COSMIN 偏倚风险评价清单

日期: 2018 年 7

联系

L.B. Mokkink, PhD

VU University Medical Center

Department of Epidemiology and Biostatistics

Amsterdam Public Health research institute

P.O. box 7057

1007 MB Amsterdam

The Netherlands

Website: www.cosmin.nl

E-mail: w.mokkink@vumc.nl

译者：施月仙[1]　尚少梅[1]　万巧琴[1]　于明明[1]　孙　萌[1]　黄亚琪[2]　韩明月[3]

单位（1 北京大学护理学院；2 天津医科大学护理学院；3 北京大学医学部人文学院）

联系：施月仙，北京市海淀区学院路 38 号北京大学医学部，北京，中国.

邮箱：nevergiveup2006@163.com

如何引用*COSMIN偏倚风险评价清单*

当使用COSMIN偏倚风险评价清单请参考以下研究：

Mokkink, L.B., De Vet, H.C.W., Prinsen, C.A.C, Patrick, D.L., Alonso, J., Bouter, L.M., et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. 在Quality of Life Research中发表.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., Vet, H. C., et al. COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures.

在Quality of Life Research中发表.

Terwee, C. B., Prinsen, C. A., Chiarotto, A., Vet, H. C., Westerman, M. J., Patrick, D. L., et al. COSMIN methodology for evaluating the content validity of Patient-Reported Outcome Measures: a Delphi study. 已投稿.

关于如何使用 COSMIN 偏倚风险清单的详细信息，见 “患者报告结局测量工具（Patient-Reported Outcome Measures，PROMs）系统评价的 COSMIN 方法学—用户手册”和“用于评估患者报告结局测量工具（PROMs）内容有效性的 COSMIN 方法—用户手册”，可从我们的网站获取： www.cosmin.nl.

缩写：

CTT – classical test theory，经典测试理论

DIF – differential item functioning，项目功能差异

IRT – Item response theory，项目反应理论

MGCFA – multi-group confirmatory factor analysis，多组验证性因素分析

MI – measurement invariance，测量不变性

NA – not applicable，不适用的

PROM – patient-reported outcome measure，患者报告结局测量工具

1PL model – 1 parameter IRT model，单参数项目反应理论模型

2PL model – 2 parameter IRT model ，双参数项目反应理论模型

说明

在需要为文章完成评价的方框中打“ √ ”

	COSMIN 偏倚风险清单
	方框1. PROM的开发
	方框2. 内容效度
	方框3. 结构效度
	方框4. 内部一致性
	方框5. 跨文化效度\测量等同性
	方框6. 信度
	方框7. 测量误差
	方框8. 校标效度
	方框9. 结构效度的假设检验
	方框10. 反应度

为了评估每个研究的方法学质量，即评估研究结果的偏倚风险，需要完成相应的COSMIN偏倚风险评估方框。为了确定每个研究的整体质量，应采用方框中任何评分标准的最低得分（即“最差分数计数”原则）。例如，对于信度研究，如果方框中的一个条目被评为“不充分的”，那么该信度研究的总体方法学质量就被评为“不充分的”。某些标准的回答存在“NA”（不适用）选项。例如，若一个关于结构效度的研究是基于CTT，那么IRT的标准是不适用的，所以对于这个特定的研究，该标准不应该被考虑在“最低分计数”中。对于不存在此选项的标准，这些单元格是灰色的，不应该使用。

方框 1. PROM 开发

1a. PROM 设计

总体设计要求		**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
1	是否清晰地描述了要测量的结构？	清晰地描述了结构			没有清晰地描述结构	
2	结构的来源清晰吗：是否使用了理论、概念性框架或疾病模型，或提供了清晰的原理，从而定义所测量的结构？	结构的来源清晰		结构的来源不清晰		
3	是否清晰地描述了所开发的 PROM 针对的目标人群？	清晰地描述了目标人群			没有清晰地描述目标人群	
4	是否清晰地描述了使用的背景？	清晰地描述了使用背景		没有清晰地描述使用背景		
5	开发 PROM 的研究是否在能够代表 PROM 开发的目标人群的样本中进行？	研究在能够代表目标人群的样本中实施	可以假定研究在能够代表目标人群的样本中实施，但是没有清晰地描述	怀疑研究是否在能够代表目标人群的样本中实施	研究没有在能够代表目标人群的样本中实施**(跳过 6-12 项)**	

概念引出（相关性和全面性）		**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
6	是否使用了恰当的定性数据收集方法以确定新的 PROM 的相关条目？	使用了广泛认可的或合理的质性方法，适用于结构和研究人群	可以假定定性的方法是恰当的，并且适用于结构和研究人群，但是没有清晰地描述	仅使用定量的（调查）方法或怀疑方法是否适用于结构和目标人群	使用的方法不恰当或不适用于结构或目标人群	
7	是否雇用了有经验的小组主持人/访谈者？	雇用了有经验的小组主持人/访谈者	小组主持人/访谈者经验有限或接受专门针对研究的培训	不清楚小组主持人/访谈者是否接受了培训，或小组主持人/访谈者没有经过培训，且没有经验		不适用
8	小组会议或小组访谈是否基于恰当的话题或访谈提纲？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有进行清晰地描述	不清楚是否使用了话题提纲，或怀疑话题或访谈提纲是否恰当，或没有使用提纲		不适用
9	小组会议或小组访谈是否录音并逐字转录？	所有小组会议或访谈均被录音记录并逐字转录	可以假定所有小组会议或访谈均被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈均被录音并逐字转录，或只有录音而没有逐字转录，或只在小组会议/访谈时做了笔记	没有录音、没有笔记	不适用
10	是否使用了恰当的方法分析数据？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	

11	是否至少有部分数据是独立编码？	至少 50%的数据由至少两名研究人员独立地编码	11-49%的数据由至少两名研究人员独立地编码	怀疑是否有两位研究者参与编码数据或仅有 1-10%的数据由至少两名研究人员独立地编码	仅一位研究者参与数据编码或没有编码数据	不适用
12	是否持续收集数据直至达到饱和？	有证据表明已经达到饱和	可以假定达到饱和	怀疑是否达到饱和	有证据表明没有达到饱和	不适用
13	对于定量研究（调查）：样本量是否合适？	≥100	50-99	30-49	<30	不适用

1b. 认知访谈研究或其它预实验					
	非常好	**合格**	**有问题的**	**不合格**	**不适用**
14 是否进行了认知访谈研究或其它预实验？	是			否（跳过 15-35 项）	
总体设计要求					
15 认知访谈研究或其它预实验是否在能够代表目标人群的样本中进行？	研究在能够代表目标人群的样本中进行	可以假定研究在能够代表目标人群的样本中进行，但是没有清晰地描述	怀疑研究是否在能够代表目标人群的样本中进行	研究没有在能够代表目标人群的样本中进行	
可理解性					
16 是否向患者问及 PROM 的可理解性？	是		否（跳过 17-25 项）	**不清楚（跳过 17-25 项）**	
17 是否所有条目都以最终形式进行了测试？	所有条目都以最终形式进行了测试	可以假定所有条目都以最终形式进行了测试，但是没有清晰地描述	不清楚是否所有条目都以最终形式进行了测试	条目没有以最终形式进行预实验或者在进行大量调整后没有重新进行测试	
18 是否使用了恰当的定性方法评价 PROM 指导语、条目、回答选项和回忆期的可理解性？	使用了广泛认可的或合理的定性方法	可以假定方法是恰当的，但是没有清晰地描述	只用了定量（调查）方法或怀疑方法是否恰当或不清楚患者是否被问及条目，回答选项或回忆期的可理解性，或患者没有被问及 PROM 指导语或回忆期的可理解性	使用方法不恰当或患者没有被问及条目或回答选项的可理解性	

19	每个条目是否在恰当数量的患者中进行了测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
20	是否雇用了有经验的主持人/访谈者？	雇用了有经验的小组主持人/访谈者	小组主持人/访谈者经验有限或接受专门针对研究的培训	不清楚小组主持人/访谈者是否接受了培训，或小组主持人/访谈者没有经过培训，且没有经验		不适用
21	访谈是否基于恰当的访谈提纲？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有清晰地描述	不清楚是否使用了话题提纲，或怀疑话题或访谈提纲是否恰当，或没有使用提纲		不适用
22	访谈是否录音并逐字转录？	所有小组会议或访谈均被录音并且逐字转录	可以假定所有小组会议或访谈均被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈均被录音并逐字转录，或只有录音而没有逐字转录，或只在小组会议/访谈时做了笔记	没有录音，没有笔记	不适用
23	是否使用了恰当的分析数据方法？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法或者怀疑方法是否恰当	方法不恰当	
24	是否至少两名研究者参与分析？	至少两名研究者参与了分析	可以假定至少两名研究者参与了分析，但是没有清晰地描述	不清楚是否有两名研究者参与分析，或只有一名研究者参与分析		

25	是否通过调整 PROM 恰当地解决了 PROM 指导语、条目、回答选项和回忆期的可理解性问题？	没有发现任何问题，或问题得到恰当地解决并且在必要时对 PROM 进行了调整并重新进行了测试	可以假定没有问题或问题得到恰当地解决，但是没有清晰地描述	不清楚是否有问题，或怀疑问题是否得到恰当地解决	问题没有得到恰当地解决，或对 PROM 进行了调整但条目在大量调整后没有重新进行测试	不适用

	全面性	非常好	合格	有问题的	不合格	不适用
26	是否向患者问及 PROM 的全面性？	是		否或不清楚（**跳过 27-35 项**）		
27	是否对条目的最终版本进行了测试？	对条目的最终版本进行了测试	可以假定对条目的最终版本进行了测试，但是没有清晰地描述	不清楚是否对条目的最终版本进行了测试，或在删除或增加条目后没有重新进行测试		
28	是否使用了恰当的方法评价 PROM 的全面性？	使用了广泛认可或合理的方法	可以假定方法恰当，但是没有清晰地描述或只使用了定量（调查）方法	怀疑方法是否恰当，或使用方法不恰当		
29	每个条目是否在恰当数量的患者中进行了测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
30	是否雇用了有经验的访谈者？	雇用了有经验的访谈者	访谈者经验有限或接受专门针对研究的培训	不清楚访谈者是否有经验，或访谈者没有经过培训且没有经验		不适用
31	访谈是否基于恰当的访谈提纲？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有清晰地描述	不清楚是否使用话题提纲，或怀疑话题或访谈提纲是否恰当，或没有使用提纲		不适用

32	访谈是否录音并逐字转录？	所有小组会议或访谈均被录音并逐字转录	可以假定所有小组会议或访谈均被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈均被录音并逐字转录，或只有录音而没有逐字转录，或只在小组会议/访谈时做了笔记，或没有录音和笔记		不适用
33	是否使用了恰当的方法分析数据？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当，或方法不恰当		
34	是否至少有两名研究者参与了分析？	至少两名研究者参与了分析	可以假定至少两名研究者参与了分析，但是没有清晰地描述	不清楚是否有两名研究者参与分析，或只有一名研究者参与分析		
35	是否通过调整 PROM 恰当地解决了 PROM 全面性的问题？	没有发现任何问题，或问题得到恰当地解决且在必要时对 PROM 进行了调整并重新进行了测试	可以假定没有问题或问题得到恰当地解决，但是没有清晰地描述	不清楚是否有问题，或怀疑问题是否得到恰当地解决，或对 PROM 进行了调整但条目在大量调整后没有重新进行测试	问题没有得到恰当地解决	不适用

方框 2. 内容效度
2a.询问患者关于相关性的问题

设计要求		**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
1	是否使用了恰当的方法询问患者每个条目是否与他们的病情体验有关？	使用了广泛认可或合理的方法	只使用了定量（调查）方法，或可以假定所使用的方法是适当的，但没有被清晰地描述	不清楚患者有无被询问过每一个条目是否是相关的，或对方法的恰当有过疑惑（怀疑方法是否是合适的）	所使用的的方法是不合适的，或患者没有被询问所有条目的相关性	
2	每个条目是否在适当数量的患者中进行了测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
3	是否雇用了有经验的小组主持人/访谈者？	雇用了有经验的小组主持人/访谈者	小组主持人/访谈者有受限的经历或被专门培训用于研究	不清楚是否小组主持人/访谈者受过培训，或小组主持人/访谈者没有受过培训，且没有经验		不适用
4	小组会议或访谈是否根据恰当的话题或访谈提纲安排？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有清晰地描述	不清楚是否话题提纲被使用，或怀疑话题或访谈提纲是否恰当，或没有使用提纲		不适用

5	小组会议或访谈是否录音并逐字转录？	所有小组会议或访谈被录音并逐字转录	可以假定所有小组会议或访谈被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈被录音并逐字转录，或只有录音而没有被逐字转录，或只在小组会议期间做了笔记	没有录音、没有笔记	不适用
分析						
6	用于分析数据的方法是否恰当？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	
7	是否至少有两名研究者参与了分析？	至少有两名研究者参与了分析	可以假定至少有两名研究者参与了分析，但是没有清晰地描述	不清楚是否至少有两名研究者参与了分析，或仅有一名研究者参与了分析		

2b 询问患者关于全面性的问题

设计要求		非常好	合格	有问题的	不合格	不适用
8	是否使用了评估 PROM 全面性的恰当方法？	使用了广泛认可或合理的方法	只有定量的（调查）方法被使用，或可以假定所使用的方法是恰当的，但没有清晰地描述	怀疑方法是否恰当	方法不恰当	
9	每个条目是否在适当数量的患者中进行了测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
10	是否雇用了有经验的小组主持人/访谈者？	雇用了有经验的小组主持人/访谈者	小组主持人/访谈者有受限的经历或被专门培训用于研究	不清楚是否小组主持人/访谈者受过培训，或小组主持人/访谈者没有受过培训，且没有经验		不适用
11	小组会议或访谈是否根据恰当的话题或访谈提纲安排？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有清晰地描述	不清楚是否话题提纲被使用，或怀疑话题或访谈提纲是否恰当，或没有使用提纲		不适用

12	小组会议或访谈是否录音并逐字转录？	所有小组会议或访谈被录音并逐字转录	可以假定所有小组会议或访谈被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈被录音并逐字转录，或只有录音而没有逐字转录，或只在小组会议期间做了笔记	没有记录、没有笔记	不适用
分析						
13	用于分析数据的方法是否恰当？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	
14	是否至少有两位研究者参与了分析？	至少有两名研究者参与了分析	可以假定至少有两名研究者参与了分析，但是没有清晰地描述	不清楚是否至少有两名研究者参与了分析，或仅有一名研究者参与了分析		

2c 询问患者关于可理解性的问题

设计要求	非常好	合格	有问题的	不合格	不适用
15 是否是一个用于评估 PROM 指导语、条目、回答选项和回忆期的可理解性的恰当的质性研究方法？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	只有定量的（调查）方法被使用，或怀疑方法是否恰当，或不清楚病人是否被询问过条目、回答方式和回忆期的可理解性，或病人没有被询问过 PROM 指导语的可理解性	所使用的方法是不合适的，或患者没有被询问过条目、回答方式和回忆期的可理解性	
16 每个条目是否在适当数量的患者中进行了测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
17 是否雇用了有经验的小组主持人/访谈者？	雇用了有经验的小组主持人/访谈者	小组主持人/访谈者有受限的经历或被专门培训用于研究	不清楚是否小组主持人/访谈者受过培训，或小组主持人/访谈者没有受过培训，且没有经验		不适用
18 小组会议或访谈是否根据恰当的话题或访谈提纲安排？	恰当的话题或访谈提纲	可以假定话题或访谈提纲是恰当的，但是没有清晰地描述	不清楚是否话题提纲被使用，或怀疑话题或访谈提纲是否恰当，或没有使用指南		不适用

19	小组会议或访谈是否录音并逐字转录？	所有小组会议或访谈被录音并逐字转录	可以假定所有小组会议或访谈被录音并逐字转录，但是没有清晰地描述	不清楚是否所有小组会议或访谈被录音并逐字转录，或只有录音而没有被逐字转录，或只在小组会议期间做了笔记	没有录音、没有笔记	不适用
分析						
20	用于分析数据的方法是否恰当？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	
21	是否至少有两位研究者参与了分析？	至少有两名研究者参与了分析	可以假定至少有两名研究者参与了分析，但是没有清晰地描述	不清楚是否至少有两名研究者参与了分析，或仅有一名研究者参与了分析		

2d. 询问专业人员关于相关性的问题					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
22 是否使用了恰当的方法来询问专业人员，以确定每个条目与感兴趣的结构相关？	使用了广泛认可或合理的方法	只有定量（调查）方法被使用，或可以假定所使用的方法是恰当的，但是没有清晰地描述	不清楚专业人员有无被询问过每一个条目是否是相关的，或怀疑方法是否恰当	所使用的方法是不恰当的，或专业人员没有被询问所有条目的相关性	
23 是否纳入来自所有相关学科的专业人员？	纳入了来自所有相关学科的专业人员	可以假定纳入了来自所有相关学科的专业人员，但是没有清晰地描述	怀疑是否纳入了来自所有相关学科的专业人员，或没有纳入相关学科的专业人员		
24 每个条目是否在适当数量的专业人员中进行过测试？ 对于质性研究 对于量性（调查）研究	 ≥7 ≥50	 4-6 ≥30	 <4 或不清楚 <30 或不清楚		
分析					
25 用于分析数据的方法是否恰当？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	
26 是否至少有两位研究者参与了分析？	至少有两名研究者参与了分析	可以假定至少有两名研究者参与了分析，但是没有清晰地描述	不清楚是否至少有两名研究者参与了分析，或仅有一名研究者参与了分析		

2e. 询问专业人员关于全面性的问题					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
27 是否是用于评估 PROM 全面性的方法恰当？	使用了广泛认可或合理的方法	只有定量（调查）方法被使用，或可以假定所使用的方法是恰当的，但是没有清晰地描述	怀疑方法是否恰当	所使用的方法不恰当	
28 是否纳入了来自所有相关学科的专业人员？	纳入了来自所有相关学科的专业人员	可以假定纳入了来自所有相关学科的专业人员，但是没有清晰地描述	怀疑是否纳入了来自所有相关学科的专业人员，或没有纳入相关的专业人员		
29 每个条目是否在适当数量的专业人员中进行了测试？					
对于质性研究	≥7	4-6	<4 或不清楚		
对于量性（调查）研究	≥50	≥30	<30 或不清楚		
分析					
30 用于分析数据的方法是否恰当？	使用了广泛认可或合理的方法	可以假定方法是恰当的，但是没有清晰地描述	不清楚使用了什么方法，或怀疑方法是否恰当	方法不恰当	
31 是否至少有两位研究者参与了分析？	至少有两名研究者参与了分析	可以假定至少有两名研究者参与了分析，但是没有被清晰地描述	不清楚是否至少有两名研究者参与了分析，或仅有一名研究者参与了分析		

方框 3. 结构效度

量表是否由效应指标构成？换句话说，它是否基于反映性模型？是/否 [1]

研究涉及到单维性还是结构效度？ [2]　　单维性 / 结构效度

统计方法	非常好	合格	有问题的	不合格	不适用
1 对于经典测试理论：是否进行了探索性或验证性因子分析？	进行了验证性因子分析	进行了探索性因子分析		没有进行探索性或验证性因子分析	不适用
2 对于项目反应理论/Rasch：选择的模型是否适用于研究问题？	选择的模型很适用于研究问题	可以假定选择的模型很适用于研究问题	怀疑选择的模型是否很适用于研究问题	选择的模型不适用于研究问题	不适用
3 分析中是否有足够的样本量？	FA：条目数的 7 倍和≥100 Rasch/1PL 模型：≥200 名被试 2PL 参数项目反应理论模型或 Mokken 量表分析：≥ 1000 名被试	FA：至少条目数的 5 倍和≥100; 或至少条目数的 6 倍，但是<100 Rasch/1PL 模型：100-199 名被试 2PL 参数项目反应理论模型或 Mokken 量表分析：500-999 名被试	FA：条目数的 5 倍，但是<100 Rasch/1PL 模型: 50-99 名被试 2PL 参数项目反应理论模型或 Mokken 量表分析：250-499 名被试	FA：< 条目数的 5 倍 Rasch/1PL 模型: < 50 名被试 2PL 参数项目反应理论模型或 Mokken 量表分析：<250 名被试	
其他					
4 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法缺陷（例如，旋转方法或没有被描述）	其他重要的方法缺陷（例如，不恰当的旋转方法）	

[1] 如果量表不是基于反映性模型，则单维性或结构效度不适用。这种情况下，该研究可以被忽略。

[2] 在系统评价中，对单维研究[（研究分别对每个（分）量表进行因子分析以评估（分）量表是否是单维的）]和结构效度研究[研究对量表中所有条目进行因子分析以评估量表中（预期）分量表数量和分量表内条目聚类]进行区分是有帮助的。

方框 4. 内部一致性

量表是否由效应指标构成？换句话说，它是否基于反映性模型？ 是/ 否 [1]

设计要求	非常好	合格	有问题的	不合格	不适用
1 是否分别为每一个单维量表或分量表计算了内部一致性？	为每一个单维量表或分量表计算了内部一致性		不清楚量表或分量表是否为单维	没有为每一个单维量表或分量表计算内部一致性	
统计方法					
2 对于连续性分数：是否计算了 Cronbach's α 或 Ω 系数？	计算了 Cronbach'sα 或 Ω 系数		只计算了条目-总体相关性	没有计算 Cronbach'sα 值，没有计算条目-总体相关性	不适用
3 对于二分类分数：是否计算了 Cronbach's α 或 KR-20 值？	计算了 Cronbach'sα 或 KR-20 值		只计算了条目-总体相关性	没有计算 Cronbach'sα 或 KR-20 值和条目-总体相关性	不适用
4 对于基于项目反应理论的分数：是否计算了标准误差 θ 值（SE(θ)）或潜在特质估计值的信度系数（（主题或条目的）分离指数）？	计算了 SE(θ)或信度系数			没有计算 SE(θ)或信度系数	不适用
其它					
5 研究的设计或统计方法中是否存在其他重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

[1] 如果量表不是基于反映性模型，内部一致性不适用。这种情况下，该研究可以被忽略。

方框 5.跨文化效度/测量不变性					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
1 样本除了小组变量外，是否具有相似的相关特征？	有证据支持样本除了小组变量外具有相似的相关特征	陈述了（但是没有证据支持）样本除了小组变量外具有相似的相关特征	不清楚样本是否除了小组变量外具有相似的相关特征	样本除了小组变量外不具有相似的相关特征	
统计学方法					
2 用于分析数据的方法是否合适？	使用了广泛认可或合理的方法	可以假定方法是合适的，但是没有被清晰地描述	不清楚使用了什么方法，或怀疑方法是否合适	方法不合适	不适用
3 分析中的样本量是否合适？	回归分析或基于 IRT/Rasch 的分析：每组 200 名被试 MGCFA*：条目数的 7 倍和≥100 名被试	每组 150 名被试 条目数的 5 倍和≥100 例被试；或条目数的 5-7 倍，但是<100 名被试	每组 100 名被试 条目数的 5 倍，但是<100 名被试	每组＜100 名被试 <条目数的 5 倍	
其它					
4 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

**MGCFA: 多组验证性因子分析*

方框 6. 信度					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
1 在测量的间隔期间，在所测量的结构方面患者是否稳定？	有证据支持患者是稳定的	可以假定患者是稳定的	不清楚患者是否稳定	患者不稳定	
2 时间间隔是否合适？	时间间隔合适		怀疑时间间隔是否合适或时间间隔没有被描述	时间间隔不合适	
3 测量时的测量条件是否相似？如：测量方式、环境和指导语	测量条件相似（提供了证据）	可以假定测量条件相似	不清楚测量条件是否相似	测量条件不相似	
统计方法					
4 对于连续性分数：是否计算了组内相关系数（ICC）？	计算了 ICC，并且 ICC 模型或公式被描述	计算了 ICC，但是未描述 ICC 模型/公式，或 ICC 模型/公式不是最佳的。 有证据支持没有系统变化出现的情况下计算了 Pearson 或 Spearman 相关系数	没有证据支持未出现系统变化，或有证据支持出现了系统变化，在此情况下计算了 Pearson 或 Spearman 相关系数	没有计算 ICC 或 Pearson 或 Spearman 相关系数	不适用
5 对于二分类/名义/有序分数：是否计算了 kappa 值？	计算了 kappa 值			没有计算 kappa 值	不适用
6 对于有序分数：是否计算了加权 kappa 值？	计算了加权 kappa 值		计算了未加权的 Kappa 值，或没有描述		不适用
7 对于有序分数：是否描述了加权方案？如，线性加权、二次加权	描述了加权方案	没有描述加权方案			不适用
其它					
8 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

方框 7. 测量误差

设计要求	非常好	合格	有问题的	不合格	不适用
1 在测量的间隔期间，在所测量的结构方面患者是否稳定？	患者是稳定的（提供了证据）	可以假定患者是稳定的	不清楚患者是否稳定	患者是不稳定的	
2 时间间隔是否合适？	时间间隔合适		怀疑时间间隔是否合适或时间间隔没有被描述	时间间隔不合适	
3 测量时的测量条件是否相似？（如：测量方式、环境和指导语）	测量条件相似(提供了证据)	可以假定测量条件相似	不清楚测量条件是否相似	测量条件不相似	
统计方法					
4 对于连续性分数：是否计算了测量标准误差（SEM）、最小可测变化值（SDC）或一致性界限（LoA）？	计算了 SEM，SDC，或 LoA	从提供的数据中可能计算 LoA		基于 Cronbach's α 计算 SEM，或基于其它人群的 SD 值	不适用
5 对于二分类/名义/有序分数：是否计算了百分比一致性（阳性和阴性）？	计算了阳性和阴性百分比一致性	计算了百分比一致性		没有计算百分比一致性	不适用
其它					
6 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

方框 8. 效标效度					
	非常好	**合格**	**有问题的**	**不合格**	**不适用**
统计方法					
1 对于连续性分数：是否计算了相关性、或受试者工作特征曲线下的面积？	计算了相关性或 AUC			没有计算相关性或 AUC	不适用
2 对于二分类分数：是否确定了敏感度和特异度？	计算了敏感度和特异度			没有计算敏感度和特异度	不适用
其它					
3 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

方框 9. 结构效度的假设检验

9a. 与其它结果测量工具的比较（聚合效度）

设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
1 对照测量工具测量的结构是否清晰？	对照测量工具所测量的结构清晰			对照测量工具所测量的结构不清晰	
2 对照测量工具的测量属性是否足够？	在与研究人群相似的人群中，对照测量工具有足够的测量属性	对照测量工具有足够的测量属性，但不确定这些属性是否适用于研究人群	有一些关于在任何研究人群中，对照测量工具的测量属性方面的信息	没有关于对照测量工具测量属性的信息，或者关于对照测量工具测量属性的证据不足够	
统计方法					
3 用于检验假设的统计方法是否合适？	统计方法是合适的	可以假定统计方法是合适的	应用的统计方法不是最佳的	应用的统计方法是不合适的	
其它					
4 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷（例如，只呈现了与测量另一个结构工具相比较的数据）	其它重要的方法学缺陷	

9b. 亚组之间的比较(区分效度或公认群体效度)					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
5 对亚组的重要特征是否提供了足够的描述?	充分描述亚组的重要特征	充分描述亚组的大部分重要特征	未描述亚组的重要特征		
统计方法					
6 用于检验假设的统计方法是否合适?	统计方法是合适的	可以假定统计方法是合适的	应用的统计方法不是最佳的	应用的统计方法是不合适的	
其它					
7 研究的设计或统计方法中是否存在其它重要缺陷?	没有其它重要的方法学缺陷		其它次要的方法学缺陷（例如，只呈现了与测量另一个结构工具相比较的数据）	其它重要的方法学缺陷	

方框 10. 反应度

10a. 标准方法（如：与金标准比较）

	非常好	合格	有问题的	不合格	不适用
统计方法					
1 对于连续性分数：是否计算了变化分数之间的相关性或受试者工作特征曲线（ROC）曲线下的面积？	计算了相关性或 ROC 曲线下的面积（AUC）			没有计算相关性或 AUC	不适用
2 对于二分类分数：是否确定了敏感度和特异度（变化 VS 不变化）？	计算了敏感度和特异度			没有计算敏感度和特异度	不适用
其它					
3 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

10b. 构想方法（如：假设检验；与其它结果测量工具对比）					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
4 对照测量工具测量的结构是否清晰？	对照测量工具所测量的结构清晰			对照测量工具所测量的结构不清晰	
5 对照测量工具的测量属性是否足够？	在与研究人群相似的人群中，对照测量工具有足够的测量属性	对照测量工具有足够的测量属性，但不确定这些属性是否适用于研究人群	有一些关于在任何研究人群中，对照测量工具的测量属性方面的信息	没有关于对照测量工具测量属性的信息，或者关于对照测量工具的证据质量差	
统计方法					
6 用于检验假设的统计方法是否恰当？	统计方法是恰当的	可以假定统计方法是恰当的	应用的统计方法不是最佳的	应用的统计方法是不恰当的	
其它					
7 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

10c. 构想方法（如：假设检验：亚组之间比较）					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
8 对亚组的重要特征是否提供了足够的描述？	充分描述亚组的重要特征	充分描述亚组的大部分重要的特征	未描述亚组的重要特征		
统计方法					
9 用于检验假设的统计方法是否恰当？	统计方法是恰当的	可以假定统计方法是恰当的	应用的统计方法不是最佳的	应用的统计方法是不恰当的	
其它					
10 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	

10d. 构想方法（如：假设检验：干预前和干预后）					
设计要求	**非常好**	**合格**	**有问题的**	**不合格**	**不适用**
11 对给予的干预方法是否提供了足够的描述？	干预的描述足够		对干预的描述不佳	没有描述干预	
统计方法					
12 用于检验假设的统计方法是否恰当？	统计方法是恰当的	可以假定统计方法是恰当的	应用的统计方法不是最佳的	应用的统计方法是不恰当的	
其它					
13 研究的设计或统计方法中是否存在其它重要缺陷？	没有其它重要的方法学缺陷		其它次要的方法学缺陷	其它重要的方法学缺陷	