations* (pp. 111–129). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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.
- 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.
- Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, 77, 32–41.
- Schuler, H. (2008). Improving assessment centers by the trimodal concept of personnel assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1, 128–130.
- Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L., Jr. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology*, 39, 565–578.
- Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88, 500–517.
- Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, 34, 397–423.
- Thornton, G. C., III, & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resources Management Review*, 19, 169–187.
- Thornton, G. C., III, & Rupp, D. R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.
- Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, 17, 71–93.
- Westring, A. J., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance*, 22, 44–63.
- 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.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189–205.
- Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, 15, 134–161.

The Construct Validity Puzzle of Assessment Centers: Are We Measuring Dimensions or Exercises?

BIAN Ran; GAO Qin; CHE Hongsheng

(Beijing Key Lab of Applied Experimental Psychology, School of Psychology,
Beijing Normal University, Beijing 100875, China)

Abstract: As high-fidelity simulations, assessment centers (ACs) are designed to measure a number of dimensions in various job-related exercises. Over the past 30 years, a great deal of research has indicated that ACs exhibit high content and criterion-related validity, but low construct-related validity. AC ratings substantially reflect the effects of the exercises and not the dimensions they were designed to reflect. This so-called AC “construct validity puzzle” drew a lot of attention from researchers. Three main research perspectives have gradually emerged: dimension-based approach, exercise-based approach, and interactionist approach, which focus on controlling factors causing measurement errors to improve the quality of dimension measurement, removing the dimensions from AC framework and focusing solely on the exercises or tasks, and taking the interaction between dimensions and exercises into account, respectively. Future research should pay more attentions to the exercise-based approach beyond the traditional dimension-based approach, and place great emphasis on the interactionist approach.

Key words: assessment centers; construct-related validity; dimension-based approach; exercise-based approach; interactionist approach