

# 情景判断测验效度的研究述评

刘润香<sup>1,2</sup>,戴海琦<sup>1</sup>,王 鹏<sup>1</sup>

(1. 江西师范大学 心理学院,南昌 330022;2. 南昌大学 心理健康教育中心,南昌 330031)

**摘 要:**文章在回顾近年来情景判断测验研究的基础上,总结了情景判断测验的效标关联效度、结构效度、增量效度和情景判断测验效度的影响因素。研究发现情景判断测验有较高的效标关联效度,是一种较好的人才选拔工具;情景判断测验是一种测量方法,可以用来测量指定的结构;情景判断测验对认知能力、人格、工作知识等变量具有增量效度;试题特性、测验开发模式、研究设计、评分方式等会影响情景判断测验的效度。

**关键词:**情景判断测验;效标关联效度;结构效度;增量效度

**中图分类号:**B841.2

**文献标识码:**A

**文章编号:**1003—5184(2010)03—0074—06

## 1 引言

情景判断测验(Situational Judgement Tests, SJTs)是近二十年越来越受欢迎的测评方法之一,被称为人才测评技术的新贵。对于什么是“情景判断测验”,不同的学者有各自独特的见解,但大家都比较认可的一般性操作定义是:给被试呈现一个与工作相关的情景,并提供这一情景条件下可能产生的行为反应,令被试针对这些行为反应进行判断、评价和选择,然后将被试的作答与标答进行比较后赋分,具体的评分方法因设问方式不同而相应地改变;根据评分结果判断被试是否适合从事这项工作的测验即情景判断测验<sup>[1,2]</sup>。作为人才测评中应用越来越广的情景判断测验,其效度问题备受研究者们的关注。该文在总结近年来情景判断测验研究的基础上,拟从情景判断测验的效标关联效度、结构效度、增量效度、效度影响因素等几方面展开探讨。

## 2 情景判断测验的效标关联效度

评判一个选拔工具是否有效,最关键的是看该工具是否能预测工作绩效,具有高的效标关联效度。自从情景判断测验这种测验形式出现以来,其效标关联效度就为研究者们所关注,已经有大量的实证研究检验情景判断测验是否有好的预测效度。不过,他们的研究结果并不一致,甚至相反,比如 Weekly 和 Jones(1997)报告情景判断测验的预测效度较低,Motowidlo 等(1990)报告情景判断测验与

工作绩效有中等程度的相关,而 Stevens 和 Campion(1999)发现情景判断测验与工作绩效相关较高<sup>[3-5]</sup>。McDaniel 等人针对这种情况,引进元分析的方法,检索了 1887~2000 年情景判断测验的实证研究,并在此基础上进行元分析,发现情景判断测验与工作绩效平均相关为 0.34<sup>[1]</sup>。这一研究结果说明情景判断测验可以与人事评价和选拔中的很多工具相媲美,它可以较好地预测个体现在和将来的工作表现。

应该说,前人的研究为情景判断测验的发展和推广起到了很大的作用。不过,这些研究很多出于选样的方便性和可行性,选取的被试是在职者而非应聘者,两者的作答动机明显不同,应聘者更可能存在作伪的现象,因此很有必要考查不同的被试群体下情景判断测验的效度。此外,在效标的选取上,有相当一部分研究采用的是业绩这一单一变量即任务绩效作为效标。但是,工作绩效分为任务绩效和周边绩效,并且近期的研究发现在评价直接上级、同级别同事、直接下级时,周边绩效对个体总体工作绩效均有显著影响<sup>[6]</sup>;更有研究表明,在人事评价与选拔中,认知能力能有效地预测任务绩效,而人格则对周边绩效有很好的预测效果<sup>[7]</sup>。因此,仅选用任务绩效作为效标所得到的研究结果并不全面,将来的研究还应该探讨情景判断测验对周边绩效的预测作用。

### 3 情景判断测验的结构效度

情景判断测验虽然被广泛用于应用研究,并被大量研究验证对工作绩效有预测作用,但为什么情景判断测验能够预测工作绩效,它又到底测量了什么呢,这一直是研究者的争论焦点。目前,主要有以下几种观点。

#### 3.1 单一结构的观点

Sternberg 等人的研究表明情景判断测验与一般认知能力、人格不相关,基于此,他们认为情景判断测验测量到的是内隐知识。这种知识是个体在没有或较少外部支持的条件下获得的,与个体价值目标密切关联的知识,一般难以用言语来表达,属于程序性知识<sup>[8]</sup>。对 Sternberg 等人的观点,有人认为他们的研究样本量太小,其结果值得怀疑;还有人指出他们的研究在探讨情景判断测验与人格的关系时,采用的效标并不是目前被广泛使用的大五人格测验,其研究结果值得商榷<sup>[9]</sup>。

#### 3.2 多维结构的观点

与单一结构观点相对立的是多维结构的观点。目前验证情景判断测验是多维结构的方法主要有两类:相关法和因素分析法。McDaniel 和 Whetzel (2005)的研究发现,同一个情景下被试在各个行为反应项上的得分与一般认知能力、人格的相关并不一致。比如有一个情景题,被试在第一个反应项上得分与一般认知能力相关为 0.10,与大五人格中的宜人性格相关为-0.13;而第二个和第五个反应项与一般认知能力的相关分别为 0.11 和 0.13;第三个反应项与宜人性格的相关为-0.16;第四个反应项与一般认知能力的相关为-0.17<sup>[9]</sup>。这个研究结果说明由此类题目组成的情景判断测验测量了多个维度,并且每个题目在多个维度上具有负荷。此外,McDaniel 等人的元分析发现,情景判断测验与一般认知能力的平均相关为 0.46,与大五人格中宜人性格平均相关为 0.25;与责任心的平均相关为 0.26;与情绪稳定性的平均相关为 0.31,但与外向性与开放性的相关较低,分别为 0.06 和 0.09<sup>[1,10]</sup>。这些研究结果可以表明,情景判断测验并不像 Sternberg 等人所认为的那样只是测量了与一般认知能力、人格

无关的内隐知识。

另一种验证情景判断测验所测为多维结构的方法为因素分析法。如 Lee 和 Kim(2002)抽取了五个因子和一个方法因子;Legree 等人(2003)对两个自动化驾驶情景判断测验均抽取了三个因子;Chan 和 Schmitt(1997)对一个用于选拔的情景判断测验做因素分析,得到四个因子<sup>[9]</sup>。这些研究没有一个说明情景判断测验所测为单一结构。

总之,从以上提供的研究证据,McDaniel 等人认为情景判断测验的确测量的不是单一结构而是多维结构。

#### 3.3 测量方法的观点

虽然 McDaniel 等人的元分析发现情景判断测验与一般认知能力和大五人格具有一定的相关,但具体的研究得出的结论并不相同。比如 Chan 和 Schmitt(2002)的研究发现情景判断测验与一般认知能力的相关为-0.20<sup>[11]</sup>;而 Weekly 和 Jones (1997)的研究则发现情景判断测验与一般认知能力的相关较高<sup>[3]</sup>。情景判断测验与人格的相关研究结果也存在不一致,Chan 和 Schmitt(2002)发现情景判断测验与责任心、外向性、宜人性格的相关为 0.23、0.24 和 0.29,与神经质的相关为-0.20<sup>[11]</sup>;Lievens 和 Coetsier(2002)则报告情景判断测验与大五人格各维度的相关没有高于 0.15 的<sup>[12]</sup>;Clevenger (2001)的研究发现在三个不同样本中,情景判断测验与大五人格中的责任心相关不一样<sup>[13]</sup>。为什么研究结果会不一致? Chan 和 Schmitt(2002)、McDaniel(2005)认为,一个可能的解释就是情景判断测验是一种测量方法,像面试一样,能够测量各种各样的结构,而这取决于测验的内容<sup>[11,9]</sup>。

情景判断测验结构的复杂性及不同学者对情景判断测验结构效度的分歧,其主要原因是测验的开发程序所致。已有的大多数情景判断测验以关键事件技术为基础,采用绩效导向的开发模式,不管测验的结构,而影响绩效的结构有时不是单一的结构,也因此不同的情景判断测验有的是单一结构,有的是多维结构;此外,对于不同的领域,影响绩效的结构也是不同的,这势必导致情景判断测验结构的复杂性。而情

景判断测验目前越来越流行的结构导向开发的模式,是在测验编制之前就已经确定欲测哪些结构,并有针对性地设计不同的情景来测量这些结构,这也支持了情景判断测验测量方法的观点<sup>[14,15]</sup>。

总的来说,情景判断测验是一种测量方法的观点,使大家能够更好地理解情景判断测验的本质,从而更好地考查和验证测验的结构及其内在关系,也能够更好地理解情景判断测验对工作绩效等的预测作用。当然,其前提是要找到那些影响工作绩效的结构,然后设计相应的情景测量这些结构,这是确保情景判断测验高预测效度的保证。

#### 4 情景判断测验的增量效度

前文已指出,情景判断测验具有一定的预测作用,那么,它对传统的和目前广泛使用的预测指标的增量效度又如何呢,这也是众多研究者所关心的一个问题。

目前,已经有一些元分析研究和实证研究表明情景判断测验对认知能力、人格、工作经验这些已证明能预测工作绩效的变量有增量效度。Weekly 和 Jones(1997)发现选拔零售员的一个情景判断测验对认知能力和工作经验的共同增量效度为 0.025;他研究中的另一个情景判断测验在两种标答策略下(实证法和理性法)对这两个预测指标的共同增量效度为 0.057 和 0.096<sup>[3]</sup>。Clevenger 等人(2001)在 Weekly 等人的研究基础上,进一步考查了情景判断测验对认知能力、工作知识、责任心、工作经验四个变量的共同增量效度,发现在政府官员、交通服务人员、工程师三个样本中的增量效度分别为 0.026、0.016、0.017<sup>[13]</sup>。此外,Weekly 和 Ployhart(2005)的研究及 McDaniel 等人(2001)的元分析研究也发现情景判断测验对认知能力有增量效度<sup>[1,16]</sup>。

但是,不同情景判断测验对认知能力、人格、工作经验这些变量的增量效度有时不一样。一个可能的原因是测验的反应指导语不一样引起的。McDaniel 等人(2007)的元分析研究发现,最佳行为指导语下的情景判断测验对认知能力、人格的增量效度以及两者的共同增量效度比典型行为指导语下的情景判断测验高<sup>[17]</sup>。另一个可能的原因是预测的

工作绩效的维度不一样。Chan 和 Schmitt(2002)的研究发现预测不同的工作绩效,情景判断测验对认知能力、人格、工作经验的共同增量效度不一样,预测任务绩效、成就动机、人际交往、总体绩效的增量效度分别为 0.05、0.08、0.03、0.04<sup>[11]</sup>。O'Connell 等人(2007)的研究也支持这一研究结果,当预测任务绩效时,情景判断测验对认知能力、人格的增量效度以及他们的共同增量效度为 0.11、0.10、0.03;而在预测周边绩效时,对应的值为 0.07、0.03、0.01<sup>[18]</sup>。

当然,引起情景判断测验增量效度不一样的另一个重要原因是这种测验结构效度的复杂性。随着情景判断测验测量方法观点的提出,可以根据需要设计对某种特质具有高增量效度的情景判断测验。

#### 5 情景判断测验效度的影响因素

##### 5.1 试题的特性对测验效度的影响

###### 1) 仿真度对情景判断测验效度的影响

情景判断测验可分为纸笔式和录像式,前者采用文字描述,后者采用录音口语描述或影视短片展现。显然地,录像式的情景判断测验仿真度更高。McDaniel 等人认为情景判断测验的仿真度既会影响测验的效标关联效度,又会影响测验的结构效度。Lievens 和 Sackett(2006)的实证研究也支持了这一观点。该研究发现录像情景判断测验的效标关联效度比纸笔式的高(效标关联效度分别为 0.34 和 0.08);进一步的研究发现录像情景判断测验与认知能力的相关比纸笔式的低,相应地,其对认知能力的增量效度比纸笔式的高,说明两种形式的情景判断测验有不同的结构效度<sup>[19]</sup>。不过,虽然录像情景判断测验较纸笔式表现出更高的预测效度,但是这种形式的测验开发成本高,施测也更不方便。因此,在实际测验开发过程中应根据实际情况选择合适的形式。

###### 2) 试题认知复杂性对情景判断测验效度的影响

情景判断测验试题的题干有的更为详细、具体、复杂,但有的更为简单笼统。McDaniel 等人(2001)认为题干越复杂、详细程度越高的情景判断测验效度会越高。其原因可能是试题越详细具体,那么试



题的表面效度就越高,进而被试的作答积极性也会相应地提高;另一个可能原因是复杂详细的试题要求更高的阅读能力,因而这些试题可能与认知能力相关更高。一般地,认知能力对工作绩效的预测效度较其它非认知能力的高,因此与认知能力相关较高的试题预测效度也相应地更高。不过,McDaniel 等人(2001)的元分析结果正好与之相反,他们分析有可能是因为元分析是基于少量研究得出的结论,不一定可信。因此,试题认知复杂性对情景判断测验效度的影响到底如何,还需要更多的实证研究。

3)反应指导语对情景判断测验效度的影响

情景判断测验的反应指导语大致可以分为两类,一类是最佳行为指导语,一类是典型行为指导语,前者要求被试对行为反应项的有效性进行评价,后者要求被试选择自己可能采取的行为反应项或评价自己采取各行为反应项的可能性大小。反应指导语可能会影响测验的结构效度,最佳行为指导语的试题测的是被试的最佳行为,因此可能测到更多的认知能力;而典型行为指导语的试题测的是被试的典型行为,可能更多测到的是人格。McDaniel 等人(2003)比较了两种反应指导语下情景判断测验与认知能力、宜人性、责任心、情绪稳定性的相关,研究结果支持了上述观点。McDaniel 等人(2007)关于反应指导语的元分析研究也证实了这一观点<sup>[17,20]</sup>。

因此,如果欲测量的是人格特质,那么使用典型行为指导语更合适;如果欲测量的是能力特质,则使用最佳行为指导语更佳。不过,Lievens 和 Sackett(2009)发现在高利害(High-stakes)情境下测试,两种指导语下的情景判断测验的效标关联效度有显著差异<sup>[21]</sup>。一个可能原因就是被试的作伪动机,在低利害情境下,被试在典型行为指导语下更会作伪<sup>[22]</sup>;而在高利害情境下,无论指导语是什么,被试都有可能具有高水平的作伪动机。这也意味着,高利害情境下,情景判断测验可以随便用哪种指导语,当然,这还需要更多的实证研究来证实。

5.2 情景判断测验的开发模式对测验效度的影响

总结而言,情景判断测验的开发模式可以分为两大类:绩效导向开发模式和结构导向开发模式。

两者的区别主要表现在三个方面:1)试题题干的编写。绩效导向的开发模式一般采用关键事件法;结构导向开发模式一般采用理论导向法<sup>[23]</sup>。2)试题行为反应项的编写。前者使用主题专家法(SMEs),后者采用结构导向的方法。3)行为反应项的标答。常用的标答方法有实证式的、理论式的、专家法<sup>[24]</sup>。绩效导向的开发模式常用实证式,而结构导向的开发模式采用理论式和专家法的标答方法。

在以往传记式(Biodata)测验开发过程中,前人认为基于理论构思的测验在内容效度和结构效度上具有较好的稳定性,测验效度也更具有外推性。但是,绩效导向的开发模式也有其优势。从以往的研究来看,基于绩效导向开发的测验从保证效标关联效度出发,能够较好地捕捉预测指标与效标间的关系。那么,情景判断测验的情况如何呢?金杨华(2003)的研究表明,结构导向开发的情景判断测验更具有稳定性,与以往的 Biodata 测验的结论一致<sup>[25]</sup>。但是,目前关于测验开发模式对情景判断测验效度的影响研究还是太少,到底哪种测验开发模式更好,还需要更多的实证研究来证实,这将是一个非常值得研究的问题。

影响情景判断测验的因素非常多,除了以上这些因素外,还有情景所涉及的内容领域、研究的设计、测验开发是否基于工作分析、评分方式等。McDaniel(2001)的元分析已经证实基于工作分析基础上的情景判断测验效标关联效度(0.38)比没有经过工作分析开发的测验效率高(0.29);他们还发现预测效度设计的情景判断测验效标关联效度(0.18)比同时效度设计的测验效率低(0.35)<sup>[1]</sup>。Bergman 等人(2006)比较了 11 种评分方法对情景判断测验效度的影响,结果发现用新手和专家作答比较的评分方法(Novice vs. expert)效标关联效度最高;只有主题专家法(SEMs)、混合法、实证法几种评分方式对认知能力、人格具有显著增量效度<sup>[24]</sup>。从这些研究结果可以知道,在开发情景判断测验时,也必须把这些因素考虑进去。

## 6 小结

通过回顾前人的研究,可以发现情景判断测验有较高的效标关联效度,是一种可以用来测量指定结构的测量方法,还对认知能力、人格、工作知识等变量具有增量效度,总之,情景判断测验是一种较好的人才选拔工具。

但是,并不是所有的此类测验都具有好的效度,这取决于测验的开发过程。如果测验采取绩效导向的开发模式,则首先要确保搜集的关键事件具有关键事件性、信息完备性、事件典型性和普遍性的性质,并且要选取合适准确的效标;如果测验采取结构导向的开发模式,最关键的是要探究清楚影响绩效的结构有哪些,然后有针对性的编制情景判断题来测量这些特质,从而保障测验的效度。此外,测验试题的特性、研究设计、评分方式等也都会影响测验的效度,在测验开发过程中还需要充分考虑这些因素。

## 参考文献

- McDaniel M A, Morgeson F P, Finnegan E B, et al. . Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 2001, 86: 730—740.
- Lievens F, Peeters H, Schollaert E. Situational judgment tests: A review of recent research. *Personnel Review*, 2008, 37: 426—441.
- Weekley J A, Jones C. Video—based situational testing. *Personnel Psychology*, 1997, 50: 25—49.
- Motowidl S J, Dunnette M D, Carter G W. An alternative selection procedure: The low—fidelity simulation. *Journal of Applied Psychology*, 1990, 75: 640—647.
- Stevens M J, Campion M A. Staffing work team: Development and validation of a selection test for teamwork settings. *Journal of Management*, 1999, 25: 207—208.
- 陈胜军. 周边绩效与总绩效的评价关系研究. *山西财经大学学报*, 2008, 30: 84—89.
- 沈峥嵘, 王二平. 关系绩效研究. *心理科学进展*, 2004, 12: 924—931.
- Insch G S, McIntyre N, Dawley D. Tacit knowledge: A refinement and empirical test of the academic tacit knowledge Scale. *The Journal of Psychology*, 2008, 142: 561—579.
- McDaniel M A, Deborah L W. Situational judgment tests research: Informing the debate on practical intelligence theory. *Intelligence*, 2005, 33: 515—525.
- McDaniel M A, Nguyen N T. Situational Judgment Tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 2001, 9: 103—113.
- Chan D, Schmitt N. Situational judgment and job performance. *Human Performance*, 2002, 15: 233—254.
- Lievens F, Coetsier P. Situational test in student selection an examination of predictive validity, adverse impact and construct validity. *International Journal of Selection and Assessment*, 2002, 10: 245—257.
- Clevenger J, Pereira G M, Wiechmann D, et al. . Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 2001, 86: 410—417.
- Mumford T V, Iddekinge C H V, Morgeson F P, et al. . The team role test: development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology*, 2008, 93: 250—267.
- Ronald B, Michael F. Situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, 2009, 62: 229—258.
- Weekley J A, Ployhart R E. Situational judgment: Antecedents and relationships with performance. *Human Performance*, 2005, 18: 81—104.
- McDaniel M A, Hartman N S, Whetzel D L, et al. . Situational judgment tests, response instructions, and validity: A Meta—analysis. *Personnel Psychology*, 2007, 60: 63—91.
- OConnell M S, Hartman N S, McDaniel M A, et al. . Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment*, 2007, 15: 19—29.
- Lievens F, Sackett P R. Video—based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, 2006, 91: 1181—1188.
- McDaniel M A, Hartman N S, Whetzel D L, et al. . Situational judgment tests, knowledge, behavioral tendency,

and validity: A Meta-analysis. Paper Presented at 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL, 2003.

21 Lievens F, Sackett P R. The effects of response instruction on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, 2009, 94: 1095-1101.

22 Nguyen N T, Biderman M D, McDaniel M A. Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, 2005, 13: 250-260.

23 Weekley J A, Ployhart R E. *Situational judgment tests: Theory, measurement and application*. San Francisco, Jossey Bass, 2006.

24 Bergman M E, Drasgow F, Donovan M A, et al. . Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 2006, 14: 223-235.

25 金杨华. 人力资源胜任特征的情景评价与效度研究. 博士论文. 杭州: 浙江大学管理学院, 2003.

A Review of Validity for Situational Judgment Test

Liu Runxiang<sup>1,2</sup>, Dai Haiqi<sup>1</sup>, Wang Peng<sup>1</sup>

(1. School of Psychology, Jiangxi Normal University, Nanchang 330022;  
2. Center of Psychology and Education, Nanchang University, Nanchang 330031)

**Abstract:** We summarized the research of criterion-related validity, construct-related validity, incremental validity and the affecting factors of the situational judgment test on the basis of reviewing the advance of studying situational judgment test for recent years, which showed the situational judgment test had high criterion-related validity, it could become a superior personnel selection tool; the situational judgment test was a method for measurement that could be used to measure the specific structure; the situational judgment test had the incremental validity for the variables such as cognitive ability, personality, job knowledge etc. ; Besides, the affecting factors of the situational judgment test's validity mainly included the characteristics of items, the developing model of tests, the research design, and scoring method and so on.

**Key words:** situational judgment test; criterion-related validity; construct-related validity; incremental validity